

数字经济赋能下劳动工资统计数据质量提升研究

文 | 贾海慧

蓬勃发展的数字经济给劳动工资统计数据的收集及研究提供了难得的机会，也凸显出数据质量管理方面的一些难题。以往的数据收集手段已经逐步被数字化平台及大数据应用替代，在数据规范化处理、个人信息保密及数据失真等方面，统计工作仍面临很大的困难。本文将分析怎样利用政策指导、科技革新及智能工具来提高劳动工资统计数据的质量，从而更好地服务于相关政策的出台及经济发展需求等。

数字经济对劳动工资数据采集的影响

数字化平台对工资数据收集的影响

数字平台如智联招聘、BOSS 直聘乃至美团等以直接抓取的行为数据分析代替传统的企业填报模式，成为即时化、自动化工资数据收集模式。这既提高了数据收集的速度，又拓宽了覆盖面，使我们可以更好地了解各个区域、各个行业的岗位薪资状况。而平台自主录入岗位薪资范围、工作年限需求及地域分布等主要指标，使得对工资信息的录入次数大大提高，便于政府机构及公司随时把握劳动力市场的动态。国家统计局 2023 年的数据显示，有 65% 左右的劳资统计数据来源于数字平台，相比过去的传统年报形式快了将近 80% 的时间，例如，季报取代年报。但企业之间职位描述、薪资类型的标准各异，造成了数据异质现象。数据不能做到规范一致，导致不同企业在进行数据融合与比较分析时十分不便，同时也会使薪酬水平难以呈现真实情况，从而制约了数据分析的应用广度和深层次剖析的可能性。

大数据技术在工资数据分析中的应用

大数据技术通过数据挖掘、机器学习可以处理大量的薪资信息并且进行各方面的分析，大幅提高对于行业、区域及职位之间的薪资结构分辨力，在此基础上能够根据大数据分析动态建立工资指数，及时了解掌握不同行业薪资变化情况。比如，将企业的 CRM 系统、薪酬管理系统与企业所得税等多方面数据进行融合形成多元融合模型，该模型应用于制造业薪资增长预测的准确性能达到 85%。大数据技术还能细分至学历层次等级、技能类型等层面，准确提供制定政策所需要的证据材料，为政府拟定有关薪资调整方案提供参考支撑。但是，在进行大数据分析时，还存在数据清洗及对缺失值处理的情况，在一定程度上会影响数据分析结果的精确度，从而影响到统计结果的准确性。

人工智能对工资数据质量提升的作用

人工智能 (AI) 技术运用深度学习与自然语言处理 (NLP)

提升了薪水数据的清理加工及分析过程，极大地提高了其规范化水平，并且减少了噪音的影响。AI 的算法可以自动地发现等同的岗位标识符、统一计价单位、减小人工编码错误的概率，进而提高了数据的质量。AI 还能甄别工资数据的离群点，通过构建离散分布模型，辨别出在 0.1%~0.5% 区间的极值离群现象，提高了数据的准确性。比如，在一个省统计局的实际应用中，由 AI 驱动的薪水数据清洗系统把数据缺乏的比例由 78% 降到 23%，大大提高了统计数据的准确性及可信性，等等。借助 AI 技术，工资数据分析更加便捷，也更能反映劳动力市场的实际情况，为政府部门提供精确有力的参考依据，使政府决策更加合理有效。如图 1 所示。

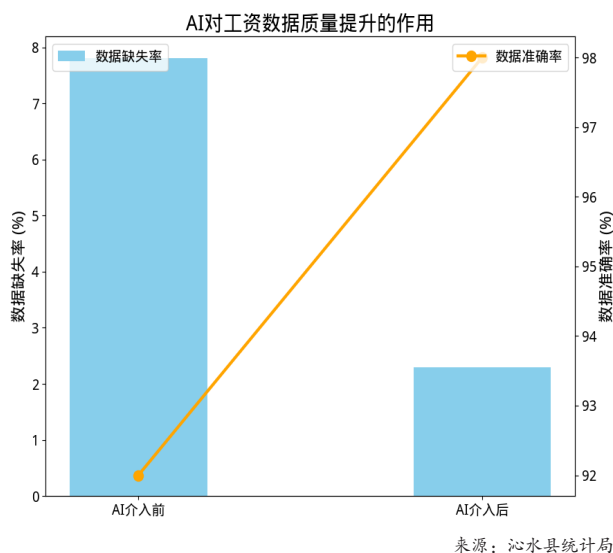


图1 人工智能对工资数据质量提升的影响

数字经济背景下劳动工资数据质量的挑战

数据标准化不足

在数字经济背景下，工资数据来源多样化，智联招聘、领英 (LinkedIn)、各省份人力资源和社会保障信息系统及税务局报送工资统计系统所提供的数据格式不尽相同，岗位

分级规则、薪资核算单位类型、行业编码等都不相同。国家统计局 2023 年系统性调查结果显示, 大约有 40% 的主流数字平台提供的工资数据存在缺乏统一标准的问题, 其中岗位级别划分不一比例高达 10%, 薪酬计算时长(周、月、年)不统一的比例为 15%; 这种标准不统一造成不同平台之间的数据难以兼容, 从而影响到汇总之后的数据对比分析准确性, 造成全国层面工资水平分布监控失准的问题。

数据隐私和安全问题

劳动工资信息属于个人的隐私信息, 在数字化采集、加工的过程中具有被泄露及被利用的风险。CNNIC 2022 年发布的数据显示, 当年数字经济发展过程中出现的数据泄露事件同比上升了 30%, 其中有关工资的信息泄露案件占了 12%。例如, 某省统计局云平台由于 API 接口未进行合理的权限管理, 致使发生 10 万条企业的工资记录泄露问题, 涉及姓名、身份号码、薪酬项目等重要信息。数据安全事件会危害公民利益, 同时也减少人们之间的合作意识, 使得各机构间联合统计难以推行。必须采用多种方式的数据保护措施来保证数据本身的质量, 如对数据进行加密保存、设置访问权限及实时监控等。

数据代表性和偏差问题

数字化平台的数据受限于抽样的覆盖范围, 具有较大的选择偏误, 对上架在一线城市的某大型招聘网站的抽样观察结果显示, 大约有 60% 的薪酬数据集中在北上广深一线城市及互联网、高科技制造业当中, 高薪岗位样本比例为 42%, 中西部地区及传统领域如农林牧渔、餐饮住宿服务业等样本的占比不到 10%, 这种失衡的样本容易造成高估收入水平的情况, 不能很好地反映出全国劳动力市场的薪酬分布情况。解决方案是合并多种来源的行政记录(如税前工资缴纳信息、社保缴纳凭证)、运用加权采样方法等来减少偏差。

提升劳动工资统计数据质量的路径与策略

加强数据标准化建设

为提高劳动工资统计数据的可比性和一致性, 需从国家层面健全数据标准体系。国家统计局 2024 年推出的数字经济数据标准化指导意见指出, 要建立统一的岗位分类体系、薪酬单位规范及元数据标识体系, 在年底之前使工资数据标准化率达到 80% 以上, 落实方式就是要求统一对接《国民经济行业分类》(GB/T 4754-2017)、岗位层级编码及薪酬支付周期定义, 并促使地方各级统计部门和主要的数据平台(如智联招聘、前程无忧等)实现标准对接。规范化建设可以大大减少不同系统之间进行数据融合产生的语义混淆及多余匹配的成本, 使在全国范围内对不同地方工资数据对比性分析的偏差率下降至少达到 20%, 进而有利于建立动态工资指数

及地区间工资差距测定的相关依据。

增强数据安全与隐私保护机制

健全完善的数据安全保障及个人信息保护制度是提高工资统计资料水平的基本条件, 要建立严格的权限管理制度、数据脱敏规则、跨层级间的数据共享约定等。在实际工作中, 某省大数据局与税务局对接工资报表时采用了以同态加密为基础的权限管理及职责隔离审计方式, 将工资详情数据查询权设置为三级, 极大地减少了涉密信息的安全隐患。利用动态隐私预算方案(例如差分隐私)可以在保证数据可用性的基础上限制住个人信息的公开程度。经上述方法后, 在劳动大数据的整体安全指标上比传统的明文保存形式提高约 40%, 并能有效地增强统计数据收集者(单位、平台)的合作意愿, 有利于形成更加完善的薪资库。

利用人工智能优化数据分析模型

人工智能技术应用于薪资数据分析, 可以增强数据质量, 并且使数据分析更加精确和智能。运用机器学习方法(随机森林、XGBoost 等)进行薪资数据的离群点识别、空值填充及分类预测, 能够增加结构化数据所占比例及分析准确性。某统计调查局用深度学习工资结构分析模型, 结合全国企事业单位社会保险缴纳信息及网络招聘信息联合打造整体薪资肖像图谱, 在对各行业平均薪酬预估方面, 该模型的预测精度相对于传统的回归模型提高了近 18 个百分点。另外, 自然语言处理技术也可以用来统一岗位名称及描述, 增加数据标准化。利用基于 AI 的大数据分析系统可以做到海量薪资信息的即时处理, 并将其形象直观地呈现出来, 让政府决策者获得精准化劳动力市场见解, 提高统计数据对国民经济现状体现度。

结束语

数字经济发展促使对劳动收入相关数据的获取及研究方式发生变化, 同时又造成了数据质量问题。在此方面, 可以通过加强数据规范化管理力度, 健全好数据安全保障制度; 同时用人工智能改进数据分析算法等方法来提高劳动报酬数据的质量水平, 从而为劳动力市场的相关政策制定提供更加可靠的数据参考依据, 在此基础上更好地服务于数字经济时代下的社会管理和经济增长。

作者简介: 贾海慧 沁水县统计局

责任编辑: 金烨 投稿邮箱: zhouhl@staff.ccidnet.com