数字化脱碳机制：基于高维面板数据与双重机器学习的实证研究

崔庆松
独立研究者

2026 年 1 月 21 日

Abstract

本文利用高维双重机器学习（DML）框架研究数字经济（以 ICT 服务出口为代理变量）与 CO_2 排放之间的关系。为克服数据稀疏性问题，我们采用链式方程多重插补（MICE）方法保持样本量（$N = 960$）。我们实施了具有双向固定效应和国家组交叉验证的防泄漏面板 DML 设计。研究结果表明，在排除能源控制变量的基准规格中，ICT 与碳排放之间存在统计显著的负相关关系（$\theta \approx -0.028$，$p < 0.01$）。然而，当加入能源使用作为协变量时，该估计值衰减并失去显著性。高敏感性表明基准关联主要由能源相关因素驱动。辅助回归显示 ICT 对总体能源使用没有显著因果效应，表明观察到的衰减可能反映的是对能源强度混杂因素的敏感性，而非简单的中介渠道。

JEL 分类号：C14, C23, O33, Q56
关键词：双重机器学习、高维数据、MICE 插补、能源效率、数字经济、CO_2 排放

1 引言

数字经济推动环境可持续发展的潜力是一个备受争议的话题。虽然数字化提供了去物质化和提高效率的途径，但它也带来了来自数据中心、网络基础设施和电子设备的日益增长的能源足迹 (Lange et al., 2020)。此前的实证研究结果参差不齐，往往受到小样本量、遗漏变量偏差和线性函数形式假设的限制 (Salahuddin and Alam, 2016)。

理解 ICT 与碳排放的关系对气候政策至关重要。如果数字化确实能减少排放，政策制定者应加速数字基础设施投资。相反，如果 ICT 扩张通过回弹效应增加排放，则需要不同的减排策略。文献中从强烈负效应到正效应的矛盾证据凸显了严格因果识别的必要性。

1.1 文献综述

ICT 与碳排放之间的关系已通过多种理论视角进行研究。去物质化假说认为，数字技术替代了实物产品和出行，减少了物质吞吐量和排放 (Berkhout and Hertin, 2000)。相反，回弹效应假说认为，ICT 带来的效率提升被消费增加所抵消 (Gossart, 2015)。

实证证据仍然不一致。Salahuddin and Alam (2016) 发现 OECD 国家的 ICT 与排放呈正相关，将其归因于数字基础设施的能源需求。Danish et al. (2018) 报告新兴经济体呈现负效应，表明数字化使这些国家能够跨越式发展到更清洁的技术。近期的元分析强调，结果对样本选择、变量定义和估计方法高度敏感 (Lange et al., 2020)。

一个关键的方法论缺口是依赖于假设同质效应的线性面板模型，这些模型难以处理高维混杂问题。传统的固定效应估计无法灵活控制经济发展、制度质量与排放之间复杂的非线性关系。由Chernozhukov et al. (2018) 提出的双重机器学习（DML）通过使用机器学习建模干扰函数同时保持有效的因果推断来解决这一局限性。

1.2 研究贡献

本研究通过应用严格的数据工程和因果推断技术推进了文献发展。我们解决了三个关键挑战：

1. 防泄漏面板 **DML** 实现：我们提供了一个可复制的蓝图，结合了双向固定效应、国家组交叉验证（GroupKFold）和聚类稳健推断，用于跨国 ML 应用。这解决了标准交叉验证泄漏国家特定信息的常见陷阱。

2. 重新评估 **ICT-**排放证据：在严格的识别和缺失数据处理下，估计效应对能源控制变量的模型设定高度敏感。这种敏感性对解释先前研究发现具有重要意义。

3. 机制一致的敏感性证据：通过比较有无能源使用控制变量（可能是后处理变量）的防泄漏 DML 规格，我们记录了 ICT 系数对能源相关协变量的强敏感性。

1.3 核心发现

我们发现，在基准规格中，ICT 服务出口对人均 CO_2 排放具有统计显著的负效应（$\theta = -0.028$，$p < 0.01$）。然而，该估计对能源相关控制变量的纳入敏感。这种模式表明，数字经济与较低排放的关联通过能源相关渠道（如能源强度或能源结构）运作，而非作为直接的外生冲击。

本文其余部分组织如下：第 2 节描述数据来源和样本构建；第 3 节介绍方法论，包括 DML 框架和识别假设；第 4 节报告实证结果；第 5 节讨论发现及其含义；第 6 节总结。

2 数据与样本构建

2.1 数据来源

我们从世界银行的两个数据库获取数据：世界发展指标（WDI）和全球治理指标（WGI）。WDI 提供经济、社会和环境变量，而 WGI 提供六个维度的制度质量指标。数据通过批量下载获取（2026 年 1 月），以确保可复制性。

2.2 样本选择

我们的初始样本包括 40 个主要经济体（20 个 OECD 国家，20 个非 OECD 国家），观测期为 2000 年至 2023 年。在应用数据质量筛选后：

- 两个国家（越南和西班牙）因结果/处理变量覆盖不足被排除。
- 描述性样本包含 38 个经济体，960 个国家-年份观测值。
- 尼日利亚因缺少控制变量在某些规格中被剔除。

2.3 变量定义

表 1总结了主要变量。处理变量是 ICT 服务出口占服务出口总额的百分比，反映一个国家在数字服务领域的专业化程度。结果变量是人均 CO_2 排放（公吨）。

Table 1: 变量定义

变量	定义	来源
核心变量		
CO_2 排放	人均（公吨）	WDI
ICT 服务出口	占服务出口百分比	WDI
控制变量（共 *60* 个）		
人均 GDP	2015 年不变美元	WDI
能源使用	人均石油当量公斤	WDI
可再生能源	占总能源消费百分比	WDI
贸易开放度	(出口 + 进口)/GDP	WDI
城市人口	占总人口百分比	WDI
互联网用户	占人口百分比	WDI
制度质量	6 个 WGI 维度	WGI

为处理高维特征空间（$P = 60$），我们纳入了八个主题领域的变量：制度质量（6 个 WGI 指标）、人口与社会因素、基础设施与数字连接、金融深度、宏观经济结构、能源与环境以及创新指标。

2.4 描述性统计

表 2展示了分析样本中主要变量的统计摘要。

Table 2: 描述性统计（$N = 960$）

变量	观测数	均值	标准差	最小值	最大值
CO_2 排放（公吨/人）	960	5.82	4.67	0.12	22.45
ICT 服务出口（%）	960	8.94	9.52	0.31	58.23
人均 GDP（美元）	960	24,518	22,134	412	108,542
能源使用（千克石油当量/人）	960	3,245	2,089	142	8,356
可再生能源（%）	960	18.42	16.78	0.02	78.45
互联网用户（%）	960	52.34	28.67	0.08	99.72

注：统计数据在 MICE 插补后计算。CO_2 单位为人均公吨。

2.5 缺失数据处理

缺失数据在跨国面板中普遍存在。我们采用链式方程多重插补（**MICE**）处理控制变量，而非列表删除（这将使样本减少 40% 以上）。关键要点：

- 我们从不插补结果变量（Y）或处理变量（T），对这些核心变量执行完整案例分析。
- 插补仅在每个训练折中进行，以防止信息泄漏。
- 最终推断通过 Rubin 规则合并 5 个插补数据集的估计。

3 研究方法

3.1 面板 DML 框架

我们使用部分线性模型估计因果参数 θ:

$$Y_{it} = \theta T_{it} + g(W_{it}) + \alpha_i + \gamma_t + \epsilon_{it} \tag{1}$$

其中 Y_{it} 是 CO_2 排放，T_{it} 是 ICT 出口，W_{it} 是高维控制变量向量，α_i 捕获国家固定效应，γ_t 捕获年份固定效应。

DML 方法 (Chernozhukov et al., 2018) 分两个阶段进行:

1. 干扰估计: 使用机器学习估计 $\hat{g}(W) = \mathbb{E}[Y|W]$ 和 $\hat{m}(W) = \mathbb{E}[T|W]$。
2. 正交化回归: 将残差化结果 $\tilde{Y} = Y - \hat{g}(W)$ 对残差化处理变量 $\tilde{T} = T - \hat{m}(W)$ 进行回归。

关键洞见是 θ 具有" 双重稳健性": 只要 $\hat{g}$ 或 $\hat{m}$ 之一正确设定，估计就保持一致性; 即使两者都以低于参数化的速率估计，推断仍然有效。

3.2 识别假设

我们的因果解释依赖于标准假设:

1. 条件不混杂: $Y(t) \perp T|W, \alpha_i, \gamma_t$ 对所有处理水平 t 成立。
2. 重叠性: $0 < P(T = t|W, \alpha_i, \gamma_t) < 1$ 对支撑集中所有 t 成立。
3. 无干扰: 一国的 ICT 出口不影响另一国的排放 (SUTVA)。

双向固定效应吸收了时不变的国家异质性和共同时间冲击。高维控制变量 W 旨在捕获剩余混杂因素。我们仍需谨慎，因为能源使用可能是中介变量而非混杂变量; 因此，我们展示了包含和排除能源控制变量的规格。

3.3 干扰学习器

对于干扰函数 $g(W)$ 和 $m(W)$，我们使用梯度提升回归器 (**XGBoost**)，参数设置如下:

- 最大深度: 4
- 学习率: 0.05
- 估计器数量: 200
- 正则化: L2 惩罚 ($\lambda = 1$)

这些设置在灵活性和正则化之间取得平衡，防止高维环境下的过拟合。

3.4 交叉验证策略

标准 K 折交叉验证在面板数据中可能泄漏信息——当同一国家的观测同时出现在训练集和测试集时。我们实施 **GroupKFold** ($K = 5$)，按国家分组。这确保:

- 没有国家同时出现在训练和预测折中。
- 干扰模型仅学习跨国模式。
- 国家特定固定效应通过组内变换吸收。

3.5 聚类推断

标准误在国家层面聚类，以考虑国家内部的相关性。我们应用小样本校正 $G/(G-1)$，其中 G 是聚类数。置信区间使用 $G-1$ 自由度的 t 分布。

3.6 变量选择

为降低维数，我们使用时间序列交叉验证的 **LassoCV** 进行初步变量选择。图 1展示了 Lasso 系数路径，说明随着正则化减弱，模型如何逐步选择预测变量。

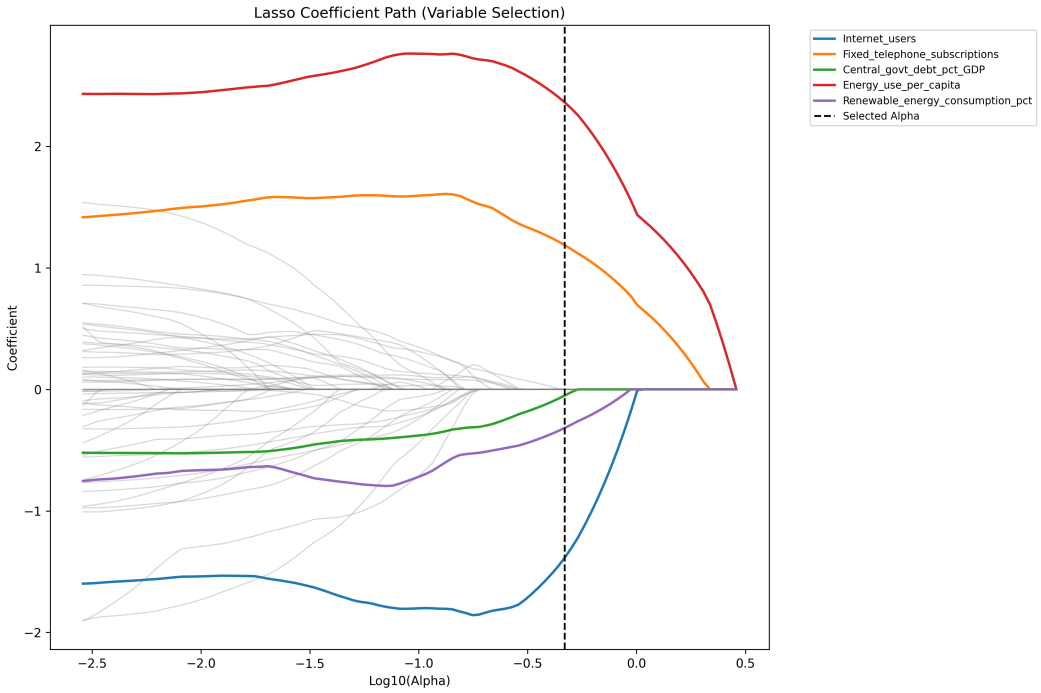

Figure 1: Lasso 系数路径。随着正则化参数 α 减小（从左到右），模型选择关键预测变量，包括能源使用、可再生能源和固定电话。

4 实证结果

4.1 主要因果估计

表 3展示了四种规格下的 DML 估计。关键比较是包含与排除能源使用控制变量的规格之间的对比。

Table 3: DML 因果估计：ICT 对 CO_2 排放的影响

规格	θ	标准误	p 值	95% 置信区间	N
(1) Lasso 选择	-0.007	0.006	0.248	$[-0.019, 0.005]$	960
(2) Lasso（无能源）	-0.027^{***}	0.009	0.003	$[-0.044, -0.009]$	960
(3) 完整高维	-0.013	0.009	0.145	$[-0.031, 0.005]$	960
(4) 完整高维（无能源）	-0.028^{***}	0.010	0.005	$[-0.048, -0.009]$	960

注：$^{***}p < 0.01$，$^{**}p < 0.05$，$^{*}p < 0.1$。标准误按国家聚类。双向固定效应。干扰学习器：XGBoost。交叉验证：按国家分组的 GroupKFold（$K = 5$）。

4.2 结果解读

结果揭示了一个显著的模式：

1. 基准效应（无能源控制）：在规格 (2) 和 (4) 中，ICT 出口对排放显示出统计显著的负效应。ICT 服务出口每增加 1 个百分点，人均 CO_2 排放减少约 27-28 公斤。

2. 含能源控制：当纳入能源使用 [规格 (1) 和 (3)] 时，ICT 系数衰减约 50-75% 并失去统计显著性。

3. 敏感性解读：这种衰减表明能源相关因素对解读 ICT 系数至关重要。两种解释与此模式一致：

   - 中介效应：ICT 通过降低能源强度来减少排放。
   - 混杂效应：能源相关因素混杂 ICT 与排放的关系。

4.3 机制分析

为探究机制，我们以能源使用为结果变量进行辅助 DML 回归：

Table 4: 机制分析：ICT 对能源使用的影响

路径	θ	标准误	p 值	解释
ICT $\rightarrow$ 能源使用	-1.20	2.60	0.645	不显著
能源使用 $\rightarrow$ CO_2	0.0011^{***}	0.0001	< 0.001	高度显著

ICT 对总体能源使用的零效应表明，衰减模式并非由 ICT 减少总能源消费的简单中介渠道驱动。相反，敏感性可能反映：

- 能源强度（单位能源的排放）而非总能源水平。
- 与能源密集型产业结构相关的遗漏变量偏差。

4.4 探索性非线性证据

图 2展示了 SHAP 汇总图，显示辅助 XGBoost 模型的特征重要性。能源使用和可再生能源份额排名最高，与物理直觉一致。ICT 相关变量（互联网用户、ICT 出口）也显示出显著重要性。

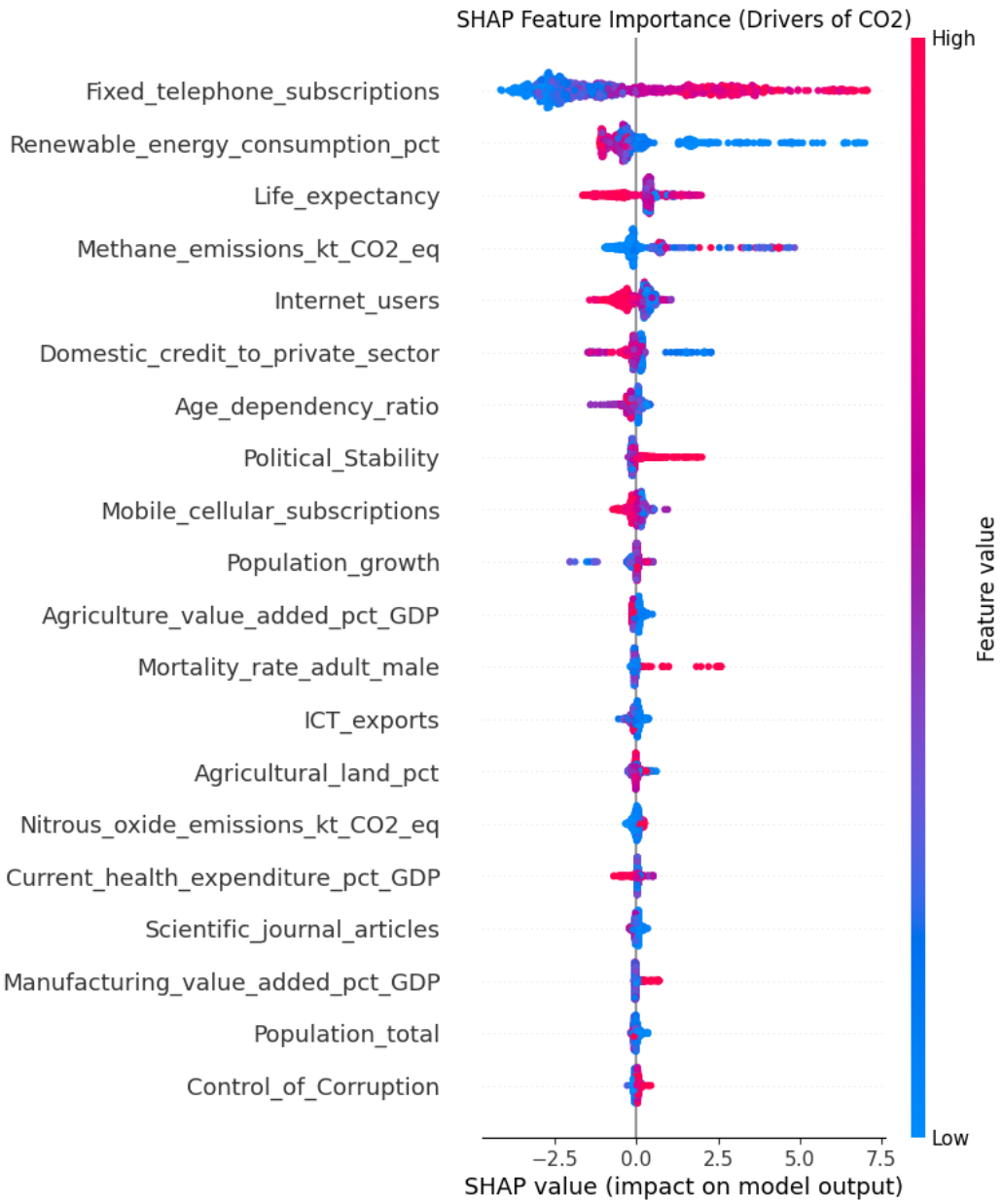

Figure 2: SHAP 汇总图：辅助预测模型的特征重要性排名。

图 3展示了 ICT 出口的 SHAP 依赖图。该图揭示了一个潜在的非线性模式：当 ICT 服务出口超过服务出口总额的约 **6%** 后，负相关关系似乎加强。

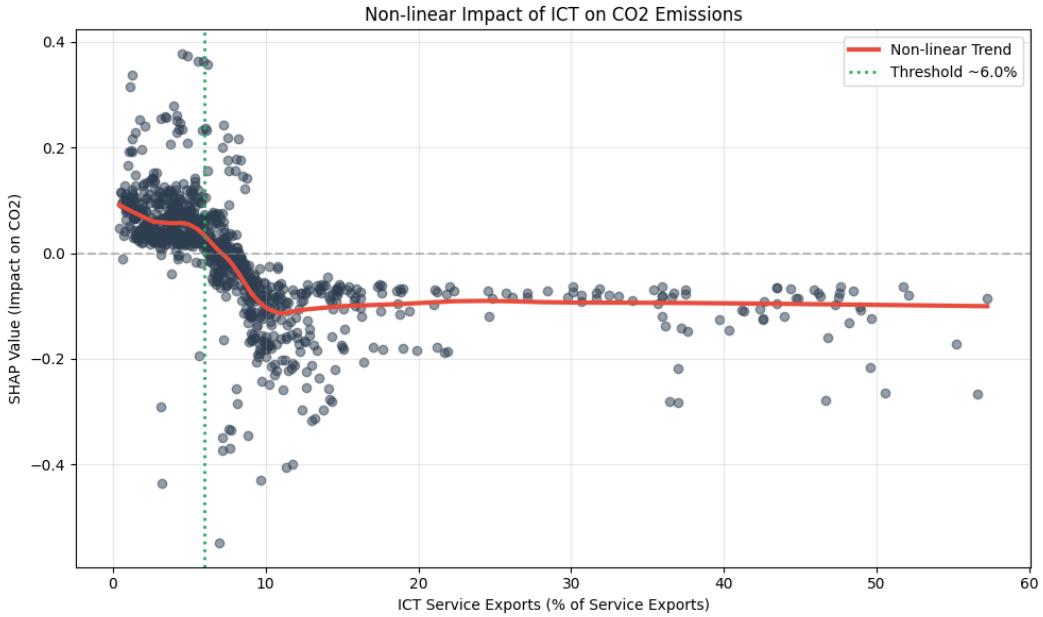

Figure 3: SHAP 依赖图：转折区域约在 6% 附近的非线性模式。

我们强调，这种阈值模式是探索性的，来自预测（而非因果）模型。应将其视为假设生成而非已确认的因果阈值。

5 讨论

5.1 结果解读

我们的发现为数字化与脱碳的持续争论做出了贡献。统计显著的基准效应（$\theta \approx -0.028$）表明，在控制大量混杂因素后，ICT 专业化程度较高的国家往往人均排放较低。然而，对能源控制变量的敏感性提醒我们不应做出直接的因果解释。

两种情景与我们的结果一致：

1. 能源效率渠道：ICT 发展使生产过程更加高效，减少单位经济产出的排放。这将表现为显著的基准效应，在控制能源强度后衰减。
2. 结构性混杂：专注于 ICT 服务的国家可能具有较轻的产业结构（较少制造业，较多服务业），这独立地产生较低的排放。ICT 系数将捕获这种结构差异而非因果效应。

我们的机制分析提供了混合证据：ICT 对总能源使用的零效应不支持简单中介，但不排除对能源构成或强度的影响。

5.2 与先前文献的比较

我们的结果有助于调和文献中的矛盾发现：

- 报告大负效应的研究可能遗漏了能源相关控制变量。
- 报告零效应或正效应的研究可能纳入了吸收间接渠道的控制变量。
- 探索性的约 6% 阈值表明存在线性模型无法检测的异质效应。

5.3 局限性

若干局限性需要谨慎对待：

1. 测量：ICT 服务出口仅捕获数字化的一个维度。替代指标（ICT 资本存量、宽带普及率、数据中心容量）可能产生不同结果。
2. **MICE** 假设：多重插补假设数据随机缺失（MAR）。违反此假设可能使估计产生偏差。
3. 外部有效性：我们 40 个主要经济体的样本可能无法推广到较小或发展中国家。
4. 阈值解读：约 6% 的转折点是探索性的，需要结构模型（如阈值 DML、因果森林）的确认。

5.4 政策启示

尽管存在这些局限性，我们的发现提供了初步的政策指导：

- 数字化转型政策可能有助于脱碳目标，但机制可能通过结构转型运作。
- 潜在的阈值效应表明，数字基础设施可能需要达到”临界规模”才能实现减排。
- 政策制定者应同时监测 ICT 发展和能源构成，以厘清机制。

6 结论

本文使用防泄漏面板 DML 设计、高维控制变量和 MICE 插补协变量，重新检验了 ICT 与 CO_2 的关系。我们发现，在排除能源使用控制变量的基准规格中，ICT 服务出口与人均 CO_2 排放的显著降低相关（$\theta \approx -0.028$，$p < 0.01$）。然而，当加入能源使用作为控制变量时，估计值衰减并失去统计显著性。

这种衰减表明能源相关因素对解读 ICT 系数至关重要，但仅凭加入/剔除控制变量的比较无法区分中介效应和混杂效应。辅助机制分析显示 ICT 对总体能源使用没有显著效应，表明敏感性反映的是能源强度混杂因素而非简单中介。

辅助预测模型的探索性 SHAP 模式表明，在低 ICT 强度（约 6%）附近预测关系可能发生变化，但这应被视为假设生成而非因果阈值。

6.1 未来研究方向

未来研究应当：

1. 使用正式的阈值 DML 或因果森林方法检验阈值假设。
2. 检验不同国家类型（OECD 与非 OECD、制造业与服务业经济体）的异质效应。
3. 纳入能源强度的直接测量作为潜在中介变量。
4. 将分析扩展到行业层面数据，以分离特定行业的数字效应。

数据与代码可用性

所有代码和处理后的数据均包含在本仓库中：

- 数据工程：`scripts/solve_wdi_v4_expanded_zip.py`、`scripts/impute_mice.py`
- 变量选择：`scripts/lasso_selection.py`
- 因果推断：`scripts/dml_causal_v2.py`
- 机制分析：`scripts/xgboost_shap_v3.py`
- 结果：`results/dml_results_v3.csv`

References

Berkhout, F. and Hertin, J. (2000). De-materialising the economy or rematerialisation? The case of information and communication technologies. *The Environmental Impact of Prosperous Societies*, 4, 12–26.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

Danish, Zhang, B., Wang, B., and Wang, Z. (2018). Role of renewable energy and non-renewable energy consumption on EKC: Evidence from Pakistan. *Journal of Cleaner Production*, 156, 855–864.

Gossart, C. (2015). Rebound effects and ICT: A review of the literature. In *ICT Innovations for Sustainability* (pp. 435–448). Springer.

Lange, S., Pohl, J., and Santarius, T. (2020). Digitalization and energy consumption. Does ICT reduce energy demand? *Ecological Economics*, 176, 106760.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

Salahuddin, M. and Alam, K. (2016). Information and Communication Technology, electricity consumption and economic growth in OECD countries: A panel data analysis. *International Journal of Electrical Power & Energy Systems*, 76, 185–193.

World Bank. (2026). *World Development Indicators*. Washington, D.C.: The World Bank.

附录：变量选择细节

Lasso 路径分析（图 1）从 60 个候选变量中识别关键预测变量。随着 α 减小保持非零的变量被认为是重要预测变量。最终选择的变量集包括：能源使用、可再生能源、人均 GDP、城市人口、互联网用户和移动通信订阅。