数量经济技术经济研究
Journal of Quantitative & Technological Economics
ISSN 1000-3894,CN 11-1087/F

《数量经济技术经济研究》网络首发论文

题目：　中国居民收入机会不平等再测算——来自机器学习的新发现
作者：　万相昱，张晨，唐亮
DOI：　10.13653/j.cnki.jqte.20231117.002
网络首发日期：　2023-11-17
引用格式：　万相昱，张晨，唐亮．中国居民收入机会不平等再测算——来自机器学习的新发现[J/OL]．数量经济技术经济研究.
https://doi.org/10.13653/j.cnki.jqte.20231117.002

DOI：10.13653/j.cnki.jqte.20231117.002 网络首发时间：2023-11-17 15:23:07
网络首发地址：https://link.cnki.net/urlid/11.1087.F.20231117.1102.004

中国居民收入机会不平等再测算

——来自机器学习的新发现

万相昱 张晨 唐亮*

摘要：缩小机会不平等，消除收入差距扩大的源生动力，从而达到分配公平，是实现共同富裕的必由途径，也是推进中国式现代化的治理抓手。本文采用基于集成回归树算法的机器学习模型，尝试克服传统方法在测算机会不平等方面普遍存在的重要缺陷，同时引入分位数回归森林将收入均值的机会不平等拓展至收入分布的机会不平等，以此提供收入机会不平等测算的全新方法和全新视角。基于2010~2021年中国综合社会调查数据的测算结果表明，以基尼系数衡量的全样本收入均值的机会不平等约为0.244~0.307，大致占总体不平等的38.1%~52.4%；这种非线性机器学习模型的测算结果明显高于依赖线性模型的传统测算方法；城镇居民收入的机会不平等高于农村，环境因素对城乡间收入差距形成的贡献度最大；个体及其父亲的可观测特征差异更倾向于拉大收入差距，而其母亲的可观测特征则相反；另外，收入分布的机会不平等测算结果表明，环境因素显著影响子代的收入风险，优良的环境基础更倾向于赋予子代收入分布的右偏优势；从分布结构看，收入下限、收入上限、偶然收入和收入风险的不平等程度都要显著高于收入均值的机会不平等；不可控的环境因素导致的子代收入下限和收入风险不平等需要被重点关注。

关键词：环境因素 机会不平等 分布不平等 机器学习

中图分类号：F124 **文献标识码：**A

一、引言与文献综述

党的十九大报告明确提出到2035年全体人民共同富裕取得更为明显的实质性进展，到2050年全体人民共同富裕基本实现。党的二十大报告进一步强调了共同富裕的战略意义，指出中国式现代化是全体人民共同富裕的现代化。李实（2022）认为实现共同富裕是在实现权利平等、机会均等基础上，人人参与共建共享发展过程中达到富裕。事实上，依据相关理论与实践，公共政策能够有效地调节收入分配格局并抑制社会财富的加速逆熵过程，但政策制定的科学性和实践效果需以精准的量化评估为前提条件，特别是要明确收入差距形成的主要动因及其关键应对方案（孙豪和曹肖烨，2022）。同样的，实现共同富裕需充分认识当前中国社会在收入差距、分配公平等方面的真实国情以及科学治理的有效抓手。基于此，本文将焦点置于在收入分配研究和推进共同富裕实现过程中备受重视的机会不平等上，对当前中国居民的收入机会不平等进行系统性测算，揭示其演进规律与发展现状，为实现共同富裕提供更为精准的数据支撑与量化决策参考。

* 万相昱，研究员，中国社会科学院数量经济与技术经济研究所、中国社会科学院大学教授，电子邮箱：wanxy@cass.org.cn；张晨（通讯作者），讲师，山东财经大学财政税务学院，电子邮箱：2777913876@qq.com；唐亮，副教授，东北师范大学经济与管理学院，电子邮箱：tangl123@nenu.edu.cn 。本文获得国家自然科学基金重大项目（71991475）、中国社会科学院大学人文社科类重大项目培育专项（02011903822004）的资助。感谢匿名审稿专家的宝贵意见，文责自负。

对于变量的精准测度需以明确其内涵为先决基础。机会均等化（Equality of Opportunity）的概念由 Roemer（1998）提出，在环境－努力二元分析框架下将影响个人结果（收入、教育等）的因素划分为两类，其一是个体可控因素，称之为努力（Effort），但不局限于学习努力、工作努力、职业选择等狭义努力，还涵盖运气、天赋等因素；其二是个体无法控制、不承担责任或不能先验选择的因素，称之为环境（Circumstances），包括性别、家庭背景（父母教育、职业、出生地等）等。理论上，如果能够将拥有完全相同外生环境的个体抽出并归类，便可将类型内部的结果差距（如收入不平等）归结为个体努力不同，这是收入差距形成的重要原因，但不应是再分配政策的主要靶向；而类型之间的差距显然源于不可控的环境因素，这是更值得引起关注并更应进行政策补偿的重点。机会均等化的施政体系就是消除或削弱环境对收入、公共服务等结果分配的净影响，这不仅是推动共同富裕的源生动力，更是治理现代化的必然要求。

在 Roemer 机会均等化分析基础之上，学者们尝试扩展机会不平等的分析框架并不断提出新的测算方法以精确研究结果（Roemer 和 Trannoy，2016）。少量文献（Kanbur 和 Snell，2019）试图针对机会不平等的存在性进行统计检验。诸多学者（Checchi 和 Peragine，2010；Almås 等，2011）则通过基尼系数、广义熵等指数对机会不平等的程度进行量化。针对后一类文献，可以进一步将其按照测算角度和测算方法分类，从机会不平等的测算角度来看，可以区分为事前（Ex-ant）估计和事后（Ex-post）估计：前者将个体分配至不同的环境类型，拥有不同类型群体间的结果差距即为机会不平等（Fleurbaey 和 Peragine，2013）；后者基于努力程度进行分组，机会不平等由相同努力程度下个体之间的结果差距所反映（Juárez 和 Soloaga，2014）。由于努力程度难以直接观察和度量，因此事前估计更受欢迎。从机会不平等的测算方法上看，可以将其划分为参数方法和非参数方法：参数方法需要针对包含环境与努力变量的结果方程进行估计，基于变量系数及取值模拟消除环境影响后的反事实结果，利用真实结果与反事实的差距度量机会不平等；非参数方法的思想就是基于环境和努力变量进行分组，组间或组内差距即为机会不平等。由于非参数方法的维度诅咒问题（龚锋等，2017；雷欣等，2018），参数方法的使用更为普遍（李莹和吕光明，2018；史新杰等，2018、2022）。

无论事前还是事后，抑或参数和非参数方法，共同存在的问题是实证结果对于变量选择和模型设定过于敏感，而这些技术细节完全由学者自行决定，人为操纵研究结果的可能性增加，由此造成的偏误难以避免（Brunori 等，2018）：首先，研究者必须针对哪些环境变量进入实证模型做出选择。Ferreira 和 Gignoux（2011）指出，可观测环境因素只是影响个体结果的外生环境变量子集，片面的对照差分将造成机会不平等估计的向下偏误。当然，也可以通过提供更为详细的个体环境数据（例如，基因数据）进行解决（Hufe 等，2017），但是样本量不足而环境变量过多将导致高维问题并过度压缩模型自由度，迫使研究人员仍需对环境因素进行事前筛选。

其次，某些环境因素对于结果的影响可能部分依赖于其他环境特征，要求研究者需要谨慎设定模型形式（Brunori 等，2018）。例如，García 等（2018）发现孩童照料安排的后续影响强烈取决于生理性别。然而现有文献普遍使用线性模型，过于简化的函数形式将低估机会不平等（Ferreira 和 Gignoux，2011）。虽然可以通过提升模型复杂度（例如，引入交互项和高阶项）来缓解偏误问题，但是由于可用自由度的限制，在样本量一定的情况下，纳入更多

环境变量或者放松线性假定均有可能放大参数估计方差，估计有效性和预测置信度明显降低。例如，Brunori 等（2019b）发现过度拟合又会导致机会不平等被显著高估。上述分析充分表明在估计机会不平等时模型设定和环境变量选择的重要性，研究者必需在多种偏误来源和估计有效性之间进行权衡。

将环境和努力因素所引致的个体结果进行分离，本质上是一个预测问题，即预测不同因素导致的收入差距。传统计量模型更关注识别问题，而计算机科学领域的机器学习和深度学习模型更加擅长预测。伴随大数据和人工智能技术的兴起，如何在计量经济模型中引入非参数方法（例如，机器学习）成为近年来的研究热点（Athey 和 Imbens，2019；姚鹏和李金泽，2023）。回溯 Roemer 关于机会均等化研究的思想本质：个体被分配到特定的环境类型，同一类型的群体拥有相同的环境特征，群体内部和群体间结果的差距分别源于努力和环境因素。这与机器学习领域中的树模型具有高度的结构相似性和逻辑吻合性，即基于某一准则（例如，组间结果差异最大化），将个体按照环境变量进行递进式分组，最终所有样本被分配至互不重叠的子组当中，同一子组内部意味着相同的环境类型，组内和组间的结果差距即可用于刻画努力和机会不平等。

相比已有测算方法，树模型至少拥有以下三点优势：（1）树模型属于非线性模型，充分考虑环境因素对结果的非线性影响以及因素间的相互作用，规避了参数方法的模型误设问题。（2）树模型根据算法自动将由环境变量构成的特征空间分割成不重叠的区域，环境类型的获得完全基于结果的可变性，而非预先假设哪些环境因素在决定个体结果中发挥了显著作用，最大限度降低环境因素选择的主观任意性。（3）一系列剪枝算法将防止树模型过度拟合，提高样本外结果预测的准确性，也可以通过集成思想（例如，随机森林模型构建大量复杂度较低的树模型进行平均预测）有效降低预测结果的方差。综合来看，前两个优势能够缓解传统方法关于环境变量选择和模型形式简化所产生的向下偏误，最后一个优势规避了过度拟合产生的高估问题，同时提升估计结果有效性。因此，本文认为尝试利用树模型测算机会不平等既是有效的研究途径，也是有益的探索方向。

与既有研究相比，本文的增量工作和边际贡献体现在研究方法和研究内容两个方面：（1）在研究方法上，鉴于关于努力因素的界定标准并无一致观点，本文依然遵循事前估计的范式，但是与传统研究依赖线性回归模型不同的是，本文采用前沿的机器学习模型并尝试依据目标不断拓展技术，以弥补现有方法理论上的主要缺陷，更为科学地测算中国居民收入的机会不平等以及环境因素贡献度，基于此，本文也能够揭示一些新的重要发现并获取实证解释的全新维度。（2）在研究内容上，已有关于机会不平等的研究普遍关注的是均值层面的不平等，而环境因素更可能导致整个分布范围的机会不平等，这一点在既有研究中被完全忽视，本文借助集成机器学习模型以复原个体的收入分布，从收入下限、收入上限、偶然收入以及收入风险四个方面较为完整地刻画环境因素所导致的子代收入分布差距，以此评估收入分布的机会不平等。本文拓宽了机会不平等这一收入分配领域重要概念的外延，并为相关公共政策调整提供了行动靶向和量化依据。

二、机会不平等测算方法

（一）树模型

决策树（Decision Tree）是机器学习领域一种基本的分类预测方法（Breiman，2001），呈现树形结构，表示基于样本特征进行分类的过程，可以被认为是 if-then 规则集合，也可以被认为是定义在特征空间与类空间上的条件概率分布，模型具有可读性、分类速度快等优点。目前有多种算法可以生成回归树模型，其中最受欢迎的是分类与回归树（Classification and Regression Trees，CART）算法，该类树模型的训练过程可以表述如下，假设X和Y分别为输入和输出变量，并且Y为连续变量，给定样本数据集：

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \tag{1}$$

一棵回归树对应着特征空间的划分以及所划分子空间上的输出值。假设已将特征空间划分为M个子空间，并且在每个子空间R_m上有一个固定的输出值c_m，于是回归树模型可以表示为：

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{2}$$

当特征空间的划分确定时，可以使用均方误差$\frac{1}{N}\sum_{x_i \in R_m}\left(y_i - f(x_i)\right)^2$来表示回归树在训练数据上的预测误差，使用均方误差最小准则求解每个子空间上的最优输出值。易知，子空间R_m上c_m的最优估计值$\hat{c}_m$是位于此空间所有输入样本x_i对应y_i的均值。

基于回归树的收入预测结果可以写为：

$$E[Income|Circumstances = R_m] = \hat{c}_m = \text{avg}(y_i | x_i \in R_m) \tag{3}$$

所谓的子空间（或树的叶子节点）R_m即为特定的环境类型，回归树采用子空间训练样本的平均收入作为拥有相同特征样本的预测收入，回归树的训练和预测过程可以很自然地移植到 Roemer 关于环境－努力的二元分析框架中。对于位于同一叶子节点（即拥有相同环境类型）的样本，回归树将他们的平均收入归结为环境因素所导致的结果，而不同叶子节点即拥有不同环境类型样本间的收入差距就是 Roemer 所定义的机会不平等。

传统树模型虽然应用广泛，但是生成算法存在四个突出问题：（1）容易过度拟合，当叶子节点上的样本数量太少时，模型的泛化能力将被严重削弱，虽然可以采用事前或者事后剪枝控制生成树的规模，但是由于缺乏客观标准导致选择的任意性较强。（2）Mingers（1987）认为 CART 属于启发式算法，回归树生成过程缺乏统计意义，选择切分变量和切分点时无法区别信息增益或者均方误差变动的显著性。（3）类似 CART 的穷尽式的搜索过程将产生选择变量偏误，容易选择拥有更多备选切分点的特征。（4）CART 由于能够同时处理回归和分类问题被深度使用，但是缺乏对于更广泛数据类型的支持，比如排序数据、截断数据。为了解决上述问题，Hothorn 等（2006）提出了条件推断树（Conditional Inference Trees），在切分变量和切分点选择上采用基于统计检验方法而非均方误差最小化准则。具体步骤如下：

（1）检验所有输入特征X与结果Y的相关性即H_0，如果不能拒绝原假设则停止树的生成。

$$H_0 = \cap_{j=1}^{m} H_0^j = 0 \tag{4}$$

其中，m为特征变量个数。采用与数据类型无关的置换检验逐一评估X与Y的相关性，判断是否拒绝H_0：

$$H_0^j：D(Y|X_j) = D(Y) \tag{5}$$

此时涉及多重假设检验（Multiple Hypothesis Test）问题，基于 Bonferroni Correction 进行 p 值调整：

$$p_{adj}^{Xj} = 1 - (1 - p^{Xj})^m$$

如果所有特征变量的显著性$p_{adj}^{Xj} > 1 - \alpha$，那么将接受$H_0$，停止切分。

（2）当拒绝H_0时，选择与Y相关性最强即最小p_{adj}^{Xj}的特征X_{j*}作为切分变量。（3）根据X_{j*}备选切分点S进行节点样本二切分，检验两组切分样本Y的差异性，选择能够产生最小 p 值的切分点。（4）重复上述算法。

相比 CART 等传统树模型算法，条件推断树赋予了机器学习模型更多的经济解释。具体到机会不平等的测算，每个假设检验本质上是对一群特定样本中是否存在平等机会的检验。如果算法没有进行任何切分，那么就不能拒绝机会均等化的原假设。只要进行切分就说明切分变量即环境因素引发了显著的结果差异，回归树长得越深就意味着更多的环境因素导致机会不平等。待整个树模型生成完毕，仍然采用位于同一叶子节点（拥有相同环境类型）的样本的平均收入作为预测收入，根据位于不同叶子节点（拥有不同环境类型）样本间的收入差距度量收入机会不平等。

（二）集成树

无论传统回归树还是条件推断树均能够通过一种非任意的方式选择环境因素并切分样本。然而，利用树模型构建反事实预测时存在两个缺陷：（1）单棵树仅能利用所有可获得环境因素的子集，模型所忽略的特征虽然在特定显著性水平上与结果的相关性不强，但是将其纳入模型显然能够提升预测能力，这一情况通常可能发生在环境因素之间存在高度相关的场景中，一旦某一特征被选为切分变量，与之紧密相关的其他环境因素后续将不再被使用。（2）单棵树将导致模型预测结果的方差较大，模型有效性和预测精度均有所下降，其原因在于二叉树的结构以及随之而来的预测分布对输入样本变化比较敏感，特别是在环境因素间的预测竞争力较为接近时（Luna 等，2019）。单棵树的缺陷可以采用集成学习算法（Ensemble Learning）进行规避，基于回归树的集成算法首推随机森林。简而言之，随机森林就是创建大量回归树，基于所有树的平均预测结果进行预测。常规意义下的随机森林以传统回归树作为元学习器。如果将元学习器设定为条件推断树，那么此时的随机森林称为条件推断森林（Conditional Inference Forests）。下文将分别使用回归树、随机森林、条件推断树和条件推断森林预测样本收入和测算收入机会不平等，对比不同模型的结果差异，分析关于收入机会不平等的新发现。

三、收入机会不平等测算结果与分析

（一）样本选择与变量度量

1.样本来源与筛选

本文实证样本来自中国综合社会调查（Chinese General Social Survey，CGSS）。作为第

一个全国性、综合性且连续性的大型社会调查项目，CGSS 涉及中国国民和社会各个方面的内容，被广泛用于政治学、经济学、社会学等领域的研究。由于入户调查时详细询问父母的年龄、教育、政治面貌以及工作信息，而环境因素的可获得性是机会不平等测算结果可靠性的基础保障，因此 CGSS 往往是研究机会不平等问题的首选数据源（李莹和吕光明，2019；汪晨等，2020）。本文选取 2010、2012、2013、2015、2017、2018 和 2021 年的调查数据[①]，并且将劳动力年龄限定为 18~60 周岁（李莹和吕光明，2018），剔除数据缺失和数据异常问卷，最终获得共计 36407 个实证样本。

2.变量选择与描述性统计

根据调查问卷，将测算机会不平等的环境变量集划分为个体特征、父亲特征和母亲特征。个体特征包括性别、年龄和民族。父代特征包括父母受教育程度、政治面貌、子女 14 岁时就业状态、职务级别以及单位类型。以 2009 年为基期，采用消费者价格指数对收入数据进行平减调整[②]，样本平均年龄为 44 岁，平均年收入为 32801 元，其余均为离散型变量。

（二）收入机会不平等测算结果

1.全样本收入机会不平等测算

我们首先在全样本层面上预测环境因素对个体收入的影响，进而利用拥有不同环境类型个体间的收入差距反映收入的机会不平等，表 1 分别基于基尼系数（Gini）和对数偏差均值指数（MLD）两种常用的不平等指标展示了不同年份、不同模型下全样本的收入机会不平等程度。首先，对比不同机器学习模型可以发现，传统回归树和随机森林的不平等测度结果要大幅高于条件推断树和条件推断森林。具体来看，回归树模型测算的 Gini 和 MLD 不仅绝对值最大，其在总体不平等中的占比明显过高，最低为 2016 年 MLD 的 75.7%，最高达到 2009 年 108.9%。仅从数值结果似乎可以认为中国居民收入不平等全部由环境因素引致，这与经济常识相违背，意味着回归树的过度拟合问题十分严重。随机森林尝试通过构造大量弱学习器来缓解单一树模型的过度拟合缺陷，根据表 1 显示的结果，无论不平等的绝对值还是总量占比均有了明显下降，其中，2009 年的 MLD 由回归树的 0.808 降至随机森林 0.305，降幅达到 62.2%。上述结果揭示出在利用环境因素对收入进行预测时，需要重点关注机器学习模型普遍存在的过度拟合问题，否则实证结果将会出现较大偏差。另外，即使随机森林模型在一定程度上缓解了过度拟合，但是根据其具体测算结果，我们仍能看到环境因素对于收入不平等较高的贡献份额，其中，Gini 最低总量占比达到 60%以上，最高为 2012 年的 75.1%，仍然支持环境因素主导收入分配格局的假设，说明预测模型仍有进一步的改进空间。

通过引入统计显著性，采用置换检验选择切分变量和切分点，条件推断树和条件推断森林能够更加有效地缓解传统树模型和森林模型的过度拟合和选择变量偏误等问题。根据表 1 可知，条件推断树将随机森林的测算结果进一步降低，过度拟合问题进一步被缓解，其中，条件推断树的 Gini 普遍在 0.3 左右，MLD 则在 0.2 左右，对应占比前者一般在 50%以上，后者则位于 20%~33%区间内[③]。我们利用条件推断森林模型将大量条件推断树的平均预测结

① 未选择年份存在的主要问题是环境变量相关问题缺失过多或者调查内容与其他年份差异较大。例如 2011 年只有父母教育和政治面貌，缺少父代工作等诸多信息。由于询问的是“去年收入”，因此实际年份为调查年份的上一年。

② 消费者价格指数来自国家统计局网站 http://www.stats.gov.cn/。

③ 很明显可以发现 Gini 和 MLD 的机会不平等占比差异很大，后者大幅低于前者。原因在于 MLD 对位于收入分布两端的样本较为敏感，而 Gini 则给予分布中间样本更大比重。但是模型往往得到较为平滑的预测分布，由此造成两者的机会不平等测算结果存在显著差异。Brunori 等（2019a）指出虽然同为评估机会不平等时的常用指标，但是 MLD 仅仅是因为便于组间和组内分解而被选择，鉴于其对于异常值过分的关注，更加推荐使用 Gini，本文同样如此。MLD 虽然在绝对量（机会不平等量值

果作为机会不平等的测度标准，最终完成对过度拟合的修正校准。事实上，由于条件推断树本身的过度拟合问题不如回归树严重，条件推断森林的测算结果相对于条件推断树并不如前述改进模型时的变化幅度明显，这也意味着通过现有技术改进消除过度拟合的效果开始收敛。因此，本文最终以条件推断森林的不平等测算结果作为评价标准：Gini 落在 0.244~0.307 范围内，MLD 大致为 0.099~0.153；环境因素引发的不平等占比约为 38.1%~52.4%（Gini）或 11.5%~21.0%（MLD）。

表 1　　全样本收入机会不平等测算结果

年份	不平等指标	总体不平等	机会不平等			
			回归树	随机森林	条件推断树	条件推断森林
2009	Gini	0.612	0.656 （1.072）	0.421 （0.688）	0.327 （0.534）	0.285 （0.466）
	MLD	0.742	0.808 （1.089）	0.305 （0.411）	0.182 （0.245）	0.128 （0.172）
2011	Gini	0.534	0.519 （0.971）	0.400 （0.748）	0.312 （0.584）	0.263 （0.492）
	MLD	0.584	0.490 （0.839）	0.279 （0.477）	0.170 （0.291）	0.110 （0.188）
2012	Gini	0.514	0.507 （0.986）	0.386 （0.751）	0.324 （0.630）	0.269 （0.524）
	MLD	0.547	0.481 （0.879）	0.263 （0.480）	0.181 （0.330）	0.115 （0.210）
2014	Gini	0.612	0.573 （0.936）	0.392 （0.641）	0.297 （0.485）	0.254 （0.415）
	MLD	0.776	0.606 （0.780）	0.272 （0.350）	0.162 （0.208）	0.100 （0.129）
2016	Gini	0.615	0.571 （0.929）	0.431 （0.701）	0.348 （0.565）	0.307 （0.499）
	MLD	0.813	0.616 （0.757）	0.335 （0.412）	0.216 （0.266）	0.152 （0.186）
2017	Gini	0.598	0.587 （0.982）	0.439 （0.734）	0.347 （0.580）	0.307 （0.513）
	MLD	0.769	0.658 （0.855）	0.352 （0.458）	0.227 （0.296）	0.153 （0.199）
2020	Gini	0.640	0.639 （0.999）	0.437 （0.683）	0.360 （0.562）	0.244 （0.381）
	MLD	0.865	0.799 （0.924）	0.348 （0.403）	0.228 （0.264）	0.099 （0.115）

注：括号内为机会不平等占总不平等的比重，采用预测收入不平等指标与原始收入不平等指标的比值进行度量。

2.城镇和农村收入机会不平等测算

依户籍将样本拆分为农村和城镇两个子集，分别测算各样本子集内部（城镇和农村）和

以及占比）上与 Gini 存在差异，但是相对指标（例如时间趋势、城乡等组间不平等程度对比）并无不同。

样本子集间（城乡）的机会不平等程度，以此量化中国居民收入机会不平等的城乡二元结构差异，数值测算结果见表 2。需要说明的是，此处不同机器学习模型导致的测算结果的差异与全样本情况相似，故不再赘述，主要重点基于条件推断森林的测算结果展开详细论述。

我们首先从样本子集内部着眼，模型测算结果显示城镇内部收入机会不平等的 Gini 绝对值为 0.200~0.267，MLD 则为 0.076~0.134，转化为环境因素引致的不平等总量占比，前者大约为 35.9%~54.0%，后者约为 13.0%~29.0%；另一方面，研究结果显示农村内部机会不平等的测算结果为0.196~0.263（Gini）和0.060-0.109（MLD），相应总量占比分别为30.2%~40.8%（Gini）和 6.8%~12.6%（MLD）。两者横向对比可以发现，尽管农村收入不平等明显高于城镇，但是机会不平等却恰恰相反，且由此导致环境因素引致的机会不平等总量占比更低，意味着城镇居民收入差距更倾向于受到个体及其父代特征的影响。

表 2　　城镇、农村及城乡间收入机会不平等测算结果

年份	不平等指标	城镇		农村		城乡间	
		总体不平等	机会不平等	总体不平等	机会不平等	总体不平等	机会不平等
2009	Gini	0.564	0.245 （0.434）	0.568	0.209 （0.368）	0.345	0.175 （0.506）
	MLD	0.594	0.110 （0.185）	0.616	0.068 （0.110）	-	-
2011	Gini	0.468	0.222 （0.474）	0.537	0.212 （0.394）	0.294	0.156 （0.531）
	MLD	0.414	0.093 （0.224）	0.576	0.069 （0.120）	-	-
2012	Gini	0.441	0.238 （0.540）	0.528	0.213 （0.403）	0.278	0.158 （0.570）
	MLD	0.366	0.106 （0.290）	0.565	0.071 （0.126）	-	-
2014	Gini	0.527	0.240 （0.456）	0.648	0.196 （0.302）	0.312	0.147 （0.470）
	MLD	0.523	0.102 （0.195）	0.874	0.060 （0.068）	-	-
2016	Gini	0.533	0.257 （0.483）	0.642	0.257 （0.400）	0.333	0.181 （0.543）
	MLD	0.546	0.125 （0.230）	0.881	0.102 （0.116）	-	-
2017	Gini	0.501	0.267 （0.533）	0.646	0.263 （0.408）	0.314	0.176 （0.561）
	MLD	0.481	0.134 （0.280）	0.895	0.109 （0.122）	-	-
2020	Gini	0.558	0.200 （0.359）	0.684	0.241 （0.353）	0.324	0.130 （0.401）
	MLD	0.590	0.076 （0.130）	1.004	0.092 （0.091）	-	

注：同表 1。

城乡间收入差距一直以来都被广泛关注，然而，相关的文献并未对环境因素引致的城乡间不平等做出有效评估，因此，本文进一步尝试填补现存空白。根据表 2 倒数第二列数值与表 1 第三列 Gini 的比值可以发现，城乡间收入不平等占全样本不平等的比重普遍超过 50%，数据表明城乡间收入差距是全样本收入不平等的主要构成。同时，表 2 最后一列与表 1 最后一列 Gini 比值大约在 60%左右，这表明环境因素引致的机会不平等中，城乡间差异仍然占据主要地位，意味着探究环境因素与城乡间收入差距的关系具有非常重要的意义。根据表 2 最后一列给出的具体量值，城乡间机会不平等的 Gini 大约为 0.130~0.181，普遍占机会不平等总量比重超过一半，约为 40.1%~57.0%，分别对比城镇和农村内部的机会不平等占比可以发现，环境因素所引发的不平等在城乡之间最为突出，也就是说，城乡间差距是中国居民收入不平等的主因，而个体和父代特征等不可控因素则是形成城乡间收入差距的主因。

3.收入机会不平等的演变趋势

进一步，我们尝试测算中国居民收入机会不平等的演变趋势[①]，图 1~图 4 给出了应用 Gini 的测算结果。当我们暂时不考虑 2020 年这一特殊时期时，无论全样本、城镇、农村还是城乡间，2009~2017 年机会不平等的演变趋势均具有 V 型特征，即先降低后增加，全样本、农村和城乡间的转折点均为 2014 年，城镇则提早至 2011 年，如果仅考虑 2009 和 2017 两年数值，则所有组别的收入机会不平等均随时间有所增加，农村样本增幅最大为 26%，城乡间最小为 1%。同时期内，环境因素所引致的不平等占比具有 N 型演变特征，表现为“增加（2009~2012 年）”— “降低（2012~2014 年）” —“增加（2014~2017 年）”的趋势，同样只考虑样本期两端，机会不平等的占比也出现明显增长，最高依然为城镇，达到 23%。因此可以发现中国居民收入机会不平等的反弹趋势和增长态势较为明显。

不过上述趋势被 2020 年的新冠肺炎疫情完全打破。仅从曲线走势来看，2020 年相比 2017 年无论机会不平等绝对值还是所占比重均出现了骤降。（1）从总体不平等的测算结果来看，相比 2017 年，2020 年的居民收入差距有所扩大，全样本、城镇、农村以及城乡间分别增长了约 7%、11%、6%和 3%，说明疫情冲击恶化了收入不平等。（2）与总体不平等完全相反的是，机会不平等出现了显著下降。从绝对值来看，全样本、城镇、农村以及城乡间分别降低了约 21%、25%、8%和 26%。从占比来看，分别降低了约 26%、33%、13%和 29%。因此，疫情冲击出人意料的缩小了不可控环境因素所带来的收入差距[②]。综合对比总体不平等和机会不平等的显著差异，似乎可以获得如此启示，隐藏于不平等中的环境因素力量根深蒂固，难以通过经济社会系统的自发性调节在短期内进行消除。当无法预期的外部冲击来临时，有可能将当前陷入僵局的收入分配局面打破，此时家庭背景、社会关系等非“努力”因素与收入的紧密关系将被削弱，机会的公平性同时得到提升。

① MLD 具有相似规律特征，不再赘述。

② 需要注意的是，2021 年有效样本共计 2902 个，占比 8%，几乎只是其他年份样本量的一半。无法排除样本量过少引致如此结果的可能性。

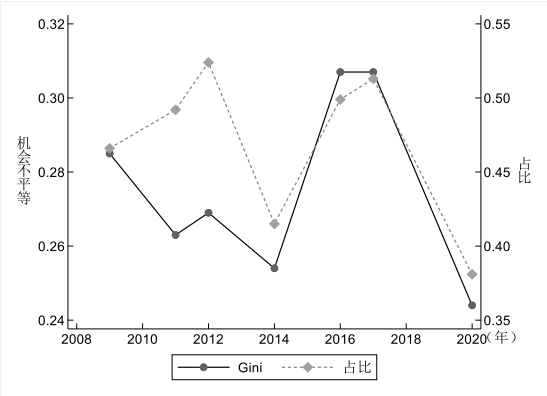

图 1 收入机会不平等变动趋势

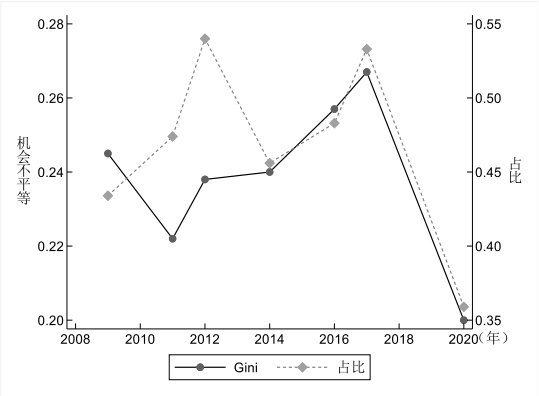

图 2 城镇收入机会不平等变动趋势

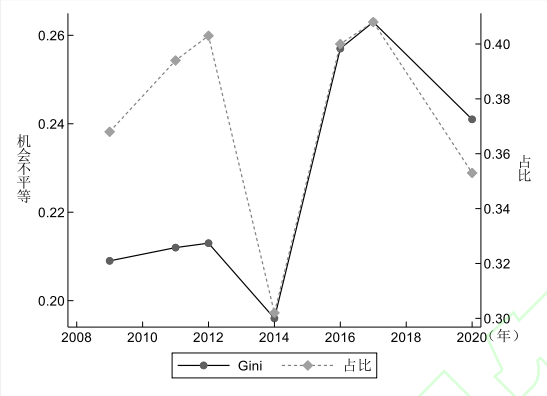

图 3 农村收入机会不平等变动趋势

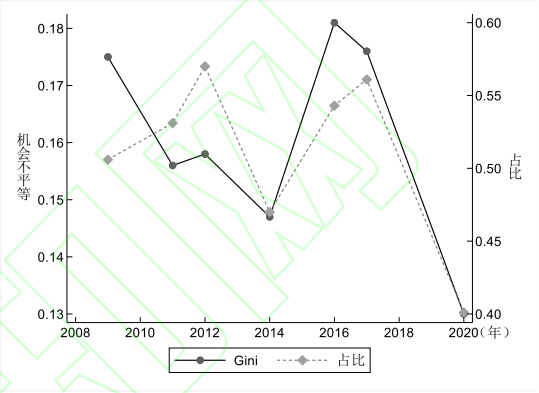

图 4 城乡间收入机会不平等变动趋势

4.与已有研究关于测算结果的比较

既有文献中，李莹和吕光明（2019）以及汪晨等（2020）使用相同数据源对中国居民的收入机会不平等进行了测算，两篇文献均使用参数方法，构建明瑟方程预测收入，前者使用 MLD 度量不平等，后者则使用 Gini，具体测算结果见表 3。对比本文结果可以发现：首先，在样本期的 4 年中，汪晨等（2020）关于全样本、城镇和农村的机会不平等平均占比分别为 30%、25%和 33%，而本文相应的测算结果平均为 47%、48%和 37%，明显高于既有文献，其中城镇样本的测算差距尤为突出，农村则较为接近。上述实证结果与 Ferreira 和 Gignoux（2011）关于函数形式简化容易低估机会不平等的观点相符。原因在于除非引入大量交互项和高阶项，否则线性模型容易产生遗漏环境变量的问题。当那些容易引致后代收入差距的因素，比如个体特征和父亲特征[①]，被部分（高阶项和交互项）置于随机扰动项中时，机会不平等自然有可能被低估。上述对比是在环境因素选择完全相同的情况下，实际上，汪晨等（2020）由于模型缺陷所选取的环境变量数量相对于本文的数量较少，因而可能是造成其收入机会不平等被低估的一个重要原因[②]。

其次，对比李莹和吕光明（2019），其在样本期的 3 年中关于全样本、城镇和农村的机会不平等占比分别为 39%、24%和 34%，而本文采用相同指标的测算结果仅为 17%、22%和

① 环境因素贡献度测算和分析结果揭示出个体和父亲特征持续拉大收入差距，虽然母亲特征具有相反影响，但是前者综合影响力要明显大于后者。所以即使所有环境因素均出现部分遗漏，也有可能产生低估机会不平等的结果。

② 汪晨等（2020）和李莹和吕光明（2019）都只选择了问卷提供的部分环境因素，我们猜测是因为涉及大量虚拟变量，如果考虑交互项，那么线性回归模型中的变量数量急剧增加，可能会对系数估计产生不良影响，因此事先进行了主观筛选。上述问题也是文献综述部分提到的关于传统线性模型缺陷的体现，而机器学习模型完全不用考虑变量数量，由算法自行筛选。

10%，明显低于既有文献，其中农村样本的结果差距最为显著。MLD 揭示出与 Gini 完全相反的差异，但是我们认为表 1 和表 2 基于 MLD 的机会不平等测算结果更为可信。原因在于，对于相同样本，MLD 和 Gini 两种不平等指标之间存在天然的差距，这样的差距并不是来自样本选择、数据处理等流程的偶然，而是具有理论上的必然性。虽然国内研究倾向于两类指标的机会不平等测算结果较为一致，但是国外文献与本文一样，发现后者明显高于前者。Brunori 等（2019a）明确指出原因：（1）机会不平等的测算是从原始收入中移除努力、运气等非环境因素（运气因素有可能引致异常收入）影响的过程，最终得到一个较为平滑的收入分布。由于 MLD 对于异常值较为敏感，而模型所得收入分布更为平滑，并无太多异常信息，因此基于 MLD 的机会不平等占比通常会低于 Gini。（2）如果根据环境因素所划分的组别数量很大，相比组别数量较少时，组间的不平等程度肯定要更低一些。而 MLD 对较低水平的组间差距表现出一定的迟钝，导致其估计结果系统性趋于零。因此，无论根据理论还是实证，对于相同样本关于 MLD 的机会不平等占比理应显著低于 Gini[①]，综上认为本文的测算结果更具可信性。

表 3　　已有研究收入机会不平等总量占比的测算结果　　（单位：%）

参考文献	不平等指标	年份	全样本	城镇	农村
汪晨等（2020）	Gini	2009	24.39	22.40	25.85
		2011	28.04	24.51	34.71
		2012	35.72	29.79	36.76
		2014	30.56	23.33	35.00
李莹和 吕光明（2019）	MLD	2009	38.68	20.98	34.69
		2011	-	-	-
		2012	42.28	27.08	36.52
		2014	34.88	24.27	31.67

资料来源：根据文献整理得到。

（三）环境因素贡献度测算和分析

上述关于机会不平等的测算结果阐释了环境因素对于收入不平等的显著影响。那么，进一步需要重点回答的问题是，究竟哪一类环境因素在其中发挥了主要作用？对环境因素的结构解析是当前关于机会不平等研究的核心重点。已有研究关于机会不平等的因素贡献度测算主要采用 Shapley 分解（李实和沈扬扬，2022）。以明瑟方程为预测模型，在控制（以样本均值替代真实取值）和不控制某一因素下观察机会不平等的变动程度，并以此度量该环境因素的贡献度。然而值得关注的是，如同利用参数方法测算机会不平等一样，评估环境因素贡献度的 Shapley 分解同样依赖于线性模型，而其简化的模型设定理论上将导致数据结果的偏误。另外传统的 Shapley 分解采用样本均值实现因素控制，我们认为更为合理的是在剔除和未剔除特定因素下计算不平等指标的变动。因此，在利用非线性模型对收入机会不平等进行测算的成果经验基础上，我们尝试对机会不平等的分解方法做出探索，以此突破传统 Shapley 分

① 从表 1 和表 2 来看，基于 MLD 的机会不平等占比要大幅低于 Gini，如此结果是合理的。正是由于 MLD 关于异常值的敏感度问题，Brunori 等（2019a）更加推荐使用 Gini 作为机会不平等的度量。虽然两者绝对量值存在差距，但是趋势理应保持一致，这也是本文同时汇报两种指标的原因。

解的线性约束，从而构建能够对树模型和集成树的测算结果作出良好解释的模型工具。具体采用基于机器学习的 Shapley 分解：SHAP（Shapley Additive Explanations）方法进行贡献度测算。另外，SHAP 方法的目的是测算每一特征对于预测值的贡献度，因此需要将加总指标（不平等指数）分解到个体层面，并作为结果变量带入 SHAP 方法，具体采用 Liao（2022）针对基尼系数的分解方法。

本文关于环境因素贡献度测算的具体步骤如下：（1）根据原始收入计算总体不平等 Gini。（2）根据上述不平等指标的个体贡献度分解方法，将全样本 Gini 按照户籍（城镇和农村）分解至个体层面获得g_{ik}、g_{wik}和g_{bik}，分别对应全样本、城乡内部和城乡间不平等贡献度。（3）以个体不平等贡献度作为结果变量，环境因素作为预测变量，使用 SHAP 方法获得每一环境变量在个体层面的预测贡献度。（4）由于 SHAP 方法测算的贡献度具有可加特征，因此可以直接从结果变量中剔除某一类环境因素的预测贡献度，随后通过全样本加总得到剔除特定环境因素后的新 Gini。（5）与第一步的 Gini 进行比较，若新 Gini 更大，则说明该环境因素能够降低机会不平等，反之则表明其会拉大收入差距。基于此，我们仍然将环境因素划分为个体特征、父亲特征和母亲特征，分别计算剔除因素前后新 Gini 及其变动比例，相关结果如表 4 所示。

根据表 4 不难发现，个体特征和父亲特征一直扮演拉大收入差距的角色，并且后者影响力持续大于前者，而母亲特征则普遍具有相反的影响。具体来看，剔除个体特征后除 2016 年外其余 6 年的 Gini 均有不同程度的下降，7 年平均变动比例为-0.019；剔除父亲特征后，除 2020 年外其余年份的变动比例均为负值，Gini 的平均下降幅度为-0.021；母亲特征的影响则展现出完全相反的变动趋势，除 2016 和 2020 年外，剩余时间点变动比例均大于零，剔除后的 Gini 的平均增幅为 0.014。

表 4　　环境因素贡献度测算结果

组别	年份	总体不平等	剔除个体特征		剔除父亲特征		剔除母亲特征	
		Gini	Gini	变动比例	Gini	变动比例	Gini	变动比例
全样本	2009	0.612	0.577	-0.056	0.577	-0.057	0.666	0.088
	2011	0.534	0.530	-0.007	0.523	-0.021	0.540	0.010
	2012	0.514	0.510	-0.008	0.508	-0.012	0.515	0.001
	2014	0.612	0.593	-0.030	0.599	-0.021	0.624	0.020
	2016	0.615	0.617	0.003	0.610	-0.009	0.606	-0.015
	2017	0.598	0.595	-0.005	0.576	-0.037	0.605	0.011
	2020	0.640	0.622	-0.028	0.647	0.011	0.628	-0.019
城乡内部	2009	0.267	0.257	-0.036	0.256	-0.041	0.279	0.047
	2011	0.241	0.240	-0.002	0.236	-0.018	0.241	0.003
	2012	0.236	0.234	-0.010	0.235	-0.005	0.235	-0.003
	2014	0.300	0.290	-0.032	0.301	0.005	0.297	-0.010
	2016	0.282	0.284	0.007	0.280	-0.009	0.277	-0.018
	2017	0.284	0.284	-0.002	0.273	-0.039	0.286	0.007
	2020	0.317	0.311	-0.016	0.315	-0.004	0.312	-0.015
城乡间	2009	0.345	0.317	-0.082	0.327	-0.053	0.384	0.114
	2011	0.294	0.291	-0.010	0.289	-0.017	0.296	0.008
	2012	0.278	0.276	-0.008	0.274	-0.016	0.279	0.003

	2014	0.312	0.304	-0.027	0.297	-0.050	0.326	0.045
	2016	0.333	0.333	0.000	0.330	-0.008	0.328	-0.013
	2017	0.314	0.313	-0.004	0.301	-0.042	0.319	0.017
	2020	0.324	0.315	-0.026	0.326	0.008	0.317	-0.021

仅从符号上观察，首先针对城乡内部不平等而言，个体特征和父亲特征的影响与全样本相似，母亲特征则略微复杂，剔除个体特征后城乡内部 Gini 平均下降了 0.013，剔除父亲特征的下降幅度稍大，约为 0.016，剔除母亲特征后 Gini 平均增加 0.002，但在各年份表现为正负各半的分化状况；其次针对城乡不平等而言，个体特征和父亲特征的差异能够拉大城乡间收入差距，母亲特征的差异则更倾向于降低不平等，其具体数值表现为，剔除个体特征后 Gini 平均变动比例为-0.022，剔除父亲特征后 Gini 平均降低了-0.025，剔除母亲特征却导致 Gini 有着 0.022 的平均增幅。

考虑不同类型环境因素关于机会不平等的贡献度，似乎可以发现个体和父亲特征多数情况下一直在拉大收入差距，而母亲特征则缓解了机会不平等。正如汪晨等（2020）所述，性别作为引致机会不平等的重要因素，女性由于生育、产假问题所面临的就业压力和职场歧视并未得到缓解。作为家庭收入的主要创造者，父亲特征很大程度上决定了子代早期家庭的生活水平，进而引致后期教育投资等方面的差距。而在本文样本中，父亲特征对于平均收入线以上群体的积极影响较大，说明父代特征对于子代收入格局形成的影响非常明显。关于母亲特征，实证样本中母亲在理论上能够创造收入优势的特征上均要弱于父亲，例如母亲未上学比例为 0.562，父亲则为 0.401，正如表 4 所示，如此规律将导致母亲关于子代收入差距的影响力度普遍弱于父亲特征。在样本中发现虽然大部分情况下母亲特征对子代收入具有负向影响，但是对于平均收入线以下群体的影响力要弱一些，最终导致其降低机会不平等的实证结果。

四、机会不平等的新视角：从均值不平等到分布不平等

到目前为止，我们应用新方法对中国居民收入机会不平等的量值进行了全新的测算，进一步，我们将尝试从分布结构这一新的视角揭示机会不平等。Callaway 和 Huang（2020）认为现有关于代际收入流动的研究通常只关注后代收入的均值，而忽视了收入分布等其他更有价值的信息。他们估计了子代收入的反事实条件（父代收入）分布，测算包括均值、方差、分位数在内的多种分布特征，发现高收入家庭的子代未来不仅能够获得较高收入（均值），而且取得高收入的风险更低（方差）。本质上，代际收入流动的大小决定了机会不平等的程度，而既有机会不平等的相关文献，只关注收入均值的机会不平等，忽略了环境因素引致的收入分布差距这一更为重要的事实。因此上述研究从另一视角提供了补充机会不平等研究内容的重要思路，基于此，我们尝试吸纳 Callaway 和 Huang（2020）关于父代收入与子代收入分布关系的研究内容，从而将收入均值的机会不平等拓展至收入分布的机会不平等，更为全面得探究环境因素对子代收入分布的影响，丰富机会不平等的外延。具体的，下文将复原环境因素所引致的子代收入分布，从收入下限、收入上限、偶然收入以及收入风险四个方面评估收入分布的机会不平等。

（一）模型介绍与指标设计

1.分位数回归森林

这里，我们希望基于环境因素预测个体收入分布：

$$F(Income|Circumstances) \tag{6}$$

其中，$F(\cdot)$为收入的累积概率分布函数。进一步可将其转化为对于分位数的估计：

$$Q_\tau = inf\{y: F(Income|Circumstances = circums) \geq \tau\} \tag{7}$$

大多数经典的机器学习模型，例如回归树和随机森林，只关注结果变量均值的预测，忽视了其他分布特征，据此，Meinshausen（2006）提出了满足一致性估计的分位数回归森林（Quantile Regression Forests，DRF）。以$T(\theta)$表示回归树，其中θ为决定树生成的随机参数向量，一组确定的θ代表一棵已生成的回归树。R_ℓ表示叶子节点ℓ（$\ell = 1, \cdots, L$）对应的特征空间。对于样本特征x在特定回归树上所处的叶子节点记为$\ell(x, \theta)$。

回归树通过平均$\ell(x, \theta)$处的观测值获得对于新数据$X = x$的预测：

$$\hat{\mu}(x) = \sum_{i=1}^{n} \omega_i(x, \theta) Y_i \tag{8}$$

其中，$\omega_i(x, \theta)$为权重。将回归树推广至随机森林时，预测公式依然为上式，只不过权重设为在所有回归树上的平均权重：

$$\omega_i(x) = \frac{1}{k} \sum_{t=1}^{k} \omega_i(x, \theta_t) \tag{9}$$

其中，k为回归树数量。

DRF 在给定$X = x$时，根据所属叶子节点的样本进行分位数估计。Y的条件分布为：

$$F(y|X = x) = P(Y \leq y|X = x) = E\left(1_{\{Y \leq y\}}|X = x\right) \tag{10}$$

引入权重向量，经验分布函数可以基于下式计算得到：

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} \omega_i(x) \times 1_{\{Y \leq y\}} \tag{11}$$

根据分位数定义：

$$Q_\tau = inf\{y: F(y|X = x) \geq \tau\} \tag{12}$$

将经验分布函数带入即可获得分位数的预测结果。DRF 与传统随机森林的关键区别在于，对于每棵树的每一节点，传统随机森林只关注落入此节点样本的结果变量均值，而 DRF 关注包括均值在内的整个结果变量的条件分布。

2.收入分布机会不平等指标设计

当获得基于环境因素的条件收入分布后，本文希望尽可能全面地呈现除均值外的收入分布信息，从而完整地刻画收入分布的横截面差距。此处设计了四种指标用以描述环境因素引致的收入分布特征。

第一，收入下限。描述个体一定概率下取得的最低收入，收入下限越低，说明收入下降空间越大。根据分位数将收入下限Y_{Low}定义为：

$$F(Y \leq Y_{Low}|X = x) = \tau_{Low}；\ F(Y \geq Y_{Low}|X = x) = 1 - \tau_{Low} \tag{13}$$

其中，τ_{Low}为收入低于Y_{Low}的概率。

第二，收入上限。描述个体一定概率下所能获得的最高收入，收入上限越高，说明收入增长空间越大。根据分位数将收入上限Y_{High}定义为：

$$F\left(Y \geq Y_{High} | X = x\right) = \tau_{High};\ F\left(Y \leq Y_{High} | X = x\right) = 1 - \tau_{High} \tag{14}$$

其中，τ_{High}为收入高于Y_{High}的概率。实证过程中设定τ_{Low}和τ_{High}为 10%，即Y_{Low}和Y_{High}分别为个体收入分布的 10%和 90%分位数。

第三，偶然收入。本文额外计算了偶然收入Y_{Accid}，即 50%可能性下个体所能够获得的收入水平：

$$F(Y \leq Y_{Accid} | X = x) = F(Y \geq Y_{Accid} | X = x) = 50\% \tag{15}$$

第四，收入风险。收入风险用于描述环境因素所产生的个体收入的整体波动程度，采用标准差进行度量。$\hat{\mu}$为样本平均收入，风险测算公式如下：

$$SD(Y|x) = [Var(Y|X = x)]^{\frac{1}{2}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2\right]^{\frac{1}{2}} \tag{16}$$

（二）收入分布的机会不平等测算

基于上述四种收入分布特征指标的估计结果，本文进一步使用 Gini 和 MLD 两种不平等指标，探究收入分布（而非收入均值）层面的机会不平等量化及其演变规律，表 5~表 8 汇报了相关测算结果，其中，为了便于对比分析，复制表 1 和表 2 关于收入均值机会不平等的测算结果至各表第三列。

1.全样本收入分布的机会不平等测算

根据表 5 测算结果，可以归纳如下发现：（1）环境因素引致的收入分布机会不平等不止体现为收入均值不平等，在四种分布特征上均展现出更高的横截面差距。具体来看，所有年份，任意一种不平等指标，收入上限、收入下限、偶然收入以及收入风险的不平等的量值均要高于收入均值。最大差距体现在收入风险上，以 Gini 为例，样本期内平均是收入均值的 1.933 倍。最小为偶然收入，但是平均也达到了 1.392 倍。MLD 差距更大，四种分布特征均超过 2 倍。（2）横向对比，收入下限和收入风险不平等明显更为严重。根据表 5 测算结果，样本收入下限和收入风险的 Gini 和 MLD 普遍高于收入上限和偶然收入。以标准差度量的收入风险刻画了环境因素对收入整体不确定性的影响，10%分位点所度量的收入下限本质上也是一种风险，描述了收入下降的空间，所以相关结果意味着环境因素导致的子代收入风险差距需要受到重点关注。

表 5　　全样本收入分布的机会不平等测算结果

年份	不平等指标	收入均值	收入上限	收入下限	偶然收入	收入风险
2009	Gini	0.285	0.436	0.437	0.396	0.562
	MLD	0.128	0.327	0.337	0.265	0.538
2011	Gini	0.263	0.341	0.452	0.366	0.343
	MLD	0.110	0.192	0.368	0.230	0.192
2012	Gini	0.269	0.322	0.461	0.364	0.357
	MLD	0.115	0.173	0.396	0.235	0.205
2014	Gini	0.254	0.377	0.468	0.359	0.596

	MLD	0.100	0.252	0.406	0.231	0.617
2016	Gini	0.307	0.365	0.512	0.394	0.629
	MLD	0.152	0.225	0.520	0.280	0.711
2017	Gini	0.307	0.443	0.514	0.399	0.550
	MLD	0.153	0.342	0.529	0.286	0.524
2020	Gini	0.244	0.463	0.484	0.395	0.668
	MLD	0.099	0.380	0.479	0.281	0.847

2.城镇和农村收入分布的机会不平等测算

城镇内部的收入分布机会不平等展现出与全样本相似的统计特征。表 6 显示收入上限、收入下限、偶然收入以及收入风险的不平等程度均要明显高于收入均值。收入风险不平等相较收入均值差距最大，Gini 平均达到了 2.160 倍，MLD 高达 4.598 倍。偶然收入差距最小，但是 Gini 平均也有 1.467 倍，MLD 为 2.139 倍。再次说明仅关注环境因素产生的收入均值不平等远远不够，源于个体特征、父亲特征以及母亲特征的子代收入分布不平等更加严重。收入下限与收入风险所揭示出的收入下降空间与收入整体波动性差距仍然相当显著。两者对应的不平等数值普遍高于其他分布特征。特别是收入风险，无论 Gini 还是 MLD，与其他分布特征的差距非常显眼，说明城镇样本间在收入整体波动性与下降空间上的差距较为明显，收入不确定性的机会不平等程度更为突出。

表 6　　城镇收入分布的机会不平等测算结果

年份	不平等指标	收入均值	收入上限	收入下限	偶然收入	收入风险
2009	Gini	0.245	0.470	0.363	0.354	0.596
	MLD	0.110	0.390	0.275	0.227	0.620
2011	Gini	0.222	0.350	0.383	0.330	0.369
	MLD	0.093	0.206	0.307	0.198	0.222
2012	Gini	0.238	0.349	0.381	0.325	0.379
	MLD	0.106	0.202	0.333	0.196	0.234
2014	Gini	0.240	0.373	0.410	0.335	0.537
	MLD	0.102	0.240	0.368	0.208	0.489
2016	Gini	0.257	0.397	0.419	0.352	0.612
	MLD	0.125	0.266	0.421	0.236	0.659
2017	Gini	0.267	0.388	0.429	0.368	0.458
	MLD	0.134	0.250	0.434	0.251	0.360
2020	Gini	0.200	0.476	0.400	0.364	0.620
	MLD	0.076	0.395	0.360	0.238	0.699

全样本和城镇收入分布的机会不平等特征也体现在农村内部和城乡间。表 7 和表 8 共同说明环境因素对子代收入分布的影响不止停留在均值层面，包括收入下降与增长空间、收入整体波动性在内的其他分布特征差距要显著高于收入均值。以农村样本为例，收入下限与收入风险的 Gini 系数平均约为收入均值的 2 倍，MLD 分别超过 6 倍和 4 倍。与此同时，收入下降空间与整体波动所代表的收入不确定性在横截面上的差距比较明显，环境因素所产生的收入风险差距说明隐藏在传统收入不平等背后的风险不平等尤其严重。

表 7　　农村收入分布的机会不平等测算结果

年份	不平等指标	收入均值	收入上限	收入下限	偶然收入	收入风险
2009	Gini	0.209	0.326	0.370	0.337	0.477
	MLD	0.068	0.177	0.223	0.187	0.376
2011	Gini	0.212	0.282	0.404	0.327	0.278
	MLD	0.069	0.129	0.272	0.178	0.125
2012	Gini	0.213	0.258	0.413	0.338	0.315
	MLD	0.071	0.110	0.295	0.199	0.160
2014	Gini	0.196	0.355	0.417	0.331	0.628
	MLD	0.06	0.238	0.306	0.195	0.694
2016	Gini	0.257	0.284	0.482	0.353	0.639
	MLD	0.102	0.137	0.412	0.222	0.748
2017	Gini	0.263	0.465	0.491	0.364	0.600
	MLD	0.109	0.395	0.440	0.236	0.634
2020	Gini	0.241	0.430	0.501	0.386	0.694
	MLD	0.092	0.337	0.479	0.265	0.943

表 8　　城乡间收入分布的机会不平等测算结果

年份	收入均值	收入上限	收入下限	偶然收入	收入风险
2009	0.175	0.243	0.262	0.228	0.302
2011	0.156	0.186	0.263	0.206	0.184
2012	0.158	0.175	0.271	0.200	0.187
2014	0.147	0.192	0.268	0.192	0.274
2016	0.181	0.199	0.299	0.222	0.304
2017	0.176	0.221	0.295	0.220	0.258
2020	0.130	0.238	0.260	0.207	0.317

3.收入分布机会不平等的演变趋势

将全样本收入分布机会不平等的测算结果绘制如图 5~图 8 所示。同样首先排除 2020 年，不难发现，2009~2017 年收入下限的机会不平等持续上升，样本期内 Gini 增加了约 18%，MLD 增幅达到 57%。收入上限与偶然收入近似呈现 V 型走势，与收入均值机会不平等的演变趋势相似，虽然样本期内波动幅度较大，但是后续基本回复至原有水平。收入风险的机会不平等则具有明显的倒 N 型波动趋势，下降—上升—下降的演变过程导致风险不平等同样基本回复至初期水平。当我们考虑 2020 年的测算结果时，可以发现收入上限和收入风险机会不平等出现了恶化，尤其是收入风险，相比 2017 年 Gini 和 MLD 分别增长了 21%和 62%，说明疫情冲击引致了更为显著的收入风险不平等，而收入下限和偶然收入的机会不平等程度与收入均值相似有所缓解[①]。

① 与前文相似，需要时刻注意样本量可能引致的潜在问题。

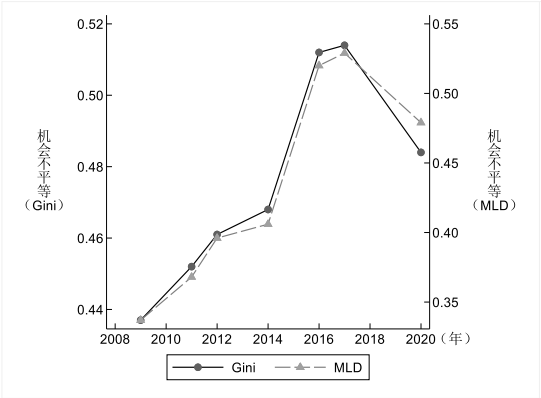

图 5 全样本收入下限机会不平等

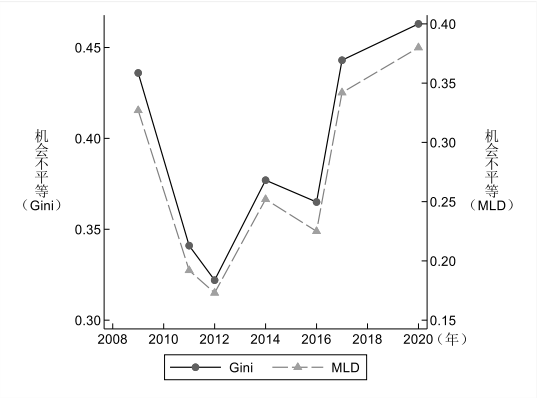

图 6 全样本收入上限机会不平等

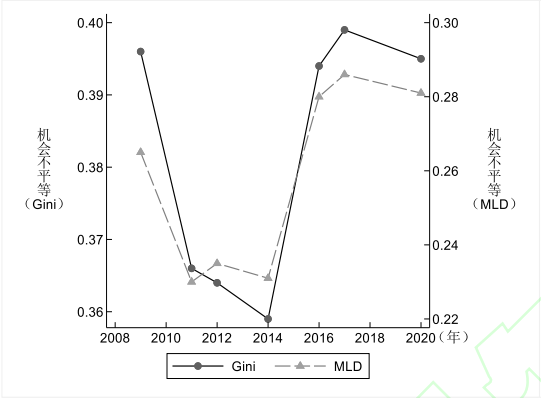

图 7 全样本偶然收入机会不平等

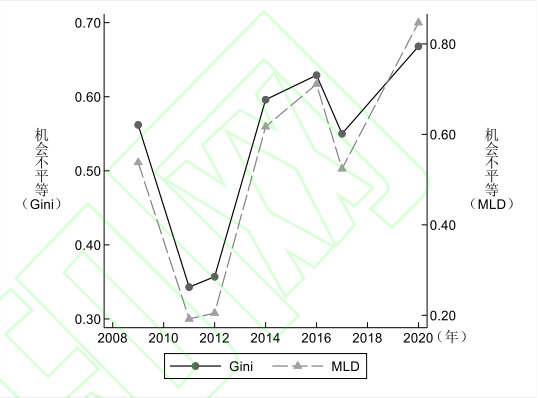

图 8 全样本收入风险机会不平等

五、研究结论与政策建议

有效解决个体层面一系列不可控环境因素所引发的机会不平等，是缩小收入差距、实现共同富裕的重要途径，更是推进中国式现代化的必需的治理抓手。鉴于现有研究在机会不平等的测算方法及其外延界定上所存在的诸多不足和重要缺陷，本文摒弃了以线性模型为基础的传统研究方法，尝试利用机器学习模型建立数据驱动的分析和预测方式，对中国当前机会不平等的量值和贡献份额进行精准量化，并从分布不平等视角拓宽机会不平等的外延。基于 2010~2021 年 CGSS 数据再次测算中国居民收入的机会不平等，研究结论特别是部分量化评价结果归纳如下：

第一，以基尼系数作为不平等指标，样本期内中国居民收入均值的机会不平等大约为 0.244~0.307，大体贡献了总体不平等的 38.1%~52.4%，其测算结果明显高于采用线性模型的传统参数测算方法。从结构上看，城镇居民的机会不平等量值（0.200~0.267）高于农村（0.196~0.263），但是在各自对总体不平等的贡献程度方面，后者（30.2%~40.8%）低于前者（35.9%~54.0%），这表明城镇居民的收入差距更倾向于受到环境因素的影响，即就组内而言，城镇居民面临较为突出的机会不平等问题。值得注意的是，城乡间机会不平等占比（40.1%~57.0%）明显高于城镇和农村，意味着环境因素是导致城乡收入差距的重要原因。时间趋势上，新冠肺炎疫情发生之前无论全样本，还是城镇和农村，抑或城乡间，机会不平等绝对量的演变具有 V 型特征，不平等占比则呈现典型的 N 型走势，环境因素引发的机会不平等处于快速上升阶段。疫情冲击虽然加剧了总体不平等，但是却逆转了原有机会不平等

持续恶化的趋势，环境因素引发的收入差距明显缩小。另外，我们将 Shapley 分解拓展至非线性模型，利用 SHAP 方法对不同环境因素关于机会不平等的贡献度进行测算，发现个体和父亲特征持续拉大收入差距，母亲特征则具有相反的影响。

第二，环境因素引致的收入分布不平等明显高于收入均值不平等。本文认为仅仅探究环境因素对子代收入均值的影响明显过于狭隘，应该从收入分布角度丰富机会不平等的外延。利用分位数回归森林模型复原环境因素所产生的子代收入分布，从收入下限、收入上限、偶然收入以及收入风险四个方面全面刻画收入分布。相关研究结果表明，样本期内环境因素导致的收入分布呈现右移趋势，收入均值、收入下限、收入上限以及偶然收入明显增加，但是收入风险也在不断攀升。收入均值与另外四种收入分布特征明显呈现正相关关系，意味着环境因素为子代创造的机会优势不仅体现在收入均值上，包括收入下限、收入上限等在内的收入分布优势也十分明显。根据收入分布测算结果评估机会不平等，发现四种收入分布特征的不平等程度均明显高于收入均值，其中以收入下限和收入风险尤为突出。以上结果说明传统机会不平等研究忽视了环境因素引发的更加严重的分布机会不平等，个体和父代特征对子代收入下降和增长空间以及收入整体波动性的影响更为突出。

结合文章研究结论，从机会不平等与共同富裕实现角度提出政策建议如下：

第一，消除环境因素引致的收入不平等是未来推动共同富裕的核心抓手。既有研究虽然关注到环境因素差异所引致的中国居民收入的机会不平等，但存在贡献度低估和单一化解读的现象。当前中国相关治理手段更多地聚焦于再分配调节，这种末段处置方式靶向可观测且可执行的收入来源进行转移支付，必然导致调节群体与政策目标的偏离，典型的，最被关注的个人所得税基本沦为对工薪（含其他劳动）所得课税而非调节财富积累的工具，因此我们建议，推动共同富裕仍然要从不平等的微观机制出发，推动前端和长效性治理，着力削弱源生性的（机会）不平等，改善财富积累和分配环境，从而促进具有明确价值导向的公平分配。

第二，将收入分布的机会不平等纳入到共同富裕的统计监测体系中。机会不平等毋庸置疑是收入分配测度的重要一环，也是共同富裕长效治理的发力关键点，但如文中所述，既有研究和社会关注点仍然主要集中于均值而非分布的差距，这导致对机会不平等的认知固化为总量或典型个体的同质性特征。然而，根据本文研究发现，个体无法控制的环境因素将在整个收入分布结构上塑造出显著的差距，忽略这种结构差异将导致不平等的监测体系出现持续潜在的致命缺陷。因此我们建议，将环境因素引致的个体收入风险、收入动态增长空间等分布特征纳入到共同富裕的测度当中，特别是要提供微观数据的获取度和代表性，并建立能够适用于大数据特征的算力体系和模型工具，根本上提高共同富裕量化监测的科学性和精准性。

第三，从收入和财富的动态分布视角完善共同富裕治理体系。提高居民收入水平与缩小收入差距始终是共同富裕治理面临的主要问题。然而，当前关于收入或者财富的理解通常被限定在静态层面，而收入的动态分布特征，包括收入增长空间及其可能性、收入整体波动性、极端收入风险以及收入的代际交叠等并没有体现在共同富裕治理体系当中。推进共同富裕实现过程中，在增加收入的同时还需要重点关注稳定收入预期、提高收入增长可能性、规避极端收入风险以及消除收入分配的阶层固化。我们建议，从治理层面建立缩小居民收入差距的一揽子系统化目标，重点从人口、收入以及外生环境的动态视角去审视微观个体所处环境因

素的差异，找准问题，找对抓手，从微观分布特征上针对性地削弱居民收入的机会不平等。

第四，基于科学测算建立居民收入分配与收入风险的预期管理体系。在能够有效监测机会不平等量化水平和收入分布结构变动态势的条件下，未来共同富裕推进工作将可能转型高质量精细化治理路径，重点结合微观主体的属性特征，进行实时或超前的治理干预，从而针对性削弱机会不平等的源生性矛盾和衍生性冲击，同时结合相关宣传体系，对居民预期作出有效管理，靶向性地消除长期性的负面情绪，这是也必将是推进国家治理体系现代化的有益探索。当然，这同样意味着中国收入分配调节工作必须从公共政策执行向系统化治理转变，因此我们建议，从学术研究、政策设计和实施等多元视角协同推进中国再分配政策体系的系统化改革，系统的视角审视收入分配问题的根源、要素、分布、趋势和微观归宿，从而前瞻性、针对性和跨周期地抑制源生性的机会不平等，消除财富要素结构性和制度性偏差，最终，根本性、长效性地削弱收入不平等，促进公平分配，实现共同富裕。

参 考 文 献

[1]龚锋,李智,雷欣.努力对机会不平等的影响:测度与比较[J].经济研究,2017,52(3):76~90.

[2]雷欣,贾亚丽,龚锋.机会不平等的衡量:参数测度法的应用与改进[J].统计研究,2018,35(4):73~85.

[3]李实,沈扬扬.中国农村居民收入分配中的机会不平等:2013—2018 年[J].农业经济问题,2022,(1):4~14.

[4]李实.因地制宜设计实现共同富裕的路径和政策[J].南方,2022,(5):16~17.

[5]李莹,吕光明.我国城镇居民收入分配机会不平等因何而生[J].统计研究,2018,35(9):67~78.

[6]李莹,吕光明.中国机会不平等的生成源泉与作用渠道研究[J].中国工业经济,2019,(9):60~78.

[7]史新杰,李实,陈天之,方师乐.机会公平视角的共同富裕——来自低收入群体的实证研究[J].经济研究,2022,57(9):99~115.

[8]史新杰,卫龙宝,方师乐,高叙文.中国收入分配中的机会不平等[J].管理世界,2018,34(3):27~37.

[9]孙豪,曹肖烨.收入分配制度协调与促进共同富裕路径[J].数量经济技术经济研究,2022,39(4):3~24.

[10]汪晨,张彤进,万广华.中国收入差距中的机会不均等[J].财贸经济,2020,41(4):66~81.

[11]姚鹏,李金泽.以水定城:资源节约型评比达标赛如何“去”资本错配[J].世界经济,2023,46(3):233~256

[12]Almås I., Cappelen A. W., Lind J. T., Sørensen E. Ø., Tungodden B., 2011, *Measuring Unfair in Equality* [J], Journal of Public Economics, 95 (7-8), 488~499.

[13]Athey S., Imbens G. W., 2019, *Machine Learning Methods That Economists Should Know About* [J], Annual Review of Economics, 11 (1), 685~725.

[14]Breiman L., 2001, *Random Forests* [J], Machine Learning, 45 (1), 5~32.

[15]Brunori P., Hufe P., Mahler D. G., 2018, *The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees* [R], World Bank Policy Research Working Paper, No. 8349.

[16]Brunori P., Palmisano F., Peragine V., 2019a, *Inequality of Opportunity in sub-Saharan Africa* [J], Applied Economics, 51 (60), 6428~6458.

[17]Brunori P., Peragine V., Serlenga L., 2019b, *Upward and Downward Bias when Measuring Inequality of Opportunity* [J], Social Choice and Welfare, 52 (4), 635~661.

[18]Callaway B., Huang W., 2020, *Distributional Effects of a Continuous Treatment with an Application on Intergenerational Mobility* [J], Oxford Bulletin of Economics and Statistics, 82 (4), 808~842.

[19]Checchi D., Peragine V., 2010, *Inequality of Opportunity in Italy* [J], Journal of Economic Inequality, 8

(4), 429~450.

[20] Ferreira F. H. G., Gignoux J., 2011, *The Measurement of Inequality of Opportunity: Theory and an Application to Latin America* [J], Review of Income and Wealth, 57 (4), 622~657.

[21] Fleurbaey M., Peragine V., 2013, *Ex Ante Versus Ex Post Equality of Opportunity* [J], Economica, 80 (317), 118~130.

[22] García J. L., Heckman J. J., Ziff A. L., 2018, *Gender Differences in the Benefits of an Influential Early Childhood Program* [J], European Economic Review, 109, 9~22.

[23] Hothorn T., Hornik K., Zeileis A., 2006, *Unbiased Recursive Partitioning: A Conditional Inference Framework* [J], Journal of Computational and Graphical Statistics, 15 (3), 651~674.

[24] Hufe P., Peichl A., Roemer J., Ungerer M., 2017, *Inequality of Income Acquisition: The Role of Childhood Circumstances* [J], Social Choice and Welfare, 49 (3/4), 499~544.

[25] Juárez F. W. C., Soloaga I., 2014, *IOP: Estimating Ex-Ante Inequality of Opportunity* [J], Stata Journal, 14 (4), 830~846.

[26] Kanbur R., Snell A., 2019, *Inequality Indices as Tests of Fairness* [J], Economic Journal, 129 (621), 2216~2239.

[27] Liao T. F., 2022, *Individual Components of Three Inequality Measures for Analyzing Shapes of Inequality* [J], Sociological Methods & Research, 51 (3), 1325~1356.

[28] Luna J. M., Gennatas E. D., Ungar L. H., Eaton E., Diffenderfer E. S., Jensen S. T., Simone II C. B., Friedman J. H., Solberg T. D., Valdes G., 2019, *Building More Accurate Decision Trees with the Additive Tree* [J], Proceedings of the National Academy of Sciences of the United States of America, 116 (40), 19887~19893.

[29] Mingers J., 1987, *Expert Systems-Rule Induction with Statistical Data* [J], Journal of the Operational Research Society, 38 (1), 39~47.

[30] Meinshausen N., 2006, *Quantile Regression Forests* [J], Journal of Machine Learning Research, 7, 983~999.

[31] Roemer J. E., Trannoy A., 2016, *Equality of Opportunity: Theory and Measurement* [J], Journal of Economic Literature, 54 (4), 1288~1332.

[32] Roemer J. E., 1998, *Equality of Opportunity* [M], Cambridge: Harvard University Press.