大数据时代的因果推断

程开明 姜山 李泗娥

维克托·迈尔·舍恩伯格在《大数据时代》一书中提出：“大数据时代，人们应放弃对因果关系的渴求，转而关注相关关系；没有必要非得知道现象背后的原因，相关关系能够帮助我们更好地了解这个世界。”图灵奖得主朱迪亚·珀尔则在最近大热的《为什么》一书中直言不讳地讲到：“数据非常愚蠢，领会因果关系才是理解世界的关键；除非由算法控制的机器能够推理因果关系，否则其效用和通用性永远不会接近人类。”显然，两人的观点大相径庭。

那么，大数据时代究竟是应该致力于相关关系测定，还是要注重因果关系的探求呢？一些人认为，对因果性的研究是小数据时代的产物，当大数据时代面对无法估量的数据时，人们无法再去寻求事物间的因果关系，只能寻找事物间的相关关系。因此，在大数据时代，相关性比因果性更重要。然而，因果思维是人类主观认知结构的核心，是人们认识世界、改造世界过程中的惯常思维模式，也是科学研究的坚实基础。故而，大数据时代，既要探测相关关系，又不能止步于相关性，还需利用因果推断方法进一步探求相关背后的因果。

区域出现信息时不应考虑该用户。

乒乓数据的处理。乒乓数据是指通信终端在多个相邻近的基站小区间出现交替切换现象而产生的数据，但用户并未发生一定距离的移动行为。理论上，当用户移动到某一基站小区时，会自动切换到该基站小区的信号。而实际上，在通信网络中由于无线电波传播过程中存在反射、绕射、散射等现象，导致在很短的时间段内信号出现异常的波动，也就是所谓的“乒乓效应”，这类数据会高估移动人口的数量。针对这种情况，可遍历所有记录，识别出在两个小区间连续多次切换，且切换间隔仅为几秒钟的数据行，并将其剔除。

漂移数据的处理。在某些情况下，信号会从邻近的基站突然切换到相对较远的基站，并在一定时间之后切换回邻近的基站小区，这种现象为信号漂移现象，由此产生的数据为漂移数据。这种漂移数据的出现会对人口动态监测产生严重影响，如在人流向统计中，很可能多统计较远基站小区的人口，而少统计邻近基站小区的人口，不仅如此，所统计的去向区域也会存在错误。因此，需要对漂移的数据进行判定。通过计算第 i 条数据和第 $i+1$ 条数据的距离和时间间隔，进而得到其速度，若该速度值过大，那么将第 $i+1$ 条数据删除即可，以此过程遍历所有的数据。

在数据清洗流程中，主要是剔除掉冗余数据，乒乓数据以及漂移数据所占比例较小。以我国最大的电信运营商为例，其原始信令数据在北京一天平均约产生 45 亿条，因各种原因所产生的冗余数据的比例约 1%，最终汇总能提取到 7 亿条可用数据。这些数据清洗步骤结束后，对清洗后的数据再检查其 CI，把 CI 异常的数据处理后，才可进行数据的人口监测分析和建模。但是在分析和建模之前，为确认上述清洗标准和流程的准确性，需要对数据进行统计检验，以保证该数据清洗流程的适用性方能凸显数据的价值，进而可以构建更准确、更合理、更有价值的人口动态监测模型，以便为人口统计制度提供可操作的数量及其政策界限，为相关决策提供强有力的支撑。

作者单位：首都经济贸易大学
中国移动通信有限公司研究院

| 大数据时代相关关系之凸显

相关关系是指一个变量发生变化时，另一个变量也随之变化，而不论这两个变量有没有必然联系。因果关系是指当一个作为原因的变量变化时，另一个作为结果的变量也一定程度上发生变化，两个变量之间存在着必然联系。在小数据世界，有限的样本数据无法反映出事物之间的普遍相关性，人们往往执着于现象背后的因果关系，试图通过有限样本数据来剖析其中的内在机理。大数据时代，建立在相关分析基础上的预测正是大数据的核心议题，人们可以通过大数据技术挖掘出事物之间隐蔽的相关关系，获得更多的认知与洞见，进而捕捉当下特征和预测未来趋势。

通过大数据关注线性相关关系及复杂的非线性相关关系，人们可以看到很多以前不曾注意的联系，掌握以前无法理解的复杂社会经济现象，甚至可以超越因果关系，成为了解这个世界的更好视角。正如舍恩伯格指出，大数据让人们关注相关关系，只需知道“是什么”，而不用知道“为什么”。人们不必非得知晓事物或现象背后的复杂深层原因，只需通过大数据分析获知“是什么”就能提供一些新颖且有价值的观点、信息和知识。在大数据时代，人们的思维方式一定程度上从因果思维转向相关思维，颠覆千百年来人类形成的传统因果思维模式。

大数据的价值在于预测，而预测正是建立于相关分析基础之上。相关关系通过识别有用的关联物来帮助人们分析一个现象，而不是通过揭示其内部的运作机制，故不具有必然性，只具有或然性。两类现象或事物之间看上去相关，实质上很可能只是巧合，但通过观察一类对象去预测另一类对象则是切实可行的。利用大数据对相关关系深入地分析和探究，可以预知将会发生什么，而一旦把因果关系考虑进来，其复杂性要求导致这些视角有可能被蒙蔽，相关关系提供的新视角可以帮助人们去发现在因果思维模式下无法知晓的新领域，这些新视角往往具有重要的商业价值。在大数据世界，由因果关系转向相关关系不仅是一个重大的思维转向，而且可以更好地利用相关关系创造出巨大的商业利润。

大数据时代的相关分析利用机器计算能力来寻找到最优的关联物，在各个领域都涌现出一些很好的应用成果，例如亚马逊的推荐系统、可视化呈现的数据新闻等，这些应用通过数据挖掘实现从数据到价值的转变，创造出经济利润和社会效益。亚马逊的推荐算法能够根据消费记录来告诉用户可能会喜欢什么，这些消费记录有可能是别人的，也可能是该用户的历史记录。虽说它不能说出你为什么会喜欢，但通过及时推荐就能实现一定概率的购买行为转化，获取经济利润，这便是相关分析优势的最强说服力！相关关系能够创造利润，表明大数据相关分析已不再是计算、统计等学科的专宠，这只王谢堂前燕正式飞入寻常百姓家，为各行各业所广泛应用，以帮助企业盈利，帮助政府决策。

大数据时代为什么人们强调相关性，而弱化因果性呢？原因可能在于，相关性更广泛，因果性更严格，相关性比较表象容易被识别，而因果性反映事物之间内在的本质关系，不容易被认识和把握（黄欣荣，2017），正如舍恩伯格所言，“有时候，影响因素成千上万，解释的理论更是多如牛毛，强找因果关系很难。”。很多日常生活与商业应用中，知晓相关关系就已足够，相关分析不提供关于世界的真相和原理，只通过知其然而不知其所以然的一些判断来创造属于其自身的价值。在许多场合，只要知道事物之间具有依随性质的相关关系，大致能够推断出与之相关的另一个现象或变量可能会发生的变化，从而抓住商业应用的机会。

| 大数据时代因果关系之必要

尽管对相关关系的探测颇具价值，但相关分析只停留于数据表面，即使相关性很强的对象之间也可能并不存在本质上的关联性。因此，当面对具体的大数据应用时，因果思维仍会不由自主地走上台前，让人们自然而然地想去寻求对象之间的因果联系。舍恩伯格提到：“在大多数情况下，一旦我们完成了对大数据的相关性分析，而又不仅仅满足于‘是什么’时，就会继续向更深层次探究因果关系，找出背后的为什么。”周涛教授也讲到：“放弃对因果关系的追寻，就是人类的自我堕落，相关性分析只是寻找因果关系的利器。”

大数据的相关性并不意味着两个变量具有因果联系，而具有因果联系的两个变量从大数据本身来看有时也并不相关。一般来说，相关关系不能确定两个变量 X、Y 之间是否存在因果关系，因为两个变量之间的相关性可能有三种解释：其一，X 是 Y 的原因或一部分原因；其二，Y 是 X 的原因或一部分原因；其三，X 和 Y 是第三个变量 Z 的原因（结果）或一部分原因（结果）。特别是第三种情况的存在，使相关分析得到的相关性很可能是“伪相关”。譬如，公鸡打鸣与太阳升起之间有相关性，却没有必然的

因果性，不能说因为公鸡打鸣太阳才出来，也不能说因为太阳出来了公鸡才打鸣（有的公鸡半夜也打鸣），它们只是共同受到时间变化的影响。这样的例子不胜枚举，珀尔在《为什么》一书中列举出了众多有趣的实例。

很多情况下，相关关系并不是大数据洞察的终结目标。因果分析是相关分析的深化，大数据的相关关系不仅没有替代因果关系，反而给因果关系的研究提供了更广泛的发展空间。譬如，医疗大数据、基因大数据的相关分析给精准医疗、药物研究等领域带来巨大变革，但仅靠相关关系往往很难找到病因，还是无法对症下药，必须进一步搞清楚内在的因果机制。阿司匹林是治疗感冒的药，经过大量临床数据分析发现阿司匹林对预防心脑血管疾病疗效显著（有相关关系），但仍不能把它作为预防心脑血管疾病的处方药，而后对阿司匹林进行药理分析，才发现阿司匹林中含有治疗心脑血管疾病的药物成分，建立了因果关系，才敢大胆服用。

史蒂夫·洛尔在《大数据主义》一书中认为相关关系可以为商业、医学等应用领域提供有效的预测工具，但不能因此而否定因果性。他借用IBM人工智能专家费鲁奇的话说："对于大量商业决策而言，有相关性就能得出令人满意的结果；但仅凭相关性是不够的，还要对因果关系产生有启发性的认识，包括理论、假设、现实世界的心理模型、事情原委等，两者必须更密切地相互配合。"如果只知道相关性而不清楚因果性，那么大数据分析的深度只达一半，一旦出现问题或疑问就无从下手。如果清楚了因果关系，则能更好地利用相关关系，更好地掌握预测未来的主动权，帮助我们更科学地进行决策（李金昌，2014）。

大数据时代因果探究之价值

人们在认识自然和社会的过程中，通常希望"知其然，又知其所以然"。几千年来，探求事物之间的因果关系是哲学、自然科学和社会科学等众多科学研究所追求的终极目标，数学家、统计学家、物理学家、医学家、哲学家、经济学家等大都将寻求自身研究领域的因果关系当作一生的追求。古希腊哲学家德谟克利特说："我宁肯找到一个因果关系，也不愿获得一个波斯王位。"哲学家培根提出"知识就是力量"，认为"真正的知识是根据因果关系得到的知识"。探索并发现因果关系的方法论，伴随人类社会的发展而不断精深，自古代哲学到现代科学，成为经久不衰的挑战。

农业时代，人们认知事物的思维模式是通过经验观察，以因果关系作为判断的依据和准则。到了工业化时代，人们积累数据的方法和手段发生了质的变化，事物因果关系的揭示过程逐渐科学化，以公理或定律的形式广泛运用于科研和实践。尽管如此，信息与数据之间的联系仍然是间接的，人们对事物的认知仍然依据于因果关系，只是经验观察的成分越来越少（何大安，2018）。休谟认为，因果关系"是我们从经验中得来的关系"，发现因果关系的必要条件包括：第一，凡被认为原因或结果的那些对象总是接近的；第二，在时间上因先于果；第三，原因和结果之间的"恒常结合"之"必然联系"。休谟奠定了科学对于因果关系的基本理解，经过密尔等人的发展，关于确立事物之间因果关系的标准就基本稳定。

因果关系是一切科学技术的基础，珀尔在《为什么》一书中给出很多因果推断的应用实例，来自社会各个领域：社会科学领域中"伯克利招生悖论"的解决以及学历和工作经验对工资水平影响的分析；公共健康领域对疫苗接种合理性的探讨，对19世纪伦敦霍乱发生原因的分析，吸烟是否致癌争论的解决；还有现实生活中诸多的悖论的解决，包括有关"运动－胆固醇水平"的辛普森悖论、有关"饮食－体重"的罗德悖论；遗传研究领域中天性与培养"研究、赖特的"豚鼠毛色"以及"出生－体重"的路径分析；气候预测领域中的异常天气事件与全球变暖的关系分析；法律领域中的罪行判定；等等。

大数据时代的丰富数据为从概率论的立场研究因果关系提供了新视角，其对因果关系的重构，使得因果关系重新焕发生命力。大数据作为总体或全部样本的数据，有助于从根本上克服由于抽样偏颇所引起的样本选择性偏误。若采用单一数据，变量遗漏问题往往非常严重，如果将不同来源的大数据匹配起来，可以克服或缓解变量遗漏问题；尽管在复杂、开放、动态的庞大系统中，因果关系的内生性问题仍难解决，但大数据对因果关系的检验比有限样本的抽样数据更为稳健和可靠，内生性问题也有所改进。另外，大数据通常表现为面板数据和分层数据，这对于确定因果效应也极为有利。

因果关系也是科学和哲学的主题，从因推导出果，找到两类对象之间的规律和相互关系一直是推动科学与哲学前进的动力之一。大数据中一个耳熟能详的说法是：大数据长于相关关系，而非因果关系，但这可能是

一个伪命题。任何科学都想追求因果解释，缺乏因果关系的解释就没有规律；反过来，希望发现规律就必然要追求因果关系。舍恩伯格强调了相关性对大数据的重要性，但并不否定因果性的存在，更没有说要用相关性完全取代因果性，相关关系能为因果关系探求创新条件，因果关系往往能够比相关关系提供更加精确的干预建议。

大数据时代因果推断之策略

如何从相关关系进一步推断出因果关系，是大数据的真正问题所在，因果推断也是当前热门的人工智能（AI）的技术基础。在大数据和人工智能时代，通常使用“数据驱动法”来设置模型和参数，用数量关系来刻画因果关系，在因与果之间架起数据连接（何大安，2018）。基于大数据的因果推断虽然只是对原因和结果关系的检验，却是一种基于相关性的因果关系量化把握，作为变量相互作用过程确定性关系的描述，因果性在更深层次关系到大数据的哲学意蕴（王天恩，2017）。

从统计意义上探讨因果关系，就不是两个变量之间的关系那么简单，因为社会生活中几乎不存在单因单果的现象。统计控制就是将可能对因变量和自变量有影响的因素纳入模型，如此因果关系的问题就转变成了因果效应。前文已提到在统计模型中准确估计因果效应主要受制于三个因素：样本选择性偏误、变量遗漏以及内生性问题。为剔除这些因素的影响，用于验证因果关系的理想办法是做随机控制实验。但现实中，尤其是对哲学社会科学来说，很难像自然科学一样对各变量进行严格控制环境下的实验。随着因果关系哲学基础的建立，基于观测性数据的因果关系发现算法或因果推断框架，逐渐成为数据科学中可能创造商业价值和进行科学发现的重要研究领域之一。

基于反事实理论框架和随机化实验思想的因果推断在政策评估中得到广泛应用。过去学者在研究政策效应时，往往只是借助定性分析，结论的科学性难以令人信服。因果推断则为政策评估提供了基于观测数据开展实证研究的新方法，可以帮助人们进一步揭示变量间的因果关系，更好地识别政策效应。从已有文献看，巧妙开展因果推断的方法主要包括多元回归、倾向值匹配、工具变量法、双重差分、断点回归等，另外还有基于时序数据格兰杰因果关系检验及基于结构方程模型的因果关系发现方法。

大数据时代需要深层次的因果分析，当前开展因果推断的两种代表性方法是以唐纳德·鲁宾为代表的结构因果模型和以朱迪亚·珀尔为代表的因果图方法。结构因果模型尝试利用结构方程模型，对潜在结果开展建模，从因果作用机制引发的数据分布特性等视角发现事物间的因果关系。因果图方法是在有向无环图（DAG）上清晰地引入因果概念，提出do算子即“干预”，进而发展出因果推断的一整套理论和方法。随着因果关系假设和因果模型的不断发展，部分学者尝试利用两类方法的特性设计混合型因果关系发现方法，一定程度上实现了高维扩展性和较强因果发现能力的优势结合，成为实现高维数据因果关系发现的有效方法（蔡瑞初 等，2018）。

总体来看，因果关系的“形式化理论”不仅解决了困扰统计学家多年的一些悖论，更重要的是利用“干预”让人类和机器摆脱了被动观察，从而转向主动地去探索因果关系，以做出更好的决策；“反事实框架”则扩展了想象的空间，从而摆脱现实世界的束缚，更为有效地探求因果关系。这两点突破带来因果革命，分别构成了因果关系之梯的第二层级和第三层级内容，沿着珀尔所构造的因果关系之梯，机器有望拥有更强的人工智能。

国家社科基金重大招标项目“大数据背景下我国新经济新动能统计监测与评价研究”(18ZDA125) 资助。

作者单位：浙江工商大学统计与数学学院

参考文献

[1] 蔡瑞初等．大数据中的因果关系发现，科学出版社，2018.

[2] 大卫·休谟．人性论（关文运译），商务印书馆，1980.

[3] 何大安．大数据思维改变人类认知的经济学分析，社会科学战线，2018(1).

[4] 黄欣荣．大数据主义者如何看待理论、因果与规律．理论探索，2016(6).

[5] 李金昌．大数据与统计新思维，统计研究，2014(1).

[6] 乔舒亚·安格里斯特、约恩－斯蒂芬·皮施克．精通计量：从原因到结果的探寻之旅，格致出版社，2019.

[7] 史蒂夫·洛尔．大数据主义，中信出版社，2015.

[8] 王天恩．大数据相关关系及其深层因果关系意蕴，社会科学，2017年第10期．

[9] 维克托·迈尔·舍恩伯格．大数据时代，浙江人民出版社，2013年版；

[10] 朱迪亚·珀尔，达纳·麦肯齐，为什么：关于因果关系的新科学，中信出版社，2019.