

广西人口结构非线性演进与生育率影响因素

黄羽梅

广西师范大学数学与统计学院, 广西 桂林

收稿日期: 2026年9月14日; 录用日期: 2026年10月5日; 发布日期: 2026年10月16日

摘要

目的: 刻画广西人口结构演进的非线性特征并识别生育率的关键影响因素, 基于省级人口数据开展实证研究。方法: 利用2005~2024年《广西统计年鉴》数据, 运用光滑样条、Chow断点检验、Bootstrap与随机森林等非参数方法, 对老龄化趋势、教育结构与初婚人数关系及出生率驱动因素进行建模分析。结果: 广西老龄化率呈分段加速上升, 年均增速由2005~2014年的0.01个百分点跃升至2015~2024年的0.44个百分点, 2016年经Chow检验确认为显著结构性转折点; 光滑样条拟合优度($R^2 = 0.958$)显著优于线性模型($R^2 = 0.700$); 大专及以上学历占比与初婚人数呈倒U型关系, 拐点为5.82% (95%置信区间: 5.33%~6.49%); 老龄化率、初婚人数与高等教育占比是影响出生率的前三大因素, 相对重要性为100%、98.9%、96.0%, 随机森林袋外 $R^2 = 0.889$ 。结论: 广西人口结构演进具有显著非线性特征, 非参数方法能更准确地刻画分段变化, 低生育率是人口老化、婚姻市场萎缩与教育水平提升共同作用的结果, 可为人口监测预警与婚育政策优化提供量化依据。

关键词

人口老龄化, 光滑样条, 随机森林, 非线性建模

Nonlinear Evolution of Population Structure and Factors Influencing Fertility in Guangxi

Yumei Huang

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: September 14, 2026; accepted: October 5, 2026; published: October 16, 2026

Abstract

Objective: To depict the nonlinear characteristics of population structure evolution in Guangxi and identify the key factors influencing fertility, and to conduct an empirical study based on provincial population data. Methods: Using data covering 2005~2024 from the Guangxi Statistical Yearbook,

this paper applies nonparametric methods including smoothing splines, the *Chow* breakpoint test, the *Bootstrap* method, and random forests to model and analyze the aging trend, the relationship between education structure and the number of first marriages, and the drivers of the fertility. Results: The aging rate in Guangxi showed segmented acceleration; its average annual growth rate jumped from 0.01 percentage points during 2005~2014 to 0.44 percentage points during 2015~2024, and 2016 was confirmed by the *Chow* test as a significant structural turning point. The smoothing spline achieved a goodness of fit ($R^2 = 0.958$) significantly better than the linear model ($R^2 = 0.700$). The proportion of the population with an associate degree or above and the number of first marriages exhibited an inverted U-shaped relationship, with an inflection point of 5.82% (95% confidence interval: 5.33%~6.49%). The aging rate, the number of first marriages, and the proportion of higher education were the top three factors affecting the fertility, with relative importance of 100%, 98.9%, and 96.0%, respectively; the out-of-bag R^2 of the random forest was 0.889. Conclusions: The evolution of Guangxi's population structure has significant nonlinear characteristics, and nonparametric methods can more accurately capture segmented changes. The low fertility rate is the joint result of population aging, the shrinking marriage market, and rising educational attainment, which can provide quantitative support for population monitoring and early warning and for optimizing marriage and fertility policies.

Keywords

Population Aging, Smoothing Spline, Random Forest, Nonlinear Modeling

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景

广西壮族自治区地处中国西南部，是人口规模较大的省级行政区。2005~2024 年间，全区 65 岁及以上人口占比由 9.57%持续攀升至 13.96%，已接近深度老龄化社会标准；人口出生率则由 14.26‰降至 8.37‰，人口自然增长率由 8.17‰断崖式下降至 0.88‰，降幅高达 89.2%，低生育率已成为中国人口发展的重要特征[1]。劳动年龄人口占比下降与老年抚养比攀升形成的“双向挤压”，使人口结构变化对经济增长和社会抚养体系构成持续压力[2]，人口老龄化对生产力发展的潜在影响亦日益受到关注[3]。在此背景下，准确刻画人口结构演进规律、识别生育率变化的关键驱动因素，对广西人口政策的精准制定具有重要的现实意义。

1.2. 文献综述

人口结构变迁及其社会经济效应是人口经济学的核心议题。已有研究多借助描述性统计或线性趋势外推等方法，刻画了区域人口年龄结构由成年型向老年型转变的基本事实。韦春北等(2023)指出，广西劳动年龄人口占比持续下降与老年抚养比不断攀升形成“双向挤压”[2]；高莹(2020)运用增长核算法发现，劳动年龄人口对广西经济增长的贡献由 17.24%降至 1.81%，人口结构变化对经济增长的拉动作用正加速衰减[4]；韦海鸣、朱笑宇(2025)识别出人口出生率、死亡率与教育水平是影响乡村老龄化程度的主要因素[5]。就老龄化趋势的刻画而言，张毅磊等(2026)预测了安徽省人口老龄化趋势[6]，武荣伟等(2025)揭示了中国县市人口老龄化地域类型的演化规律[7]，肖嵩等(2022)对广西少数民族自治县的人口老龄化水平进行了分析[8]；唐星善(2020)则分析了资源丰富地区人口老龄化的经济效应[9]。

在人口分布影响因素识别方面,史莎娜等(2019)发现,桂西北喀斯特山区人口分布受地形起伏度与石漠化程度显著制约[10];方德泉(2020)进一步发现,距 POI 距离、夜间灯光指数、路网密度等社会经济因素是广西人口密度空间分异的主导因子[11];蒋慧萍(2010)较早运用多元统计模型对广西人口增长进行了分析[12]。

在婚姻与生育行为方面,周思思(2026)揭示了非正规就业对青年婚姻动机的影响及其机制[13];潘泽泉和黄燕(2026)基于“孩次递进”分析框架发现,婚姻教育匹配对生育行为具有显著影响[14];陈启航等(2026)则发现家庭禀赋与生育决策之间存在 U 型关系[15]。结合中国低生育率的趋势与特征[1],这些研究表明,婚姻与生育行为的影响机制日益复杂,线性假设可能不足以刻画其内在规律。

然而,上述研究大多采用线性回归或相关分析方法,暗含人口指标与影响因素之间呈线性关系的假设,可能掩盖真实的非线性驱动机制。近年来,非参数方法因无需预设函数形式、能够自适应数据特征的优势而逐渐被引入人口研究领域,巩永丽等(2007)基于局部线性非参数估计构建了中国人口增长率的非参数自回归模型,预测精度显著优于参数模型[16]。但现有非参数应用仍以全国尺度的时序预测为主,尚未有研究将光滑样条与随机森林两类方法系统结合并应用于广西人口结构分析。

1.3. 研究问题与思路

既有研究的线性假设可能掩盖人口结构变迁中的阈值效应与分段变化,从而造成估计偏差。以下三个问题尚待回答:一是广西老龄化率呈匀速上升还是分段加速,线性模型是否低估了近年来的老化速度;二是高等教育普及与初婚人数下降之间是线性关系还是先升后降的非线性形态;三是影响出生率的关键驱动因子及其相对重要性。针对上述问题,本文构建光滑样条与随机森林相结合的非参数分析框架,一方面弥补线性模型对分段变化识别不足的缺陷,另一方面通过变量重要性排序厘清低生育率的多维驱动结构,为广西养老服务资源配置与婚育配套政策优化提供量化依据。

2. 数据与方法

2.1. 数据来源与变量定义

本文数据来源于历年《广西统计年鉴》¹,时间跨度为 2005~2024 年,共 20 个年度观测样本。主要变量及其定义见表 1。

Table 1. Variable definitions and descriptions
表 1. 变量定义与说明

变量类别	变量名称	符号	单位	定义
年龄结构	65 岁及以上人口占比	aging_ratio	%	65 岁及以上人口/常住人口 × 100
年龄结构	15~64 岁人口占比	work_ratio	%	劳动年龄人口/常住人口 × 100
年龄结构	0~14 岁人口占比	child_ratio	%	少儿人口/常住人口 × 100
教育结构	大专及以上人口占比	edu_college	%	6 岁及以上人口中大专及以上学历占比
教育结构	高中(含中职)人口占比	edu_high	%	6 岁及以上人口中高中学历占比
婚姻生育	初婚人数	first_marriage	万人	年度初婚登记人数
婚姻生育	人口出生率	birth_rate	‰	出生人口/总人口 × 1000

注:数据来源于历年《广西统计年鉴》,经整理计算所得。

¹<http://tjj.gxzf.gov.cn/tjsj/tjnj/index.shtml>

样本期内，广西人口结构呈现“老龄化加速、少子化延续、劳动年龄人口收缩”的典型特征：65岁及以上人口占比由9.57%升至13.96%，其中2020~2024年年均增速约为全期平均增速的两倍；劳动年龄人口占比由66.67%降至65.33%；少儿人口占比由23.76%降至20.71%。教育结构持续升级，大专及以上学历人口占比由3.96%升至12.64%，增长约3.2倍。婚姻与生育方面，初婚人数由2009年峰值97.05万人降至2024年的32.38万人，人口出生率由2015年的14.05‰降至2024年的8.37‰，人口自然增长率由8.17‰降至0.88‰，2015~2024年间初婚人数与出生率的相关系数达0.90。上述趋势表明，广西人口结构正处于深刻转型期，为后文的非线性建模提供了现实基础。

2.2. 研究方法

2.2.1. 光滑样条

光滑样条通过最小化惩罚残差平方和来估计回归函数 $f(x)$ ：

$$RSS(f, \lambda) = \sum_{i=1}^N \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt \quad (1)$$

其中 $\lambda \geq 0$ 为平滑参数，用于权衡拟合度与光滑度：当 $\lambda = 0$ 时，式(1)退化为普通插值问题；当 $\lambda \rightarrow \infty$ 时， $f''(t) \rightarrow 0$ ， $f(x)$ 退化为普通线性最小二乘拟合。式(1)的解是以所有唯一观测点 x_i 为节点的自然三次样条。本文采用广义交叉验证(GCV)自动选择最优平滑参数，并以有效自由度 $df = tr(S_\lambda)$ 度量模型复杂度，其中 S_λ 为光滑矩阵。实证分析中，以年份为自变量拟合老龄化率的演进趋势，以大专及以上学历人口占比为自变量拟合初婚人数的非线性响应。

2.2.2. 随机森林

随机森林通过对回归树进行自助聚合来降低方差，属于集成学习方法。设有 B 棵回归树 $T_b(x)$ ，则集成预测为：

$$\hat{f}_{RF}(X) = \frac{1}{B} \sum_{b=1}^B T_b(X) \quad (2)$$

每棵树在分裂节点时，从全部 p 个变量中随机抽取 $m \approx \sqrt{p}$ 个候选变量，选择使节点均方误差下降最大的变量进行分裂。变量 X_j 的重要性定义为该变量在所有节点上引起的均方误差减少总量：

$$\mathcal{I}_j = \sum_{m \in \mathcal{M}(j)} \Delta MSE_m \quad (3)$$

其中 ΔMSE_m 为节点 m 分裂前后均方误差的减少量， $\mathcal{I}_j$ 值越大表明变量 X_j 对预测的贡献越强，且无需预设函数形式。袋外(OOB)误差提供泛化误差的无偏估计。本文以人口出生率为因变量，以老龄化率、少儿人口占比、劳动年龄人口占比、大专及以上学历占比、高中占比、初婚人数为预测特征，设置 $B=5000$ 、 $m=2$ 。

2.2.3. Chow 断点检验

为验证老龄化趋势中是否存在结构性转折点，本文采用Chow断点检验，该方法用于检验线性回归模型在某一特定时间点前后是否存在显著的参数变化。设定分段回归模型如下：

$$y_t = \alpha_1 + \beta_1 x_t + \varepsilon_t, \quad t \leq \tau \quad (4)$$

$$y_t = \alpha_2 + \beta_2 x_t + \varepsilon_t, \quad t > \tau \quad (5)$$

其中 τ 为假设的断点时间，原假设为 $H_0: \alpha_1 = \alpha_2, \beta_1 = \beta_2$ ，即断点前后模型参数无显著差异。对应的检验统计量为：

$$F = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n-2k)} \quad (6)$$

其中 RSS_R 为约束模型的残差平方和, RSS_U 为非约束模型的残差平方和, k 为参数个数, n 为样本量。该统计量服从自由度为 $(k, n-2k)$ 的 F 分布。若 $p < 0.05$, 则拒绝原假设, 表明该时间点为显著的结构性断点。本文以 2016 年为假设断点, 对老龄化率的时间序列趋势进行检验, 以验证 2016 年前后老龄化增速是否发生了统计上显著的跃升。

2.2.4. Bootstrap 方法

Bootstrap 方法是由 *Efron* 提出的一种重抽样技术, 用于估计统计量的抽样分布和置信区间, 无需依赖正态分布假设。其基本思想是: 从原始样本中有放回地重复抽取 B 次, 每次抽取与原样本相同样本量的样本, 对每个重抽样样本计算目标统计量, 从而获得统计量的经验分布。本文采用百分位 *Bootstrap* 法估计倒 U 型曲线拐点的置信区间。设重抽样得到的 B 个拐点估计值按升序排列为 $\theta_{(1)}^* \leq \theta_{(2)}^* \leq \dots \leq \theta_{(B)}^*$, 则其 $100(1-\alpha)\%$ 置信区间为:

$$[\hat{\theta}_{(\alpha/2 \cdot B)}^*, \hat{\theta}_{(1-\alpha/2 \cdot B)}^*] \quad (7)$$

其中 $\alpha = 0.05$, 即 95% 置信区间取重抽样分布的第 2.5 百分位数与第 97.5 百分位数。本文重复抽样 2000 次, 对倒 U 型曲线的拐点进行估计并计算其 95% 置信区间, 以验证倒 U 型关系的统计可靠性。

3. 实证分析

3.1. 老龄化趋势的非线性特征

以年份为自变量、65 岁及以上人口占比为因变量, 拟合光滑样条模型。基于广义交叉验证选择的有效自由度为 $df = 5.41$, 调整后 $R^2 = 0.958$, 模型解释了老龄化率 95.8% 的变异, 拟合效果优良; 相比之下, 普通线性回归模型的 R^2 仅为 0.700, 光滑样条的解释力提升了约 36.9%, 充分印证了老龄化趋势非线性特征的存在。拟合结果如图 1 所示。

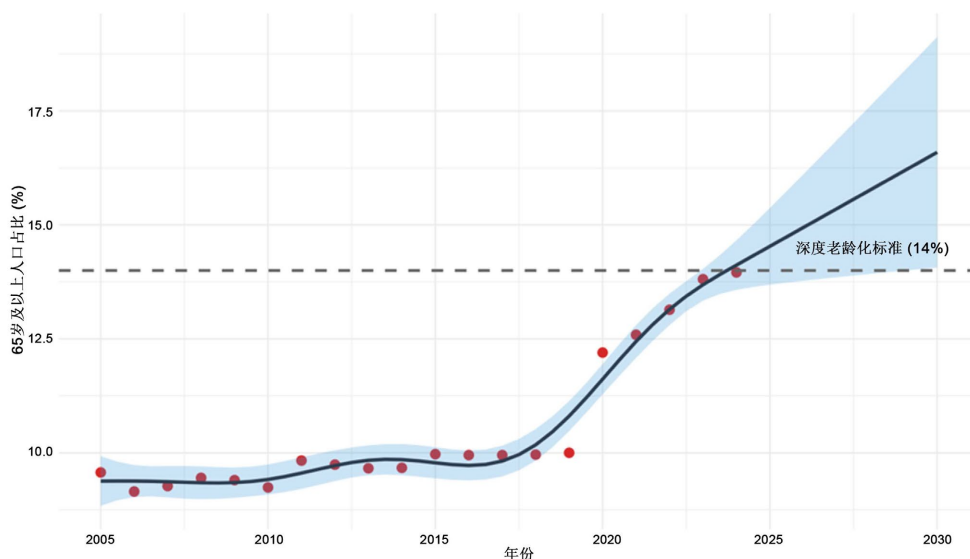


Figure 1. Smoothing spline fitting of the aging trend in Guangxi

图 1. 光滑样条拟合广西老龄化趋势

图 1 呈现出以下三个特征。第一，非线性加速趋势。老龄化率在 2005~2014 年间增长平缓，年均约 0.01 个百分点，2015 年起年均增速升至约 0.44 个百分点，呈非线性加速态势；模型对 2024 年的拟合值为 14.04%，与实际值 13.96% 高度吻合。第二，深度老龄化时点确认。实际数据显示，广西 65 岁及以上人口占比 2024 年已达 13.96%，距 14% 的深度老龄化门槛仅差 0.04 个百分点；光滑样条模型的拟合值为 14.04%，已越过该门槛。综合实际值与模型拟合值可以判断，广西正处于进入深度老龄化社会的临界阶段。第三，模型优越性。光滑样条的拟合优度显著优于线性模型，若采用线性模型，将严重低估近年来的老龄化速度。

为验证 2016 年是否为结构性转折点，本文进一步采用 Chow 断点检验。结果显示，检验统计量对应的 $p < 0.01$ ，确认 2016 年为显著结构性断点。从预测误差角度看，线性模型在 2005~2015 年间的平均绝对误差约为 0.48 个百分点，2016~2024 年间扩大至 1.06 个百分点，整体为 0.74 个百分点；光滑样条模型在两个时期的平均绝对误差分别为 0.18 与 0.24 个百分点，整体仅为 0.21 个百分点，预测精度约为线性模型的 3.5 倍，进一步说明非线性建模在处理分段变化趋势时的显著优势。

3.2. 教育结构与初婚人数的非线性关系

用光滑样条拟合大专及以上学历人口占比与初婚人数之间的关系，拟合曲线呈明显的倒 U 型，拟合优度 $R^2 = 0.884$ ，而普通线性回归的 R^2 仅为 0.635，光滑样条的解释力提升了约 39.2%，进一步证实了非线性关系的存在。该倒 U 型曲线可分解为三个阶段：第一阶段由收入效应主导(教育占比 3.96%~6.5%)，高等教育初期的普及提升了受教育人口的收入水平，婚姻的经济基础得到改善，初婚人数由 56.53 万人持续上升至 2009 年前后的 97.05 万人；第二阶段由替代效应主导(教育占比 6.5%~9.2%)，学制延长使毕业与参加工作年龄推迟，结婚年龄相应后移，初婚人数由峰值降至 2019 年前后的约 50 万人；第三阶段为补偿效应显现期(教育占比大于 9.2%)，高学历人群虽普遍晚婚，但收入预期更高、职业更稳定，最终进入婚姻的比例并不低，初婚人数降速明显放缓。

为精确估计倒 U 型曲线的拐点位置，本文采用 Bootstrap 方法估计拐点的 95% 置信区间。结果显示，拐点精确估计为教育占比 5.82%，95% 置信区间为 5.33%~6.49%，对应初婚人数约 95 万人。实际数据中，2009~2010 年教育占比处于 5.06%~6.58% 区间，初婚人数均达 96 万人以上，拐点估计值 5.82% 恰落于该区间，与观测数据高度吻合，验证了倒 U 型关系的统计可靠性。光滑样条拟合曲线及其 95% 置信带如图 2 所示。

图 2 展示了光滑样条拟合的大专及以上学历人口占比与初婚人数关系曲线及其 95% 置信带。从拟合效果来看，光滑样条模型在各个教育水平区间内的表现均优于线性模型：在教育占比低于 5% 的低水平区间，光滑样条捕捉到初婚人数随教育水平提高而上升的趋势，而线性模型则错误地预测初婚人数持续下降；在教育占比处于 6%~8% 的中等水平区间，光滑样条准确拟合了初婚人数达峰后开始回落的转折过程；在教育占比超过 10% 的高水平区间，光滑样条体现出降速趋缓的特征，线性模型则仍按固定斜率预测下降，导致预测值持续偏离实际观测值。作为对照，普通线性回归将二者的关系拟合为一条单调递减直线：大专及以上学历人口占比每提高 1 个百分点，初婚人数平均减少约 5.87 万人，模型拟合优度 R^2 仅为 0.635。该模型在统计上显著，但只能给出单调递减的结论：按该模型推算，教育水平最低的 2005 年应对应初婚人数最高的年份，但该年初婚人数实际仅为 56.53 万人，远低于 2009 年的 97.05 万人；线性回归对 2009 年的拟合值比实际值低约 10 万人，对 2024 年的拟合值又比实际值高约 10 万人。这一系统性偏差源于线性模型将倒 U 型关系强行压缩为直线，从而掩盖了真实的非线性机制。

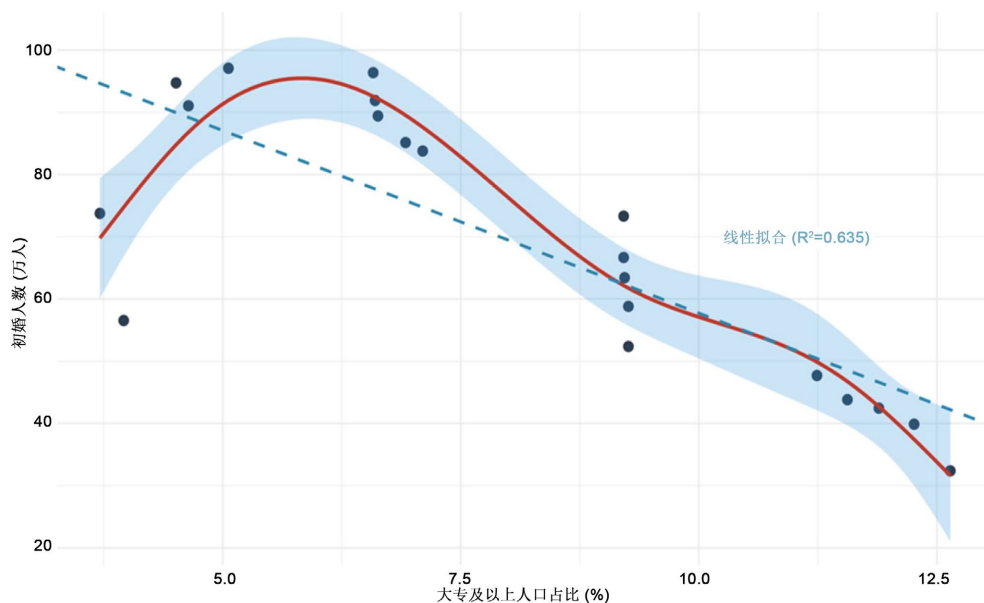


Figure 2. Nonlinear relationship between education structure and the number of first marriages
图 2. 教育结构与初婚人数的非线性关系

3.3. 出生率驱动因素的随机森林分析

以人口出生率(‰)为因变量,以老龄化率、少儿人口占比、劳动年龄人口占比、大专及以上占比、高中占比、初婚人数 6 个变量为预测变量,构建随机森林模型。模型的袋外误差率为 0.8435‰,袋外 $R^2 = 0.889$,拟合效果优良。作为参照,线性回归模型在全体样本上的 R^2 为 0.965,但由于样本量有限,线性回归可能存在过拟合现象,其较高的 R^2 未必反映真实泛化能力;随机森林的袋外 R^2 则提供了更稳健的泛化能力估计。各自变量的相对重要性排序结果见表 2。

Table 2. Variable importance ranking of the random forest

表 2. 随机森林变量重要性排序

排名	变量	原始重要性得分	相对权重/%
1	老龄化率	33.58	100.0
2	初婚人数	33.23	98.9
3	大专及以上人口占比	32.24	96.0
4	劳动年龄人口占比	29.95	89.2
5	高中人口占比	5.76	17.1
6	少儿人口占比	3.83	11.4

注:数据来源于随机森林模型输出(ntree = 5000, mtry = 2);相对权重以最高得分为 100 计算。高中占比与少儿占比得分接近,两者相对排序在不同随机种子下不稳定,仅作参考。

表 2 展示了随机森林变量重要性排序的结果。排名前三的变量得分差距均在 4 个百分点以内:老龄化率以 33.58 分居首,初婚人数以 33.23 分居次,大专及以上占比以 32.24 分位列第三;劳动年龄人口占比得分为 29.95,仅落后前三名约 2~4 个百分点;高中占比与少儿占比得分仅 5.76 和 3.83,与前四名不在同一量级,且两者得分接近,其相对排序在不同随机种子下并不稳定,不宜据此推断两者对出生率影

响的强弱。这一排序表明，老龄化率、初婚人数与高等教育占比是影响广西出生率的三个核心变量。其中，老龄化率反映人口年龄结构的顶部压力，初婚人数反映婚姻市场的人口流量，高等教育占比反映生育主体行为模式的转变，三者解释力高度接近，说明低生育率是人口结构老化、婚姻市场萎缩与教育水平提升三股力量共同作用的结果。

从具体机制来看，老龄化率对出生率的影响在 11%附近存在加速区间：老龄化率在 9.5%~11%期间预测出生率仅由约 7.1%缓慢降至约 6.8%；11%~11.5%期间出现陡降，预测出生率降至约 5.9%；此后继续缓慢下降，13%时约为 5.3%并趋于平稳。当前广西老龄化率已达 13.96%，已越过陡降区间、进入下降趋缓阶段。2015~2024 年间初婚人数与出生率的相关系数达 0.90，初婚人数由 97 万人降至 32 万人的过程中，出生率由 14.05‰降至 8.37‰，整体上看，初婚人数每减少 10 万人，出生率约下降 0.9 个千分点。大专及以上学历占比的重要性得分与老龄化率仅差 4 个百分点，教育通过推迟初婚年龄、改变生育观念、提高生育机会成本三条渠道共同压低出生率，其叠加效果与人口老化的影响相当。三个核心变量对出生率的局部依赖关系如图 3 所示。

图 3 展示了随机森林模型中三个最重要预测变量对出生率的边际效应，纵轴为预测出生率，横轴为各变量取值。三条曲线均呈现明显的非线性形态：老龄化率在 11%~11.5%附近下降斜率显著加剧、之后再度趋缓，表明老龄化对生育率的抑制存在阶段性加速效应；初婚人数与出生率在 40 万~60 万人区间内呈明显正相关，60 万人以上趋于饱和，印证了婚姻市场活跃度对出生率的直接影响；大专及以上学历占比与出生率的关系呈先平、后降、再缓的三段式形态。在老龄化率维度，当老龄化率低于 11%时，预测出生率约在 6.8‰~7.1‰之间小幅波动，11%~11.5%区间内出现陡降、降至约 5.9‰，说明该阶段老龄化对生育率的边际抑制效应明显增强；在高等教育占比维度，教育占比低于 10%时预测出生率保持相对稳定，10%~12%区间内持续下降，超过 12%后下降趋势趋于平缓。这一形态与教育影响婚育行为的分段特征基本吻合，交叉验证了教育影响的阶段性。

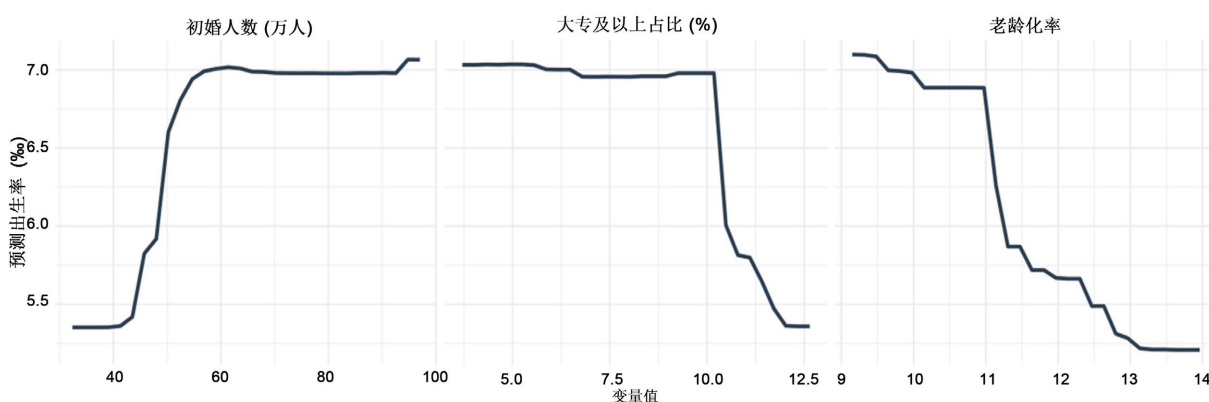


Figure 3. Partial dependence plots of the top three important variables in the random forest

图 3. 随机森林前三重要变量的局部依赖图

4. 结果讨论

本文的实证结果表明，广西人口结构演进具有显著的非线性特征。第一，老龄化率的“分段加速”特征具有明确的政策含义。2005~2014 年间广西老龄化年均增速仅约 0.01 个百分点，2015 年后跃升至约 0.44 个百分点，Chow 断点检验确认 2016 年前后发生了统计上显著的结构变化。线性模型只能将整个过程描述为均匀上升的直线，从而系统性低估了近期老化速度；基于线性趋势外推的养老资源规划可能滞后于实际需求，养老服务体系供给压力将加速显现，养老产业与长期照护体系建设亟需提前布局[17]。

第二，教育结构与初婚人数的倒 U 型关系，揭示了高等教育普及与婚姻市场之间的动态权衡。在收入效应主导阶段，教育发展有助于提升婚姻的经济基础；在替代效应主导阶段，学制延长推迟了结婚时点；而在补偿效应显现阶段，高学历群体的收入优势部分对冲了晚婚的影响。这一三阶段机制与婚姻教育匹配对生育行为的影响[14]、家庭禀赋与生育决策的非线性关系[15]等研究结论相互印证，说明教育对婚育行为的影响并非简单的单调抑制作用，相关政策设计需考虑其阶段性差异。同时，非正规就业等劳动力市场因素对青年婚姻动机的影响[13]，也提示就业保障与婚育支持政策需要协同。

第三，低生育率是多维因素共同作用的结果。老龄化率、初婚人数与高等教育占比三者重要性高度接近，共同构成影响出生率的“三足鼎立”格局：人口老化压缩了生育基数，婚姻市场萎缩减少了生育人口，教育水平提升则改变了生育主体的行为模式。在区域经济结构持续调整的背景下[18][19]，广西低生育率问题的治理需要人口政策、教育政策与就业政策统筹推进。

第四，从方法论角度看，线性回归在全体样本上虽然具有较高的拟合优度，但存在过拟合风险；随机森林通过袋外误差提供了更稳健的泛化能力估计，光滑样条则能够自适应地捕捉分段变化。需要说明的是，本文基于 20 个年度观测样本，样本量相对有限，非参数方法的估计精度可能受到一定制约；未来可结合更长的时间序列数据并引入更多社会经济变量，进一步检验结论的稳健性。

5. 结论与政策建议

5.1. 主要结论

本文基于广西 2005~2024 年人口统计数据，运用光滑样条与随机森林两种非参数统计学习方法，得出以下主要结论。第一，广西老龄化率呈分段加速增长，年均增速由 2005~2014 年的约 0.01 个百分点跃升至 2015~2024 年的约 0.44 个百分点，*Chow* 断点检验确认 2016 年为显著结构性转折点；光滑样条的拟合优度 ($R^2=0.958$) 显著优于线性模型 ($R^2=0.700$)，普通线性模型的平均绝对误差约为光滑样条模型的 3.5 倍。第二，高等教育占比与初婚人数呈倒 U 型曲线而非直线关系，光滑样条拟合优度达 0.884，远优于线性回归的 0.635；*Bootstrap* 方法估计拐点为教育占比 5.82%，95% 置信区间为 5.33%~6.49%。第三，老龄化率、初婚人数与高等教育占比对出生率的影响力高度接近，相对重要性分别为 100%、98.9% 与 96.0%，共同构成驱动广西低生育率的三个重要因素；随机森林袋外 R^2 为 0.889，在小样本下提供了比线性回归更稳健的泛化能力估计。

5.2. 政策建议

基于上述结论，提出以下政策建议。第一，建立基于非参数方法的老龄化监测预警系统。鉴于广西老龄化率已逼近 14% 的深度老龄化门槛、即将进入深度老龄化社会，建议统计部门采用光滑样条等非参数方法实时跟踪老龄化速度，当检测到斜率显著变化时，如 2016 年 *Chow* 检验所识别的结构性断点，应及时预警，为养老资源布局提供决策支持。第二，精准识别婚姻与生育政策的协同干预窗口。初婚人数的重要性与老龄化率几乎持平，说明鼓励结婚与鼓励生育需协同推进；针对拐点附近至 9.2% 的婚姻抑制区间，可考虑提供住房保障、托育服务、税收优惠等综合性措施。初婚人数由 97 万人降至 32 万人是出生率断崖式下降的重要解释因素，恢复婚姻市场活力应成为人口政策的重要着力点。第三，重视教育结构与人口政策的协同。在推动高等教育发展的同时，应关注其对婚育行为的间接影响，通过完善社会支持体系缓解教育对婚育的抑制作用。

致 谢

感谢广西师范大学数学与统计学院老师在论文写作过程中给予的悉心指导；本文数据来源于历年《广西统计年鉴》，谨向编制单位致以诚挚谢意。

参考文献

- [1] 陈卫, 欧阳柏慧, 刘金菊. 中国的低生育率: 趋势和特征[J]. 山东大学学报(哲学社会科学版), 2026(4): 199-213.
- [2] 韦春北, 刘红宇, 杨冬妮, 等. 广西人口结构演变与产业转型升级研究[M]. 成都: 西南财经大学出版社, 2023.
- [3] 裴守兴, 孙鑫, 倪希望. 人口老龄化对新质生产力的影响探究——基于产业结构高级化的中介效应[J]. 中国市场, 2026(21): 1-4.
- [4] 高莹. 广西人口变动对经济增长的影响研究[D]: [硕士学位论文]. 桂林: 广西师范大学, 2020.
- [5] 韦海鸣, 朱笑宇. 边疆民族地区乡村人口老龄化: 现状, 影响和对策——以广西壮族自治区为例[J]. 南宁师范大学学报(哲学社会科学版), 2025, 46(4): 26-38.
- [6] 张毅磊, 韦伟, 李欣, 等. 安徽省人口老龄化趋势预测及影响分析——基于人口变化模型[J]. 统计科学与实践, 2026(5): 18-22.
- [7] 武荣伟, 张佳丽, 周亮, 等. 人口流动视角下中国县市人口老龄化地域类型演化及影响因素[J]. 地理科学进展, 2025, 44(6): 1213-1226.
- [8] 肖嵩, 莫春宝, 何开连, 等. 广西某少数民族自治县人口老龄化水平分析及预测[J]. 大众科技, 2022, 24(10): 172-176.
- [9] 唐星善. 资源丰富地区人口老龄化经济效应分析——以广西为例[J]. 中国金属通报, 2020(17): 118-120.
- [10] 史莎娜, 谢炳庚, 胡宝清, 等. 桂西北喀斯特山区人口分布特征及其与自然因素的关系[J]. 地理科学, 2019, 39(9): 1484-1495.
- [11] 方德泉. 基于随机森林模型与地理探测器的广西人口分布影响因素研究[D]: [硕士学位论文]. 南宁: 南宁师范大学, 2020.
- [12] 蒋慧萍. 基于多元统计模型的广西人口增长分析[J]. 企业科技与发展, 2010(10): 190-192.
- [13] 周思思. 情感抑或理性: 非正规就业对青年婚姻动机的影响及其机制[J]. 青年研究, 2026(4): 133-154.
- [14] 潘泽泉, 黄燕. 婚姻教育匹配的生育效应研究——基于“孩次递进”分析框架的实证发现[J]. 中国人口科学, 2026, 40(4): 21-37.
- [15] 陈启航, 杨雪燕, 李婉鑫, 等. 夹心层的生育困境: 家庭禀赋与生育决策的 U 型关系[J]. 青年研究, 2026(4): 199-220.
- [16] 巩永丽, 张德生, 武新乾. 人口增长率的非参数自回归预测模型[J]. 数理统计与管理, 2007(5): 759-764.
- [17] 李丹. “十五五”规划背景下桂台养老产业创新融合与面向东盟协同发展路径研究[J]. 商展经济, 2026(14): 38-41.
- [18] 谭娴, 黄春晓. 基于统计年鉴数据的广西区域经济空间格局演变研究[J]. 智能城市, 2025, 11(11): 5-9.
- [19] 莫宸媛, 谢春美, 唐泽华. 国内外经济政策对广西经济的影响及经济开放背景下的应对措施[J]. 商场现代化, 2025(22): 128-132.

附 录

Table A1. Summary of key analytical results
表 A1. 关键分析结果汇总

分析项目	指标	数值
老龄化趋势	光滑样条拟合优度 R^2	0.958
	线性模型 R^2	0.700
	<i>Chow</i> 断点检验	$p < 0.01$, 2016 年为结构性断点
	2024 年拟合值	14.04%
教育结构与初婚人数	光滑样条拟合优度 R^2	0.884
	线性回归 R^2	0.635
	拐点估计值	5.82%
	95%置信区间	[5.33%, 6.49%]
随机森林分析	袋外 R^2	0.889
	袋外误差率	0.8435%
模型对比	线性回归全体样本 R^2	0.965 (存在过拟合风险)
	随机森林袋外 R^2	0.889 (代表真实泛化能力)