新疆伊犁青少年短跑成绩影响因素研究*

胡文萍[1]，林 兢[2]，张辉国[1†]

(1. 新疆大学 数学与系统科学学院，新疆 乌鲁木齐 830046；2. 新疆大学 体育教学研究部，新疆 乌鲁木齐 830046)

摘　要： 本文以新疆伊犁500位13~17岁青少年男子50米短跑及相关体质项目测试数据为基础，建立随机森林模型进行估计精度分析及变量评估的研究，发现随机森林模型对于复杂数据有较好的适应性和较高拟合效果，根据青少年短跑成绩影响因素研究，为非专业青少年短跑训练提供建议.

关键词： 新疆伊犁; 短跑; 随机森林; 影响因素分析

DOI： 10.13568/j.cnki.651094.651316. 2020.04.13.0001

中图分类号： O212　**文献标识码：** A　**文章编号：** 2096-7675(2021)04-0425-06

引文格式： 胡文萍, 林兢, 张辉国. 新疆伊犁青少年短跑成绩影响因素研究[J]. 新疆大学学报(自然科学版)(中英文), 2021, 38(4): 425-430.

英文引文格式： HU W P , LIN J, ZHANG H G. Study on the influencing factors of junior sprint results in Yili area of Xinjiang[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2021, 38(4): 425-430.

Study on the Influencing Factors of Junior Sprint Results in Yili Area of Xinjiang

HU Wenping[1], LIN Jing[2], ZHANG Huiguo[1]

(*1. School of Mathematics and Systems Sciences, Xinjiang University, Urumqi Xinjiang 830046, China;*

2. Physical Education Teaching and Research Department, Xinjiang University,Urumqi Xinjiang 830046, China)

Abstract： This paper takes the Xinjiang Yili region 500 teens aged 13 to 17 men's 50 m sprint and related physical project based on the test data, a random forest model estimation precision analysis and evaluate the variables, found that the random forest model for complex data has good adaptability and high fitting effect, according to the research to the influential factors of teenage sprinters, provide suggestions for non-professional teenagers sprint training.

Key words： Yili area of Xinjiang ; sprint; random forest; analysis of influence factors

0　引 言

短跑项目是体能竞速类的运动项目，也是田径运动中最具影响力的项目. 影响短跑成绩的因素绝非单一的某个因素，就生物学基因角度而言，影响运动基因的类型已非常多[1]. 从竞技体育的角度而言，学校体育训练起点较低，训练虽无法改变基因结构和细胞类型，但对学生短跑成绩还是有一定作用. 同时，训练的好坏或是否到位也在一定程度上影响学生是否会在今后的职业生涯中选择成为专业运动员. 因此加强专项训练及锁定重点考核对象就变得举足轻重. 近年来，研究者在研究短跑成绩影响因素时开始引入了其他学科知识，在定量研究影响因素方面作出科学尝试. 比如，陈及治[2]利用逐步回归和主成分分析研究了不同年龄组下影响男生短跑成绩的因素；张金峰等[3]利用目标离差计算法研究决定短跑成绩的相关因素，以实际经验结合相关性分析得出训练建议；王东阳等[4]根据地理信息系统，总结出青年男子短跑成绩与地理因素间的相关性，并给出了适宜陕西男子短跑项目的评价图.

* **收稿日期：** 2020-04-13

基金项目： 新疆维吾尔自治区自然科学基金(2019D01C045)；教育部人文社会科学研究规划基金项目(19YJA910007)；国家自然科学基金(11961065).

作者简介： 胡文萍(1994-)，女，硕士生，从事统计与空间数据分析研究，E-mail: 1030869695@qq.com.

† **通讯作者：** 张辉国(1978-)，男，副教授，从事空间统计学研究，E-mail: xjstat@qq.com.

根据董德龙[5]的经验, 由于人体各子系统的不确定性, 运动员的运动能力与各影响因素之间非单一化线性关系, 普通的线性模型无法对其进行有效建模, 甚至得出错误的结论. 同时, 由于现有的评价方法对多元指标体系缺乏较有科学的量化分析, 从而使短跑成绩影响因素的评价工作缺乏科学、客观依据. 因而, 我们迫切需要一种能够对个体的运动能力做出准确、客观的评价方法. 随机森林[6]作为现代非线性科学的重要科学方法, 具有极强的自适应、自学习能力, 有效地克服了传统模型识别方法的非线性、交互作用、维数灾难等问题. 随机森林在医学方面关于慢性症状要素的选择[7]、环境领域的$PM_{2.5}$预测[8]、生物方面蛋白质相互作用[9]、电力消费与经济影响分析[10]、智能技术领域的Android恶意软件检测[11]等领域应用广泛.

本文所研究的新疆伊犁青少年男子短跑体测数据具有维数大, 数据变量间存在复共线性的特性, 直接应用线性回归模型会使结果不准确. 故本文应用随机森林方法对短跑成绩的影响关系进行建模, 并得到估计模型, 获得短跑成绩相对重要的影响因素, 为非体育专业青少年短跑提供专项训练建议.

1 研究区域概况

伊犁地处新疆维吾尔自治区西部天山北部的伊犁河谷内, 因生活习惯及地理位置的原因, 该地区居住者的体质也与中国其它省份居住者有些许差异, 比如喝奶茶、吃烤肉的饮食习惯, 以及赛马、摔跤等娱乐项目. 本文仅针对身体素质、身体形态、生理机能三个方面对该地区青少年的短跑成绩及其影响因素进行研究, 为有发展潜力的新疆伊犁非专业青少年的日常短跑训练提供参考建议.

2 数据来源与研究方法

该数据由课题组提供, 数据样本共500个, 其中50米短跑成绩为因变量, 脉搏、身高等为自变量. 对各变量的观测数据做标准化处理, 以消除变量不同量纲单位和数量级对分析结果的影响. 图1为该数据中变量取值分布箱线图[12], 由图1可知，标准化后的变量整体都相对集中. 本文的插图绘制以及模型计算均使用R软件.

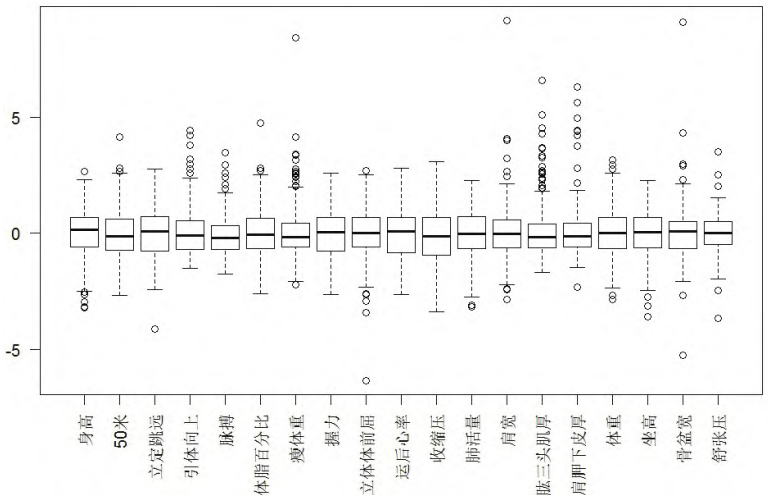

图 1 变量箱线图

Fig 1 Variable box line diagram

影响短跑成绩的因素种类多样, 本文基于身体素质(立定跳远、引体向上、握力、立体体前屈)、身体形态(身高、坐高、体重、瘦体重、骨盆宽、肩宽、体脂百分比、肩胛下皮褶厚度、肱三头肌皮褶厚度)及生理机能(肺活量、舒张压、收缩压、运后心率、脉搏)三个方面进行研究. 身体素质是影响短跑成绩的最基本因素之一; 立定跳远代表爆发力, 其在短跑运动员专项力量指标中占有很大比重[13]; 而引体向上成绩与握力反映上肢肌肉群发达程度, 一定程度上与短跑成绩存在一致性[14]; 立体体前屈代表了柔韧性, 其对短跑成绩有一定影响[15]; 身高与坐高比可构成下肢长指标(下肢长/身高×100), 该值越大, 对运动越有利; 身高与体重可构成克托莱指数(体重/身高×100), 该值反映人体发育匀称程度[16]; 骨盆宽是评价髋关节灵活度和是否具有爆发力的标准, 肩宽在一定程度上表示身体的均匀程度[17]; 体脂百分比与皮褶厚度显示了全身脂肪含量, 越小则对短跑成绩越好[16,18]; 呼吸系统和心血管系统的优劣对田径运动员而言极为重要, 良好的心肺功能可提高运动员的综合实力[19]. 在影响因

素中不仅有单向的因果关系(例如身高), 也有双向的互为因果关系(例如立定跳远能影响短跑成绩, 短跑成绩反过来也能影响立定跳远); 虽然有些因素(如身高和体重)存在线性特征, 但在影响短跑成绩方面体现了不同的侧重点, 允许它们同时存在. 通过使用随机森林模型筛选出主要影响因素, 从而找出新疆伊犁青少年短跑成绩的影响因素.

由Breiman Leo提出的随机森林（Random Forest）方法是一种机器学习方法，其在思想上结合Bagging思想和随机自变量思想[20], 整体框架以统计学理论为基础，方法上应用Bootstrap抽样方法[21]从原始样本中抽取多个样本，结合CART算法[22]在每个Bootstrap样本上构建无剪枝的决策树$h(X,\theta_k)$, $k=1,2,\cdots,K$. 在决策树模型的生成过程中，从所有影响因素中随机抽取m个影响因素，并基于均方误差MSE评价指标选出最优变量及其最优分割值进行节点分裂形成单个决策树，随后将多个决策树集成为随机森林，这种集成模型为多元非线性回归模型，其中θ_k为独立同分布的随机向量，K为随机森林中决策树棵数，随机森林以K棵树$h(X,\theta_k)$的平均值作为最终预测值. 其基本流程见图2.

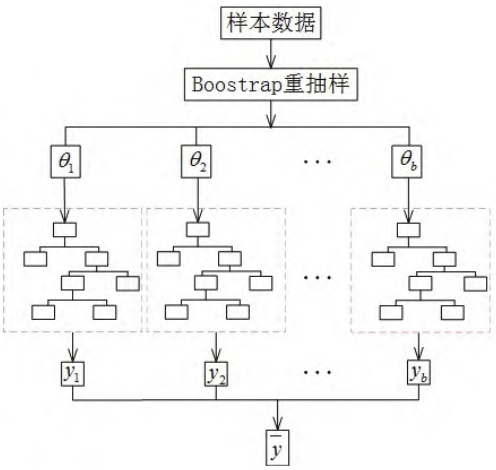

图 2　随机森林流程

Fig 2　Random forest flow chart

首先由R软件得到50米短跑成绩与各影响因素间的相关关系, 再利用随机森林方法解释自变量对50米短跑成绩的影响程度. 影响随机森林最终预测能力的参数主要有两个: 生成决策树的棵数K和随机变量分割时备选变量的个数m. 决策树数量决定了森林的大小, m确定建立决策树时用以确定各节点随机选取的自变量个数.

由图3(a)可知，决策树数量在600时误差趋于稳定, 故将数量取值定为600. 在此基础上, 采用十折交叉验证方法对子空间个数进行筛选, 图3(b)表明自变量个数取9时平均误差最小, 故将模型中自变量个数设定为9.

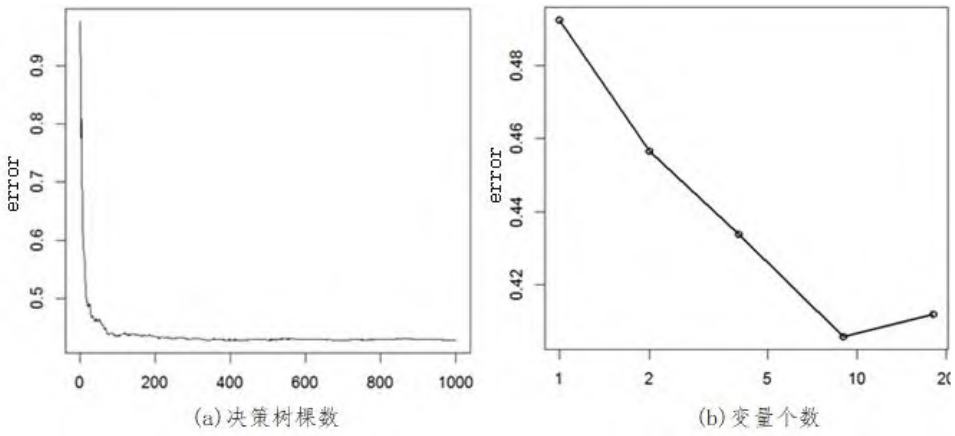

图 3　决策树数量(a)及变量个数(b)的误差曲线

Fig 3　The error curve of the number of decision trees(a)and the number of variables(b)

在R语言环境下, 采用决策树为600, 每分支备选变量个数为9的最优组合对训练集进行训练, 建立短跑成绩随机森林模型. 在此使用MSE来衡量估计模型拟合效果, 该值越小, 说明模型估计效果越好.

$$MSE=n^{-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{1}$$

式中：n为观测样本量; y_i为第i个50米短跑成绩的真实值; $\hat{y}_i$为第i个50米短跑成绩的预测值.

3 结果分析

图4为用十折交叉验证方法计算随机森林模型下的MSE. 散点数据越靠近直线$y=x$, 表明真实值和估计值越接近, MSE越小, 估计精度越高, 拟合效果越好. 由图4可知随机森林回归模型下的估计均方误差为0.152, 由此可见, 随机森林的短跑成绩估计精度较高, 可靠性强.

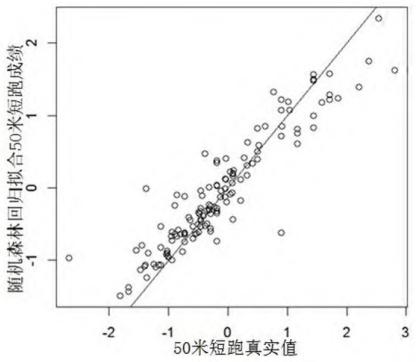

图 4 50米短跑成绩验证散点图

Fig 4 50-meter sprint performance verification scatter plot

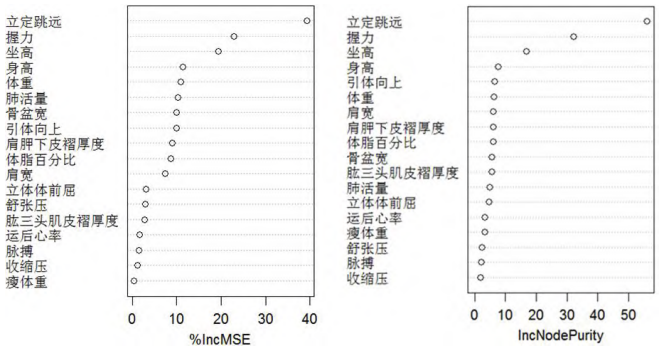

图 5 变量重要性评估图

Fig 5 Variable importance assessment graph

图5为随机森林的两种变量重要性评价指标[23]: 均方误差减小量($Inc\ \ MSE$)和模型精度减小量($In\ \ Node\ \ Purity$), 两个指标均是以通过给各变量加以不同干扰, 观察两个指标对应模型的变化, 并度量指标降低的幅度, 以此来评价变量重要性. 若指标降低幅度大, 模型的精确度上升, 则说明该变量的重要性较高. 图5中横轴为所对应准则的得分, 排序越靠前, 变量得分越多, 对短跑成绩影响越强. 以$Inc\ MSE$机制, 分析前6个影响因素对短跑成绩的影响强度, 依次为: 立定跳远40分、握力24分、坐高20分、身高13分、体重12分、肺活量11分.

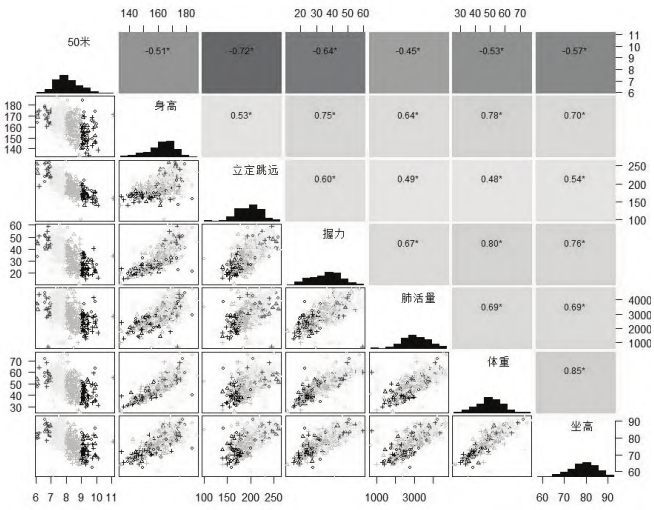

图 6 立定跳远、握力、坐高、身高、体重、肺活量对短跑成绩影响散点矩阵图

Fig 6 Scatter matrix plot of standing long jump, grip strength, sitting height, height, weight, vital capacity on sprint performance

图6为以上6个变量与50米短跑成绩的散点矩阵图, 图中左下角为6个变量间的相关关系散点图, 50米短跑的成绩以颜色深浅分别显示, 颜色越深表示50米短跑成绩越好; 中间对角线部分表示变量分布直方图; 右上角表示变量间相关系数值, 值越大表明两变量间相关程度越深. 以散点图的集中程度及相关系数来看, 立定跳远及握力与50米短跑成绩的散点更集中, 表明立定跳远及握力对50米短跑成绩影响程度越深, 结合相关系数来看, 以上两个因素对50米短跑成绩呈强负相关, 表明立定跳远及握力的值越大, 50米短跑用时越短, 成绩越好. 身高、肺活量、体重与50米短跑成绩的散点图相对较分散, 表明这些因素对50米短跑成绩影响程度相对较弱, 结合相关系数分析, 以上三个因素对50米短跑成绩呈相对较弱的负相关影响. 由变量间的相关系数可见, 这6个变量相互间也存在共线性, 故50米短跑成绩是由多个变量共同作用而决定的.

由于本文主要针对非体育专业青少年短跑运动进行研究, 故排除专业技术的考量, 从身体素质、身体形态及身体机能三个方面对短跑成绩进行分析. 身体素质方面: 立定跳远体现下肢肌群爆发力水平, 主要发力肌群为: 臀大肌、大腿屈肌肌群、小腿三头肌, 在短跑过程中下肢也是以上三种肌群的协同发力作用; 握力表现小臂上肢肌群的发达程度, 强而有力的摆臂是协调跑技术和提高跑速的重要环节, 此因素的重要作用是维持跑步时

身体动力平衡, 加快摆臂节奏及积极有力地前摆送臂有利于增大步频和保持较大步幅[24]; 身体形态方面: 身高、体重、坐高因素占比较大, 对短跑项目而言, 身高与坐高反映下肢长结构的比例, 体重反映人体身体形态, 而下肢长、下肢力量好、体脂含量低的运动员更易入选一线专业运动员[15]; 身体机能: 肺活量水平反映一个人的心肺水平及呼吸系统的潜在能力, 强有效地换气和呼气可使短跑运动更高效, 同时, 肺活量与短跑又是相辅相成的互相促进作用[18].

综上所述, 对新疆伊犁青少年男子短跑成绩的影响因素的分析中, 相较于身体形态及身体机能, 身体素质占据更多的重要性. 因此, 对于非体育专业在校学生日常的短跑训练中, 可侧重加强立定跳远及肢体协调度训练, 同时, 对于短跑训练选材方面, 可注重身高与坐高间的比例筛选.

4　结论与展望

本文在研究的青少年短跑数据存在复共线性及维数较大情况下, 引入随机森林方法进行建模分析, 发现随机森林对存在复共线性数据有一定适应性, 估计结果保持较高精度, 并较客观全面地分析了影响短跑成绩的因素, 主要结论为:

(1)利用随机森林方法对短跑成绩影响因素进行建模分析, 得到较高估计精度. 随后进行影响因素重要性排序, 重要性排序依次为: 立定跳远、握力、坐高、身高、体重、肺活量. 这些因素均对短跑呈现负向影响关系, 且越靠前因素影响程度越强.

(2)通过随机森林方法的变量排序分析, 可将筛得的6个影响因素, 围绕身体素质、身体形态、身体机能三个方面做定性分析, 这三个方面对短跑成绩影响排名依次为身体素质、身体形态和身体机能. 结合因素间相关关系图可看出, 各影响因素对短跑成绩的影响并非单一, 因素之间也存在相关性, 各因素共同作用影响短跑成绩, 比如身高协同立定跳远、坐高、体重间的相关联系同时对短跑成绩起作用.

由于短跑成绩的影响因素是复杂的, 目前大部分关于影响短跑因素的研究更多以定性分析为主, 科学定量地分析还在探索阶段. 对于本文来说, 仅仅对数据浅层定量分析, 没有进行深入的定性研究, 需要在以后的研究中加以深入探讨: (1)细化体能数据测试项目的采集, 加入技术性因素的考虑, 完善短跑成绩影响因素探究, 对后续运动员选材提供更准确全面的建议; (2)增大数据研究广度, 根据不同年龄段、性别等分类研究会使得分析结果更具针对性.

参考文献:

[1] 陈伟民, 赵广才, 彭鑫, 等. ACTN3-C1747T、ACE-I/D基因多态性在运动员选才中的应用研究[J]. 医学综述, 2013, 19(9): 1679-1681.
CHEN W M, ZHAO G C, PENG X, et al. Application of ACTN3-C1747T and ACE-I/D gene polymorphism in athletes' talent selection[J]. Medical review, 2013, 19(9): 1679-1681. (in Chinese)

[2] 陈及治. 从体质测试资料分析影响男生短跑成绩的因素[J]. 上海体育学院学报, 1990, 14(2): 62-65.
CHEN J Z. Analysis of the factors affecting the male sprint performance from the physical test data[J]. Journal of Shanghai Sports University, 1990, 14(2): 62-65. (in Chinese)

[3] 张金峰, 张武军. 运用目标计算法对决定短跑成绩相关因素的计算与分析[J]. 西安体育学院学报, 2004(S1): 48-54.
ZHANG J F, ZHANG W J. Calculation and analysis of relevant factors determining sprint performance by using objective calculation[J]. Journal of Xi'an Sports University, 2004(S1): 48-54. (in Chinese)

[4] 王东阳, 谢奇, 史兵, 等. 基于地理信息系统的青年男子短跑成绩与地理因素相关分析与评价：以陕西省为例[J]. 陕西师范大学学报(自然科学版), 2011, 39(5): 104-108.
WANG D Y, XIE Q, SHI B, et al. Analysis and evaluation of the correlation between young men's sprint performance and geographical factors based on geographic information system–a case study of Shannxi province[J]. Journal of Shannxi Normal University(natural science edition), 2011, 39(5): 104-108. (in Chinese)

[5] 董德龙. 运动员选材相关指标分析与体系构建[J]. 湖北体育科技, 2004, 23(2): 173-174.
DONG D L. Analysis and system construction of relevant indexes of athlete selection[J]. Hubei Sports Science and Technology, 2004, 23(2): 173-174. (in Chinese)

[6] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[7] 洪燕珠, 周昌乐, 张志枫, 等. 基于随机森林法的慢性疲劳证候要素特征症状的选择[J]. 中医杂志, 2010, 51(7): 634-638.
HONG Y Z, ZHOU C L, ZHANG Z F, et al. Selection of characteristics of chronic fatigue syndrome elements based on random forest method[J]. Journal of Traditional Chinese Medicine, 2010, 51(7): 634-638. (in Chinese)

[8] HU X, BELLE J H, MENG X, et al. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach[J]. Environmental Science & Technology, 2017, 51(12): 6936-6944.

[9] YAN J Q, KLEIN S J, BAR J Z. Random forest similarity for protein-protein interaction prediction from multiple sources[J]. Pacific Symposium on Biocomputing, 2005(10): 531-542.

[10] HAN Y, SHA X, GROVERSILVA E, et al. On the impact of socio-economic factors on power load forecasting[C]. Washington DC: IEEE International Conference on Big Data, 2014: 742-747.

[11] 吴非, 吴向前, 陈晓燕. 改进随机森林算法在Android恶意软件检测中的应用[J]. 新疆大学学报(自然科学版), 2017, 34(3): 322-327.
WU F, WU X Q, CHEN X Y. Application of improved random forest algorithm in detection of Android malware[J]. Journal of Xinjiang University(Natural Science Edition), 2017, 34(3): 322-327. (in Chinese)

[12] FRIGGE M, HOAGLIN D C, IGLEWICZ B. Some implementations of the boxplot[J]. The American Statistician, 1989, 43(1): 50-54.

[13] 张成. 对男子100米短跑运动员实施身体核心力量训练实验性的研究[D]. 北京: 北京体育大学, 2010.
ZHANG C. Experimental study on core strength training for men's 100-meter sprinters[D]. Beijing: Beijing Sport University, 2010. (in Chinese)

[14] 刘宏, 王永兴, 许燕. 云南省大学生体质健康与短跑的关系[J]. 中国学校卫生, 2013, 34(8):955-957.
LIU H, WANG Y X, XU Y. The relationship between college students' physical health and sprinting in Yunnan province[J]. School Health in China, 2013, 34(8): 955-957. (in Chinese)

[15] 高炫. 江苏省不同项目运动员身体形态与专项素质的特征研究: 以田径、游泳、摔跤、蹦床项目为例[D]. 南京: 南京体育学院, 2013.
GAO X. Characteristics of body shape and special qualities of athletes in different sports in Jiangsu province-a case study of track and field, swimming, wrestling and trampoline[D]. Nanjing: Nanjing Institute of Physical Education, 2013. (in Chinese)

[16] 张保顺. 上海市二线男子短跑运动员选材与育才影响因素的研究[D]. 上海: 上海师范大学, 2009.
ZHANG B S. Research on influencing factors of talent selection and talent cultivation of second-tier male sprinters in Shanghai[D]. Shanghai: Shanghai Normal University, 2009. (in Chinese)

[17] 陈杰, 巢晓春. 骨盆运动在短跑教学中的作用[J]. 南京体育学院学报(社会科学版), 1998(1): 62-65.
CHEN J, CHAO X C. The role of pelvic movement in sprint teaching[J]. Journal of Nanjing University of Physical Education(social science edition), 1998(1): 62-65. (in Chinese)

[18] 季成叶, 廖文科, 邢文华, 等. 中国11省市大学生皮褶厚度测定与体成分分析[J]. 体育科学, 2000, 20(2): 60-64.
JI C Y, LIAO W K, XING W H, et al. Thickness determination and body composition analysis of college students in 11 provinces and cities of China[J]. Sports Science, 2000, 20(2): 60-64. (in Chinese)

[19] 苏晓乾. 山西省中长跑运动员选材指标体系的研究[D]. 山西: 山西大学, 2016.
SU X Q. Research on index system of middle-distance runner selection in Shanxi province[D]. Shanxi: Shanxi University, 2016. (in Chinese)

[20] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[21] JOHNSON R W. An introduction to the bootstrap[J]. Teaching Statistics, 2001, 23(2): 49-54.

[22] BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees[M]. Monterey, CA: Wadsworth, 1984.

[23] STROBL C, BOULESTEIX A L, KNEIB T, et al. Conditional variable importance for random forests[J]. Bioinformatics, 2008, 9(1): 1-11.

[24] 张仲华. 短跑摆臂技术初探[J]. 上海体育学院学报, 1983(3): 31-32.
ZHANG Z H. Preliminary study of sprinting swing arm technique[J]. Journal of Shanghai Institute of Physical Education, 1983(3): 31-32. (in Chinese)

责任编辑：赵新科