附录：中国 2026—2100 年人口预测

陈磊 陈松蹊 何婧

附录 1：对 2000—2023 年总和生育率对数 $\ln f_t$ 建立 AR(1)模型

选取 2000—2023 年的总和生育率对数 $\ln f_t$ 的时间序列，逐步验证如下步骤：

（1） 平稳性检验；

（2） 利用 BIC 信息准则选择 ARMA(p,q)模型的参数 p 和 q；

（3） 参数估计及显著性检验；

（4） 检验残差是否为白噪声。

首先，文本使用 ADF 和 KPSS 两种检验方法，对我国 2000—2023 年总和生育率的对数序列 $y_t = \ln f_t$ 进行平稳性检验，工具使用 python 的软件包 statsmodels.tsa。检验结果如下：

附表 1　平稳性检验结果

ADF 检验
（1）模型：$\Delta y_t = \alpha + \pi y_{t-1} + \sum_{j=1}^{k} \psi_j \Delta y_{t-j} + \varepsilon_t$. $H_0: \pi = 0$（存在单位根） vs $H_1: \pi < 0$ 结果：检验统计量 = -2.2214，p 值 = 0.1986。
（2）模型：$\Delta y_t = \alpha + \beta t + \pi y_{t-1} + \sum_{j=1}^{k} \psi_j \Delta y_{t-j} + \varepsilon_t$. $H_0: \pi = 0$（存在单位根） vs $H_1: \pi < 0$ 结果：检验统计量 = -4.3452，p 值 = 0.0027。
KPSS 检验
（1）模型：$y_t = \alpha + r_t + \varepsilon_t$，$r_t = r_{t-1} + u_t$. $H_0: \mathrm{Var}(u_t) = 0$ vs $H_1: \mathrm{Var}(u_t) > 0$（存在单位根） 结果：检验统计量 = 0.2476，p 值 > 0.10 。
（2）模型：$y_t = \alpha + \beta t + r_t + \varepsilon_t$，$r_t = r_{t-1} + u_t$. $H_0: \mathrm{Var}(u_t) = 0$ vs $H_1: \mathrm{Var}(u_t) > 0$（存在单位根） 结果：检验统计量 = 0.0612，p 值 > 0.10。

综上，ADF 检验和 KPSS 检验表明时间序列 $\ln f_t$ 不存在单位根，是趋势平稳的，可能存在与时间相关的线性趋势。进一步，对时间序列 $\ln f_t$ 建立如下带有时间趋势项的ARMA(p,q)模型，

$$\left(1 - \rho_1 L - \cdots - \rho_p L^p\right)\left(\ln f_t - \mu_0 - \beta t\right) = \left(1 + \theta_1 L + \cdots + \theta_q L^q\right)\varepsilon_t\,,\ \varepsilon_t \sim N(0, \sigma^2) \tag{A1}$$

其中，L是滞后算子。利用 BIC 准则选取参数 p 和 q，结果如下，

附表 2　基于 BIC 准则选择式（A1）模型滞后阶结果

BIC 准则（前 5）		
p	q	BIC 值
1	0	-30.0
2	1	-29.3
0	1	-27.5

2	0	-27.1
1	1	-26.0

综上，基于 BIC 准则选取的模型为 AR(1)，即

$$(1-\rho L)(\ln f_t-\mu_0-\beta t)=\varepsilon_t,\ \varepsilon_t\sim N(0,\sigma^2). \tag{A2}$$

在式(A2)模型设定下，参数估计结果如下，

附表 3　模型式（A2）参数估计结果

	系数	标准差	z值	p值	95%置信区间
μ_0	0.4399	0.242	1.819	0.069	[-0.034, 0.914]
β	-0.0138	0.014	-0.989	0.323	[-0.041, 0.014]
ρ	0.5656**	0.212	2.662	0.008	[0.149, 0.982]
σ^2	0.0092***	0.003	3.431	0.001	[0.004, 0.014]

由于线性趋势项参数 μ_0 和 β 均不显著，故去除时间趋势 βt，调整为如下式(A3)模型

$$(1-\rho L)(\ln f_t-\mu_0)=\varepsilon_t,\ \varepsilon_t\sim N(0,\sigma^2). \tag{A3}$$

模型式(A3)参数估计结果如下，

附表 4　模型式（A3）参数估计结果

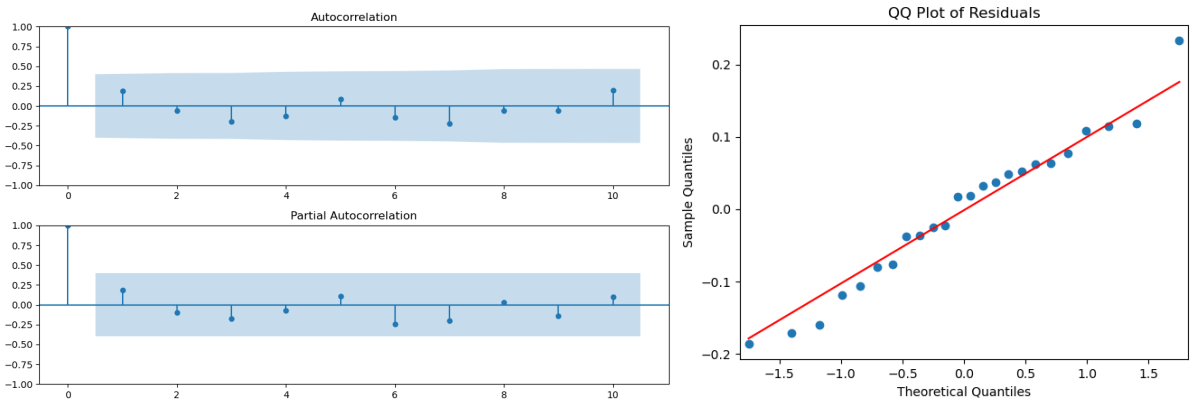
	系数	标准差	z值	p值	95%置信区间
μ_0	0.2370***	0.068	3.509	0.000	[0.105, 0.369]
ρ	0.6674**	0.238	2.801	0.005	[0.200, 1.134]
σ^2	0.0106**	0.004	2.950	0.003	[0.004, 0.018]

模型系数全部显著。

进一步对残差进行 Ljung-Box 白噪声检验（附表 5），画出残差的自相关和偏相关系数图（附图 1 左）、残差的 QQ 图（附图 1 右）。结果均支持：残差为白噪声。具体如下，

附表 5　对模型式（A3）残差的 Ljung-Box 检验

m	统计量$Q(m)$	p值
10	7.647	0.663
20	14.494	0.805



附图 1　残差的自相关和偏自相关图（左），以及 QQ 图（右）

综上，对于 2000—2023 年的总和生育率的对数序列 $\ln f_t$，最终建立如下 AR(1)模型

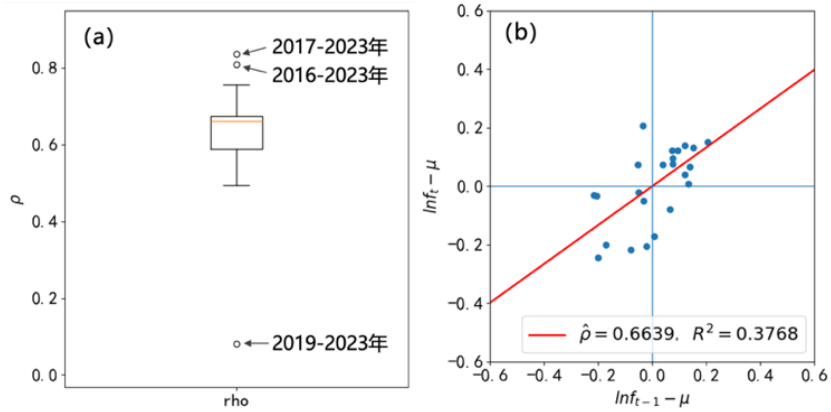
$$\ln f_t-\mu_0=\rho(\ln f_{t-1}-\mu_0)+\varepsilon_t,\ \varepsilon_t\sim N(0,\sigma^2).$$

参数估计如附表 4 所示。

附录 2：不同时间窗口选择下，基于方法(B)对模型式(8)中参数 ρ 估计的稳健性评估

附表 6　不同时间窗口选择下，模型式(8)中参数 ρ 的估计值、RMSE 及 R^2

窗口 起始年份	窗口 终点年份	样本量	$\hat{\rho}$ 估计值	RMSE	R^2
2000	2023	24	0.6639	0.1056	0.3768
2001		23	0.6820	0.1055	0.3984
2002		22	0.6746	0.1078	0.3916
2003		21	0.6618	0.1100	0.3818
2004		20	0.6391	0.1118	0.3646
2005		19	0.6502	0.1145	0.3673
2006		18	0.6381	0.1171	0.3573
2007		17	0.6073	0.1186	0.3382
2008		16	0.5576	0.1202	0.2958
2009		15	0.5455	0.1239	0.2682
2010		14	0.5890	0.1257	0.3013
2011		13	0.5718	0.1234	0.3138
2012		12	0.6652	0.1239	0.3646
2013		11	0.6656	0.1305	0.3656
2014		10	0.6676	0.1381	0.3671
2015		9	0.6831	0.1314	0.4347
2016		8	0.8102	0.1326	0.4978
2017		7	0.8366	0.997	0.6902
2018		6	0.7561	0.1048	0.6007
2019		5	0.4954	0.1006	0.3908
2020		4	0.0809	0.0744	0.0164



附图 2 (a)不同时间窗口下参数 ρ 估计值的箱线图；(b)时间窗口选为 2000—2023 年时，模型式(8)对应的散点图$(\ln f_{t-1}-\hat{\mu}, \ln f_t-\hat{\mu})$及 OLS 拟合曲线

附录 3：不同时间窗口选择下，模型式(8)和式(11)中参数 σ_0^* 和 μ_0的估计稳健性评估

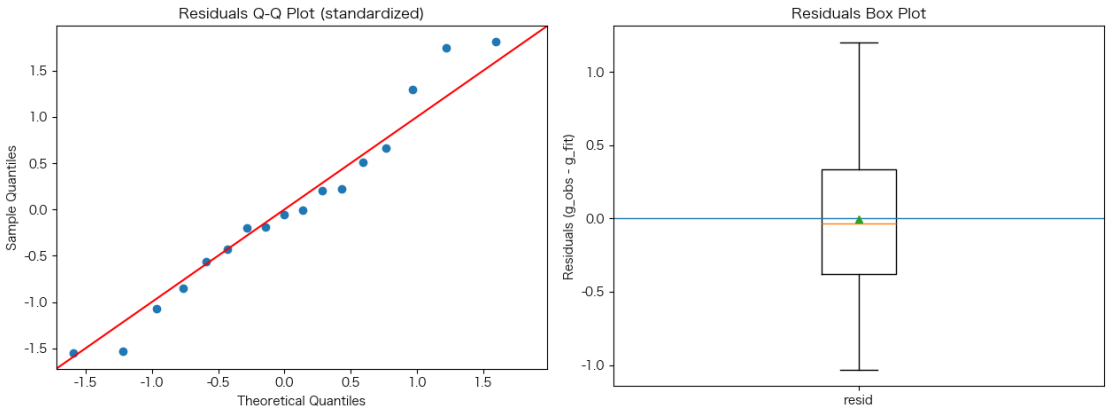
附表 7　不同时间窗口选择下，模型式(8)和式(11)中参数 σ_0^* 和 μ_0的估计值、对应总和生育率的 90%置信区间及覆盖历史观测的比例

窗口 起始年份	窗口 终点年份	$\hat{\sigma}_0^*$ 估计值	$\hat{\mu}_0$ 估计值	90%置信区间	覆盖数/样本量	覆盖率
2000	2023	0.1662	0.0239	[0.71, 1.48]	22 / 24	0.9167
2001					21 / 23	0.9130
2002					20 / 22	0.9091
2003					19 / 21	0.9048
2004					18 / 20	0.9000
2005		0.1722	0.0221	[0.70, 1.50]	18 / 19	0.9474
2006					17 / 18	0.9444
2007					16 / 17	0.9412
2008					15 /16	0.9375
2009					14 / 15	0.9333
2010					13 / 14	0.9286
2011					12 / 13	0.9231
2012					11 / 12	0.9167
2013					10 / 11	0.9091
2014					9 / 10	0.9000
2015		0.2017	0.0121	[0.65, 1.58]	9 / 9	1
2016					8 / 8	1
2017					7 / 7	1
2018		0.1722	0.0221	[0.70, 1.50]	6 / 6	1
2019		0.1628	0.0249	[0.72, 1.47]	5 / 5	1
2020		0.0994	0.0399	[0.83, 1.29]	4 / 4	1

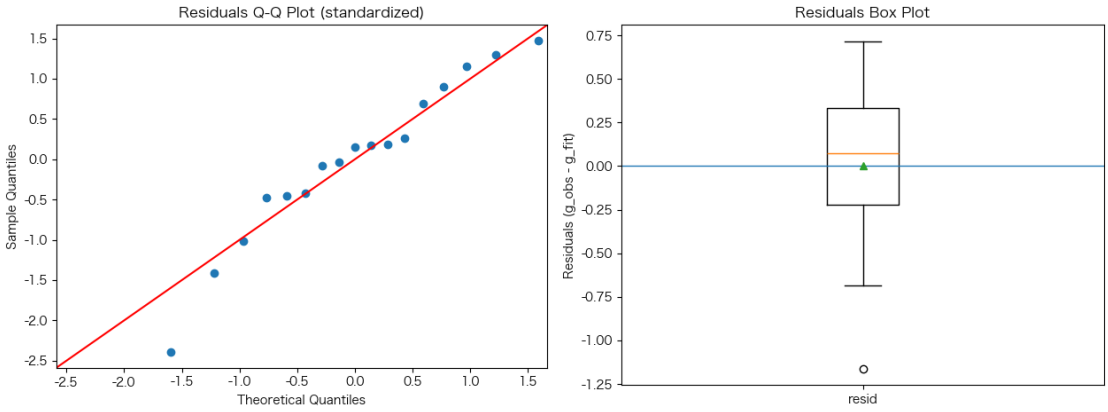
附录 4：式(18)和式(19)模型中，男性、女性双逻辑函数的参数估计和残差诊断

附表 8　男性和女性预期寿命预测模型的双逻辑函数参数估计结果

	男性	女性
d	3.62	3.58
Δ_1	35.0	16.7
Δ_2	97.9	93.6
Δ_3	1.00	3.23
Δ_4	65.6	67.9



附图 3　男性预期寿命模型中双逻辑函数拟合的残差 QQ 图（左）和箱线图（右，其中三角形为均值，橙色横线为中位数）



附图 4　女性预期寿命模型中双逻辑函数拟合的残差 QQ 图（左）和箱线图（右，其中三角形为均值，橙色横线为中位数）

附录 5：基于插值方法构造预期寿命为 e_0^* 的模型生命表

设有两张出生时预期寿命分别为 e_0^1 和 e_0^2 （均为整数，且不失一般性令 $e_0^1 < e_0^2$）的模型生命表，在年龄区间$[x, x+1)$上的死亡概率分别为 q_x^1 和 q_x^2。给定目标预期寿命 e_0^*（满足 $e_0^1 < e_0^* < e_0^2$），我们借鉴美国人口普查局 PAS 文档（Arriaga E E，2012）推荐的方法：以两张生命表的死亡概率为基础，在对数尺度下向内插值，从而构造预期寿命恰好为 e_0^* 的目标生命表。

具体而言，令权重$w \in [0,1]$，则目标生命表在年龄区间$[x, x+1)$上的死亡概率 q_x^* 满足

$$\log(q_x^*) = (1-w)\log(q_x^1) + w\log(q_x^2),$$

等价地，

$$q_x^* = (q_x^1)^{1-w}(q_x^2)^w .$$

对于目标生命表中年龄组$[x, x+1)$的分离因子a_x^*（刻画了该年龄区间死亡人数的平均存活年数），本文采用与死亡概率一致的权重 w 进行线性插值

$$a_x^* = (1-w)a_x^1 + wa_x^2,$$

其中，a_x^1 和 a_x^2 分别是模型生命表 1 和模型生命表 2 在年龄区间$[x, x+1)$的分离因子。给定权重 w 后，由插值得到的q_x^* 和 a_x^* 即可生成相应的目标生命表，并据此计算出生时的预期寿命 $e_0(w)$。一个直观的权重是

$$w_0 = \frac{e_0^* - e_0^1}{e_0^2 - e_0^1}.$$

然而，由于预期寿命 e_0 是所有年龄区间死亡概率 q_x 的非线性函数，权重 w_0 一般并不能保证 $e_0(w_0)$ 严格等于目标预期寿命 e_0^*。美国人口普查局 PAS 文档建议通过方程求解来得到权重 w，使得由插值得到生命表的预期寿命恰好达到 e_0^*。为此，定义

$$f(w) = e_0(w) - e_0^* .$$

注意到 $e_0(w)$关于 w 单调，并满足 $e_0(0) = e_0^1$ 和 $e_0(1) = e_0^2$。当权重 w 趋向于 0 时，预期寿命 $e_0(w)$ 趋向于 e_0^1；反之当权重 w 趋向于 1 时，预期寿命 $e_0(w)$ 趋向于 e_0^2。因此，可以通过二分法求解方程 $f(w) = 0$，得到精确的权重 w^*。

具体算法流程如下

1. **初始化**。记 $w_L = 0, w_U = 1$，令预期寿命 $e_0^L = e_0(w_L)$ 和 $e_0^U = e_0(w_U)$；
2. **迭代**。令 $w_M = (w_L + w_U)/2$，按照上述插值公式计算对应的死亡概率 q_x^* 和分离因子 a_x^*，生成生命表并计算预期寿命 $e_0(w_M)$；
3. **更新**。若 $e_0(w_M) < e_0^*$，则令 $w_L = w_M$；否则令 $w_U = w_M$；
4. **收敛判定**。当 $|e_0(w_M) - e_0^*| < \varepsilon$ 时停止迭代（例如，取$\varepsilon = 10^{-4}$），此时 w_M 即为目标权重 w^*。

附录 6：2000—2024 年我国净迁移人数规模情况

附表 9　2000—2024 年我国净迁移人数规模及其占同期总人数、出生人数、死亡人数的比例

年份	总人数（万）	出生人数（万）	死亡人数（万）	净迁移			
				规模（万）	占总人数比重（%）	占出生人数比重（%）	占死亡人数比重（%）
2000	126743	1778	817	-6	0.005	0.34	0.73
2001	127627	1702	818	-18	0.014	1.06	2.20
2002	128453	1647	821	-76	0.059	4.61	9.26
2003	129227	1599	825	-61	0.047	3.81	7.39
2004	129988	1593	832	-77	0.059	4.83	9.25
2005	130756	1617	849	-10	0.008	0.62	1.18
2006	131448	1584	892	-76	0.058	4.80	8.52
2007	132129	1594	913	-104	0.079	6.52	11.39
2008	132802	1608	935	-76	0.057	4.73	8.13
2009	133450	1615	943	-18	0.013	1.11	1.91
2010	134091	1596	949	-18	0.013	1.13	1.90
2011	134916	1604	960	-42	0.031	2.62	4.38
2012	135922	1635	966	-6	0.004	0.37	0.62
2013	136726	1640	972	-19	0.014	1.16	1.95
2014	137646	1687	977	-37	0.027	2.19	3.79
2015	138326	1655	975	-65	0.047	3.93	6.67
2016	139232	1786	977	-15	0.011	0.84	1.54
2017	140011	1723	986	-9	0.006	0.52	0.91
2018	140541	1523	993	-20	0.014	1.31	2.01
2019	141008	1465	998	-66	0.047	4.51	6.61
2020	141212	1203	998	-9	0.006	0.75	0.90
2021	141260	1062	1014	-38	0.027	3.58	3.75
2022	141175	956	1041	-29	0.021	3.03	2.79
2023	140967	902	1110	-57	0.040	6.32	5.14
2024	140828	954	1093	-32	0.023	3.35	2.93
平均				-39	0.029	2.72	4.23

数据来源：国家统计局，联合国人口司《2024 年世界人口展望》

附录 7：比例年龄生育率的预测

在本文中，采用比例年龄生育率 （Proportionate Age-Specific Fertility Rate, PASFR）来刻画生育的年龄模式。PASFR 定义为分年龄生育率除以当年的总和生育率，即

$$p_{a,t} = \frac{f_{a,t}}{\sum_{a=15}^{49} f_{a,t}},$$

其中，$f_{a,t}$ 是第 t 年年龄为 a 岁育龄女性（15~49 岁）的生育率，$p_{a,t}$ 是对应的第 t 年年龄为a岁的 PASFR，反映了该年龄的生育率对当年 TFR 的贡献比例。因此，PASFR 直接刻画了生育的年龄分布。

本文借鉴 Ševčíková 等（2016）提出的 PASFR 预测框架，对第 t 年年龄为 a 岁（15~49 岁）女性的比例年龄生育率 $p_{a,t}$ 构建如下线性模型

$$\text{logit}(p_{a,t}) = \text{logit}(p_{a,t_0}) + v_a \cdot (t - t_0), \tag{A4}$$

其中，p_{a,t_0}为初始年份 t_0 年龄为 a 岁的 PASFR，$\text{logit}(p_{a,t})$ 为对 $p_{a,t}$ 进行 logit 变换，v_a为 p_a 在 logit 尺度下的增长速度。Logit 变换定义为

$$p^*_{a,t} = \text{logit}(p_{a,t}) = \log\frac{p_{a,t}}{1 - p_{a,t}}.$$

采用 logit 变换的优势在于：在 logit 尺度下得到的预测值 $p^*_{a,t}$ 经逆变换

$$p_{a,t} = \text{logit}^{-1}(p^*_{a,t}) = \frac{1}{1 + e^{-p^*_{a,t}}},$$

可以确保 $p_{a,t} \in (0,1)$，从而满足比例变量的取值约束。

为了降低1‰人口变动调查数据在PASFR上的年度波动，在估计增长速度时，本文先对PASFR按时期取均值。具体地，按照 2000—2004、2005—2009、2010—2014、2015—2019、2020—2023 划分为 5 个时期，并计算各时期各年龄的平均比例年龄生育率 $\bar{p}_{a,s}$，其中 $s = 0$ 对应 2020—2023，$s = -1$对应 2015—2019，以此类推。由相邻两个时期 $\text{logit}(\bar{p}_{a,s})$ 的差可得到该时期平均增长速度

$$\bar{v}_{a,(s-1,s)} = \frac{\text{logit}(\bar{p}_{a,s}) - \text{logit}(\bar{p}_{a,s-1})}{\tau_{(s-1,s)}},$$

其中，$\tau_{(s-1,s)}$表示两个时期中点的时间间隔。除 $\tau_{(-1,0)} = 4.5$ 外（2020—2023 年中点为 2021.5，2015—2019 年中点为 2017，相差 4.5 年），其余间隔均为 5。在预测阶段，年龄为 a 岁的 $\text{logit}(p_{a,t})$ 的增长速度 v_a 取过去 T 个时期的平均，并假定在预测期内保持不变，即

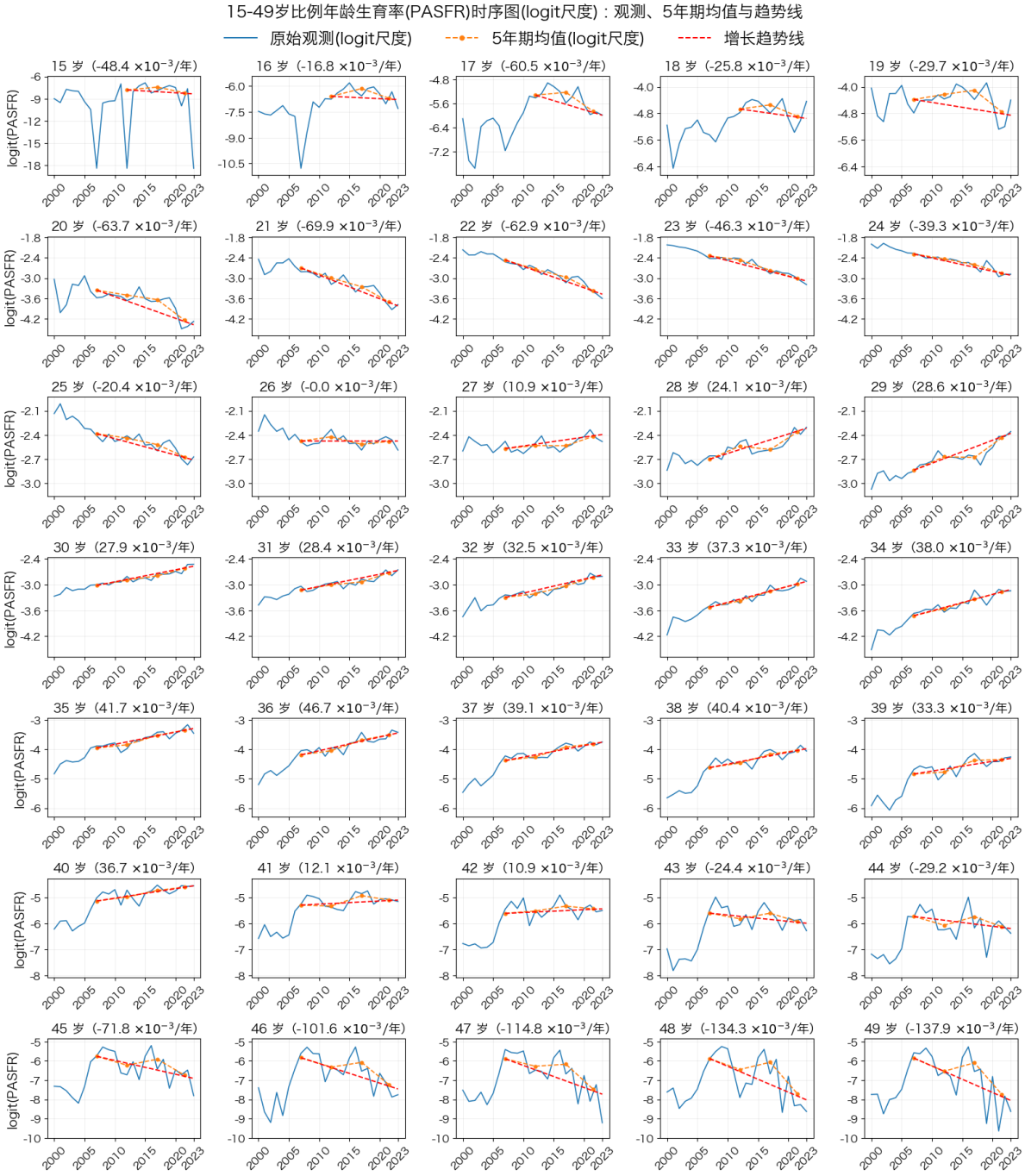
$$v_a = \frac{1}{T}\sum_{s=1}^{T} \bar{v}_{a,(-s,-s+1)}.$$

关于 T 的取值，Ševčíková 等（2016）取 $T = 3$。结合我国 PASFR 的历史变化特征，本文对 15~18 岁年龄组取 $T = 2$（该年龄段近 10 年来 PASFR 呈下降趋势（见附图 5），若取 $T = 3$ 会产生与实际相反的趋势），对 19~49 年龄组取 $T = 3$（见附图 4）。预测起点的初始状态 p_{a,t_0} 取 2021—2023 年 PASFR 的平均值。据此，可由式(A4)模型线性外推得到在任意年份 t、年龄为 a 岁的 $\text{logit}(p_{a,t})$，再经逆 logit 变换便可得 $p_{a,t}$ 的预测。由于是对年龄 a 岁的 $p_{a,t}$ 分别单独预测，在年份 t，预测结果在数值上未必严格满足求和为 1。因此，本文进一步采用 softmax 函数对预测的 $p_{a,t}$ 做归一化处理，以满足 $\sum_{a=15}^{49} p_{a,t} = 1$的约束。最终得到的预测为

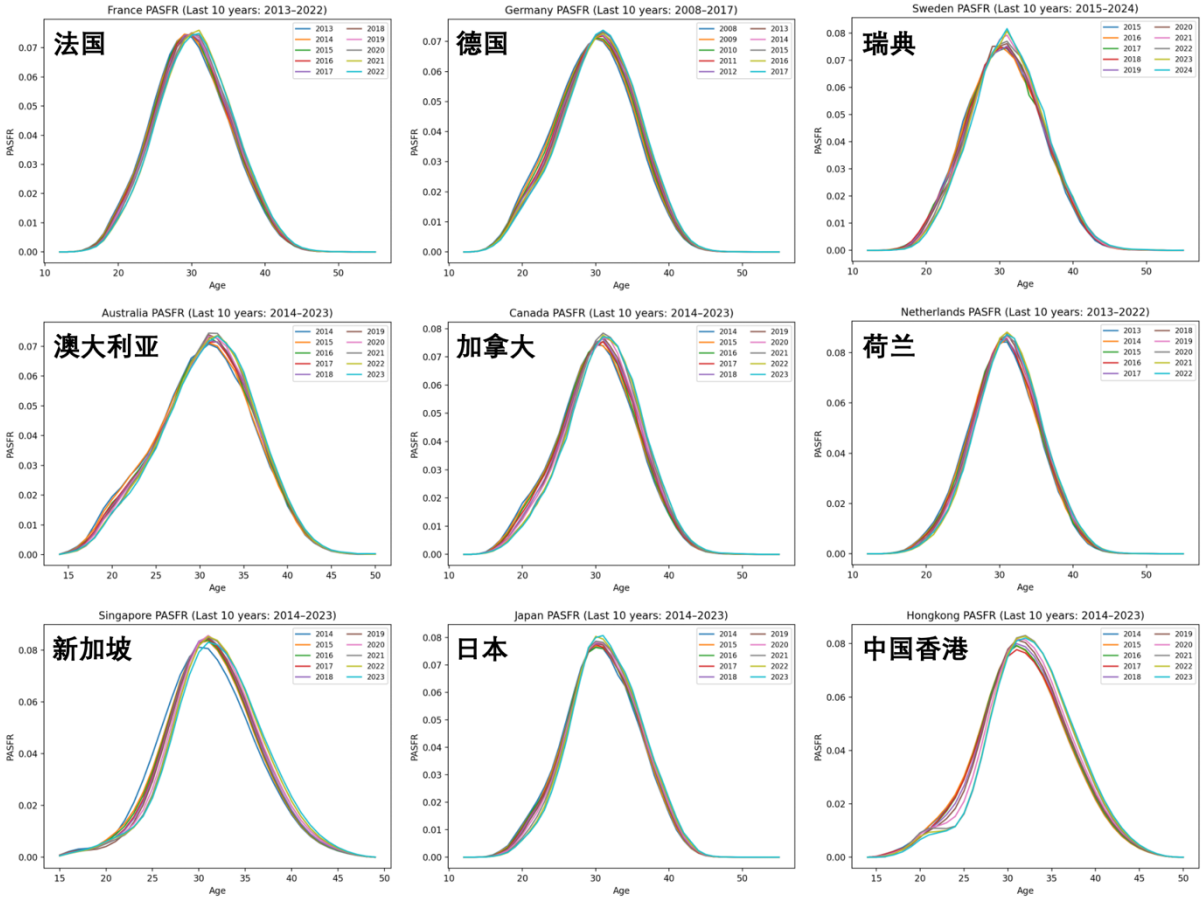
$$p^{pred}_{a,t} = \frac{\exp\left(p_{a,t} - \max_a[p_{a,t}]\right)}{\sum_{a=15}^{49}\exp\left(p_{a,t} - \max_a[p_{a,t}]\right)}.$$

其中，减去 $\max_a\{p_{a,t}\}$ 的目的在于改善归一化过程中的数值稳定性，避免在 $p_{a,t}$ 较小时出现精度损失。

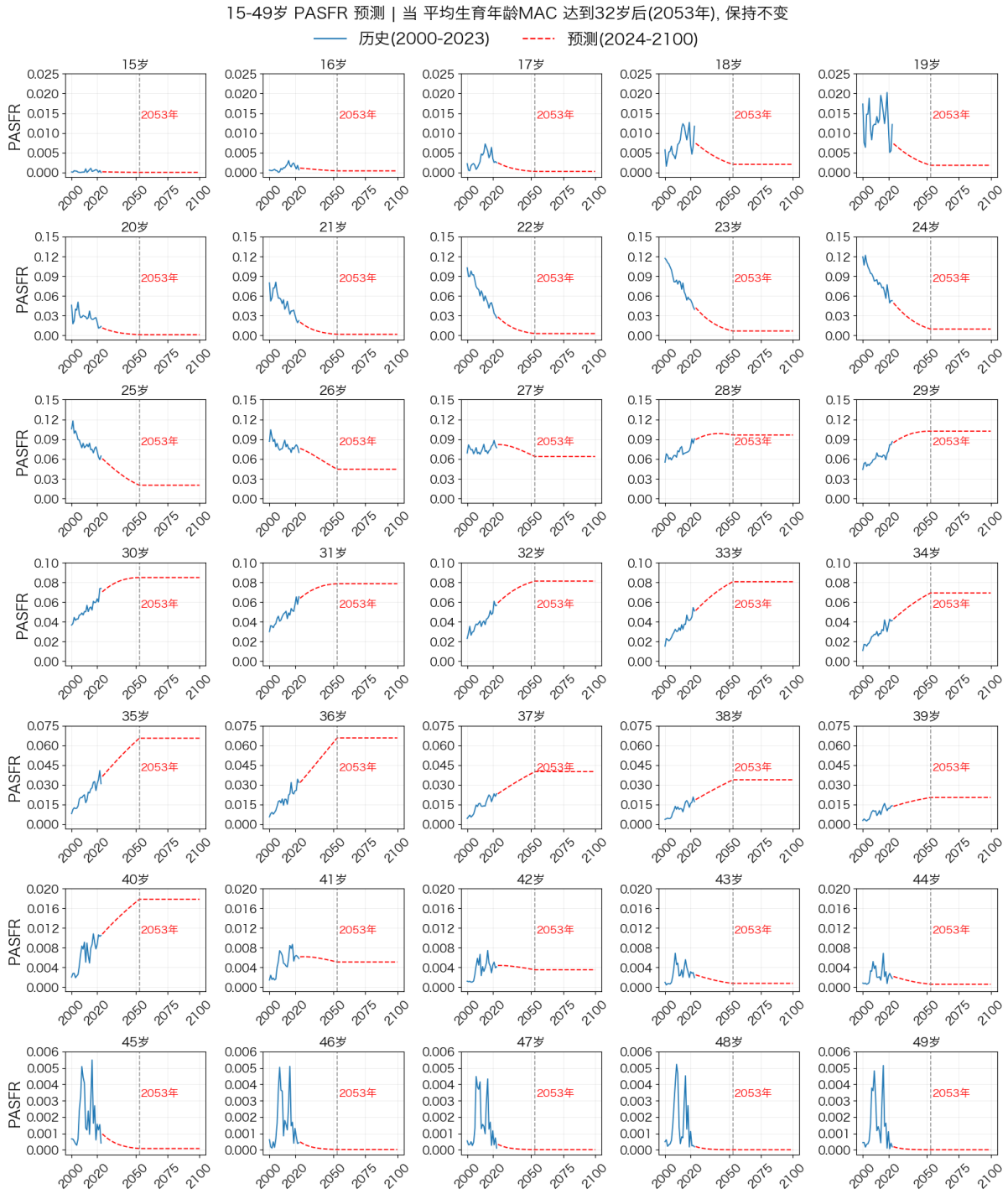
在各国比例年龄生育率的观测中，比例年龄生育率 $p_{a,t}$ 不会无限增长。本文参照联合国人口预测的方法：当平均生育年龄首次达到阈值 32 岁后，便固定在阈值年份的水平不变（United Nations，2024）。其依据来源于对世界主要国家生育模式演化的经验观察：PASFR 推迟到某一阶段后将逐渐收敛保持相对稳定（见附图 6）。本文采用同联合国相同的处理方法：当预测年份的平均生育年龄首次达到阈值 32 时，之后预测年份的 PASFR 就保持在阈值年份的水平不再变化。各年龄的 PASFR 预测见附图 7 和附图 8。



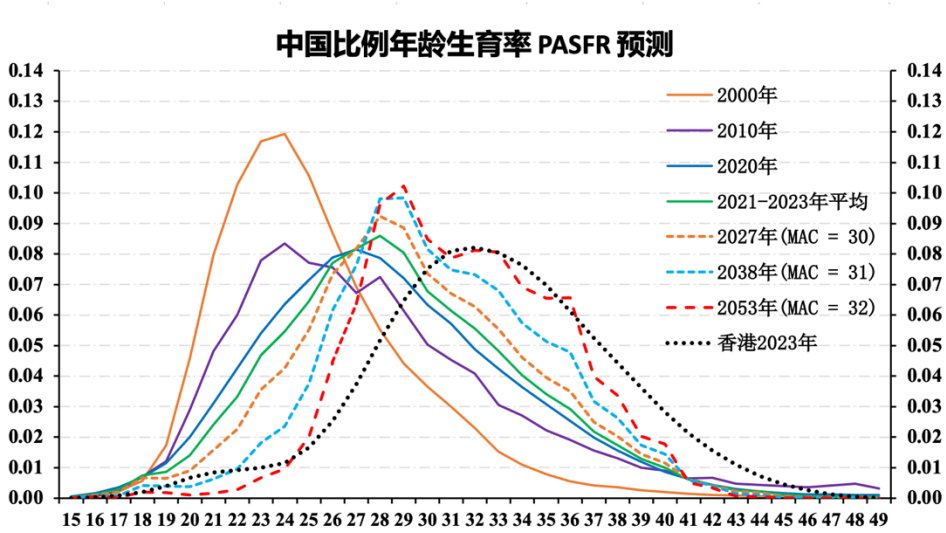
附图 5 比例年龄生育率 $p_{a,t}$ 经 logit(·) 变换后的时间序列（蓝色）、5 年均值（橙色）及趋势（红色虚线）。每幅小图标题的括号内为 logit 尺度下 $p_{a,t}$ 的增长速度 v_a



附图 6　部分国家或地区的比例年龄生育率曲线趋势（数据来源：Human Fertility Database）



附图 7　每 1 岁年龄的比例年龄生育率 $p_{a,t}$ 的预测。在 2053 年平均生育年龄首次达到 32 岁后，设定 $p_{a,t}$ 保持不变



附图 **8**　中国比例年龄生育率曲线：2000—2023 年历史趋势及预测，并与香港 2023 年比例年龄生育率曲线对照（点虚线）。预计 2027 年平均生育年龄推迟至 30 岁，2038 年推迟至 31 岁，2053 年推迟至 32 岁

参考文献

[1] Arriaga E E. Population Analysis with Microcomputers: Volume II (Extract A): Software and Documentation (Population Analysis Spreadsheets)[R]. U.S. Census Bureau, 1994 (Revised June 2012.).

[2] Ševčíková H, Li N, Kantorová V, et al. Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections[C]. In Dynamic Demographic Analysis, pp. 285-310. Cham: Springer International Publishing, 2016.

[3] United Nations Department of Economic and Social Affairs, Population Division. World Population Prospects 2024: Methodology of the United Nations Population Estimates and Projections [R]. 2024.