大数据背景下统计数据质量控制方法研究

■ 徐玉武
（安徽省滁州市全椒县统计局，安徽 滁州，239500）

统计工作是一项非常重要的基础性工作，是整个国民经济运行的主要监测手段。统计信息的及时性和准确性至关重要。利用大数据开展统计工作有利于缩短调查时间，提升统计效率，加强数据质量控制。

一、大数据背景下统计数据质量内涵

在数字经济背景下，“数据”已经成为新型生产要素，并在社会经济的发展进程中发挥着至关重要的作用。在这一过程中，统计数据的质量关系到数据的价值，并对数字社会产生深远的影响。实际上，数据质量是指在特定业务环境下，数据要符合使用者的目的，满足不同业务场景下的需求。这是数据自身拥有的属性，具有真实性与完整性的特点。根据历史时期的不同，各个业务领域与用户对数据质量的要求不同，而数据质量之间也有一定的差异。随着数字化发展进程不断加速，与数据统计有关的场景逐渐扩大，统计数据的使用方法、使用范围以及指标不断改变，不同地区对于统计数据质量提出了不一样的要求，这是一个综合性的概念，可从多角度反映各种特征因素，比如及时性与可比性。

统计数据质量将会对政府的公信力产生影响。自党的十八大以来，各级党委和部门都对统计数据质量愈加重视，要求各部门在掌握统计数据质量的数字质量属性之外，还要根据用户要求提出更多大数据时代下的新要求。由于统计数据质量管理和技术发展、外界环境之间存在着紧密的联系，且统计数据质量具有鲜明的时代特征，因此，加强对统计数据质量的研究，以及采用相应的质量控制策略，对提高统计质量与统计效率有着重要意义，能够为大数据技术的应用创造有利条件。

二、大数据在统计工作中的应用现状

（一）应用情况

大数据的发展经历了一个漫长的过程，随着大数据应用的深入发展，相应的技术已经被用于社会各个领域。随着对“大数据量”统计研究的逐渐深入，如今，人们已经开始将大数据应用于各个业务领域内的统计环节。2013 年，国家统计局提出，有必要确立大数据意识，加强对大数据技术的应用，将大数据用于政府统计工作，并联合阿里巴巴和百度等企业签订战略合作协议。与此同时，大数据在政府统计方面产生了实质性的应用成果，具体体现如下。第一，政府统计大数据平台建设。随着统计大数据中心的建立，相应的网络体系逐渐完善，数据管理平台也逐渐为政府统计工作提供技术支撑。第二，政府政务工作方面。政府相关部门应以大数据为前提高效统计经济运行指标，对经济运行情况进行预测，使宏观经济运行监测水平得到显著提升。与此同时，大数据用于公安人口信息统计、人社数据库，将使人口统计信息更加精准。第三，在价格统计工作中，相关部门应充分利用互联网和电子商务交易数据，联合商品结算信息，使价格指数的编制更加快捷准确，从而降低数据统计成本，使价格指标发布频率稳定提高。

（二）存在问题

在统计工作中，应用大数据有利于实现政府政务信息资源的共享，保证统计数据的质量，能够在降低统计成本的同时提高统计效率，突出统计工作的时效性，扩大统计范围。但目前，大数据在统计工作中的应用仍然存在一定的问题，具体体现如下。第一，源头数据质量的控制难度较大。在大数据统计过程中，统计数据的来源之一就是用户的交易信息数据，而这类数据也未必符合实际，因此源头数据的质量无法被有效控制，且因缺少校验机制，数据准确性难以得到保证。第二，数据缺乏统一标准，且获取比较困难。大数据背景下，为保证数据的全面性，各部门纷纷搭建了数据中心。在数据的采集与整合工作中，不同来源的数据缺乏统一的格式与标准，且数据技术与系统接口标准难以统一，这不利于数据采集与交换，也会对数据共享造成影响，导致统计数据的获取难度增加。第三，数据安全性缺乏保障。数据资源库可以为各个领域和部门提供服务，但会受到基础设施的限制，且当资源存储方式和数据库出现问题时，将有可能造成数据丢失。在互联网大数据环境下，信息丢失或者数据损坏会对数据统计工作造成负面影响。除此之

外，统计大数据的利用会对敏感数据、隐私数据的使用造成安全隐患。

三、大数据背景下统计数据质量的影响因素

（一）统计大数据采集

一般情况下，数据源和数据采集方式主要有人工导入、系统数据交换、网络爬虫抓取与传感器采集几种。其中，人工统计报表导入方式的效率比较低，报表导入期间可能会由于表格规范性和导入程序接口之间存在着兼容性问题，导致数据在导入期间出现错误，例如报表内的“0 值”与“空值”。采用网络爬虫技术抓取数据则能够完成对数据的高效采集。技术人员首先需要找出爬取数据的 url 地址，再向这个地址发起请求，随后获取 url 服务器发来的响应数据（网页源代码），并利用 python 数据解析库在源码中获得想要的数据，而后根据获得的数据对其清洗保存，保存至数据库、Excel 等地。传感器采集方式主要用于数据质量检验与动态数据统计跟踪，由于传感器自身可以准确地采集数据，因此数据安全能够得到有效保障。这些都是目前数据采集环节的影响因素。总的来讲，数据采集的时效性与数据质量会因为采集方法与采集工具的特点而受到影响，且数据来源十分广泛，数据之间存在矛盾，而这在一定程度上会影响数据统计效率。

（二）统计大数据的预处理

由于数据存在多方面来源，因此不管使用哪种方式，所采集到的数据都不能直接用于统计分析，而要使用相应的预处理方式来保证数据质量。技术人员应采用数据清洗，在一定技术应用下，及时找出数据中的重复或者遗漏信息，实现对数据的规范化处理。面对数据格式不同或字段数据匹配不相符的问题，技术人员需要通过数据转换来管控数据质量。数据清洗与数据转换都是常用的数据预处理技术，将直接对数据质量产生影响。

（三）统计数据存储

目前，大数据中最常见的存储方式是分布式存储，这与存储介质与数据管理方式有着直接的关联。针对大数据的不同特征，使用的存储技术也会不同，技术人员应按照存储介质使用相应的存储方式。基于统计内部网络，建立网络存储系统，能够为统计系统内部提供高性能存储空间，实现统计核心数据流转、存储和共享全流程的畅通。

（四）统计数据处理

分布式处理技术与数据类型有关，也与数据的存储形式存在一定关联。以 Java 技术为前提的 Hadoop 体系架构存在着批处理能力，可用来实现对数据的批量处理。但是该技术的时效性不太理想，无法对大规模数据进行集中且快速的处理。技术人员应以拓扑结构为基础，应用 Storm 技术完成数据流的转换。该技术适合用在数据集群中，凭借技术的时效性与强大的容错能力，实现对数据的统计处理。基于用户内存情况，为保证数据处理的灵活性，技术人员需将数据流转化为超低量秒级数据集，以此实现对数据的分类和聚类，加强数据关联分析，完成数据深度学习，进一步彰显大数据的应用价值。

（五）数据展示与统计大数据应用

大数据具有可视化的特点，这是对前期数据处理与分析结果的体现，可直观地向用户展示结果，完成数据交互处理。经过预处理与处理分析之后，所得的数据应在相应模型下进行统计分析。从某种程度来讲，大数据应用也是数据价值的重要体现，能够反映统计大数据从采集到预处理，再到成果输出的准确性。

四、大数据背景下统计数据质量控制方法

（一）强化对大数据应用的认知与制度保障

要想提高大数据的质量，就应分别从管理与技术两方面入手，掌握影响统计数据质量的因素，并通过各类影响因素的有效控制保证统计大数据质量。各级统计机构和数据管理部门应积极迎合大数据发展的新形势，提高对大数据应用的敏感度，积极探索相关统计指标口径的有效衔接，实现由“信息资源”向“信息资产”的完美转变，并掌握小样本数据与大数据统计之间的一致性，兼顾小样本数据在统计中的精准性要求，以及大数据统计中的高效性要求，努力营造出更加宽容的数据统计环境。

相关部门应健全与大数据统计有关的制度，并结合相应工作场景，制订相关管理办法。此外，各部门应做好统计业务的分类，科学设置相应指标，不断优化工作流程，确保统计人员更好地理解各项指标内涵，采用“事前”“事中”“事后”三阶段数据质量检验方法，不断提高源头统计质量，确立和大数据应用相互协调的数据采集体系、经费保障体系、技术支持体系，规范数据统计工作当中大数据的应用流程，优化技术应用路线。

（二）搭建统计大数据信息资源平台

相关部门应根据大数据时代发展特征，立足于“顶层设计”的发展理念，挖掘关于大数据的应用需求，做好统筹规划，建立大数据资源平台。平台可

采取“1 中心 +1 节点”的架构模式，确立“公有云与私有云”“互联网、专网联合政务网”的跨行业跨部门数据。平台不仅要包含各种专项调查数据与普查数据，还应包含第三方商业等数据。相关部门应采集与整合所有数据资源，建立统一的大数据信息资源平台，实现对数据的网络管理与安全管理，统一管理数据资源，完善基础设施，建立数据资源库，以此完成统计大数据在信息交换与共享等方面的应用。

平台包含数据资源整合、数据管理、安全运维、系统与应用四部分内容。系统监控体系主要有系统运行监控、数据入库监控、数据库运行监控以及数据资源监控等内容。数字资源服务涵盖了数据服务API、智能检索、数据共享等功能，同时，外部可与统计人员管理系统、统计数据生产管理系统、统计业务绩效考核系统有效对接。相关部门应遵循易用性、扩展性与安全性的平台搭建原则，采用区块链技术管理数据操作日志，及时上传并修改调查单位的原始凭证信息，将该部分信息转为区块链信息，并采用 HASH 算法生成索引信息，再将索引信息存放于区块链链头，确保信息最终能够成功溯源。

平台以大数据统计工作为支撑，对数据采集效率的提升有着重要意义。统一数据采集和数据共享平台，对数据采集范围予以统一规范，防止不完整或质量较差的数据被纳入统计数据源，将数据质量控制落实在数据管理的全过程，实现对多源异构数据的合理转换与快速清理，完成数据采集端的质量把控。综合应用云计算和区块链技术，能够提高数据存储效率，确保数据访问过程的安全。有关部门可依靠平台完成综合统计业务管理，以统计数据的生产流程为主线，依靠联网直报系统建立统计人员管理系统和数据生产管理系统，以满足各级统计机构的工作需求，使数据生产达到透明化与直观化的目的，进而完善对统计业务的考核评价，全方位保障统计数据质量的稳步提升。

（三）加强统计大数据深度挖掘分析

由于大数据具有 4V 特点，即规模性、高速性、多样性和价值性的特点，因此，在反映大数据价值的同时，也能为大数据的开发利用提出更加明确的要求。面对海量数据信息，相关人员需利用大数据的应用优势，凭借数据挖掘等技术加强数据整合，再使用 SAS 多元化数据统计工具，从大量数据中寻找有价值的内容，并基于用户多样化需求完成传统数据统计方法的补充和改进，为大数据时代下的统计工作的顺利开展创造有利条件。

一般情况下，数据挖掘主要包含数据清洗、转换以及挖掘等内容。在大数据统计工作中，技术人员应掌握统计业务模型的影响因素，选择合适的处理技术，科学设计统计模型，完成数据挖掘分析，再按照业务需求对不同维度和发布频次的数据输出形式予以明确。

（四）完善信息化体制，创新调查手段

根据大数据资源平台的建设情况，相关人员需了解大数据技术的应用要求，加强统计工作信息化建设，完善相应管理机制。随着数字化技术的逐步应用，政务信息化建设逐渐呈现出云端集中的发展趋势。5G 与人工智能技术的应用使数据采集与处理逐渐朝着智能化与移动化方向发展。因此，有必要建立相应的数据资源平台，完善相配套的运维管理机制。除此之外，随着互联网 + 业态的发展，各类平台都将成为统计大数据应用的“生态圈”。通过对大数据开发应用和共享，为统计数据的质量控制营造良好的生态空间。

为进一步提高统计数据质量，相关部门应及时创新数据调查手段。对此，以下建议可供参考。第一，确立统一的统计数据调查制度，从调查手段入手，完善相应管理制度，确立数据衔接与经济核算制度，优化统计数据抽样调查制度，保证各个地区的统计数据的抽样方式一致，加强统计数据整理，尽可能地满足用户数据需求；第二，搭建统一处理平台，完善统计数据综合处理平台，迎合大数据时代的需要，实现各个专业与领域内数据的高效处理，提高数据处理能力，完善数据评估机制，从而保证数据质量。

五、结语

总而言之，在大数据环境下，统计工作一般是对数据信息进行采集、汇总、挖掘分析，以此掌握事物发展规律，这将对经济与社会的发展会产生深远的影响。随着数字化发展进程的加快，大数据像一把双刃剑影响着当前的统计工作，统计效率的提升得益于大数据的发展，同时也面对大数据质量的挑战。因此，如何提升统计数据质量至关重要，各级部门和地区要相互配合，建立完善的统计数据综合平台，以此实现数据质量的管控。

【作者简介】徐玉武（1986 —），男，安徽淮南人，本科，中级统计师，研究方向为财务报表分析、产业集群统计分析