存储的家庭分配失效速度：来自北卡罗来纳州选民名册的证据

PRASHANT CHAUDHARY，美国独立研究员

评估家庭文件和身份图谱时，人们通常看它们是否仍能找到某个人。这样的检验忽略了它们所存储的信息。即使其中每条个人层面的关联都依然正确，家庭分配也可能失效，因为此人周围的人会发生变化。我们以选民登记册推导出的家庭为对象，通过持久标识符追踪北卡罗来纳州的登记者，衡量家庭分配失效的速度。追踪数据涵盖2008年11月至2026年11月的40个不同快照（2016年构建时，60个县共有360万名客户）。2016年构建完成八年后，对于仍列在名册中的登记者，图谱有0.831的概率找到其地址，但在找到的这些人中，家庭归属完全正确的仅占0.322（若将过时名录记录视为迁出，则该比例为0.309）。在距构建完成两年至八年的每个时间点，迁居者存储的家庭归属平均每五次仅有一次正确，因此时效性的损失既源于留居者，也源于留存家庭的变化。采用更严格的概率姓名地址匹配器阈值，准确率仅提升至约等于留居者的时效性，主要是因为筛选出了留居者。按构建时特征编制的时效性查找表，在留出县和后续文件上的布里尔分数，比仅使用证据年龄降低16%至19%，比使用证据年龄和家庭规模降低6%至10%。使用相同输入的家庭变更多状态模型，表现不及查找表；而查找表在选择应刷新文件中哪四分之一的记录时，并不比证据年龄和家庭规模更准确。在俄亥俄州和华盛顿州，八至九年后，查找表经过样本内截距偏移后仍保持原有排序；会计处理的数值水平在华盛顿州成立，在俄亥俄州则不成立。模拟的5%至21%构建误差会使八年时效性降低2%至10%。

CCS概念：•**信息系统** → **实体解析**；数据清理。

其他关键词：数据质量、数据时效性、家庭构成、记录链接、身份图谱、选民登记记录

arXiv:2611.32091v1 [cs.DB] 24 Nov 2026

1 引言

假设某个文件记录了每个人与哪些人共同生活，而且支撑这些记录的证据已有八年历史。这个系统仍然能够找到大多数人。但它返回的家庭，还是原来的家庭吗？

两个群体都会存储家庭信息。供应商出售家庭文件和身份图谱。研究人员和统计办公室根据行政地址构建家庭键，并将其关联到人员
多年来如此。两者都以匹配率（即系统仍能找到的人员所占比例）来评判结果。匹配率无法说明某人周围的人员情况，而这些人员会发生变化。成员会离开、去世、成年，或搬入一个始终未曾搬迁的人的家庭。本文衡量一个存储的家庭能够保持准确多久，以及一个根据旧证据作答的系统在无法继续准确时会返回什么：正确的家庭、错误的家庭，还是空无一物。

北卡罗来纳州的选民名册让我们能够回答名册所登记的家庭情况。以2016年登记册为基础，构建一张覆盖60个县、包含360万名客户的家庭关系图；客户是指系统需要查询其家庭的人员。八年后，对于仍已注册的客户，该图仍能通过地址找到其中83.1%。在它找到的客户中，它返回的已注册家庭与实际家庭完全一致的比例为32.2%（如果将登记册中的过时名录记录视为提前一个选举日离开，则为30.9%；见第6.1节）。按匹配率来看，这张图仍然有效；但从它返回的家庭来看，三分之二左右的答案已不再完全准确。若使用提前十个月的证据重建该图，则在各个评分日期上，它对匹配结果的准确率为75.5%至85.8%（八年时为80.0%；按同样的解读则为69.3%）。

有三类研究与这一问题较为接近。数据质量研究将存储值的时效性定义为：给定其存储时长，该值仍然有效的概率。该研究根据拟合或假定的寿命分布{v1}，估计某一实体的一个

作者联系信息：Prashant Chaudhary，独立研究员，圣克拉拉，加利福尼亚州，美国。

属性的这一概率 [43]，并将此类估计与观测有效性进行评估 [1, 8, 18]。链接研究将匹配率视为与错误链接之间权衡的一方 [57, 58, 84]。人口学研究则利用个人层面面板，以及将选民名册与消费者数据相结合的关联年度登记册，测量家庭构成如何变化。存储的家庭归属介于两者之间。它是一种群组值记录，而我们没有发现任何研究按照系统存储并返回该记录的方式来衡量其时效性：以地址或人员为锚点，以一个通过持久标识符跟踪人员的参照进行评分，并根据系统所回答的问题进行细分。

我们研究的系统通过在其证据中找到一个**锚点**，并返回构建日期时该锚点周围记录的家庭来作答。地址锚点解析客户当前的地址，并返回记录在该地址的人员。个人锚点找到客户本人在构建时的记录，并返回其周围记录的家庭。匹配率衡量该锚点是否仍可找到。准确性取决于锚点周围的家庭在评分日期是否仍是同一家庭。我们将其称为**时效性**。对于2016年构建，在两年时，仍在登记册中的登记者有0.954的地址可被找到；在这些地址匹配中，有0.701的家庭正确；八年时的数值如上所述。

我们做出四项贡献：

1. **针对群组值记录的纵向基准，以及据此评分的时效性**。我们通过持久标识符，追踪北卡罗来纳州登记者在41个快照文件（40份不同的提取数据）中的登记共同居住情况，不使用匹配器。退出情况按登记册中的移除原因分类，并以五种方式解读其过时名录记录。基于该参照，我们评估可追踪家庭中登记者的时效性，即系统锚点周围的家庭仍完全正确的概率；每位客户有三种结果（第3、4和6节）。

2. **时效性函数及其损失来源**。迁居者的存储的家庭在从两年至八年的每次建造年份中，约有五分之一是正确的。因此，损失取决于留居客户的占比，以及留居家庭内部的变化。在对过时名录记录的任何解读下，时效性都会随着证据年龄的增加而下降。该会计处理在留出县上进行了描述性复现，并且在华盛顿州的登记册中达到了其水平和分解容差，但在俄亥俄州的登记册中没有达到（第5、6和9节）。

3. **系统能做什么，以及不能做什么**。收紧我们所测量的概率姓名地址匹配器的阈值，只能将其匹配结果的精确匹配占比提高到约等于其所覆盖总体中留居者的时效性（公式（3））。这就是链接规则所施加的筛选的时效性形式 [86]。在其匹配的客户中，用完美的个人链接替换其个人链接，只会使精确匹配占比变化约0.001，这是因为匹配器的许多错误链接最终返回的仍是该客户自己的构建时家庭。年度重建对精确匹配占比的提升远大于任何阈值调整。模拟构建误差在较长证据年龄下造成的时效性损失小于其表面数值（第6.4和7节）。

4. **在留出数据上检验时效性预测**。 基于构建年份特征的时效性查找表、Heinrich 和 Klier [41] 的补充数据设计（采用集合值目标），在留出县以及后续文件（包括其中不属于开发集的县）上，相较于仅使用证据年龄，以及相较于同时使用证据年龄和家庭规模，都能降低布里尔分数。家庭变更多状态模型根据一步转移推导出时效性及其留存者和迁居者的划分，正如 Zak and Even [96] 对一个属性所做的那样。该模型优于较粗粒度表，但不及使用相同共同成员年龄的表格；其推导出的迁居者时效性也未能匹配测得的迁居者时效性。当只能刷新文件的四分之一时，两个表格均未能满足相对于证据年龄和家庭规模的标准（第8节和第9节）。

方法。 北卡罗来纳州为每位登记人分配一个持久标识符（NCID）。通过追踪2008年11月至2026年11月期间的41个公开快照文件（40个不同的提取文件）中的NCID，可以了解：对于某一日期拥有一条登记记录的每位登记人，谁与其共享登记的居住单元，以及每条记录何时发生变更或从名册中移除。我们的模拟供应商从相同的快照中构建家庭关系图，但不使用NCID。这使供应商在构建日期上达到精确，并能够分离出此后发生的情况。我们在随机抽取的20个县上开发了定义、策略和预测变量。我们在另外60个县上检验主要假设，在剩余20个县以及2026年11月文件上检验预测变量，并将其迁移到俄亥俄州和华盛顿州的选民档案上（第4.4节）。第2节回顾相关工作，第3节定义时效性及其分解，第4节介绍数据和设计，第5至7节报告测量结果，第8节介绍预测变量，第9节讨论其他登记册，第10节讨论其影响和局限。

2 相关研究

链接研究会在链接建立当天对其进行评分。历史人口普查的自动链接器描绘了匹配率与错误链接之间的前沿：Bailey 等人 [8] 发现，常用方法会产生大量错误链接，并主张不应最大化匹配率；Abramitzky 等人 [1] 表明，保守规则能够提高个人链接的精确率。健康数据链接将链接率与链接质量区分开来，并表明已链接子集是经过选择的 [38]。簇级评分 [2, 7,64], 对 F1 [21, 34] 的批评，以及群组链接的评估类别 [71] ，都在询问链接或聚类在形成时是否正确。我们要问的是：围绕一个正确链接人员的家庭，多年后是否仍然正确。

时效性是长期以来的数据质量维度 [9, 10, 12, 13, 81, 94]。时效性指标给出了在已知存储时长的情况下，存储值仍然有效的概率。该概率既可以根据带有外部速率的指数衰减推断 [44] ，也可以根据对数据拟合得到的寿命分布推断[42]。研究还会根据值被存储时已知的补充数据对其进行条件化 [41] ，并以观测有效性进行评估 [43]。Zak and Even [96] 将存储值背后的现实状态建模为连续时间马尔可夫链，根据拟合转移估计时效性，并选择重新获取数据的时机。

相关研究确定多个值中哪个是当前值 [27, 28], 设定更新频率 [83],追踪来源滞后和复制情况 [25], 清理过时值 [66]并将企业家庭归并视为数据质量任务 [62]。上述时效性指标评估的是单个实体存储的值。家庭归属是一组围绕某个实体的实体集合，即使每个人员层面的链接都正确，家庭归属仍可能过时。我们按照评分日期的有效性进行评分，这一做法沿袭了事件对齐会计处理的早期研究 [14], ；而我们对未评分退出的界限，则是针对缺失二元结果的最坏情况界限 [63]。

时间链接模型展示单个实体的属性如何随时间变化，从而将其记录跨时间链接起来 [15, 16, 20, 33, 59, 60], ；集体链接和群组链接则共同解析同时出现的参照 [11,77]。研究者已通过图匹配和时间聚类，将不同人口普查波次中的家庭关联起来[31, 32, 55, 70]。

与本文最接近的是 Christen 等人 [22] 跨越六次人口普查波次链接家庭，分类每个家庭的演变情况（保持不变、分裂、合并、迁移），并统计家庭在10年至50年间的保持情况。在该统计中，当一对一群体链接保留了某个家庭至少两名成员时，该家庭即被视为保持不变；统计按家庭加权，并承袭链接误差。本文新增的是客户侧视角：在年内分辨率下，针对持久标识符参照对精确集合进行评分，同时考察一个存储该集合的系统会返回什么结果。北卡罗来纳州选民档案是个人级链接和重复检测的标准基准 [19, 52, 78]。这些基准标记的是个人。我们标记的是从2008年至2026年间40个不同快照中的已注册共同居住情况。

研究人员和统计办公室也会存储家庭。家庭由共同地址行政记录构建 [36, 88], ，这些记录来自健康记录中的匿名化地址键 [50, 85] ，也可以由在选定事件日期分配的财产标识符构建 [35, 37]。Cuccaro-Alamin等人 [24] 综述了相关策略，统计办公室则以人口普查规模建立行政名册 [53, 67, 68, 87]。这样的键相当于供应商家庭文件的行政对应物：它在一个日期形成，却在另一个日期使用。行政地址变更滞后于实际迁移 [56], ，而健康登记册中的记录级地址与人口普查不一致，且涉及相当大比例的人员 [30], ，因此，登记册衡量的是登记共同居住情况。

在政治科学研究中，同一地址的共同登记是衡量家庭情况的成熟代理指标 [45]。选民档案已被用于验证调查、关联个人以及追踪记录的变化 [3, 4, 26, 54],，而选民登记对搬迁的反应存在滞后，且滞后时间因年龄而异 [5]。皮尤研究中心 [80]将一个全国面板与五个商业选民档案进行了匹配；匹配率介于 50% 至 79% 之间，而过去一年内搬过家的面板受访者中，只有 14% 能在全部五个档案中找到。消费者文件中的家庭字段 [79], 、第三方受众细分 [72, 73] 以及线下经纪商属性 [93] 往往不准确。商业居住历史记录已通过自我报告得到验证 [47],，IP到家庭分配则已通过运营商记录得到验证 [23]。这些研究都是在某个日期验证记录。没有一项研究追踪存储的分配随其证据年龄增长而发生的变化。

家庭构成变化也曾被直接测量。Richards 等人 [84] 对全国面板中不同家庭类型之间的转变拟合了竞争风险模型。Hays 等人 [39]发现，在北卡罗来纳州五个县的老年人面板中，四年间有 35% 的家庭构成发生了变化。Raley 等人 [82] 通过比较全国面板中儿童从出生到 18 岁期间家庭名册各轮调查之间持续共同成员标识符的集合变化，统计其家庭构成变化。本文使用的正是这种精确集合比较；他们将其用于调查名册，而我们将其用于存储的分配。整合选民名册与消费者数据的英国年度关联登记册显示，一年后有 73.7% 的家庭构成保持不变 [57], ，并提供小区域家庭变化指数 [58] 以及居住迁移估算 [91]。像这些登记册研究那样按姓名匹配居民，会将家庭变化与链接误差混为一谈。

人口预测模型通过转移率对家庭结构进行建模 [92, 97] ，并用后续人口普查数据验证预测总量，因此并不关注某个已识别家庭是否仍然完整。第8.2节使用多状态马尔可夫模型分析具有区间删失转移的面板数据 [49, 51, 89]。补充材料I介绍文献检索。

3 衡量家庭分配的时效性

3.1 锚点与时效性

系统在日期 t 建立证据，并在 $s = t + k$ 处理查询。对于锚点类型为 X（地址 A 或人员 P）的客户，**时效性** $\kappa_X(\mathrm{k})$ 表示这样的概率：在 t 围绕锚点记录的家庭，与在 s 的家庭完全相同，且两个集合中都不包括该客户。对于返回集合为 R_i、真实集合为 T_i 的客户 i，记作 $R_i' = R_i \setminus \{i\}$ 和 $T_{i}' = T_i \setminus \{i\}$；结果为

$$E_i = \mathbf{1}\left[R_i' = T_i'\right]. \qquad (1)$$

如果某客户在 t 被记录为独居，且其在 s 的真实家庭也是空集，我们就将其视为当前：即使客户已经搬家，“独居”仍是对家庭构成的正确描述。第5.3节将单独报告这些客户。这一约定适用于返回共同居住成员集合的系统；对于返回持久住户键的系统，则需要另行规定如何处理已经搬出且独居的客户。在 s 的人口名册中缺失的返回成员会导致匹配失败。

地址锚点的匹配率 P_A，下文称为**持续性** ，是指当前地址在 t时有登记人的客户所占比例。我们假设客户提供了正确的当前地址，因此 P_A

这是地址解析的最佳情形。个人锚点在此处通过构造得出，因为客户已经掌握了自身在建造时的身份信息。家庭和参照均来自同一登记册：我们衡量的是登记册衍生家庭的时效性，并以同一登记册后续的登记记录作为评分依据。除非表格另有说明， κ_A 和 κ_P 均针对仍列在 s 的客户进行评分， κ_A 则针对地址已找到的客户进行评分。

3.2 每位客户的三种结果

每位客户会得到三种结果之一：匹配到正确的家庭、匹配到错误的家庭，或没有答案。**有效性**指得到答案的客户中匹配到完全相同家庭的占比；全文件视角则给出所有客户中每种结果所占的比例（第7.4节）。我们将这三种结果分开处理。将前两种结果取平均并给予部分分数，或将没有答案计作此人单独成户的评分方式，会把它们混为一谈；两种常见惯例（自我包含评分，以及将未匹配记录计为 0）还可能颠倒哪项政策看起来更好（补充材料 G）。

3.3 留居者与迁移者

我们在 s 将客户分为**留居者**和**迁移者**：留居者的当前地址在 t 对应其所属的家庭，迁移者则是仍列在名册中的其他所有人。“对应”指系统可以进行的地址重新匹配（地址重新匹配规则）：将 s 记录的客户地址匹配到构建图中，得到该客户所属的构建时家庭。若客户所在单元未变，但 s 记录的地址格式不为构建图所包含（格式变更或县重新标记），则此处将其视为迁移者。第8.2节还采用另一条规则，即选民单位规则；该规则遵循参照中的选民单元，并将此类客户计为留居者。若 σ 为留居者占比， κ_S 为留居者的时效性， μ 为迁移者的时效性，则根据总概率公式可得

$$\kappa_P(k) = \sigma(k)\,\kappa_S(k) + \big(1 - \sigma(k)\big)\,\mu(k). \qquad (2)$$

方程（2）将登记册记录的居住流动 σ 与留居家庭内部的变化 κ_S 区分开来。乘积 $\sigma\kappa_S$ 表示存储的家庭和地址都仍然有效的客户占比。如果 μ 在不同构建年份和年龄段之间大致保持不变，那么个人锚点的下降就源于 σ 和 κ_S。第5节将说明事实确实如此。

3.4 筛选

更严格的接受规则 θ 回答的客户更少。用 M_θ 表示规则回答的客户，用 $V(k,\ \theta)$ 表示规则返回给他们的家庭中精确匹配的占比，用 σ_θ 表示其中留居者的占比。$\kappa_S(k \mid M_\theta)$ 和 $\mu(k \mid M_\theta)$ 分别表示已回答的留居者和迁居者中，返回的家庭精确匹配的占比，因此也包含错误个人链接的影响（第7.2节）。在 M_θ 内的总概率给出

$$V(k,\theta) = \sigma_\theta\,\kappa_S(k \mid M_\theta) + \big(1 - \sigma_\theta\big)\,\mu(k \mid M_\theta). \qquad (3)$$

两个条件可将方程（3）转化为一个上界。条件（i）：在留居者中，家庭是否发生变化与是否得到回复相互独立，且返回的家庭是客户自己的家庭，因此 $\kappa_S(k \mid M_\theta) = \kappa_S(k)$。条件（ii）：向匹配的迁居者返回的家庭，其完全准确的频率不高于留居者家庭的这一频率， $\mu(k \mid M_\theta) \le \kappa_S(k)$。在这两个条件下， $V(k,\theta) \le \kappa_S(k)$，当 $\sigma_\theta = 1$ 时取等号。要求客户姓名在同一单元中的规则会使 σ_θ 接近 1，因此 V 趋近于 κ_S：阈值趋于平台期。在实践中，条件（i）仅近似成立，因为婚姻或离婚导致的姓氏变化会降低姓名得分，并标记出发生变化的家庭。该上界适用于规则所回复的任何人群。仅回复客户

所属的某一种家庭类型的规则，因此会趋近于该类型留居者的时效性，这一数值可能超过κ_S对所有类型汇总后的结果。第7节对两者都进行了测量。

相较于 Heinrich 等人 [40] 对数据质量指标提出的要求，时效性具备以下性质：它是一个最小值为 0、最大值为 1 的概率，并且采用区间尺度。其配置包括五项明确规定的选择：证据年龄、精确匹配规则、单独迁移者规则、规定缺失的返回成员在何种情况下 s 会中断匹配的规则，以及规定哪些客户可被追踪的规则。除可追踪规则外，我们报告了每项规则的替代方案；可追踪规则的排除情况在补充材料 E 中统计。该指标从客户聚合到文件时采用按人数加权的平均值，并通过方程（2）从留居者和迁移者聚合到所有客户。它所需要的是纵向基准，而系统很少具备这一基准。因此，第8节检验了系统可以根据自身建造年份特征应用的预测变量，以及 Heinrich 和 Klier [41] 针对群组型目标提出的补充数据设计。

4 数据与研究设计

4.1 数据

北卡罗来纳州选举委员会发布选民登记快照，采用一个包含 90 列的架构，涵盖活跃登记者、非活跃登记者和已移除登记者，其中包含年龄但不含出生日期，也没有家庭字段 [75]。NCID 自 2008年11月起填充。我们使用了从 2008年11月到 2026年11月的 41 个快照文件：9 个偶数年选举日文件、18 个 1月1日文件、8 个奇数年 11 月文件、另外 5 个 2019 年文件（其中两个是同一提取文件），以及 2026 年 11 月文件，因此共有 40 个不同的提取文件。全州选民历史文件提供了第6.2节所使用的构建选举投票记录，而北卡罗来纳州公共地址点（NC OneMap）则用于核查该节中的单元键。构建以其年份命名。2026 年 11 月文件提供十年期数值、在后续文件上对预测变量的评估，以及第6.1节中需要 2026年1月之后文件的校正后的参照。第 9 节使用俄亥俄州（2017 年 7 月和 2026 年 10 月）及华盛顿州（2017 年 6 月、2018 年 6 月和 2026 年 10 月）的选民档案；这些档案的标识符在九年间保持不变，并且仅列出活跃登记者和非活跃登记者。密歇根州重新签发的标识符无法形成可用配对，因此我们不报告密歇根州的结果（第 9.1 节）。在进行任何比较之前，所有姓名和地址字段均转换为大写。

4.2 个人、家庭与客户文件

一个**人员** 即一个 NCID。在 s 的**登记人** 在 s有一条活跃或非活跃记录；登记人会被评分，而 s 客户文件中的其余人员则属于**退出者**。在建造年份有多条活跃或非活跃记录的人员不属于客户；在之后的快照中，少数拥有多条此类记录的登记人会按照其家庭键中最小的那个进行评分，并保留在家庭集合中。北卡罗来纳州文件不记录1至17岁的年龄；未满18岁或没有出生日期的登记人显示年龄为0。以下两种情况中的人员不属于客户，二者合起来排除了建造年份所有未满18岁的人员。第一种情况是，建造年份记录的年龄不合理（0岁或超过110岁），且最近一次记录了合理年龄的快照表明其在建造年份未满18岁。第二种情况是，记录的年龄为0，且没有任何快照记录合理年龄。此类人员仍计为其他客户家庭中的成员。在 s，一个**家庭** 是指在 s共享同一选民单元的登记人集合，家庭键依据登记册中的居住地址，将其规范化为建筑和单元；县对地址进行重新标记时，会根据居民的 NCID 将其映射到同一规范化建筑，因此重新标记不算搬家。此映射仅适用于参照；根据单个快照构建的图无法进行这种映射（第6.1节）。建造年份的家庭地址被标记为可能的独栋住宅地址或多单元建筑中的一个单元，且家庭中有一至五名登记人时，该家庭即为**可追踪**家庭。建造年份 t 的**客户文件** 包含 t所有可追踪家庭中的符合条件人员，并在之后的每个快照中追踪他们，无论他们迁往何处。

可追踪总体涵盖了2016年构建时60个确认县中379万名登记者的95.0%；其余登记者属于较大家庭、集体居住场所、记录中未标明单元的多单元建筑，以及移动房屋公园和大型建筑。

我们根据全州留存率选择了这条规则（补充材料C），因此按设计排除了成员更替最频繁的住房。客户文件也未纳入未登记者，且其年龄结构较成年人口更老（构建时年龄为18至34岁、在构建两年后仍列于客户文件中的客户占25.8%，而北卡罗来纳州成年人口中的这一比例为29.2%）。排除这些人群、年龄结构较老以及未纳入未登记者，都会推高测得的时效性；而将达到投票年龄、居住在该地址的成员计作家庭变化，则会拉低时效性。第6.2节对能够估算的部分进行了量化，而净选择效应的方向无法确定。

4.3 模拟供应商及其策略

在构建日期 t，供应商可以看到全州每位登记人的姓名和地址，并根据选民单元键划分家庭。客户在 s 提供其姓名和地址，并询问自己所属的家庭。供应商始终看不到 NCID。由于供应商和参照数据基于同一快照、使用相同的键构建，因此供应商在构建当天的结果完全准确。每位客户都包含在构建图中，因此我们报告的匹配率无法与商业匹配率相提并论，后者还包括图中从未出现过的人。策略如下：

- **地址锚点（从不重建的图）：** 将客户的当前地址与在 t 构建的图重新匹配，并返回该单元对应的家庭。
- **个人锚点（最后已知家庭）：** 返回客户在 t所属的家庭；这相当于完美的人员链接。
- **年度重建：** 与最新的1月1日图重新匹配；在每个评分日期，该图约有十个月的历史。
- **同日期重建：** 针对评分日期构建的图重新进行匹配；仅用作参照（第6节）的同日基线。
- **概率匹配器：** Fellegi-Sunter模型 [29]在Splink中的实现 [61]。层级包括名字（精确匹配、前三个字母、Soundex）、姓氏（精确匹配、Soundex）、中间名首字母、后缀，以及地址（同一单元、同一建筑、同一街道和邮政编码、同一邮政编码）。阻塞取以下条件的并集：同一单元、全州同一完整姓名，以及同一邮政编码内姓氏和名字的Soundex相同。由于不使用年龄信息，匹配器仅使用客户查询所提供的姓名和地址。u概率，即不同人员记录在某一层级上一致的可能性，来源于经过词频调整的取值频率。先验概率来自构建图。m概率，即同一人员记录在某一层级上一致的可能性，来自开发县上的期望最大化（EM）；其中三条阻塞规则分别估计其他规则保持不变的参数。默认情况下，客户会与其匹配权重 $W \geq 0$ （后验概率0.5）最高的候选项建立关联，并运行从无阈值到 $W \geq 25$的扫描。

另外两种策略仅用于第5.7节的假设检验，并在补充材料A中定义：一种是每两年使用滞后一份快照的输入进行一次重建（即滞后重建）；另一种是理想成员跟踪器，它通过 NCID 跟踪各成员在不同构建之间的变化，并返回多数承载者所属的家庭。

4.4 研究设计

从北卡罗来纳州100个县中随机抽取的20个县为开发县：我们在这些县上构建了定义、策略、匹配器训练以及两个预测变量。60个确认县用于检验主要假设，并给出第5至7节的统计数据；这些统计数据在两个划分上以完全相同的方式计算。“确认”指的是县划分，而第5至7节的时效性核算属于描述性分析。20个留出县用于评估预测变量并复现核算结果；这些县的记录还会纳入所有全州或80县计算。在抽样前进行的全州计数

（总体规模、按候选可追踪规则划分的留存率、名单维护、NCID持续性与重新签发、重新填写地址、登记滞后期和未成年人）塑造了可追踪规则，并且未按策略计算任何准确率统计数据（补充材料C）。表1列出了数据的各个部分分别评估的内容。

每项评估的预测和评估标准均在运行计算前记录；每项评估均与其所对应的结果一并报告，未报告评估标准的分析属于描述性分析，补充材料 B 列出了所有评估标准（表S2）。五项评估使用的数据，其汇总结果此前已经过检视。其中三项的评估标准是在此之后设定的：针对后续文件的政策检查（其全州时效性已计算）；刷新分析；以及其他登记册中的模型，该模型采用第9.2节所述的移除风险重拟合，其水平和划分评估标准可根据已有数值计算得出。留出县的家庭变更模型，是在查看留出集得分后，使用相同的开发数据对同一模型重新拟合所得，因此未将其评估标准列为预先设定。留出县的会计处理是在对其进行表格检验后开展的，没有评估标准，属于描述性分析。表S2标出了附有评估标准的四项评估。确认集数值为主要结果，开发集数值则明确标注为开发集结果。

表1. 数据各部分及其评估内容。

Part	Data	评估内容
开发集	北卡罗来纳州20个县的所有快照	定义、策略、匹配器训练，以及两种预测变量的拟合预测变量
确认集	60个县	主要假设（第5.7节）；会计处理和政策比较（第5至7节）
留出	20个县	时效性表格和家庭变更模型（第8节）；会计处理的复现（第5.6节）
后续文件	2026年11月：全部100个县，以及80个开发集之外的县用于表格；60个确认县用于十年货币，2018 年构建在八年时以及政策检查	表格在较晚日期（第8.1节）；十年货币；2018 年构建在八年时（第5.7节）；政策检查（第 7.3 节）
刷新	开发集之外的 80 个县，决策时间为 2024 年选举日	选择要刷新的内容（第 8.3 节）
全州	北卡罗来纳州，2018 年构建	构建错误实验（第 6.4 节）
其他寄存器	俄亥俄州，2017 年 7 月至 2026 年 10 月；华盛顿州，2017 年 6 月及 2018 年 6 月至 2026 年 10 月，以及 2017 年 6 月至 2018 年 6 月	会计处理和预测变量的转移（第9)

4.5 推断

除非另有标注，否则对于每个决策统计量以及所展示的每个区间，区间均为县聚类百分位自助法区间（2,000 个重抽样样本），并以该人员在建造年份所在的县作为聚类单位。这些区间描述的是将县视为县总体中的抽样结果时，县之间的变异，不进行有限总体修正，且为名义区间。最大的县占确认构建日期客户文件的 18.6%，按个人加权的有效县数为 60 个县中的 16.5 个（开发集分别为 39.3% 和 20 个县中的 5.3 个），因此开发阶段区间的证据力较弱。补充材料 B 针对主要假设增加了逐县删除刀切法检验，第8.2节则针对模型比较增加了相应检验。文中的差异和范围均根据全精度输出计算。

4.6 参照作为测量模型

参照记录的是登记共同居住情况，即登记册所列的家庭构成。我们将其视为一种会以已知方式产生误差的测量：

- **过时名录记录**。 离开某个地址的人可能仍会以有效或无效状态登记在那里，直到登记册记录其迁移或将其移除。如果存储的家庭仍包含此人，那么原本错误的记录看起来就是正确的。如果某人在建造年份之后搬入、并在 s之前搬离，那么实际情况正确的存储家庭看起来却是错误的。因此，精确集合评分与过时名录记录之间并非单调关系。第6.1节从五个方面解读这些记录。

- **滞后、重新签发和单元标识符**。 变更发生后才会被记录，个人可能使用新的 NCID 返回，单元编号也可能缺失或不一致。第6.2节分别对这些情况进行测量。

- **未记录的迁移**。 客户的迁移若从未被登记册记录，可能会导致 σ 偏高；共同居住成员未记录的迁移或未记录的搬入，也可能会导致 κ_S 偏高。第6.3节仅着重考察 σ。

5 存储的分配过时的速度

5.1 持续性与时效性

表2。持续性与时效性，2016年建造年份，确认集， s时的登记者； $k = 10$ 根据2026年11月文件评分。 P_A 为地址匹配率； κ_A 为地址匹配中的精确匹配占比； κ_P 为最后已知家庭的精确匹配占比； σ、 κ_S 和 μ 分别为留居者占比、留居者时效性和迁居者时效性； $\sigma\kappa_S$ 为家庭和地址均为当前状态的占比。方括号内为99%区间。

k	P_A	κ_A	κ_P [99%]	σ	κ_S	μ [99%]	$\sigma\kappa_S$
2	0.954	0.701	0.671 [0.654, 0.695]	0.884	0.732	0.203 [0.179, 0.222]	0.647
4	0.896	0.464	0.449 [0.436, 0.469]	0.740	0.535	0.205 [0.187, 0.221]	0.396
6	0.863	0.380	0.379 [0.367, 0.397]	0.665	0.469	0.200 [0.185, 0.214]	0.312
8	0.831	0.322	0.329 [0.319, 0.344]	0.600	0.422	0.189 [0.175, 0.202]	0.254
10	0.806	0.285	0.300 [0.291, 0.313]	0.554	0.393	0.185 [0.172, 0.197]	0.218

持续性缓慢下降，而时效性快速下降（表2、图1）。对于2016年构建，存储的家庭对于到第四年仍在名单上的客户而言，准确率低于一半；到第八年时，这一比例约为三分之一。与登记册相比，在五个构建版本中， κ_P 从两年到八年下降了42%至53%（2016年为51%）；第6节展示了两年水平中有多少源于登记册自身的过时性。到第八年，只有0.254的客户同时拥有仍然准确的存储家庭和存储地址。

两个锚点给出的时效性几乎相同：在20个建档与证据年龄单元中， $\kappa_A - \kappa_P$ 介于 -0.008和0.035之间。这种一致性源于地址锚点的分母。对于迁居者，地址锚点返回的是记录在迁居者新地址上的人员；在八年时，对于匹配的迁居者，其中0.059对应的是正确家庭，而个人锚点为 $\mu = 0.189$。两种时效性之所以一致，是因为地址锚点回答的迁居者更少：如果迁居者的新地址在 t 没有登记人，该迁居者就无法匹配，并从其分母中剔除。

这种下降在八年之后仍以较慢速度持续。十年时 κ_P 为0.300，比八年低0.029 [0.027、低0.032]。根据方程（2）的中点分解，留居者的时效性造成了其中0.017的下降，留居者占比造成了0.010，迁居者时效性造成了0.002； μ 下降了0.004 [0.002、下降了0.006]。

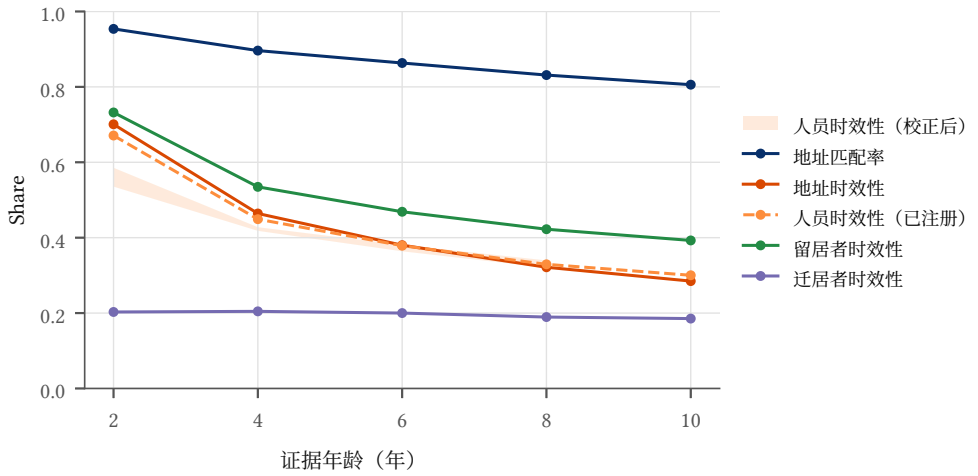

图1. 按证据年限划分的地址匹配率及各锚点的时效性（留存者和迁居者），2016年构建，使用 $k = 10$ 在2026年11月文件上评分；阴影年龄段表示在第6.1节的五项过时名册修正下，个人锚点的时效性（$k = 2$ 至8；两次选举前瞻至 $k = 6$）。确认县。

5.2 迁居者时效性

迁居者时效性几乎不随建造年份或证据年龄而变化。在2008年到2016年间五个偶数年份选举日的构建，以及四个偶数证据年龄中，μ 在确认集上的取值介于0.180和0.209之间（开发集为0.168至0.208）：在每个构建和证据年龄单元格中，迁居者的最后已知家庭约有五分之一的概率完全正确。这些单元格的99%半宽为0.008至0.045，因此数据无法区分恒定的 μ 与仅漂移几个百分点的 μ，但可以排除一种像 σ 和 κ_S 那样随证据年龄增长而下降的迁居者时效性。按自记录迁移以来的年数计算，对于固定的一批早期迁移者，μ 始终接近五分之一。例外是构建后第一年记录的迁移（0.112），此时迟登记会使构建时家庭已经过时（附录D）。

因此，个人锚点的下降源于另外两个项。对于2016年构建，σ 从两年时的0.884降至八年时的0.600，而 κ_S 则从0.732降至0.422，这是因为成员离开、去世、达到年龄门槛，或迁入仍有客户留居的家庭。根据方程（2）进行中点分解，从两年至八年间 κ_P 下降的0.342中，留居者的时效性占0.230，留居者占比占0.108，迁居者时效性占0.003。由于其他项下降而 μ 没有下降，迁居者的存储家庭在正确结果中的占比不断上升：$(1 - \sigma)\,\mu / \kappa_P$ 在八年时占精确结果的23.0%。

5.3 独居的迁移者

这五分之一中，约有一半来自独居迁移者。在 t 时独居、且到 s 时仍独居的迁居者被视为当前，因此独居迁移者会计入 μ。在这20个单元格中，他们占迁移者的22%至28%，贡献了 μ 的41%至55%。按家庭规模来看，迁移者时效性处于较窄的范围内：登记为独居的迁移者为0.336至0.407，登记为与他人同住的迁移者为0.124至0.157（开发集分别为0.332至0.397和0.108至0.136；表 S7 按家庭规模列出了2016年构建的数据）。因此，μ 是两个范围都较窄的数值的混合结果，其中第一个数值较高，是因为单独迁移者规则将仍然独居的独居迁移者计为当前。

5.4 构建年份到构建年份

这一水平随构建年份而变，也随选举日历而变。两年期留居者占比在2008年至2016年的各个构建中分别为0.916、0.867、0.908、0.842和0.884，随选举年份奇偶性而波动；而 κ_P 分别为0.731、0.634、0.710、0.601和0.671（表 S5）。美国社区调查（ACS）的迁移率既未显示这种波动，也未显示如此幅度的流动性变化（第6.3节），因此这种模式是登记册本身的特性。按照已注册参照，总统选举年的两年期时效性在2008年、2012年和2016年逐年下降；采用过时名册修正后，降幅更大（表 S8）。采用第6.1节的一次选举前瞻后，这种波动有所减弱：2008年至2016年各个构建的两年期 κ_P 分别为0.706、0.610、0.663、0.575和0.585。

5.5 前两年

损失在构建完成后不久最快。我们在最多四个日期评估2008年到2016年每个偶数年构建的最后已知家庭：构建完成后的1月1日、下一份11月名册、再下一年的1月1日，以及两年后的选举日。这样，对于每个构建，在最初两年内都有四个观测值，但2010年除外，因为没有2011年11月的名册文件（表 S18）。构建完成两个月后 κ_P，数值为0.922至0.956；一年后为0.805至0.831；两年后为0.601至0.731。最初两个月每月下降0.022至0.044，是第一年其余时间下降速度（每月0.011至0.014）的两至四倍。第二年若以总统选举结束，下降幅度会更大：从第十四个月到两年 κ_P，2010年和2014年构建的数值分别下降0.183和0.184，而2008年、2012年和2016年构建下降0.082至0.119，这就是第5.4节所述的均等性变化。早期损失的一部分可能来自构建日期滞后（第6.2节）。一种尚未在此检验的解释是：人员重新登记时，登记册才记录其迁居；而重新登记会在选举前后集中发生，因此损失出现的时间也会反映登记册记录迁居的时间。

5.6 留出县

在20个留出县中，会计处理结果相同。在全部20个建档与证据年龄单元格中，留出值与确认集的差异分别不超过：κ_P为0.012，σ为0.014，κ_S为0.010，μ为0.026，κ_A为0.007。对于2016年构建，概率匹配器在 $W \geq 20$ 处的精确匹配占比与确认集相差不超过0.008（表S6），按家庭规模划分后的差异也不超过0.008。此次复现没有设定判定标准，而这些幅度的差异远在留出区间之内（在有效县数为10.4时，κ_P 的99%区间宽度为0.07至0.09），因此无法检测到几个百分点的差异。

5.7 主要假设

我们陈述三项主要假设，并分七部分进行检验；检验指标为**共同成员 F1 值**，这是一种部分得分：对于客户 i，若 R_i'和 T_i均为空，则得分为1；若恰有一个为空，则得分为0；否则得分为 $2|R_i' \cap T_i'| / (|R_i'| + |T_i'|)$；未匹配的客户则按独居者计分。在 s的登记者中，地址锚点的匹配率与其共同成员平均F1之间的差值G(k)，衡量部分得分的时效性下降速度比持续性快多少。这些假设预测：家庭准确率下降得比匹配率更快；理想成员跟踪器的剩余损失出现在拆分家庭中（二者均定义于补充材料A）；证据越新，损失越小。

在2016年构建版本中，每个部分都得到支持，且没有任何决策不同于开发集（表S2列出了每个部分，表S3列出了其开发集数值及刀切区间）。两年时差距为正（0.130 [0.118，0.139]），从两年到八年扩大了0.225 [0.214，0.234]，而阈值为0.10。F1的下降幅度比匹配率的下降幅度高出0.102 [0.082，0.122]的两倍。从2008年到2016年的五个构建版本中，差距扩大均达到阈值，其中2010年最小（0.172 [0.165，0.179]）。拆分家庭中的迁移者至少承担了理想跟踪器的损失的0.631 [0.621，0.641]，而阈值为0.50。滞后两年的重建使差距在

四年至八年间保持至少为0.311 [0.117，0.319]„，并使相对于从不重建的图的损失至少减少0.311 [0.304，0.319]„，而阈值为0.25。

这些决策以 s的登记人为条件，其中八年时占文件的72.8%。在允许每个退出者的得分保持自由的情况下，扩展范围为 [0.066、0.399] ，超过了其阈值（方程（S1），补充材料 F），因此第7节.4给出了整个文件的处理方式。补充材料 B给出了接受规则下以及十年时的差距。

我们在2018 年构建上重新检验了第一个假设的三个部分，该构建根据2026 年 11 月文件进行评分；在 s的登记人中，这三个部分均成立。两年时差距为0.193 [0.180、0.200] ，并扩大了0.158 [0.148、0.167]。F1的下降幅度比匹配率的下降幅度高出两倍，差值为0.055 [0.037、0.073]。这些结果对退出者的评分方式较为敏感：如果将每个退出者的得分都记为1，扩展幅度为0.067 [0.061、0.073]，，低于其阈值，并且F1的裕度为 −0.004 [−0.015、0.009] ，这一结论不受支持。在2016年构建上，采用这种评分方式时，两者均成立。

6 参照可能有多大错误

6.1 过时名录记录

登记册本身已经过时，因此测得的衰减可能部分源于参照，而非存储的家庭。五种替代参照会根据登记人的非活跃状态或后续记录，将部分登记人视为在 s时缺失。一次选举前瞻和两次选举前瞻分别向前查看一个和两个选举日：如果在较晚日期，该人员因不活跃而被移除，或在另一个县登记且登记日期不晚于 s，则该成员在 s 时缺失。排除非活跃者的解读会剔除在 s时被列为非活跃的成员。两种客户端解读分别在一次选举前瞻和排除非活跃者的解读基础上加入客户端信息。对于同一规则标记出的客户，其在 s 时的真实集合被视为空（第3.1节的单独迁移者约定），因此只有在其单独存储时，这种处理才是准确的。这两种解读都是建立在该假设上的情景。每种解读都可能出错：一些非活跃登记者仍居住在其登记地址，一些前瞻性移除是在 s之后迁居所导致的，而前瞻性解读只能捕捉登记册后来记录的迁出情况。

表 3。登记参照和五个校正后的参照下的人员锚点时效性与同日基线，确认集，2016年构建。同日期重建是基于同一登记册、在评分日期构建的图；对于校正后的参照，其得分就是新图的构建日期准确性。两次选举前瞻需要选举日之后两次选举日的文件 s，并且在八年时为空白。

参照	κ_P，k = 2	k = 4	k = 6	k = 8	同日，k = 2	同日，k = 8
已注册	0.671	0.449	0.379	0.329	0.999	1.000
一次选举前瞻	0.585	0.421	0.364	0.321	0.832	0.835
两次选举前瞻	0.567	0.419	0.364		0.798	
删除非活跃项	0.565	0.428	0.377	0.334	0.806	0.891
客户端前瞻	0.557	0.418	0.366	0.326	0.771	0.776
客户端非活跃项	0.535	0.427	0.381	0.340	0.734	0.850

两年时的修正幅度较大，八年时则较小（表 3）。两年时，对于 2016年构建，修正使 κ_P 降低 0.085（一次选举前瞻），至 0.135 [0.120、0.149] （客户端非活跃项）；对于其他四个构建，修正幅度为 0.024 至 0.097（表 S8）。八年时，各种变体在每个建造年份的变化幅度均为 κ_P −0.017 至 +0.013。留出县也呈现出相同的幅度。修正作用于σ和 κ_S，而 μ则上升：两年时，在客户端前瞻下，σ从 0.884 降至 0.755，κ_S 从 0.732降至 0.665，而 μ 从0.203升至0.224。

新图继承了 s处的每一条过时名录记录，而存储的分配仅会校正它仍保有的过时名录记录。同日期重建与登记册的匹配度为0.999至1.000（略低于1，因为参照将重新标记的建筑物中的各单元合并，而根据单个快照构建的图无法做到这一点），与校正后的参照的匹配度为0.734至0.891。

在一次选举前瞻下，其校正率在两年时为0.167，在八年时为0.165，而对 κ_P 的校正率则从0.085降至0.009。两者的比率只有在存储集合在最新集合出错的地方均正确时，才能衡量证据时效性的损失；但这种情况确实会发生，因此我们不采用该比率。地址锚点的变化方式相同：其八年时效性相对于登记册为0.322，在一次选举前瞻下为0.309。

由此得出三点结论：

1. **在所有参照下，时效性都会随着证据年龄的增加而下降**。 κ_P 从两年到八年，针对登记册下降了 0.342；针对八年期数值的四种校正后参照，则下降了 0.196 至 0.265。2. **两年时的水平取决于参照；八年时的水平对这些读数相对不敏感**。 κ_P 两年时，针对登记册为 0.671，针对校正后的参照为 0.535 至 0.585；八年时，针对登记册以及具有八年期数值的四种校正后参照为 0.321 至 0.340。过时条目在早期影响更大，此时存储的家庭仍包含登记册所列的大多数人；到了八年时，大多数存储家庭因其他原因而不再准确。3. **不同策略之间的比较顺序保持不变，差距则有所缩小**。年度重建相较于存储的家庭，在八年时的优势针对登记册为 0.471，针对校正后的参照为 0.326 至 0.387。每个证据年龄下各策略的排序在所有参照中都相同，但有两种策略在登记册下六年时已并列（补充材料 E）。

登记参照仍然是主要依据：本文中的每项标准都以此为基准表述，登记册在 s 的登记记录是对目标的直接测量，而每次校正后的读数都可能在两个方向上产生误差。我们列出主要水平，并在其旁标注一次选举前瞻值。

6.2 其他登记册敏感性

其他基于登记册的敏感性分析中， κ_P 在两个方向上的变动均小于四个百分点（表S9）。从存储的集合和真实集合中同时剔除非活跃登记者，会使 κ_P 提高0.014至0.029；这也会将已离开人员从存储的家庭中一并移除，因此抵消了第6.1节所测量的误差。在 s之后两个月对同一批客户进行评分，这一短期前瞻与一次选举前瞻存在重叠，使 κ_P 降低0.002至0.021，其中两年时的降幅最大。将范围限定为在构建选举中投过票的客户，会使 κ_P 提高0.025至0.036，这一结果混合了构建日期滞后与选民流动性较低的影响。NCID重新签发交叉对照表使 κ_P 降低0.002至0.004，而剔除位于混合单元标识符建筑中的客户，则使其提高0.006至0.010；在2016年构建中，选民单元键与北卡罗来纳州公共地址点在2.0%的匹配建筑中不一致，而这些建筑容纳了7.6%的登记者。按照成年年龄构成对客户文件重新加权，会使 κ_P 降低0.003至0.007。在开发县中，这些项的方向和幅度相同。

目标定义中的两个选择在较长的证据年龄下更为重要。一名在 t时未满18岁、之后在客户地址登记的成员，会被计为客户家庭发生了变化；将这些成员从两组中同时剔除，会使 κ_P 在两年时提高0.031，在八年时提高0.071。登记册无法区分原地达到投票年龄的儿童与迁入后登记的成年人，因此这反映了对目标定义的敏感性。将一名在 s 时缺失的返回成员视为未知，而不是匹配失败，会使八年时的 κ_P 从0.329提高到0.384（表4）。综合第6.1至6.3节中的目标定义选择和压力情景，八年时的 κ_P 范围为0.26至0.40；已注册值0.329是主要估计值。

6.3 流动性：登记册未记录的内容

登记册只有在人重新登记时才能识别迁移，因此 σ 是登记册记录的留居者占比。记录的迁移率约为调查所得所有成年人的一半。在美国社区调查2010年至2019年及2021年至2024年的年份中（2020年没有标准发布），成年登记者记录的地址变更率为每年0.063（计入迁出县或州后为0.070，计入名册维护后为0.073），而美国社区调查中北卡罗来纳州成年人的相应比例为0.138（表B07001） [90]；死亡率为每年0.010，不计为迁移。差距的一部分源于登记册延迟记录或从未记录的迁移，另一部分源于登记者与所有成年人之间的差异。作为压力情景，我们将全部未记录迁移率累积施加于留居者，并将每一名此类留居者计为迁居者。这会使 κ_P 在每个证据年龄下均降低0.056至0.067（八年时降至0.264至0.271）。该情景夸大了影响，因为美国社区调查中的迁移集中于重复迁移者；它只迁移 σ，因此共同成员的未记录迁移和未记录迁入在此处均未被涵盖；并且它与过时名册修正存在重叠，所以二者从不相加。美国社区调查序列显示没有均等性变化。

6.4 构建误差

模拟供应商的构建日期完全准确；实际供应商会错误地建造部分家庭。我们对 2018 年全州范围的构建进行扰动，并在三个组别中以 1.15 年和 7.99 年的滞后进行评分：

- **滞后**。 每个人的地址取自构建日期（40%），或取自早 10 至 24 个月的快照；未列在其中的人员会被延迟纳入或被遗漏。
- **噪声**。 在 10% 的建筑中，所有单元都合并为一个家庭；另有 2% 的地址无法匹配，导致仅留下个人信息。
- **滞后和噪声**。

构建准确率，即在构建日期上其构建家庭信息完全准确的、按客户加权的占比，在截至7.99年的仍被列出的客户中为0.785至0.949（表S22）。

八年时，构建误差的代价低于其表面值。各组保留了准确构建时效性的0.905至0.978（相对于0.320而言为0.290至0.313）。这高于其构建准确率：时效性是构建准确率乘以准确构建时效性的1.031至1.153倍，处于预测的0.95至1.25范围内。过时证据主要对原本已经稳定的家庭是正确的，因此，滞后供应商的正确构建更常保持正确：其时效性比准确构建高0.026[0.024，0.027], 如预测所示，而随机噪声几乎独立于后续变化（0.004）。一些错误构建也会变得正确，例如遗漏的加入者离开，或错误拆分的家庭重新合并：八年时，错误构建中有0.061至0.103是正确的，高于预测上限0.05。

在1.15年时，成本更高：各组别保留了 κ_P的0.823至0.958。对于这三种误差类型，在每个组别中，精确构建曲线都高于供应商的曲线；八年时高出2%至10%。

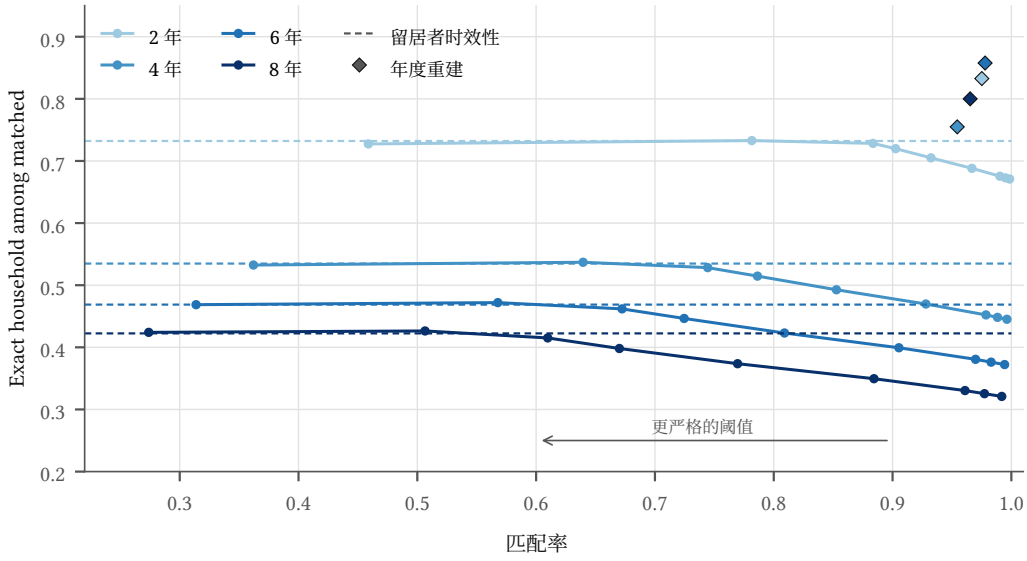

图 2. 确认县中，概率匹配器阈值扫描下的匹配率与匹配客户中的精确家庭占比；虚线表示各证据年龄对应的留居者时效性 κ_S，菱形表示年度重建。

7 匹配规则、新证据与完整文件

表4。 $k = 8$处、2016年构建的有效性前沿。精确匹配占比以匹配客户为分母（在 s 缺失的返回成员会导致匹配失效），并附有其99%区间；下一列则将此类成员视为未知。F1是对 s 所有登记者计算的共同成员 F1 值（第5.7节）。人员链接精确率是指概率链接指向客户自身建造年份记录的占比。

Rule	匹配率	精确匹配占比 [99%]	缺失为未知	F1	人物链接精确率	开发-ment 精确份额
概率式，无阈值	0.992	0.321 [0.310，0.337]	0.375	0.491	0.912	0.295
概率式，W ≥ 0（默认）	0.884	0.350 [0.338，0.365]	0.407	0.499	0.978	0.323
概率式，W ≥ 10	0.670	0.398 [0.386，0.414]	0.460	0.483	0.995	0.371
概率式，W ≥ 20	0.506	0.426 [0.413，0.443]	0.489	0.434	0.998	0.400
概率式，W ≥ 25	0.274	0.424 [0.409，0.445]	0.484	0.328	0.998	0.396
地址锚点	0.831	0.322 [0.307，0.344]	0.382	0.477		0.293
个人锚点	1.000	0.329 [0.319，0.344]	0.384	0.503		0.304
年度重建	0.965	0.800 [0.790，0.814]	0.804	0.891		0.789
留存者时效性 κ_S（无匹配器）		0.422 [0.410，0.436]				0.397

7.1 严格阈值下的筛选

提高概率匹配器的阈值，会提升其匹配结果中的精确匹配占比，直到达到留居者的时效性，此后便不再上升（图2）。在 $W \geq 20$，各个证据年龄下的精确匹配占比都比 κ_S 高 0.001 至 0.004（开发集为 0.000 至 0.003）；而 $W \geq 25$ 则略低于$W \geq 20$，同时匹配率再次减半。其机制正是第 3.4 节中的界限：严格的

阈值通过选择要回答哪些对象来发挥作用。在 $W \geq 20$ 和八年时，留居者占匹配客户的 0.984。匹配留居者的时效性为 0.430，比 κ_S高 0.008；相比之下，规则未能匹配的留居者为 0.385，因此条件 (i) 仅在 $W \geq 20$大致成立。成立的界限是匹配留居者的时效性，它接近合并后的 κ_S，因为该规则匹配了大多数留居者。与匹配的迁居者（0.195）混合后，$W \geq 20$ 的精确匹配占比为 0.426。在 $W \geq 25$，已匹配与未匹配留居者的时效性分别为 0.426 和 0.420。历史人口普查链接也呈现出相同的筛选现象：链接规则要求家庭或地理一致 [86], ，而链接后的人口普查样本与其来源总体存在系统性差异 [6]。

相较于合并后的 κ_S，这一小幅优势来自构成差异：在单独记录的客户和与他人共同记录的客户中，$W \geq 20$ 与各组的 κ_S之差均在 −0.003 至 +0.002 之间（见表 S12；配对区间见表 S11）。一条只对单独记录的客户作答的规则，在八年时对其中 0.497 的客户作答正确，高于合并后的 κ_S = 0.422，因为它针对的是家庭变化空间较小的群体；而在这些客户中，严格阈值同样只能达到留居者的时效性。阈值的代价是覆盖率。在同一客户文件上从 $W \geq 0$ 移至 $W \geq 20$，精确匹配占比从 0.350 升至 0.426，匹配率则从 0.884 降至 0.506；所有客户的共同成员 F1 值从 0.499 降至 0.434，因为未匹配的客户会被计为独居。

7.2 人员链接精度与家庭有效性

匹配器造成的错误个人链接几乎不会改变结果。在默认阈值下，概率匹配器将其匹配结果中的 0.978 链接到客户自身的构建记录；这一精确率反映了在两侧采用相同规范化方式，将登记册与其自身进行链接的情况。在所有客户中，其精确匹配占比在每个证据年龄下都比 κ_P 高 0.017 至 0.021，这一差距源于筛选。对于匹配到的客户，完美人员链接在两个切分中、在 $W \geq 0$ 和 $W \geq 20$ 下的每个证据年龄对应的精确匹配占比，差异均不超过 0.0011（八年时为 0.350）。在 $W \geq 0$ 下，剩余差异随证据年龄改变正负；在 $W \geq 20$ 下，差异始终约为 0.0001（表 S11）。

八年时，错误链接占匹配结果的 2.2%，因此它们最多可使精确匹配占比改变这么多。但实际影响约为 0.001，因为其中许多链接指向客户构建单元中的另一人，例如亲属或同名者，而此人的构建时家庭就是客户自己的家庭（在开发县中，两年时的大多数错误链接以及八年时约一半的错误链接都属于这种情况）。其余大多数是地址已变更客户的仅凭姓名链接。跨越家庭的错误可能使精确匹配占比最多改变其发生频率那么多；其影响取决于哪些客户被错误链接，因此结果反映的是该匹配器所产生的错误。

7.3 新鲜证据

更新鲜的证据对时效性的影响远大于任何阈值。年度重建在八年时，对 0.800 的匹配找回了完全相同的家庭，同时匹配了 0.965 的客户；在四个评分日期上的数值则为 0.755 至 0.858。在一次选举前瞻下，按两年、四年、六年和八年计算，数值分别为 0.711、0.668、0.759 和 0.693。补充材料 G 给出了家庭链接存续率，以及仅限于构建时候选项的方法所能达到的支持上限。

这些比较在 2026 年 11 月文件中、在 2018、2020、2022 和 2024 年的构建年份上均成立，但有其适用范围。$W \geq 20$ 与 κ_S的差距在 0.004 以内（点估计；上限为 0.02）。默认阈值与019 [0 0.018，021] 的差距在 0. κ_P 以内（上限为 0.035），地址锚点与个人锚点的差距在 0.029 [00.027，031]以内（上限为 0.05）。年度重建得分至少为 0.847 [00.835，0.864] （下限为 0.760）。第一项比较和地址锚点比较都接近恒等关系，因为同一单元内的全名匹配按构造即属于留存者（表 S14）。

7.4 整个客户文件

2016 年客户文件中，超过四分之一的人员在八年内已离开名册，且对于 s，没有观察到家庭。登记册记录了每名人员离开的原因，因此我们根据其在 s或之前的最后一条记录中的移除原因，对每次离开进行评分。已故客户的正确答案是没有家庭，因此返回一个家庭属于错误。对于迁出州外的客户，如果返回的成员仍在北卡罗来纳州登记，则返回的家庭是错误的。对于在州内迁移的客户，如果返回的成员仍位于返回地址，则属于错误。对于任何离开，如果返回的成员已故，或返回成员在 s处于两个或更多地址，则属于错误。因一段时间不活跃而被移除的客户，在主要界限中属于未知；该界限先将每个未知离开判为错误，再判为正确 [63], , 并在某一情景中将其解读为可能发生了迁移；其他代码均为未知。不返回任何内容的政策得分为无答案。补充材料 F 按原因给出了离开的会计处理（表 S10），代码发布包包含用于重新计算这些结果的输入。

表 5. 2016年构建的确认集客户文件中各结果所占比例（N = 3,603,082）。对于“正确”这一项，界限范围从将所有未知退出都判为错误，到将其判为正确；“上限：将不活跃视为搬迁”是将这些退出解读为搬迁时的上限。未知是个人锚点中未知部分的占比。若将每次退出都判为 0 或 1，所得界限范围会与退出所占比例一样宽。

k	退出	个人锚点正确	上限，不活跃作为搬迁	地址锚点正确	地址锚点，没有回答	未知
2	0.061	[0.630，0.661]	0.650	[0.627，0.659]	0.043	0.031
4	0.157	[0.378，0.471]	0.438	[0.351，0.442]	0.089	0.092
6	0.230	[0.292，0.422]	0.388	[0.253，0.381]	0.110	0.131
8	0.272	[0.240，0.385]	0.355	[0.195，0.336]	0.132	0.146

八年时，个人锚点为完整文件中 0.240 至 0.385 的客户返回了正确的家庭（表 5）；如果排除已故者而不是对其评分，最高可达 0.420。将不活跃后的移除视为迁居，会使上限降至 0.355。按原因识别在八年时留下 0.146 的未知部分，而 0/1评分为 0.272。不活跃后的未解决移除占剩余未知部分的 74%；由于迁居者的移除标记本身容易出错，这些部分在主要界限中仍被视为未知 [46]。地址锚点能回答的客户更少（文件中有 0.043 至 0.132 的客户无法得到回答），并且在八年时对 [0.195，0.336] 的客户是正确的。

8 通过构建时特征预测时效性

如果时效性是存储的分配所具有的建造年份特征的稳定函数，那么一个持有自身建造年份特征的系统就能预测每个存储的分配的时效性。我们在开发县上构建了两个预测变量，并在留出数据上对其进行了评估：按构建时单元划分的时效性查找表（第 8.1 节），以及一个家庭变化模型，并由此推导时效性（第 8.2 节）。第 8.3 节探讨这些表是否能选择更适合刷新的分配。

8.1 查找表

目标是个人锚点的精确结果（公式 (1)），条件是客户在 s 被列名；κ_P 列名占比乘以建档日期客户的列名且精确占比（第 9.1 节）。**粗粒度表**按证据年龄（2、4、6、8）、构建版本、构建时家庭规模（1 至 5）、建筑类型（地址包含单元标识，或建筑内有多个单元时为多单元建筑）以及客户在构建时的年龄（分为七个年龄段）划分单元格：共 1,400 个单元格。**含共同成员年龄的表**进一步按 18 至 25 岁的共同成员人数、分为六个年龄段的共同成员平均年龄，以及客户与共同成员的年龄差拆分每个单元格：共 21,788 个单元格。种族、族裔、性别和党派不是特征，

且每项输入都来自系统。单元格的取值方法是：将其证据年龄对应的汇总精确占比加到该单元格的精确结果数上，再除以该单元格的客户数加一；该取值是在 20 个开发县的全部五个构建版本上拟合得出的。表中缺失的单元格采用该年龄对应的汇总占比（在 $k = 2$ 至 8 时分别为 0.641、0.463、0.375 和 0.317），这也是**仅依据年龄**的基准。对于拟合时未见过的构建版本，我们还以构建奇偶性（总统选举年或中期选举年）替代构建日期，拟合每张表；这种奇偶性形式在开发集上的布里尔分数损失为 0.0008。

我们根据四项评估标准，在20个留出县以及2026年11月文件（后续文件）的全部100个县上评估这些表格。第一项标准是：留出布里尔分数不得比开发集高出超过0.005。第二项标准是：在20个建档与证据年龄单元上，预测精确比例的平均绝对误差（单元格误差）不得超过0.010。第三项标准是：预期校准误差（ECE）不得超过0.010。第四项标准针对后续文件中的均等性表，要求在四个构建与证据时长配对上的单元格误差不得超过0.030。

表6。时效性表和两个基线，个人锚点：在20个留出县上（18,975,925个客户行：每个人针对每个建造年份和证据年龄各贡献一行；五个构建版本，四个年龄），以及在后续文件的全部100个县上针对四个构建与证据年龄配对进行评估（开发集之外的80个县和20个留出县单独列于表S16）。仅年龄是按证据年龄汇总的精确匹配占比；年龄 × 规模是一个由证据年龄和构建时家庭规模组成的20单元格表，在开发县上拟合。技能是1减去预测变量的布里尔分数与基线布里尔分数之比，并附带其99%区间。可靠性和分辨率对布里尔分数进行分解 [69] （补充材料H）；不确定性为0.2496至0.2499。单元格误差是各建档与证据年龄单元上的平均绝对误差。ECE将每个并列预测值区块保留在同一个分箱中（补充材料H）。

预测变量	布里尔	技能随年龄变化 [99%]	技能随年龄变化 × 规模 [99%]	可靠性-ity	分辨率-tion	单元格误差	ECE
留出县							
粗粒度表	0.1976	0.159 [0.145, 0.180]	0.068 [0.059, 0.078]	0.0003	0.0526	0.0106	0.0107
共同成员年龄	0.1913	0.186 [0.172, 0.207]	0.098 [0.089, 0.109]	0.0005	0.0591	0.0090	0.0091
年龄 × 规模	0.2120	0.098 [0.088, 0.113]		0.0007	0.0386	0.0342	0.0236
仅年龄	0.2350			0.0009	0.0157	0.0377	0.0292
后续文件，*100* 个县							
粗粒度表	0.1963	0.164 [0.159, 0.171]	0.061 [0.058, 0.065]	0.0011	0.0544	0.0127	0.0122
共同成员年龄	0.1896	0.193 [0.188, 0.199]	0.093 [0.090, 0.097]	0.0017	0.0617	0.0147	0.0164
年龄 × 规模	0.2091	0.109 [0.106, 0.115]		0.0006	0.0411	0.0153	0.0168
仅年龄	0.2348			0.0003	0.0151	0.0142	0.0142

在留出县中，按点估计，包含共同居住成员年龄的表格满足全部三个评估标准：其布里尔分数比开发集低0.0001，单元格误差为0.0090，预期校准误差为0.0091。粗粒度表满足布里尔分数标准，比开发集高0.0004；其余两项未达标，但差距均小于千分之一（单元格误差为0.0106，预期校准误差为0.0107）。区间无法确定哪一项结果成立。在20个县、10.4个有效县的情况下，两张表的单元格误差和预期校准误差区间均包含0.010（表S13）。在后续文件中，两张均等性表均满足单元格误差标准（分别为0.0127和0.0147）。

表格的优势在于分辨率。在留出县和后续文件中，粗粒度表相较于仅年龄的技能提升为 0.16，加入共同居住成员年龄后则为 0.19；所有预测变量的可靠性误差至多为 0.002，而仅年龄的分辨率为 0.015 至 0.016，表格则为 0.053 至 0.062。通过与年龄 × 规模表进行序贯比较可知，提升的一半至三分之二来自家庭规模；表格相较于该表的技能提升，其所有 99%区间均不包含零。尚不能确认表格与仅年龄之间存在校准差异：在留出县中，仅年龄的 ECE 为 0.029 [0.003，0.073]，该区间与表格的区间重叠。

在留出县中，两张表对全部 20 个单元格的预测都偏低：粗粒度表平均低估 0.011，加入共同居住成员年龄后平均低估 0.009。问题在于县级变异，而表格并未纳入这一因素。按县计算的开发偏移量标准差约为 0.024，因此，包含 10.4 个有效县的区块的平均偏移量标准差约为 0.007，低于 0.010 的判定标准。

这些表格会随着日期推移而发生漂移，最新的构建版本尤为明显。在涵盖全部 100 个县的后续文件中（其中包括用于拟合汇 pooled年龄份额的开发县），仅年龄的单元格误差为 0.0142，低于包含共同居住成员年龄的表格（0.0147）。在开发集之外的 80 个县中，两种表格的表现都优于仅年龄（误差分别为 0.0109 和 0.0130，而仅年龄为 0.0216；表S16）。后续文件中的最大单元格误差出现在 2024 年构建版本的两年时点，分别高估了 0.037 和 0.043（仅年龄的偏差为 0.009；表S15）。均等性表汇总了 2008年、2012 年和 2016年总统选举的构建版本；在开发县中，它们两年的时效性依次从 0.712 降至 0.681，再降至 0.637。在相同县中，2024 年构建版本的数值更低，为 0.625。

这些表格拟合的是已注册参照，却会高估校正后的参照。在 20 个留出的县上，使用一次选举前瞻进行评分且不重新拟合时，这些表格在两年时最多高估 0.059 和 0.060，在四年时最多高估 0.034 和 0.036，在六年时最多高估 0.025 和 0.027，在八年时最多高估 0.013 和 0.014。这涉及 19 个单元格：2026年11月之前的 40 份文件中，在 2024年11月之后没有任何选举日，因此其中的 2016年构建在八年时没有更晚的选举日，而表 3 则从 2026 年 11 月文件中取得该值。补充材料H给出了根据这一解读计算的布里尔分数和单元格误差。

8.2 家庭变化模型

一个表格单元格包含精确匹配占比的平滑估计，而不包含家庭为何发生变化的任何信息。它不会说明拟合未能涵盖的证据年龄，也不会说明损失如何在 σ、κ_S 和 μ 之间分配。一个描述改变家庭的事件模型可以提供这三项信息，前提是它推导出的时效性在其拟合从未读取过的县上仍然成立。

该模型是一个具有隐藏状态 [49, 51]的多状态马尔可夫模型。每次快照都会观察名册客户的两个事实：客户所在的家庭是否恰好是构建时家庭，以及客户是否位于构建单元。κ_P、σ、κ_S 和 μ 分别是这些事实的边际概率和条件概率。一个潜在连续时间链在完整、破裂和重组家庭之间转移；每种家庭状态又分为位于构建单元或其他地点，并与名册内和名册外两种状态交叉组合，因为时效性仅针对名册客户进行评分。破裂家庭会经历两个潜在阶段，因此其返回率可能随时间下降 [89]。事件率取决于客户及每位共同居住成员的年龄、家庭规模、建筑类型、距构建时间以及日历年份，共涉及205个系数；共同成员年龄表以不同方式编码相同的年龄，共包含21,788个单元格。与Zak and Even [96], 一样，时效性是根据拟合转移推导得出的，但此处有效性的目标是一组人。

我们通过开发县在2008年至2026年间的195,288,247对连续快照对的复合似然来拟合模型，并纳入建造年份后最多122个月的状态转变。随后，根据拟合得到的速率推导最长十年的各年龄时效性；这项推导在拟合窗口内进行，并非作为目标进行拟合。它表示在仍列于名册的客户中，家庭与建造年份时的家庭完全一致的概率；未列于名册的客户则按名册客户的速率改变家庭（补充材料K说明了状态、速率和似然）。

我们在开发县上选择模型，并在留出县上依据两项评估标准进行评估。第一项是布里尔分数低于粗粒度表，且差值的99%上限低于零。第二项是在模型所拟合的选民单位留存规则下，σ，κ_S 和 μ 各自与20个建档与证据年龄单元中的观测值之间的平均绝对误差不超过0.02（第3.3节）。

表7。家庭变更模型与留出县表格：在20个留出县中，个人锚点，$k = 2$ 至8，共18,975,925条客户行。证据年龄是每个预测变量的输入。“模型减去预测变量”表示布里尔分数之差及其99%县聚类区间；负值表示模型表现更优。年龄 × 规模和仅年龄是表6的基线，使用相同客户进行评分，不含配对区间。

预测变量	其他建造年份输入	Size	布里尔分数	模型减去预测变量 [99%]
家庭变更模型：客户年龄、每位共同居住成员的年龄	年龄段、建筑类型、建造年份	205 个系数	0.19631	
包含共同居住成员的表格 ages	相同年龄的汇总	21,788 个单元格	0.19126	0.00505 [0.00412, 0.00591]
粗粒度表	客户年龄、家庭规模、建筑类型、建造年份	1,400 个单元格	0.19760	−0.00129 [−0.00198, −0.00076]
不含建筑类型	客户年龄、家庭规模、建造年份	700 个单元格	0.19925	−0.00294 [−0.00414, −0.00189]
年龄 × 规模	家庭规模	20 个单元格	0.21201	
仅年龄	none	4 个单元格	0.23501	

模型优于粗粒度表，但不及能够获取相同年龄信息的表。模型在留出县上的布里尔分数为0.19631（表7），粗粒度表为0.19760，差值为 −0.00129 [−0.00198,−0.00076]，，且模型在全部20个县的平方误差均较低，每位客户低0.00022至0.00269（单侧符号检验，p值小于百万分之一；逐县删除刀切法t区间为 [−0.00214, −0.00045]）。模型相对于仅年龄的技能为0.165，介于粗粒度表的0.159与更精细表的0.186之间；相对于年龄 × 规模的技能为0.074。粗粒度表是较弱的比较对象，因为模型还能获取每位共同居住成员的年龄段。更贴切的比较对象是包含共同成员年龄的表，它以另一种编码方式使用相同的年龄信息，表现优于模型，领先0.00505 [0.00412, 0.00591]（刀切法t区间为 [0.00387, 0.00623]），领先幅度是模型相对于粗粒度表优势的四倍，而且在每个县都胜过模型。

推导出的分解只有在选民单位规则下才符合标准，σ、κ_S 和 μ的平均绝对误差分别为 0.0137、0.0093 和 0.0187。与第3.3节的分解相比——本文其他所有分解均采用该分解——这些误差分别为 0.0126、0.0090 和 0.0309：推导出的迁居者时效性与测得值相差 0.031，超过了 0.02 的上限（补充材料K.5）。推导出的 κ_P 也偏低。在 20 个单元格中，有 19 个低于观测值，平均低 0.014，因此其单元格误差为 0.0138，而粗粒度表为 0.0106。在表格未涵盖的十年时点，对于 2008年至2014年的建造年份，其数值低了 0.007 至 0.017。

该模型需要 205 个系数，而更精细的表格需要 21,788 个单元格；它还能在拟合窗口内的任意证据年龄上给出时效性。但根据本文的留存者规则，其对损失的分解尚未得到验证。

8.3 选择要刷新的内容

时效性估计应能帮助确定，在预算只能覆盖部分档案时，哪些存储的分配需要刷新。在预算限制下选择刷新对象，就是 [17, 83, 96] 所研究的时效性问题。在决策日期——2024 年选举日，存储的档案为每位人员保存了一个家庭，来自 2018、2020 或 2022 年的构建；若该人员在多个构建中都有记录，则通过固定哈希选定其中一个。刷新某人的分配，就是将其替换为 2024 年图返回的家庭；每项分配均在 2026年11月评分。每条规则都会对档案进行排序，并刷新排名靠前的一定比例 b：随机选择、优先选择最早的构建、依据预测增益、依据最低预测时效性（朴素规则），或依据实际增益（预言机）。预测增益是指，如果现在刷新，预测变量赋予此人的时效性（评分时证据已有两年历史，且使用相同的构建时特征）

减去赋予存储的家庭的时效性；预测变量为年龄、 × 规模和两张奇偶性表。每张表相对于年龄 × 规模的判定标准包含三部分。在 $b = 0.25$ 处，精确匹配占比的增益，其 99%界限下界必须至少为 0.002。在 $b = 0.10$ 和 0.50 处，99%界限上界必须至少为零。在 $b = 0.25$ 处，对于拥有全部三个构建版本的人员，点估计增益必须为正。

表8。2024 年选举日的刷新分析，于2026年11月进行评分：开发集之外的80 个县，每人刷新成本为1，文件中共有4,898,857名个人。Exact表示：在按照规则规定的顺序刷新文件前 b 部分（预算）后，截至评分日期仍在名册中的4,485,983名个人中，其家庭归属完全正确的比例；未进行刷新时该比例为0.4126，全部刷新时为0.6695。D是在相同预算下，Exact减去年龄小于 × size者所占的精确匹配比例，并附有其99%县聚类区间。年龄 × size以及两个表格均按预测增益排序；“预测值最低者优先”依据包含共同居住成员年龄的表格所预测的存储时效性进行排序；预言机则按实际增益排序，代表上限。

预算	Rule	准确的	D [99%]
10%	随机	0.4382	−0.0262 [−0.0273, −0.0246]
	年龄最大者优先	0.4488	−0.0157 [−0.0167, −0.0144]
	年龄 × 大小	0.4644	
	粗粒度表	0.4631	−0.0013 [−0.0023, −0.00003]
	共同成员年龄	0.4626	−0.0019 [−0.0035, 0.0002]
	预测值最低者优先	0.4579	−0.0065 [−0.0104, −0.0034]
	预言机	0.5218	0.0574 [0.0563、0.0584]
25%	随机	0.4768	−0.0426 [−0.0449, −0.0395]
	年龄最大者优先	0.5029	−0.0165 [−0.0187, −0.0137]
	年龄 × 大小	0.5194	
	粗粒度表	0.5184	−0.0010 [−0.0029, 0.0010]
	共同成员年龄	0.5199	0.0005 [−0.0015, 0.0029]
	预测值最低者优先	0.5156	−0.0038 [−0.0079, −0.00033]
	预言机	0.6856	0.1662 [0.1637, 0.1683]
50%	随机	0.5411	−0.0366 [−0.0382, −0.0348]
	年龄最大者优先	0.5695	−0.0081 [−0.0088, −0.0074]
	年龄 × 规模	0.5776	
	粗粒度表	0.5851	0.0075 [0.0035，0.0111]
	共同成员年龄	0.5887	0.0110 [0.0076, 0.0140]
	预测值最低者优先	0.5931	0.0155 [0.0107, 0.0193]
	预言机	0.6907	0.1130 [0.1103, 0.1153]

两张表都未达到各自的标准（表8）。在 $b = 0.25$，包含共同居住成员年龄的表提升了0.0005[−0.0015, 0.0029] 高于年龄 × 规模，而粗粒度表则下降了0.0010 [−0.0029，0.0010]；在完整案例中，即在三个建造年份都有记录的个人，两者都下降（分别为−0.0013和 −0.0039）。在 $b = 0.10$，两者都下降，分别为0.0019和0.0013；包含共同居住成员年龄的表，其区间包含零，而粗粒度表的区间终点略低于零（−0.00003）。在 $b = 0.50$，两者都提升，分别为0.0110和0.0075，区间均明显高于零，但较简单的规则在此处提升更多（0.0155），在较小预算下则会下降。年龄 × 规模这一最简单的预测增益规则，优于年龄最大者优先，幅度为0.0165 [0.0137，0.0187]，在 $b = 0.25$。所有规则的表现都远不及预言机（0.6856 对比 0.5194）。

该结果不取决于分层或日期。在全部100个县以及完整案例中，两张表在 $b = 0.25$ 处的增益均不显著（区间不跨越零），并且在 $b = 0.50$处均有增益。在第二个决策日期——2022年选举日——两张表在 $b = 0.10$ 处均有损失（区间低于零），在 $b = 0.25$ 处均无增益（区间不跨越零），并且在 $b = 0.50$ 处均有增益（补充材料J，表S17）。预测增益是同一张表的两项预测之差：刷新后的预测和存储的预测；二者均基于存储的构建时特征生成，其中没有任何部分是针对增益本身拟合的；针对增益拟合的模型留待后续研究（第10.3节）。

9 两个其他登记册

9.1 会计处理

俄亥俄州和华盛顿州保留了可持续九年的标识符。密歇根州重新签发了许多标识符，而用于连接其长期间隔的姓名与出生年份交叉表未能满足其偏差标准：在俄亥俄州和华盛顿州分别为−0.032和 −0.021，均超过0.02的限值；表S2。因此，我们不报告密歇根州的结果。俄亥俄州和华盛顿州的名单仅包含活跃登记者和非活跃登记者，因此无法按退出原因进行分类。与北卡罗来纳州一样，在构建日期未满18岁的客户将被排除，并仅作为共同居住成员保留。配对样本为：俄亥俄州从2017 年 7 月至2026 年 10 月（9.26年），以及华盛顿州从2017 年 6 月和2018 年 6 月至2026 年 10 月（分别为9.25年和8.25年），另有一组间隔一年的华盛顿州配对。评估标准来自北卡罗来纳州的全州数值。北卡罗来纳州的拟合是：根据2026年11月评分的五组北卡罗来纳州构建配对，拟合全州时效性与间隔之间的最小二乘直线（补充材料L.2）。κ_P在相同间隔处必须与该拟合值相差不超过0.05。该州自身的留居者占比与北卡罗来纳州的 κ_S 及 μ 结合后，必须在0.025以内得到 κ_P。此外，μ 必须处于0.15至0.25的年龄段内。

华盛顿州的水平和分解结果符合相对于北卡罗来纳州拟合结果所规定的容差，而俄亥俄州则不符合。在华盛顿州，间隔为 9.25 年（2017 年 6 月至 2026 年 10 月）和 8.25 年（2018 年 6 月至 2026 年 10 月）时，κ_P 分别为 0.314 和 0.325，与北卡罗来纳州拟合结果相差不超过 0.007，且分解关系成立，偏差分别为 0.014 和 0.013。在俄亥俄州，间隔为 9.26 年时，κ_P 为 0.377 [0.361，比拟合结果高出 0.397], 0.071；俄亥俄州较高的留居者占比只能解释其中 0.019 的差距（偏差为 0.051）。在两个州，长间隔下的 μ 均位于年龄段内（俄亥俄州为 0.242，华盛顿州为 0.212 和 0.207），而华盛顿州的一年间隔组合为 0.253，恰好位于该年龄段之外。排序关系 $\kappa_S > \kappa_P > \mu$以及单独记录的客户比与他人共同记录的客户具有更高时效性的结论，在所有情况下都成立；两个锚点之间的差异不超过 0.032（表格 S2 和 S20）。

俄亥俄州较高的水平，部分源于仍留在其名单上的人。κ_P 以在 s仍列于名单上为条件，而俄亥俄州保留在名单上的客户较少：在建档日期客户中，9.26年后仍列于名单上的比例为0.615；相比之下，华盛顿州为0.718，北卡罗来纳州2016年至2018年构建的数据在2026年11月评分时为0.666至0.703。建档日期客户中列名且精确占比 κ_P（即列名占比乘以精确匹配占比）在俄亥俄州为0.232，在华盛顿州两个长期配对中分别为0.225和0.238，在北卡罗来纳州为0.196至0.225。它将每一项退出都视为不精确，因此只是一个下限，并非完整文件的时效性，也无法对各登记册的全文件有效性进行排序。按这一指标计算，俄亥俄州相对华盛顿州的差距从0.063降至0.007；相对北卡罗来纳州，在9.26年处对八年时的2018年构建（0.225）与十年时的2016年构建（0.196）进行插值，得出0.207，因此仍有0.026的差距。一种解释是，名单剔除更多家庭情况已发生变化的客户，从而提高了留存客户的时效性；第9.2节将检验其中一种形式。

9.2 预测变量与模型

各登记册之间的水平不必保持一致，因此我们在整体重新校准后对排名进行评分。对于每个预测变量和配对，我们在被评分的配对本身上拟合一个截距 α，$\operatorname{logit}(p') = \alpha + \operatorname{logit}(p)$，其中 p 是该预测变量的概率，p' 是重新校准后的概率，且 $\operatorname{logit}(p) = \log\big(p/(1-p)\big)$。因此，水平比较是在样本内进行的。这些比较针对三个长期配对，采用表格中的八年单元格；如果建造年份不是偶数年 11 月，则取中期选举单元格和总统选举单元格的平均值。排名具有可迁移性。含共同居住成员年龄的表格表现优于年龄 × 规模，在俄亥俄州高出 0.0247，在华盛顿州高出 0.0181 和 0.0186，均超过预测的最低值 0.010（北卡罗来纳州的差值为 0.022 至 0.023）；各表格之间的每项比较也都保持了其在北卡罗来纳州的排序，且置信区间不跨越零（表 S21）。

我们在相同配对上应用第8.2节的家庭变更模型，只重新拟合名录流程的移除风险率，并据此估算每组配对的留存率：建档日期客户中仍然

在 s仍列入名录的比例，按单元格和建筑类型划分（补充材料M）。该重新拟合的简化单偏移版本显示，在两组长期配对中，俄亥俄州的移除率约为北卡罗来纳州的1.4倍，华盛顿州的约为0.86倍。重新拟合时未纳入任何家庭结果，只使用了评分配对在 s 的留存率，而系统在建档时无法获知该数据。其评估标准应用于三组长期配对，共有四项。 κ_P 在每个州都必须与观测值相差不超过0.025，且俄亥俄州的误差最多为不重新拟合时的一半。 σ、 κ_S 和 μ 的差值都必须不超过0.03。其原始布里尔分数必须优于每张表中的分数，且上界低于零。此外，俄亥俄州的重新拟合必须降低原始布里尔分数（表S2）。华盛顿州的一年期配对结果未设评估标准。

该模型捕捉到了华盛顿州的水平，却未能捕捉俄亥俄州的水平。在华盛顿州，长期配对的预测值与 κ_P 相差不到0.015和0.017，处于误差范围内；在一年期配对中，预测值高出0.036，而其 μ 0低了0.063。在俄亥俄州，重新拟合后的预测值低了0.054，未重新拟合时低了0.053（0.3236和0.3241，相比之下为0.3773；表S23）；重新拟合后俄亥俄州的布里尔分数略微变差（+00.0001），因此，按该模型设定的名录流程无法解释俄亥俄州的水平。

在两个州中，模型都未能捕捉到分布差异。按第3.3节的规则， σ 在俄亥俄州偏差为 −00.070，在华盛顿州偏差为0.059和0.053； μ 低了0.041至0.060，而 κ_S 的偏差在0.025以内。按选民单位规则（仅俄亥俄州和华盛顿州有格式变更；补充材料M）， σ 的偏差最高达0.081。按第3.3节规则计算的观测值 μ 本身，在两州的八至十年期（0.182至0.191）均高于北卡罗来纳州的所有数值（表S20）。按原始布里尔分数计算（表S24），该模型在华盛顿州优于粗粒度表（−0.0015 [−00.0027， −0.0008] 以及 −0.0011 [−00.0020， −0.0006]），但在俄亥俄州并非如此（+00.0004[−00.0002， +0.0013]）；与包含共同成员年龄的表相比，该模型在每一组配对中都落后0.005至0.010。它在每一组长期配对中都比年龄 × 规模的方案高0.012至0.017；经过相同的样本内截距偏移后，它在三组配对中均优于粗粒度表（俄亥俄州 −0.0024 [−00.0030， −0.0019]）。

因此，该模型在保留俄亥俄州的水平方面并不比表格做得更好，而且其划分也无法迁移。在拟合结果中，较高的移除率会移除仍登记在建造年份地址的客户，并降低 σ，而俄亥俄州观测到的 σ 更高。若要保留俄亥俄州的水平，就需要一条登记册特定规则来决定移除哪些人，而仅凭留存率无法识别这一点。

10 讨论

10.1 应向家庭系统提出什么要求

如果你购买家庭数据，请向系统提出三个问题，并在自己的文件中核查第三个问题。每项存储的分配所依据的证据有多旧——从上次刷新算起？当客户的姓名和地址与该证据不再一致时，系统会返回什么？系统能为多少比例的客户给出答案？匹配率只回答了最后一个问题。

这里的曲线适用于这样一种系统：仅当每个分配的存续时间分别从其上次刷新开始计算时，才单独刷新各个成员，并且刷新不会由这些变化本身触发。人员链接精度仅衡量人员链接本身（第 7.2 节）；在这里影响较小，是因为对于错误链接很少或大多仍留在客户家庭内部的匹配器而言，这种影响有限。对于这里测量的匹配器，更严格的阈值可以使有效性维持到大约等于该系统所服务人口中留居者的时效性，但代价是覆盖率下降（第 7.1 节）。以大约十个月前的证据为基础进行年度重建，在八年时，其匹配结果中有 0.800 是正确的（第 7.3 节）。

如果你建造这样的系统，那么你自己的构建时特征可以让你针对存储家庭的各个分组，在仍列出的客户中说明其时效性。在留出县上，校准误差接近一个百分点；对于较新版本的建造年份，在更晚的日期，漂移约为四个百分点。这些是针对各分组的概率，并不能说明哪个客户的家庭仍然正确。

只能刷新部分文件的系统，应根据证据年龄和家庭规模所预测的收益进行排序：这一规则比优先刷新年龄最大者的做法高出 0.0165，而刷新范围仅占文件的四分之一。表格并未显示按该排序在四分之一范围内达到所需的改进，这也是其评判标准。在 2022年决策日期（第 8.3 节），刷新一半时收益约提高一个百分点，刷新十分之一时则有所下降。使用另一登记册拟合结果的系统，应根据自身文件设定水平（第 9 节）。

研究人员若将某一日期生成的家庭键与多年后观察到的结果关联起来，其处境与买家相同。这个键仍能建立关联，但在此处测量的登记册中，八年后，它所归为一组的人群中，只有约三分之一的客户恰好属于同一个家庭。键的年龄应作为一项测量得到的误差来源纳入分析，与链接率并列。

家庭文件下游一步的情况也是如此。许多家庭层面定位和测量会通过家庭键连接两个文件：一方持有家庭成员的属性，另一方持有能够触达这些成员的设备、地址或账户。连接依据是地址；即使过了八年，仍能在 0.831 的客户中找到该地址。该地址所关联人员的属性，时效性不会超过其背后家庭信息的时效性。根据构建日期记录的人群信息推导出的属性（例如年龄段或生命阶段），只有在家庭未发生变化时，才能确定描述的是当前居住在那里的人。两年后，找到的地址中有 0.701 符合这一情况；即便证据已有十个月之久，这一比例也达到 75.5% 至 85.8%。住宅本身的属性不会以这种方式随时间变旧。我们测量的是登记册衍生图，而非商业文件，因此这些比率针对的是已登记家庭。任何持有家庭键的人都可以进行这项测试：记录每项分配所依据的证据日期，并按证据年限报告精确家庭时效性，同时列出匹配率。

10.2 局限性

参照对象是选民文件在某个单位中记录的人员。未登记者不可见。在一项与多个选民名单和商业名单进行匹配的全国样本中，至少有 11% 的成年公民未出现在其中任何一份名单上，另有 12% 的人员信息被错误列出 [48]。因此，家庭看起来比实际规模更小，而年满 18 岁后登记的孩子则会被视为家庭发生了变化。

参照信息过时是对有效性的主要威胁。过时名册修正使2016年构建的两年时效性降低0.085至0.135（各次构建中为0.024至0.135），并使八年时效性变动 −0.009至 +0.010（各次构建中为−0.017至 +0.013；第6.1节）。构建日期滞后则产生相反影响（第6.2节）。未记录的迁移，以及客户和共同成员的迁入情况，都未能被完整观测；本文没有对其净影响进行界定（第6.3节）。这些部分彼此重叠，不作累加。

从精确的构建版本出发，可以将衰减与构建时链接误差区分开来。第 6.4 节对模拟误差进行了映射；我们没有测量真实供应商的误差。

这项测量涵盖2008年至2026年的一个州，以及从登记册中识别出的可追踪家庭中的登记者。迁移分析涵盖俄亥俄州的一个构建版本和华盛顿州的两个构建版本。确认是在同一州、同一时期内进行的县外复现，迁入者则跨越县界。推断为名义推断（第4.5节）。每项策略都是系统根据过时证据作答的一种风格化版本。五项评估使用的数据，其汇总结果此前已接受检视（第4.4节）。

10.3 开放问题

1. 实际供应商的构建时误差：第 6.4 节将模拟误差映射出来，而只有用登记册对供应商文件进行评分，才能测量实际误差。2. 不活跃后的移除：登记册中的后续记录能否指出这些人员去了哪里，并缩小最大的一项剩余未知？3. 大规模击败 κ_S：一种方法若要返回人员的建造年份地址和当前地址中预期共同成员 F1 值最高的人员子集，或选择不作判断，就需要制定自己的评估标准，并使用本文未用过的留出集，例如后续文件或另一个州的数据。4. 构建间漂移：2018 年构建中，年份间隔从两年扩大到八年时的差距增幅为 0.158，是六个构建中最低的；总统选举两年时效性在 2008年、2012年、2016年和2024年构建中持续下降（确认县中从 0.731 降至 0.660，见表格 S5 和 S14）。随着 2020 至 2024 年构建逐渐达到八年，完整年龄曲线是否会继续变化，仍有待研究。5. 不同登记册的水平：针对移除哪些客户制定的登记册特定规则能否达到俄亥俄州的水平，以及如何在不使用该规则旨在预测的结果的情况下估计这项规则。6. 按预测增益刷新：拟合增益本身的模型，而不是对两个结果预测值求差（第 8.3 节）。

11 结论

存储的家庭归属失去有效性的速度，远快于其锚点变得难以找到的速度。在北卡罗来纳州的登记册中，对于那些八年后仍能找到地址的客户而言，2016年家庭归属仍有0.322的比例完全准确（个人锚点为0.329）。无论如何解读登记册中的过时名录记录，这一比例都约为三分之一；在第6节所考察的目标替代方案和压力测试下，该比例为0.26至0.40，这一范围适用于登记共同居住，而不是实际家庭的上限。

对于迁居者，在从两年到八年的每个评估年龄段，其存储的家庭信息只有五分之一左右是正确的，因此信息失准不仅涉及留居者，也涉及留居家庭的变化。该匹配器更严格的阈值主要筛选出留居者，因为他们的家庭最有可能保持完整；而

新鲜证据对时效性的提升最大。可以根据系统自身的构建时特征预测各分组的时效性；家庭变更多状态模型则推算出最长达十年的时效性。在留出县的 20 个两年至八年单元格中，平均低估约 0.014。两种方法都能给出各分组的概率，但无法确定哪位客户的家庭信息仍然准确；而且该模型未能将其水平迁移至俄亥俄州。

家庭文件应记录每项分配所依据证据的证据年龄；报告找到多少客户的系统，也应报告其返回的家庭中有多少仍然准确。多年来，找到人员或地址一直很容易。但他们所在的家庭并不会一直准确：八年后，约三分之一仍然完全准确。

补充材料。 补充材料中包含补充材料 A 至 M；表 S1 给出了符号说明。

声明

数据与伦理。 本研究使用北卡罗来纳州州选举委员会根据州法律发布为公共记录的选民登记快照 [74],、全州选民历史文件和北卡罗来纳州公共地址点；对于第 9 节，还使用俄亥俄州和华盛顿州的选民档案。这些数据由作者以独立研究员身份，按照各州公布的条款获取：俄亥俄州公开提供其全州档案供下载 [76], ，华盛顿州向任何提出请求的人员提供其登记名单，并禁止将其用于商业广告或招揽 [95]。本研究遵守这些条款：不出售、共享或再分发任何文件，也不使用任何记录向任何人投放广告、进行招揽或联系任何人。密歇根州的档案通过该州依据《信息自由法》提出的合格选民档案数据请求获得 [65], ，仅用于检查标识符持续性，不作其他用途。研究未与任何人员进行互动或干预，也未链接至任何个人层级来源；除州政府自身的登记和选民历史文件外，地址点不包含个人数据。姓名和地址仅用于构建家庭键并在私有环境中模拟匹配，绝不发布或再分发。不识别或联系任何个人，任何结果也不用于对任何人进行画像或定位。仅报告或发布汇总数据：在北卡罗来纳州覆盖 20 个或更多县，在其他州覆盖全州范围；除县划分记录每个样本属于哪个县外，发布内容不包含县级数据或个人层级记录；发布的单元格中人数少于 10 人的会合并；在建造年份被编码为未成年人的人员不作为客户纳入（第 4.2 节）；不报告按种族或族裔进行的分析。作者是一名没有机构审查委员会的独立研究员，判断对公共记录进行且不涉及互动或干预的二次分析无需审查；美国共同规则将其归为可豁免的、针对公开可用信息的二次研究（45 CFR § 46.104(d)(4)）；未作出任何机构层面的认定。

数据可用性。 北卡罗来纳州的快照、全州选民投票历史文件和地址点均可从州选举委员会和 NC OneMap 公开下载；代码发布包列出了每个文件及其校验和。俄亥俄州和华盛顿州根据上述条款提供当前文件；本文使用的文件在代码发布包（补充材料 L.1）中通过文件名、大小和 SHA-256 标识，且不予再分发。不共享任何人员级数据。

代码可用性。 代码使用 Python 编写，位于 https://github.com/methodtrace/2026-stale-household-assignments，本文发布时已公开，并存档于https://doi.org/10.5281/zenodo.24026471。使用 41 份公开的北卡罗来纳州快照文件，一次端到端运行即可按文中打印的精确度复现本文及其补充材料中的所有北卡罗来纳州数值。例外是建造年份选举投票限制，以及地址点审计（第6节第2小节和表格 S9）；这两项需要使用同一份每周选民历史文件副本和地址点服务副本；发布包记录了所用副本的校验和。俄亥俄州和

华盛顿州的结果需要使用发布包中列明日期、大小和 SHA-256 校验和的提取文件。各州提供其当前文件，而较早的提取文件可能无法获取；使用其他文件时，代码仍可运行，但会得出不同的数值。发布包包含代码、输入架构和汇总结果说明，包括拟合后的表格、家庭变更模型的设定及拟合系数，以及县划分；但不包含任何输入记录。运行过程会拟合并评分未合并的表格；发布的表格文件会将少于 10 人的单元格合并，因此仅根据这些文件进行评分会略有差异。

资助。 本研究未获得外部资助；费用由作者承担。

利益冲突。作者声明不存在利益冲突。作者是美国专利 10,419,883 和 11,223,926 的共同发明人，这些专利涉及将移动和非移动设备与地理区域进行统计关联：对设备的位置观测进行聚类，为每个聚类评估设备用户居住在该处某个家庭的可能性，并将设备分配至得分最高的区域。本研究未使用任何获专利保护的方法。

AI 工具的使用。 在准备本研究期间，作者使用 Claude Opus（Anthropic）审查计算代码、检索文献并协助起草文本。使用这些工具后，作者根据需要审阅并编辑了内容，并对已发表文章的内容承担全部责任。

参考资料

[1] R. Abramitzky、L. Boustan、K. Eriksson、J. Feigenbaum 和 S. Pérez。2021。历史数据的自动链接。《经济文献杂志》 59, 3 2021 865 918 10 1257 20201599 ( ), –。https://doi.org/。/jel。E. Amigó、J. Gonzalo、J. Artiles 和 F. Verdejo。2009。基于形式约束的外部聚类评估指标比较。《信息检索》 12, 4 2009 461 486 10 1007 10791 008 9066 8 ( ), –。https://doi.org/。/s---S. Ansolabehere 和 E. D. Hersh。2012。验证：大数据揭示了哪些关于调查误报和实际选民群体的信息。《政治分析》 20, 4 2012 437 459 10 1093 023 ( ), –。https://doi.org/。/pan/mps S. Ansolabehere 和 E. D. Hersh。2017。ADGN：一种使用地址、出生日期、性别和姓名进行记录链接的算法。《统计与公共政策》 4, 1 2017 1 10 10 1080 2330443 2017 1389620 ( ), –。https://doi.org/。/ X.。S. Ansolabehere、E. D. Hersh 和 K. A. Shepsle。2012。迁移者、留居者与登记：为什么年龄与美国的登记存在相关性。《政治科学季刊》 7, 4 2012 333 363 10 1561 100 00011112 ( ), –。https://doi.org/。/。L. Antonie、K. Inwood、C. Minns 和 F. Summerfield。2020。系统链接历史人口普查记录时遇到的选择偏差。《社会科学史》 44, 3 2020 555 570 10 1017 2020 15 ( ), –。https://doi.org/。/ssh。。A. Bagga 和 B. Baldwin。1998。共指链评分算法。载于《LREC'98 语言学共指研讨会会议录》，563–566。[8] M. J. Bailey、C. Cole、M. Henderson 和 C. Massey。2020。自动链接方法的表现有多好？来自美国历史数据的启示。《经济文献杂志》 58, 4 2020 997 1044 10 1257 20191526 ( ), –。https://doi.org/。/jel。D. Ballou、R. Wang、H. Pazer 和 G. K. Tayi。1998。构建信息生产系统模型以确定信息产品质量。《管理科学》 44, 4 1998 462 484 10 1287 44 4 462 ( ), –。https://doi.org/。/mnsc。。C. Batini、C. Cappiello、C. Francalanci 和 A. Maurino。2009。数据质量评估与改进的方法论。《ACM计算调查》 41, 3, 16 2009 10 1145 1541880 1541883 文章( )。https://doi.org/。/。I. Bhattacharya 和 L. Getoor。2007。关系数据中的集体实体消歧。《 ACM 数据知识发现汇刊》 1, 1, 5 2007 10 1145 1217299 1217304 文章( )。https://doi.org/。/。M. Bouzeghoub。。数据新鲜度分析框架。载于2004年信息系统信息质量国际研讨会会议录（IQIS 2004 59 67 10 1145 1012453 1012464 ），–。https://doi.org/。/。C. Cappiello、C. Francalanci 和 B. Pernici。2003。多渠道信息系统中数据质量的时间相关因素。《管理信息系统杂志》20，3（2003），71–92。https://doi.org/10 1080 07421222 2003 11045769 。/。。P. Chaudhary 和 K. Khandelwal。。动态广告插入的触发时机、截止期限就绪度与事件对齐会计。arXiv: 2609 19899 10 48550 2609 19899 。。https://doi.org/。/arXiv。。Y.-H. Chiang、A. Doan 和 J. F. Naughton。2014。为时序记录匹配建模实体演变。载于2014年ACM SIGMOD数据管理国际会议会议录，1175–1186。https://doi.org/10 1145 2588555 2588560 。/。Y.-H. Chiang、A. Doan 和 J. F. Naughton。2014。追踪动态世界中的实体：一种用于匹配时序记录的快速算法。《VLDB基金会会议录》7，6（2014），469–480。https://doi.org/10.14778/2732279.2732284

[17] J. Cho 和 H. Garcia-Molina。2003。网络爬虫的有效页面刷新策略。《ACM数据库系统学报》 28, 4 2003 390 426 10 1145 958942 958945 ( ), –。https://doi.org/。/。P. Christen。。《数据匹配：记录链接、实体解析与重复检测的概念和技术》。柏林：Springer。https://doi.org/ 10 1007 978 3 642 31164 2 。/- ----P. Christen。。为记录链接与重复检测研究准备真实的时间性选民数据集。技术报告，澳大利亚国立大学，最后更新于2014年6月29日。https://users.cecs.anu.edu.au/~Peter.Christen/publications/ncvoter-report-29june2014.pdf[20] P. Christen 和 R. W. Gayler。2013。动态数据库上的自适应时间实体解析。载于《知识发现与数据挖掘进展》（PAKDD 2013），《计算机科学讲义》，第7819卷，558–569。Springer。https://doi.org/ 10 1007 978 3 642 37456 2_47 。/----P. Christen、D. J. Hand 和 N. Kirielle。。F度量综述：历史、性质、批评与替代方案。《ACM计算调查》 56, 3, 73 2024 10 1145 3606367 文章( )。https://doi.org/。/ V. Christen、A. Groß、J. Fisher、Q. Wang、P. Christen 和 E. Rahm。。人口普查数据的时间群体链接与演化分析。载于第20届数据库技术扩展国际会议会议录（EDBT 2017 620 631 10 5441 002 2017 83 ），–。https://doi.org/。//edbt。。CIMM、Go Addressable 和 Truthset。。使用IP地址进行受众识别与测量所面临的挑战。创新媒体测量联盟。2025年11月。https://cimm-us.org/wp-content/uploads/2025/11/CIMM-_-GoAddressable-_-Truthset-Challenges-of-Using-IP-Address-for-Audience-Identification-and-Measurement-November-2025.pdf[24] S. Cuccaro-Alamin、A. L. Eastman、R. Foust、J. McCroskey、H. T. Nghiem 和 E. Putnam-Hornstein。2021。利用链接的行政记录构建家庭与家族单元的策略。《儿童与青少年服务评论》 120 2021 105706 10 1016 2020 105706 ( ), 。https://doi.org/。/j.childyouth。。X. L. Dong、L. Berti-Équille 和 D. Srivastava。。动态世界中的真值发现与复制检测。《VLDB基金会会议录》 2, 1 2009 562 573 10 14778 1687627 1687691 ( ), –。https://doi.org/。/。T. Enamorado、B. Fifield 和 K. Imai。。使用概率模型辅助合并大规模行政记录。《美国政治学评论》 113, 2 2019 353 371 10 1017 0003055418000783 ( ), –。https://doi.org/。/ S W. Fan、F. Geerts、N. Tang 和 W. Yu。。利用数据时效性与一致性进行冲突消解。《数据与信息质量期刊》 5, 1 2, 6 2014 10 1145 2631923 –文章( )。https://doi.org/。/ W. Fan、F. Geerts 和 J. Wijsen。。确定数据的时效性。《ACM数据库系统学报》37，4，文章 25 2012 10 1145 2389241 2389244 ( )。https://doi.org/。/。I. P. Fellegi 和 A. B. Sunter。。记录链接理论。《美国统计学会会刊》64，328（ 1969 1183 1210 10 1080 01621459 1969 10501049 ），–。https://doi.org/。/。。B. Foley、I. Shuttleworth 和 D. Martin。。行政数据质量：调查北爱尔兰健康登记册中的记录级地址准确性。《官方统计学报》34，1（2018），55–81。https://doi.org/ 10 1515 2018 0004 。/jos---Z. Fu、H. M. Boot、P. Christen 和 J. Zhou。2014。历史人口普查数据中个人与家庭的自动记录链接。《人文与艺术计算国际期刊》8，2（2014），204–225。https://doi.org/ 10 3366 2014 0130 。/ijhac。。Z. Fu、P. Christen 和 J. Zhou。。历史人口普查家庭链接的图匹配方法。载于《知识发现与数据挖掘进展》（PAKDD 2014），《计算机科学讲义》，第8443卷，485–496。Springer。https://doi.org/ 10 1007 978 3 319 06608 0_40 。/----A. Gruenheid、X. L. Dong 和 D. Srivastava。。增量记录链接。《VLDB基金会会议录》7，9 2014 697 708 1014778 2732939 （），–。https://doi.org。/。D. Hand 和 P. Christen。。关于使用F度量评估记录链接算法的说明。《统计与计算》 28, 3 2018 539 547 10 1007 11222 017 9746 6 ( ), –。https://doi.org。/s---G. Harper、N. Firman、M. Wilk、M. Marszalek、P. Simon、D. Stables、R. Fry、K. Smith 和 C. Dezateux。2024。根据患者地址和唯一房产参考编号确定家庭：全科医生电子健康记录中的应用。《人口数据科学国际期刊》9，1，文章2379（2024）。https://doi.org/ 10 23889 9 1 2379 。/ijpds.v i。G. Harper 和 L. Mayhew。。利用行政数据通过地方应用统计和分类家庭。《应用空间分析与政策》 9, 4 2016 433 462 10 1007 12061 015 9162 2 ( ), –。https://doi.org。/s---G. Harper、D. Stables、P. Simon、Z. Ahmed、K. Smith、J. Robson 和 C. Dezateux。。评估ASSIGN开源确定性地址匹配算法：为全科医生记录的患者地址分配唯一房产参考编号。《人口数据科学国际期刊》6，1，文章1674（2021）。https://doi.org/ 10 23889 6 1 1674 。/ijpds.v i。K. L. Harron、J. C. Doidge、H. E. Knight、R. E. Gilbert、H. Goldstein、D. A. Cromwell 和 J. H. van der Meulen。2017。评估链接质量以分析链接数据的指南。《国际流行病学杂志》46，5（2017），

1699 1710 10 1093 177 – . https://doi.org/ . /ije/dyx J. C. Hays、G. G. Fillenbaum、D. T. Gold、M. C. Shanley 和 D. G. Blazer。1995年。老年人家庭构成稳定性的黑人与白人、城市与农村差异。《老年学杂志：B辑》50B，5（1995），S301–S311。https://doi.org/ 10 1093 50 5 301 。/geronb/b。 .s B. Heinrich、D. Hristova、M. Klier、A. Schiller 和 M. Szubartowicz。2018年。数据质量指标的要求。《数据与信息质量期刊》 9, 2, 12 2018 10 1145 3148238 文章（）。https://doi.org/ 。/ B. Heinrich 和 M. Klier。 。考虑补充数据的及时性新型数据质量指标。载于《第17届欧洲信息系统会议论文集》（ECIS 2009），2701–2713。[42] B. Heinrich 和 M. Klier。2011年。评估数据时效性：一种概率方法。《信息科学杂志》37，1（ 2011 86 100 10 1177 0165551510392653 ），–。https://doi.org/ 。/ B. Heinrich 和 M. Klier。 。基于指标的数据质量评估：开发和评估基于概率的时效性指标。《决策支持系统》 72 2015 82 96 10 1016 2015 02 009 （），–。https://doi.org/ 。/j.dss。 。 . B. Heinrich、M. Klier 和 M. Kaiser。 。开发时效性指标及其在CRM中的应用流程。《数据与信息质量期刊》 1, 1, 5 2009 10 1145 1515693 1515697 文章（）。https://doi.org/ 。/ 。 E. D. Hersh 和 Y. Ghitza。 。美国的混合党派家庭与选举参与。《PLOS ONE》 13, 10 2018 0203997 10 1371 0203997 （），e。https://doi.org/ 。/journal.pone。 G. A. Huber、M. Meredith、M. Morse 和 K. Steele。 。选民名单维护错误造成的种族负担：来自威斯康星州补充迁移选民名册的证据。《科学进展》7，8（2021），eabe4498。https://doi.org/ 10 1126 4498 。/sciadv.abe J. L. Ish、M. Daniel、P. Ringwald、N. M. Niehoff、R. R. Jones 和 A. J. White。2026年。Sister Study队列中源自LexisNexis的回溯性地址历史的准确性。《暴露科学与环境流行病学杂志》 36, 2 2026 244 250 10 1038 41370 025 00802 1 （），–。https://doi.org/ 。/s--- S. Jackman 和 B. Spahn。 。在美国政治上隐形。《PS：政治科学与政治》54，4（2021），623–629。https://doi.org/ 10 1017 1049096521000639 。/S C. H. Jackson。 。面板数据的多状态模型：R的msm软件包。《统计软件杂志》38，8（ 2011 1 28 10 18637 038 08 ），–。https://doi.org/ 。/jss.v.i R. D. Johnson、L. J. Griffiths、J. P. Hollinghurst、A. Akbari、A. M. Lee、D. A. Thompson、R. A. Lyons 和 R. J. Fry。2021年。利用人口规模电子健康记录数据推导家庭构成：一种可复现的方法。《PLOS ONE》 16, 3 2021 0248195 10 1371 0248195 （），e。https://doi.org/ 。/journal.pone。 J. D. Kalbfleisch 和 J. F. Lawless。 。马尔可夫假设下的面板数据分析。《美国统计学会会刊》 80, 392 1985 863 871 10 1080 01621459 1985 10478195 （），–。https://doi.org/ 。/ 。 . A. Kaplan、B. Betancourt 和 R. C. Steorts。 。对链接数据进行适当推断的实用方法。《美国统计学家》 76, 4 2022 384 393 10 1080 00031305 2022 2041482 （），–。https://doi.org/ 。/ 。 . A. D. Keller。 。人口普查中行政记录的使用：建模与处理。载于《JSM论文集：调查研究方法分会》。美国统计学会。[54] S. S. Kim、S. Schneider 和 R. M. Alvarez。2020年。评估选民登记数据库中变更的质量。《美国政治研究》 48, 6 2020 670 676 10 1177 1532673 19870512 （），–。https://doi.org/ 。/ X N. Kirielle、P. Christen 和 T. Ranbaduge。 。复杂实体的无监督基于图的实体解析。《ACM数据知识发现汇刊》 17, 1, 12 2023 10 1145 3533016 文章（）。https://doi.org/ 。/ J. Lam、M. Cortina-Borja、P. Christen、R. Thomas、R. Aldridge、R. Blackburn、R. Gilbert、A. Boyd 和 K. Harron。2025年。确定行政数据中地址变更时间的策略。《人口数据科学国际期刊》 10, 1, 2995 2025 10 23889 10 1 2995 文章（）。https://doi.org/ 。/ijpds.v i 。 G. Lansley、W. Li 和 P. A. Longley。 。利用行政登记册和消费者登记册表征人口动态。载于《英国GIS研究第25届会议论文集》（GISRUK 2017）。https://discovery.ucl.ac.uk/id/eprint/1556511/ [58] G. Lansley、W. Li 和 P. A. Longley。2019年。创建用于细粒度人口统计分析的链接消费者登记册。《皇家统计学会会刊：A辑》 182, 4 2019 1587 1605 10 1111 12476 （），–。https://doi.org/ 。/rssa。 F. Li、M. L. Lee、W. Hsu 和 W.-C. Tan。 。用于实体画像的时序记录链接。载于《2015年ACM SIGMOD数据管理国际会议论文集》， 593 605 10 1145 2723372 2737789 –。https://doi.org/ 。/ 。 P. Li、X. L. Dong、A. Maurino 和 D. Srivastava。 。链接时序记录。《VLDB基金会会议录》 4, 11 2011 956 967 10 14778 3402707 3402733 （），–。https://doi.org/ 。/ 。 R. Linacre、S. Lindsay、T. Manassis、Z. Slade 和 T. Hepworth。2022年。Splink：大规模概率记录链接免费软件（会议摘要）。《人口数据科学国际期刊》7，3，文章1794（2022）。https://doi.org/ 10 23889 7 3 1794 。/ijpds.v i 。 S. Madnick、R. Wang 和 X. Xian。 。设计和实现企业家庭归并知识处理器以提升数据质量。《管理信息系统杂志》20，3（2003），41–70。https://doi.org/10.1080/07421222.2003.11045772

[63] 2003 10 1007 97478 C. F. Manski。 。
[64] 2010 概率分布的部分识别。纽约：Springer。https://doi.org/ 。/b D. Menestrina、S. E. Whang 和 H. Garcia-Molina。 。评估实体解析结果。《VLDB基金会会议录》 3, 1 2010 208 219 10 14778 1920841 1920871 （），–。https://doi.org/ 。/ 。密歇根州务院。n. d. 。合格选民档案数据申请表及说明。[65] [ ] https://www.michigan.gov/-/media/Project/Websites/sos/02lehman/FOIA_FORM.pdf（访问于2026年11月）。[66] M. Milani、Z. Zheng 和 F. Chiang。2019年。CurrentClean：陈旧数据的时空数据清洗。载于《第 35 2019 172 183 10 1109 2019 00024 届IEEE国际数据工程会议论文集》（ICDE），–。https://doi.org/ 。/ICDE。 。[67] 2023 V. T. Mule、A. Keller、M. H. Mulry 和 S. Konicki。 。美国2020年十年一次的人口普查中的行政记录建模使用情况。载于《第64届国际统计学会世界统计大会论文集》，IPS论文1122。[68] M. H. Mulry、C. J. Tello-Trillo、V. T. Mule 和 A. Keller。2024年。行政记录名册与人口普查自填答复及NRFU家庭成员答复的比较。《国际官方统计协会统计期刊》40，1（2024），41–52。https://doi.org/ 10 3233 230048 。/SJI- A. H. Murphy。 。[69] 1973 概率评分的新向量划分。《应用气象学杂志》12，4（1973），595 600 10 1175 1520 0450 1973 012 3 0595 3 2 0 2 –。https://doi.org/ 。/- （） % C :ANVPOT% E 。 .[70] 2019 CO; C. Nanayakkara、P. Christen 和 T. Ranbaduge。 。用于群组记录链接的稳健时间图聚类。载于《知识发现与数据挖掘进展》（PAKDD 2019 11440，526 538 101007 978 3 030 16145）3, 41《计算机科学讲义》，第–卷。Springer。https://doi.org/ 。/---[71] 2020 C. Nanayakkara、P. Christen、T. Ranbaduge 和 E. Garrett。 。基于群组的记录链接评估指标。《人口数据科学国际期刊》 4, 1, 1127 2020 10 23889 4 1 1127 文章（）。https://doi.org/ 。/ijpds.v i 。[72] 2024 N. Neumann、C. E. Tucker、L. Kaplan、A. Mislove 和 P. Sapiezynski。 。数据荒漠与黑箱：社会经济地位对消费者画像的影响。《管理科学》70，11（2024），8003–8029。https://doi.org/ 10 1287 2023 4979 。/mnsc。 。[73] N. Neumann、C. E. Tucker 和 T. Whitfield。2019年。前沿：第三方消费者画像有多有效？来自实地研究的证据。《营销科学》 38, 6 2019 918 926 10 1287 2019 1188 （），–。https://doi.org/ 。/[74] mksc。 。北卡罗来纳州议会。n. d. 。[ ] N.C.G.S. §-。选民登记官方记录。163 82 10, https://www.ncleg.gov/EnactedLegislation/Statutes/HTML/BySection/Chapter_163/GS_163-82.10.html（访问于2026年11月）。[75] 北卡罗来纳州选举委员会。 [n. d.]。2005年至2026年的选民登记快照。https://dl.ncsbe.gov/?prefix=data/Snapshots/（访问于2026年11月）。[76] 俄亥俄州务卿。 [n. d.]。选民档案下载页面。https://www6.ohiosos.gov/ords/f?p=VOTERFTP:HOME（访问于2026年11月）。[77] B.-W. On、N. Koudas、D. Lee 和 D. Srivastava。2007年。群组链接。载于《第23届IEEE国际数据工程会议论文集》（ICDE *2007* 496 505 10 1109 2007 367895 ），–。https://doi.org/ 。[78] /ICDE。 。[79] 2021 F. Panse、A. Düjon、W. Wingerath 和 B. Wollmer。 。利用历史选民数据生成用于大规模重复检测的真实测试数据集。载于《第24届数据库技术扩展国际会议论文集》（EDBT *2021* 570 581 10 5441 002 2021 67 ），–。https://doi.org/ 。//edbt。 。 J. Pasek、S. M. Jang、C. L. Cobb III、J. M. Dennis 和 C. DiSogra。2014年。营销数据能否助力调查研究？考察消费者档案数据的准确性与完整性。《公众舆论季刊》78，4（2014），889–916。https://doi.org/ 10 1093 043 。/poq/nfu 皮尤研究中心。 。商业选民档案与美国政治研究。2018年2月15日。[80] 2018 https://www.pewresearch.org/methods/2018/02/15/commercial-voter-files-and-the-study-of-u-s-politics/ [81] L. L. Pipino、Y. W. Lee 和 R. Y. Wang。2002年。数据质量评估。《ACM通讯》45，4（2002），211 218 10 1145 505248 506010 –。https://doi.org/ 。/ 。 R. K. Raley、I. Weiss、R. Reynolds 和 S. E. Cavanagh。2019年。利用纵向家庭名册数据估计儿童从出生至18岁期间的家庭不稳定性。《人口学》56，5（2019），1957–1973。https://doi.org/ 10 1007 13524 019 00806 1 。/s--- S. Razniewski。 。优化衰减信息的更新频率。载于《第25届ACM信息与知识管理国际会议论文集》（CIKM 2016），1191–1200。https://doi.org/ 10 1145 2983323 2983719 。/ 。[82] [83] 2016 T. Richards 、 M. J. White 和 A. O. Tsui。1987 年。居住安排变迁：家庭类型之间转变的风险模型。《人口学》 24, 1 1987 77 97 10 2307 2061509 （），–。https://doi.org/ 。/ S. E. Rodgers、R. A. Lyons、R. Dsilva、K. H. Jones、C. J. Brooks、D. V. Ford、G. John 和 J.-P. Verplancke。2009年。住宅匿名链接字段（RALFs）：研究环境与个人健康相互作用的新型信息基础设施。《公共卫生杂志》31，4（2009），582–588。https://doi.org/ 10 1093 041 。/pubmed/fdp S. Ruggles、C. A. Fitch 和 E. Roberts。2018年。历史人口普查记录链接。[84] [85] [86] 《社会学年度评论》44（2018），19–37。https://doi.org/10.1146/annurev-soc-073117-041447

[87] 新西兰统计局。2022年。行政数据枚举：2023年人口普查的规划方法。惠灵顿：新西兰统计局。https://www.stats.govt.nz/methods/admin-enumeration-planned-approach-for-the-2023-census/ [88] K. S. Tingay、A. Anastasiou 和 C. Musselwhite。2018年。使用链接的威尔士行政数据建立家庭的纵向历史模型（会议摘要）。《人口数据科学国际期刊》3，4，文章962（ 2018 10 23889 3 4 962 ）。https://doi.org/ . /ijpds.v i . A. C. Titman 和 L. D. Sharples。 。具有相位型驻留时间分布的半马尔可夫模型[89]。2010。《生物统计学》66，3（2010年），742–752。https://doi.org/10.1111/j.1541-0420.2009.01339.x [90] 美国人口普查局。 [n. d.]。《美国社区调查》1年期估计，北卡罗来纳州：表B07001，按年龄划分的地理流动性，2010年至2019年及2021年至2024年，以及2008年至2021年的全年龄迁移率（访问于2026年11月）。https://data.census.gov/ [91] J. T. van Dijk、G. Lansley 和P.A. Longley。2021年。使用链接的消费者登记册估算英国的居住迁移。《皇家统计学会会刊：A辑》184，4（2021年），1452–1474。https://doi.org/ 10 1111 12713 。/rssa。E. van Imhoff 和 N. Keilman。1991年。LIPRO 2.0：动态人口预测模型在荷兰家庭结构中的应用。阿姆斯特丹：Swets and Zeitlinger。[92] [93] G. Venkatadri、P. Sapiezynski、E. M. Redmiles、A. Mislove、O. Goga、M. L. Mazurek 和 K. P. Gummadi。2019年。通过脸书广告平台审计离线数据经纪商。载于《万维网大会论文集》（WWW ’ *19* 1920 1930 10 1145 3308558 3313666 ），–。https://doi.org/ . / 。R. Y. Wang 和 D. M. Strong。 。超越准确性：数据质量对数据消费者意味着什么。《管理信息系统杂志》[94] 1996 12, 4 1996 5 33 10 1080 07421222 1996 11518099 （ ），–。https://doi.org/ . / 。 。华盛顿州议会。n. d.。RCW A.。：登记、投票记录作为公共记录；所提供的信息；限制。[95] [ ] 29 08 720 https://app.leg.wa.gov/rcw/default.aspx?cite=29A.08.720（访问于2026年11月）。[96] Y. Zak 和 A. Even。2017年。开发和评估用于检测和处理数据时效性下降的连续时间马尔可夫链模型。《决策支持系统》103 2017 82 93 10 1016 2017 09 006 （ ），–。https://doi.org/ . /j.dss。 。 。Y. Zeng、K. C. Land、Z. Wang 和 D. Gu。 。美国家庭户的惯性与动态：ProFamy方法的扩展与应用。《人口研究与政策评论》25，1（2006年），1–41。https://doi.org/10.1007/s11113-006-7034-9 [97] 2006