doi:10.3772/j.issn.2095-915x.2024.01.007

基于 ARIMA、GM（1,1）模型的高校 ESI 学科发展预测研究

开放科学（资源服务）标识码（OSID）

柳佳彤[1] 康榆晨[1] 秦丽岩[1] 曹芳[2]

1. 新疆医科大学公共卫生学院 乌鲁木齐 830017；
2. 新疆医科大学图书馆 乌鲁木齐 830017

摘要：[目的 / 意义] 学科建设是高校提升教育质量的关键环节，对科学研究起着重要的支撑作用。采用数学统计建模探索一种科学有效的方法，实现潜力学科入围 ESI 前 1% 的时间预测，对于机构学科发展规划有着重要指导意义。[方法 / 过程] 基于 ESI 数据库，获取目标机构 4 个潜力学科的被引频次和 ESI 入围阈值，建立时间序列并创建预测模型：先引入转换系数来去除不同数据库的差异，使其可比，然后分别拟合 GM（1,1）模型、ARIMA 模型，预测目标学术机构学科被引频次和 ESI 入围阈值，找到目标机构学科被引频次赶上 ESI 入围阈值的时间，即预测的入围时间。通过采用平均绝对百分比误差（MAPE）、平均绝对误差（MAE）和均方根误差（RMSE）对模型的拟合预测效果进行评估和比较，根据 MAPE、MAE 和 RMSE 三个指标来评价模型拟合及预测效果，以此为学校的学科建设及长远发展规划提供参考依据。[局限] 本研究仅局限于目标机构 4 个学科的数据，尚需获取其他机构、更多学科的数据进行模型预测效果验证。[结果 / 结论]ARIMA 模型的拟合效果和预测效果优于 GM（1,1）模型。目标机构的生物学与生物化学学科可能于近期入围 ESI 前 1%；免疫学科有入围 ESI 前 1% 学科的潜力，但入围时间可能会稍微滞后；分子生物学与遗传学和神经科学与行为学学科，离入围还有较大差距。

关键词：ESI；Incites；潜力学科；灰色模型；ARIMA 模型

中图分类号：G35; TP391

Research on the Prediction of the Development of ESI Disciplines in Universities Based on ARIMA and GM (1,1) Models

LIU Jiatong[1] KANG Yuchen[1] QIN Liyan[1] CAO Fang[2]

1. Public Health of Xinjiang Medical University, Urumqi 830017, China;
2. Xinjiang Medical University Library, Urumqi 830017, China

Abstract: [Objective/Significance] Subject construction is a key aspect for universities to enhance the quality of education

作者简介 柳佳彤（2000-），本科生，主要研究方向为预防医学；康榆晨（2001-），本科生，主要研究方向为预防医学；秦丽岩（1990-），通讯作者，硕士，讲师，主要研究方向为流行病学；曹芳（1985-），硕士，副研究馆员，主要研究方向为图书竞争情报分析、数据挖掘，E-mail：Joanna567@xjmu.edu.cn。

引用格式 柳佳彤，康榆晨，秦丽岩，等 . 基于 ARIMA、GM（1,1）模型的高校 ESI 学科发展预测研究 [J]. 情报工程，2024, 10(1): 85-95.

doi:10.3772/j.issn.2095-915x.2024.01.007

and plays an important role in supporting scientific research. This article adopts mathematical statistical modeling to explore a scientifically effective method for predicting the time it takes for a potential subject to enter the top 1% in ESI rankings. This has significant guidance implications for institutional subject development planning. [Methods/Processes] Based on the ESI database, this paper obtains the citation frequency and ESI shortlisting threshold of the four potential disciplines of the target institution, establishes the time series, and creates a prediction model: first introduce the conversion coefficient to remove the differences between different databases and make them comparable, and then fit the GM(1,1) model and ARIMA model respectively to predict the citation frequency and ESI shortlisting threshold of the target academic institution, and find the time when the citation frequency of the subject of the target institution catches up with the ESI shortlisting threshold, that is, the predicted shortlisting time. By using mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE) to evaluate and compare the fitting and prediction effect of the model, the model fitting and prediction effect were evaluated according to the three indicators of MAPE, MAE and RMSE, so as to provide a reference basis for the discipline construction and long-term development planning of the school. [Limitations] The limitation of the study is data from only four disciplines in the target institution. Additional data from other institutions and more disciplines are needed to validate the predictive performance of the model. [Results /Conclusions] The fitting effect and prediction effect of ARIMA model are better than those of GM(1,1) model. The biology and biochemistry disciplines of the target institution will be in the top 1% of ESI in the coming months; Immunology has the potential to be shortlisted in the top 1% of ESI, but the shortlisting time may be slightly delayed; The disciplines of molecular biology and genetics and neuroscience and behavior are still far from being shortlisted.

Keywords: ESI；Incites; Potential Discipline; Gray Model; ARIMA Model

引言

学科建设是各大高校走内涵式发展道路的重要基础，亦是各高校从量的扩张到质的提升的关键因素。“双一流”建设是国家重大战略决策，为我国学科建设指明了新的方向，代表着我们国家为推动一流大学、一流学科建设的坚定决心。2022 年 1 月，党中央、国务院发布了《关于深入推进世界一流大学和一流学科建设的若干意见》，突出强调要培养一流人才、服务国家战略需求、争创世界一流的导向 [1]。

近年来，广大发展中国家将基本科学指标数据库（ESI）作为评价不同国家、不同地区、不同高校及不同科研机构学术能力及影响力的衡量指标。ESI 根据世界各国或研究单位不同分类的学科，将所有学科分为 22 个种类。ESI 数据库每两个月更新一次，统计的数据范围是近 10 年的论文总被引频次，在我国 ESI 评价体系已被教育部门和各级评价部门广泛用以高校绩效的评价。由于 ESI 数据库仅能提供机构入围学科的发文量和被引频次等信息，因此预测未入围学科晋级 ESI 前 1% 的时间是高校学科建设工作的重要目标。

灰色系统理论是由邓聚龙 [2] 在 1982 年创建的解决信息不完全和不确定性问题的学习和建模方法。灰色预测模型通过对系统内部灰色因素的分析和建模，对未来趋势进行预测，由于灰色预测模型具有简单易用、适用时序短、数据样本量较小等优点，被广泛应用于各个领域，如经济、环境、医疗等，为决策者提供了重要的参考依据。ARIMA 模型其全称是自回归移动平均模型，属于统计模型中最常见的一

种，是利用时间序列多个历史时刻对应的值预测未来时刻对应的值的一种方法。朱文佳等[3]用 ARIMA 模型拟合目标机构 ESI 被引频次预估值的时间序列，用时间序列模型拟合 ESI 入围阈值时间序列，以复旦大学经济与商学为例进行实证研究，结果预测模型在两个时间序列上都有较高的拟合度，得出的入围时间预测值可信度较高。本文在研究灰色预测模型 GM（1，1）及 ARIMA 模型的基础上，通过实证分析，对目标机构 4 个学科进行 ESI 前 1% 排名入围时间进行预测，以期助力该校进一步提升学科建设质量。将笔者所在机构视为目标机构，研究目标机构的学科建设及发展。目标机构现有医学、理学、工学、管理学、法学等学科门类，其中有一门学科进入 ESI 全球排名前 2.63‰，还有一门学科位列 ESI 全球排名前 1%。4 个学科入选 “十四五”重点学科建设序列。

1 研究方法

1.1 GM（1,1）模型

这是一维一次求导的灰色预测模型，即自变量对自身进行预测分析[4]。虽然可以选择各种类型的灰色模型，但由于计算效率高，以往的研究大多集中在预测 GM（1,1）模型[5]。

已知参考数据列为：$X^{(0)}=\{X^{(0)}(1), X^{(0)}(2), \cdots\cdots, X^{(0)}(n), \}$，做一阶累加生成序列（AGO）：$X^{(1)}=\{X^{(1)}(1), X^{(1)}(2), \ldots\ldots, X^{(1)}(n)\}$，其中

$$X^1\ \mathrm{k}=\sum_{\mathrm{i}=1}^{\mathrm{k}} X^0\ \mathrm{i} \qquad (1)$$

k=1,2,3，……n，

求累加序列的均值：$Z^{(1)}(k)=\{0.5X^{(1)}(k)+0.5X^{(1)}(k-1)$，k=2,3，……n，则 GM（1,1）对应的白化方程为：

$$\frac{\mathrm{dX}^{(1)}}{\mathrm{dt}}+\alpha X^{(1)}(\mathrm{t})=\mu \qquad (2)$$

其中 α 为发展灰数，μ 为内生控制灰数，利用最小二乘法得：

$$\begin{bmatrix}\alpha\\ \mu\end{bmatrix}=(B^TB)^{-1}B^TY \qquad (3)$$

其中，$B=\begin{vmatrix}-Z^{(1)}(2) & 1\\ -Z^{(1)}(3) & 1\\ \vdots & \vdots\\ -Z^{(1)}(\mathrm{n}) & 1\end{vmatrix}$，$Y=\begin{vmatrix}X^{(0)}(2)\\ X^{(0)}(3)\\ \vdots\\ X^{(0)}(\mathrm{n})\end{vmatrix}$，可得到预测方程：

$$\hat{X}^{(1)}(\mathrm{k}+1)=(X^{(0)}(1)-\frac{\mu}{\alpha})\mathrm{e}^{-\alpha k}+\frac{\mu}{\alpha} \qquad (4)$$

k=1,2，……n-1，

原始预测值为：

$$\hat{X}^{(0)}(\mathrm{k}+1)=\hat{X}^{(1)}(k+1)-\hat{X}^{(1)}(k) \qquad (5)$$

k=1,2，……n-1，

模型的检验：$\mathrm{e}=X^{(0)}(\mathrm{k})-\hat{X}^{(0)}(\mathrm{k})$，

$S_1=\sqrt{\sum_{\mathrm{i}=0}^{\mathrm{k}}(X^{(0)}(\mathrm{i}+1)-\hat{X}^{(0)})^2/\mathrm{k}\text{-}1}$，$S_2=\sqrt{\sum_{\mathrm{i}=0}^{\mathrm{k}}(\mathrm{e}^{(0)}(\mathrm{i}+1)-e^{(0)})^2/k-1}$ 计算后验差比值、小误差概率：

$$C=\frac{S_2}{S_1}\qquad p=P\left\{|e(k)-\bar{e}|<0.6745\mathrm{S}_1\right\} \qquad (6)$$

GM（1,1）模型精度等级评价见表 1 所示。

表 1 GM（1,1）模型精度等级表

模型精度等级	p	c
1 级（好）	$p \geq 0.95$	$c \leq 0.35$
2 级（合格）	$0.80 \leq p < 0.95$	$0.35 \leq c < 0.5$
3 级（勉强）	$0.70 \leq p < 0.80$	$0.5 < c \leq 0.65$
4 级（不合格）	$p < 0.70$	$c > 0.65$

doi:10.3772/j.issn.2095-915x.2024.01.007

1.2 ARIMA模型

ARIMA 是一种著名的时间序列预测方法，它强调分析时间序列数据的随机性和概率性，被广泛用于预测传染病的出现和演变。该模型无需考虑各种影响因素，可以将这些综合效应纳入时间变量中，只需分析历史数据即可实现建模和定量预测，在医疗卫生领域具有广阔的应用前景[6]。其模型核心部分为自回归（AR）、差分（d）和移动平均（MA)。当模型为平稳序列时，时序问题可用 ARMA 模型进行预测。

（1）ARMA（p,q）模型

假定时间序列中包括自回归与移动平均两部分，ARMA 模型就可以表达为：

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \quad (7)$$

其中，自回归模型 AR（p）为：

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + e_t \quad (8)$$

移动平均模型 MA（q）为：

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \quad (9)$$

（2）ARIMA（p,d,q）模型

假设数据一般模型为：

$$Y_t = \mu_t + X_t \quad (10)$$

其中 μ_t 是时变均值，X_t 是零均值平稳序列。如果一个时间序列 $\{Y_t\}$ 的 d 次差分 $W_t = \nabla^d Y_t$ 是一个平稳的 ARMA 过程，则称 $\{Y_t\}$ 为差分自回归移动平均模型。

ARIMA（p,d,q）模型[7]，其中 AR 是自回归，p 为自回归项；d 是需要对数据进行差分的阶数；MA 为移动平均，q 为移动平均项数。建模基本步骤：（1）验证数据的平稳性；（2）模型识别与定阶；（3）参数估计；（4）预测并确定最优模型[8-9]。建模流程见图 1 所示。

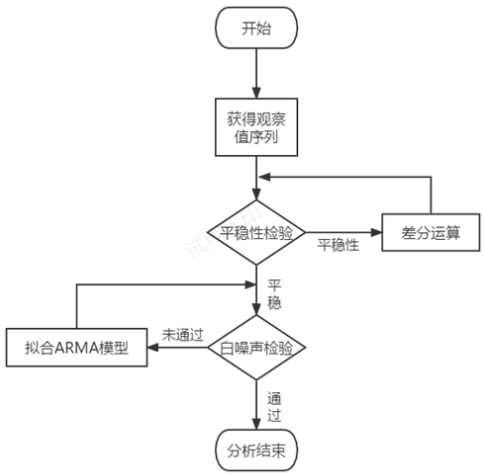

图 1 建模流程图

模型的检验流程如下：

（1）单位根检验

单位根检验有许多方法，较为常用的是 ADF 检验，是 DF 检验的扩展，其应用的是在建模前对数据进行假设检验来判断数据是否为平稳数据，其中 H0：数据不平稳，H1：数据平稳，检验水准为 0.05。若 P>0.05，则存在单位根，表明数据不平稳，需要进行差分处理使之平稳；反之，则数据平稳，可以进行后续建模。

（2）白噪声检验

对数据进行白噪声检验，若为白噪声，则过去的行为对未来没有影响，不能进行预测，反之则可进行预测。常见的白噪声检验为 Ljung-Box 检验。其中 H0：数据为白噪声序列，H1：数据不为白噪声序列，检验水准为 0.05。经检验，上述数据并非白噪声序列，可以对其进行建模。

（3）模型识别与定阶

对平稳后的时序绘制自相关图（ACF）、偏自相关图（PACF），ARIMA 模型的定阶原则如表 2 所示。

表 2　ARIMA 模型的定阶原则

模型	ACF	PACF
AR(p)	拖尾	p 阶截尾
MA(q)	q 阶截尾	拖尾
ARMA(p,q)	拖尾	拖尾

但仅自相关图（ACF）、偏自相关图（PACF）不能较为精确进行定阶，故我们基于 AIC 准则和 BIC 准则，参考各 ARIMA 模型的 AIC 函数值和 BIC 函数值选出最优的模型。

赤池信息量（AIC）：$AIC=-2\ln(L)+2k$。其中，k 表示模型参数个数，L 表示似然函数。贝叶斯信息量（BIC）：$BIC=-2\ln(L)+k\ln(n)$。其中，k 表示模型参数个数，n 表示样本数量，L 表示似然函数。AIC、BIC 值越小的模型效果越好。

2　实证分析

2.1　数据来源

本研究基于 ESI 数据库，数据来自团队定期采集的历史数据，本研究使用 2019—2023 年更新的 23 期数据作为历史样本。ESI 数据库最近一次更新为 2023 年 5 月 11 日，本期 ESI 统计数据覆盖时间范围为 2013 年 1 月 1 日至 2023 年 2 月 28 日，整 10 年。每年 5 月 ESI 都会剔除最早一年的数据，数据覆盖的时间范围在 10-11 年间波动。一般年底与年初更新的数据覆盖范围都接近 11 年，本研究将每年年底作为时间序列的关键时间节点，因此为保持一致，将 InCites 与 WoS 采集的数据时间范围也设定为近 11 年。

2.2　时间序列的建立

ESI 数据库每年更新 6 次，最新一次是 2023 年数据的第 3 次更新。将最新一期的 ESI 被引频次作为转换系数的计算依据，各学科 23 个历史时间节点的入围机构最低被引频次形成的时间序列，将作为预测未来入围阈值的依据。

WoS 数据库不提供 ESI 学科分类，要通过 InCites 数据库确定 WoS 数据库论文的 ESI 学科分类。以 2022 年底为例，先从 InCites 数据库中获得目标机构在目标 ESI 学科 2012—2022 年发表的论文集合，然后在 WoS 论文合集中检索，使用 WoS 的引文分析功能，即可得到近 11 年的 WoS 总被引频次，再根据转换系数得到 2022 年底时的近 11 年 ESI 总被引频次的估计值。

同理，可得到截至 2022 年底的近 11 年 ESI 总被引频次的估计值，从而预测目标机构 2023 年底时的近 11 年 ESI 总被引频次，再与 2023 年底 ESI 入围机构最低被引频次的预测值进行比较，如果后者小于前者，该时间就是 ESI 前 1% 学科预测的入围时间。

朱文佳等学者的研究 [3] 表明，只要有一定数量的样本，使用转换系数或者修正因子的平均值，来推测未入围机构的 ESI 数据，是具有一定准确性的。笔者选择了修正因子的平均值进行了实证研究。

修正因子的计算方法如下：

$$ESI/\text{WOS修正因子}_i\text{平均值}=Average\left(\frac{ESI\text{中第}i\text{个学科某个机构的被引频次}}{WoS\text{第}i\text{个学科某机构的被引频次}}\right) \quad (11)$$

基于以上原理，我们选择目标机构的四个最有潜力的学科，分别计算其转换系数，并根据计算所得的转换系数预估目标机构的 ESI 总被引频次的估计值，得到所选四个学科的 23 期学科阈值的时间序列（见表 3）。

doi:10.3772/j.issn.2095-915x.2024.01.007

表 3　4 个学科 ESI 入围机构最低被引频次(入围阈值)(篇)

更新时间	分子生物学与遗传学	免疫学	神经科学与行为学	生物学与生物化学
2019-09	14174	5141	6581	6538
2019-11	14628	5253	6818	6728
2020-01	14716	5327	6904	6747
2020-03	14681	5419	6795	6823
2020-05	14132	5149	6426	6316
2020-07	14275	5204	6456	6348
2020-09	14621	5281	6545	6441
2020-11	14743	5357	6618	6584
2021-01	14817	5401	6715	6603
2021-03	14990	5492	6793	6769
2021-05	14208	5201	6480	6397
2021-07	14704	5368	6527	6602
2021-09	14615	5462	6650	6694
2021-11	14957	5552	6856	6890
2022-01	15100	5586	6986	6895
2022-03	15205	5668	7080	6986
2022-05	14830	5417	6668	6624
2022-07	15085	5428	6807	6673
2022-09	14849	5525	6889	6812
2022-11	14918	5631	6813	6910
2023-01	15227	5679	6951	7065
2023-03	15437	5699	7059	7126
2023-05	13496	5431	6914	6732

2.3　数据处理与预测

2.3.1　灰色模型预测结果

研究过程中，通过 GM（1，1）模型建立，利用 MATLAB，对样本数据进行计算分析及预测。GM（1,1）模型预测相关参数指标见表 4。

（1）建立预测公式：

基于以上相关参数指标可以建立相关学科的预测方程。

基于 ESI 的预测方程为：

表 4　基于 GM(1,1) 模型预测相关参数指标

学科		α	μ	c	p
分子生物学与遗传学	ESI	0.0266	15966	0.8071	0.8
	Wos	-0.0042	7870	0.1729	1
免疫学	ESI	0.0102	5781.9	0.7484	0.8
	Wos	-0.1168	1896.4	0.33	1
神经科学与行为学	ESI	-0.0059	6812.5	0.8323	0.8
	Wos	-0.0527	1329.8	0.2074	1
生物学与生物化学	ESI	0.0067	7097.4	0.863	0.8
	Wos	-0.0737	3988	0.2552	1

分子生物学与遗传学：

$$X^{(1)}(k+1) = -585376.56e^{-0.0266k} + 600225.56 \quad (12)$$

其中，k=1,2，……n-1。

免疫学：

$$X^{(1)}(k+1) = -561327.94e^{-0.0102k} + 566852.94 \quad (13)$$

其中，k=1,2，……n-1。

神经科学与行为学：

$$X^{(1)}(k+1) = 1161550.02e^{0.0059k} - 1154661.02 \quad (14)$$

其中，k=1,2，……n-1。

生物学与生物化学：

$$X^{(1)}(k+1) = -1052501.43e^{-0.0067k} + 1059313.43 \quad (15)$$

其中，k=1,2，……n-1。

基于 WOS 的预测方程为：

分子生物学与遗传学：

$$X^{(1)}(k+1) = 1880063.52e^{0.0042k} - 1873809.52 \quad (16)$$

其中，k=1,2，……n-1。

免疫学：

$$X^{(1)}(k+1) = 18097.30e^{0.1168k} - 16236.30 \quad (17)$$

其中，k=1,2，……n-1。

神经科学与行为学：

$$X^{(1)}(k+1) = 26375.40e^{0.0527k} - 25233.40 \quad (18)$$

其中，k=1,2，……n-1。

生物学与生物化学：

$$X^{(1)}(k+1)=57789.26e^{0.0737k}-54111.26 \quad (19)$$

其中，k=1,2，……n-1。

（2）预测结果

$$\hat{X}^{(0)}(k+1)=\hat{X}^{(1)}(k+1)-\hat{X}^{(1)}(k) \quad (20)$$

通过上述公式可以计算出原始序列的预测值，结果见表 5、表 6。

表 5　目标机构 4 个学科 ESI 入围机构最低被引频次预测结果及残差（灰色模型）

时间	分子生物学与遗传学		免疫学		神经科学与行为学		生物学与生物化学	
	预测值	残差	预测值	残差	预测值	残差	预测值	残差
2022-09	14849	0	5525	0	6889	0	6812	0
2022-11	15364	-446	5696.3	-65.3	6873.2	-60.2	7028.2	-118.2
2023-01	14960	267	5638.3	40.7	6913.7	37.3	6981.4	83.6
2023-03	14567	870	5581	118	6954.5	104.5	6934.8	191.2
2023-05	14184	-688	5524.2	-93.2	6995.5	-81.5	6888.5	-156.5
2023-07	13811		5468		7036.7		6842.6	
2023-09	13448		5412.4		7078.2		6797	
2023-11	13095		5357.4		7120		6751.6	
2024-01	12750		5302.9		7162		6706.6	

表 6　目标机构 4 个学科 ESI 被引频次预估值预测结果及残差（灰色模型）

时间	分子生物学与遗传学		免疫学		神经科学与行为学		生物学与生物化学	
	预测值	残差	预测值	残差	预测值	残差	预测值	残差
2008-2018	6254	0	1861	0	1142	0	3678	0
2009-2019	7912.1	-147.1	2242.2	-212.2	1427.3	-13.3	4420.1	-203.1
2010-2020	7945.1	203.9	2520	192	1504.6	-18.6	4758	187
2011-2021	7978.2	31.8	2832.2	168.8	1586	73	5121.8	187.2
2012-2022	8011.5	-88.5	3183.1	-146.1	1671.9	-39.9	5513.4	-165.4
2013-2023	8044.9		3577.4		1762.3		5935	
2014-2024	8078.4		4020.6		1857.7		6388.7	

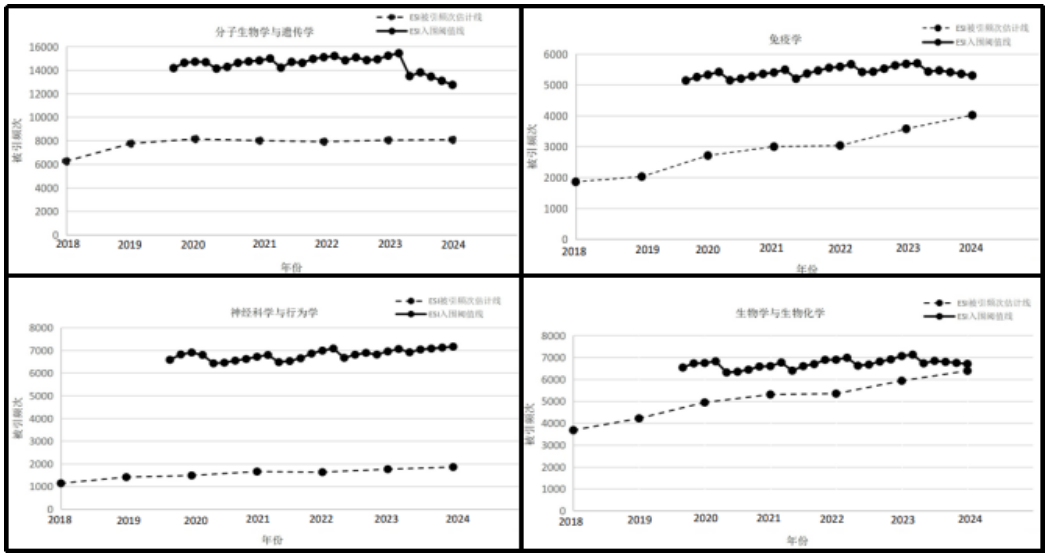

图 2　目标机构 4 个学科 ESI 前 1% 入围时间预测模型（灰色模型）

doi:10.3772/j.issn.2095-915x.2024.01.007

2.3.2 ARIMA模型预测结果

我们分别选取2019年9月到2023年5月的ESI数据和2002年到2023年的ESI被引频次估计值，分别选取后三组数据测试，其余为训练集，进行模型建立。预测结果见表7、表8。

表7 目标机构4个学科ESI入围机构最低被引频次预测结果及残差（ARIMA模型）

时间	分子生物学与遗传学		免疫学		神经科学与行为学		生物学与生物化学	
	预测值	残差	预测值	残差	预测值	残差	预测值	残差
2022-07	15336.2	-251.2	5514.8	-86.8	6704.8	102.2	6744.0	-71.0
2022-09	14986.3	-137.3	5504.0	21.0	6909.7	-20.7	6811.2	0.8
2022-11	15166.6	-248.6	5608.0	23.0	7015.4	-202.4	6967.6	-57.6
2023-01	15022.7	204.3	5673.3	5.7	6923.2	27.8	6927.4	137.6
2023-03	15089.4	347.6	5753.8	-54.8	7007.7	51.3	7033.4	92.6
2023-05	14676.1	-1180.1	5475.8	-44.8	6654.9	259.1	6651.9	80.1
2023-07	14892.8		5544.5		6738.9		6780.3	
2023-09	14618.6		5626.4		6845.4		6869.1	
2023-11	14649.3		5713.0		6938.2		7010.4	
2024-01	14715.7		5751.4		7060.3		6999.1	

表8 目标机构4个学科ESI被引频次预估值预测结果及残差（ARIMA模型）

时间	分子生物学与遗传学		免疫学		神经科学与行为学		生物学与生物化学	
	预测值	残差	预测值	残差	预测值	残差	预测值	残差
2018年	7128.3	-874.1	1714.2	147.2	1227.8	-85.4	3781.7	-104.0
2019年	6922.8	841.7	2195.0	-165.0	1287.0	126.7	4130.9	85.9
2020年	9274.8	-1126.0	2198.5	513.7	1580.2	-94.6	4755.9	189.0
2021年	8533.1	-523.2	3394.6	-393.3	1690.9	-31.4	5673.0	-363.6
2022年	7871.0	51.9	3290.2	-253.3	1765.1	-133.0	5673.9	-325.6
2023年	7732.1		3579.2		1916.4		6038.4	
2024年	7593.2		3868.2		2037.2		6403.0	

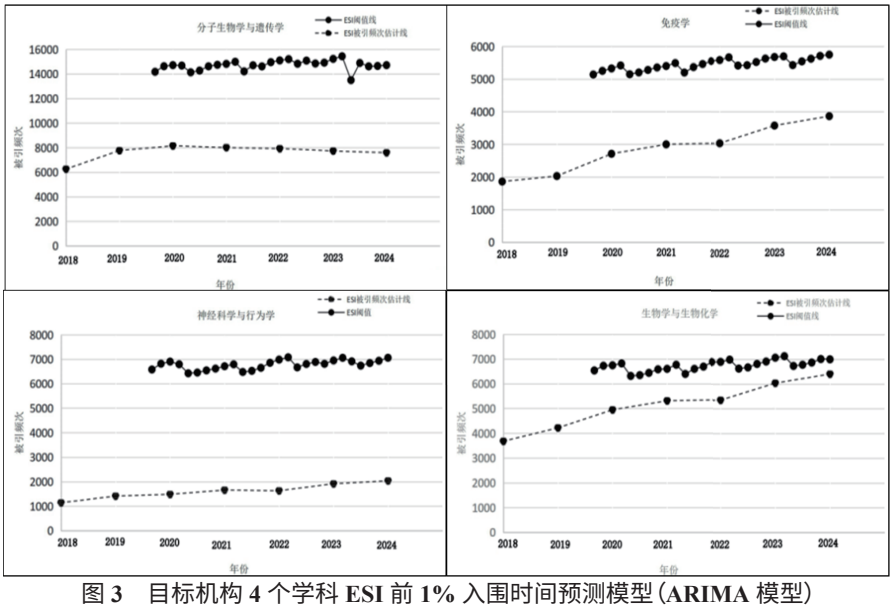

图3 目标机构4个学科ESI前1%入围时间预测模型（ARIMA模型）

2.4 模型评价

对于上述两种模型采用平均绝对百分比误差（MAPE）、平均绝对误差（MAE）和均方根误差（RMSE）三个指标来评价模型拟合及预测效果，其中 MAPE、RMSE 越小，模型的预测精度越高 [10-12]。计算公式如下：

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{21}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{22}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{23}$$

通过对上述三种模型预测的精度比较，从 MAPE 来看，拟合效果 ARIMA 优于 GM(1，1)，MAE、RMSE 指标拟合效果与 MAPE 基本一致。综合比较来看 ARIMA 模型拟合效果和预测效果相对于 GM（1,1）模型较好，不同模型拟合及预测效果见表 9、表 10。

此外，基于上述三种模型对 ESI 阈值和 ESI 被引频次估计值的预测，我们发现，目标机构的生物学与生物化学学科会于未来几个月入围 ESI 前 1%；免疫学科有入围 ESI 前 1% 学科的潜力，但入围时间可能会稍微滞后；分子生物学与遗传学和神经科学与行为学学科，离入围还有较大差距。

表 9 3 种模型对 ESI 阈值拟合及预测效果

学科	模型	拟合值			预测值		
		MAPE	MAE	RMSE	MAPE	MAE	RMSE
分子生物学与遗传学	GM（1，1）	1.58%	237.67	300.11	5.37%	779	784.3
	ARIMA	1.03%	153.04	213.25	4.11%	577.33	720.02
免疫学	GM（1，1）	0.63%	35.33	44.42	1.89%	105.6	106.33
	ARIMA	0.33%	17.79	31.4	0.63%	35.06	40.96
神经科学与行为学	GM（1，1）	0.47%	32.5	40.89	1.33%	93	93.71
	ARIMA	0.73%	49.83	76.91	1.62%	112.72	153.32
生物学与生物化学	GM（1，1）	0.96%	67.27	83.59	2.50%	173.85	174.71
	ARIMA	0.53%	35.17	50.91	1.48%	103.44	106.35

表 10 3 种模型对 ESI 被引频次估计值拟合及预测效果

学科	模型	拟合值			预测值		
		MAPE	MAE	RMSE	MAPE	MAE	RMSE
分子生物学与遗传学	GM（1，1）	1.47%	117	145.16	0.76%	60.15	66.5
	ARIMA	7.59%	463.83	596.52	0.65%	51.86	51.86
免疫学	GM（1，1）	5.84%	134.73	165.22	5.22%	157.45	157.86
	ARIMA	11.71%	184.21	239.32	8.34%	253.3	253.3
神经科学与行为学	GM（1，1）	0.73%	10.63	13.2	3.42%	56.45	58.83
	ARIMA	7.00%	60.58	75.06	8.00%	132.96	132.96
生物学与生物化学	GM（1，1）	2.87%	130.03	159.39	3.31%	176.3	176.64
	ARIMA	6.82%	172.7	219.58	6.00%	325.62	325.62

doi:10.3772/j.issn.2095-915x.2024.01.007

3 总结与建议

在“双一流”建设背景下，学科建设对高校发展至关重要。如何通过学科建设赋能高校发展无疑是我国高校值得探究的重要事项[13]。本研究基于 ESI 采集的历史数据，发现 ESI、InCites 与 WoS 数据库之间被引频次存在差异，我们引入转换系数来去除不同数据库的差异，使其可比。其次，我们基于采集的历史数据和从不同数据库收集到的数据，结合时间序列数据的时序性特征，分别采用 GM（1,1）模型、ARIMA 模型拟合，预测目标机构潜力学科入围 ESI 前 1% 排名的时间。以目标机构的生物学与生物化学、免疫学、分子生物学与遗传学、神经科学与行为学四个学科为例进行实证研究，采用绝对百分比误差（APE）、平均绝对百分比误差（MAPE） 和均方根误差（RMSE）比较模型拟合预测效果，根据 APE、MAPE 和 RMSE 最小原则选择最优预测模型。通过本研究提出以下建议。

3.1 深入了解自身基础，强化潜势学科培育

目标机构需要明确学校发展定位，深入了解自身基础。夯实基础，加强调查研究，充分结合学校情况，分析各学科的优势与不足，找出优势学科与潜势学科，强化潜势学科培育。重视生物学与生物化学学科和免疫学科的培育，提升学科质量，增强科研产出能力。

3.2 营造良好学科环境，保持优势学科特色

高校应积极加强学科科研工作，完善学科科研创新管理机制，营造良好的科研环境。优势学科是学科生态系统生长发育的基石，高校应该充分利用特色学科的优势。对于进入 ESI 全球排名前 2.63‰，以及位列 ESI 全球排名前 1% 的优势学科，继续提升学科优势，带动相关学科发展，能够更好地营造良好学科环境。同时要遵循学科发展规律，保持学科发展的灵活性、成长性等特色[14]。

3.3 增加学科建设投入，加强师资队伍建设

增加学科建设投入经费，提高高校自主获取经费的能力，促使其通过多渠道获取经费，争取更多中央财政资金和地方财政资金，经费投入向国家战略需求的学科适度倾斜。在人才培养增加投入，加强对高层次人才和团队的建设，给学者提供准确及时的信息，以提高学者的科学研究效率。

总之，基于对 ESI 数据库的利用和发掘，能够促进高校高质量内涵式发展。各高校应该更加重视学科建设，为世界一流高校的建设强基固本。

参 考 文 献

[1] 周建国，王教志，陈丽，等．“双一流”建设背景下提升研究生学位授予质量的路径探索——以温州医科大学为例 [J]. 温州医科大学学报，2023, 53(4): 340-344.

[2] JU-LONG D. Control problems of grey systems[J]. Systems and Control Letters, 1982, 1(5): 288-294.

[3] 朱文佳，朱莉．基于时间序列分析法的 ESI 前 1% 学科入围时间预测模型 [J]. 情报理论与实践，2019, 42(10): 137-145.

[4] 肖玲．基于灰色预测模型的 B2B 电子商务交易规模预测研究 [D]. 武汉：华中师范大学，2022.

[5] WEI W, WANG G, TAO X, et al. Time series prediction for the epidemic trends of monkeypox using the ARIMA, exponential smoothing, GM (1,

1) and LSTM deep learning methods[J]. Journal of General Virology, 2023, 104(4): 001839.

[6] MALKI Z, ATLAM ES, EWIS A, et al. ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound[J]. Neural Computing & Applications, 2021, 33(7): 2929-2948.

[7] 李田田 . 基于时间序列分析与神经网络模型的股票价格预测 [D]. 大连 : 东北财经大学 , 2021.

[8] 王燕 . 应用时间序列分析 [M]. 北京 : 中国人民大学出版社 , 2022: 134-140.

[9] 杜亚萍 , 魏骅 , 陶群山 . 基于 ARIMA 和 GM(1, 1) 模型的中药材价格指数预测研究 [J]. 广东药科大学学报 , 2022, 38(5): 53-58.

[10] 刘天 , 姚梦雷 , 黄继贵 , 等 . 组合预测模型在丙型病毒性肝炎发病率预测中的应用 [J]. 中国疫苗和免疫 , 2018, 24(6): 674-679.

[11] 孙义 , 周陇陇 . 基于 Python 的金融时间序列 ARIMA 模型教学 [J]. 现代信息科技 , 2021, 5(10): 192-195.

[12] WEI W, JIANG J, GAO L, et al. A new hybrid model using an autoregressive integrated moving average and a general-ized regression neural network for the incidence of tuberculosis in Heng County, China[J]. The American Journal of Tropical Medicine and Hygiene, 2017, 97(3): 799–805.

[13] 巫芯宇 , 商润泽 . 以学科建设赋能我国高校发展研究——基于对 43 所高校一流学科建设方案的共现频谱分析 [J]. 西南师范大学学报 (自然科学版), 2023, 48(5): 102-110.

[14] 荆林波 , 杨佳乐 . 哲学社会科学学科建设与人才培养：成绩、问题及建议 [J]. 北京大学学报 (哲学社会科学版), 2022, 59(5): 150-158.