基于区间统计分析的迭代式价格信息采集模型评价信息采集点质量

田文炳　张梦川　王　竞

郑州市烟草公司卷烟配送中心 河南 郑州 450000

【摘　要】：本文搭建起一种以区间统计分析为基础的迭代式价格信息收集模型，将卷烟价格相关数据作为研究目标且融合大数定律、供求规律等理论，借助多轮“问卷发放-数据修改”操作实现对价格数据的迭代优化。同时从真实性判别、可靠度计算、偏离度标识三个维度拓展出信息采集点质量评估体系，结合 754 户商户实际数据验证可知，该模型能精确修正价格，评估体系可切实有效筛选出高质量采集点，进而为烟草市场监督管理给予数据支持。

【关键词】：区间统计；迭代模型；价格采集；信息采集点；质量评价

DOI:10.12417/2705-1358.25.18.047

1 引言

随着信息技术不断进步，卷烟价格数据在决策及监管工作中发挥着愈发突出的作用。然而国家局直接采集、营销子系统等多种渠道的数据存在采集时间不一致、标准有差异等情况，致使数据准确性与一致性不达标。本文选取国家局选定点位的商户作样本，通过“发放问卷-回收数据-处理修改”的反复循环构建起基于区间统计分析的价格采集模型并扩展形成信息采集点质量评估体系，旨在为烟草行业提供精确的数据支撑以巩固数字化服务根基。

2 理论基础和技术路线

（1）理论基础。大数定律揭示大量随机事件背后蕴含的必然性，如抛硬币时正面朝上频率会逐渐趋近于 50%，这为价格采集提供依据，即当样本量足够大时认知误差可相互抵消从而形成较稳定的市场价格预期。商品供求规律反映供求与价格间的相互影响作用，卷烟在不同地区的价格差异及非法流通现象很好地证实了该规律，商品价格由生产成本、流通费用、税金及利润共同组成，零售客户对价格的判断实际是对流通费用和利润的一种权衡，在非合作博弈理论中零售客户会调整销售价格以追求利润最优状态，最终形成市场均衡价格，为理解价格形成过程提供思路。

（2）技术路线。基于区间统计分析及迭代理念将技术路线设计成多轮交互修改流程，以省级采集的价格作为最初判断价格并结合上一轮修正后的平均价格生成新一轮有待判断的价格，接着以问卷形式向选定商户发放。商户需针对价格给出“过低、较低、合理、较高、过高”评价并分别对应触发“大幅上调、小幅上调、保持原值、小幅下调、大幅下调”修正动作，系统对回收的修正数据计算平均值后判断迭代轮次是否达到预先设定标准，若未达到则重复上述流程直至完成规定轮次，最终输出校正后的价格数据，此循环过程通过低代码平台实现自动化处理涵盖问卷发放、数据回收、修正计算等各环节，有效提高价格数据精确程度及采集效率。

3 假设论证

（1）基于选点原则提出的假设。选点的科学性对样本在市场中的代表性起决定性作用，国家局作为烟草行业权威性管理机构。其选点标准全面考量诸如商户所处位置、经营规模大小、市场覆盖程度等宏观与微观方面因素，经多个层级筛选工作确保样本分布具均衡特性，实际操作过程里选取的商户既囊括城市核心商业区域又涵盖乡镇销售网点，以全方位展现不同消费场景下价格特点，且统一选点标准使各样本在数据收集口径及统计维度上保持一致。

（2）基于产品流通提出的假设。卷烟作为特殊商品其价格仍受市场供求规律约束，对于未订满商品规格商户可通过调整订购数量灵活应对市场需求变化。市场需求呈上升态势时增加订购数量以抑制价格波动，已订满商品规格因有限供应数量促使价格上涨，价格升高又激励商户释放淡季库存进而形成供需自我调节机制，季节方面因素也影响价格走势。

（3）基于价格感知提出的假设。零售客户对价格的认知是商品流动性与价值性的综合反映，从价格构成角度看生产成本及税金相对稳定，流通费用体现商品流动性。利润展现商品价值性，客户判断价格实际是对这两者权衡，同时专卖体制下

作者简介：田文炳（1997-），男，汉族，河南省永城市，无职称，本科，郑州市烟草专卖局航空港区直属分局。

卷烟具低风险特点致使高流动性与高收益性无法同时实现，客户需在资金周转效率与盈利空间之间抉择。

（4）基于模型构建提出的假设。采用分段式修正办法有可能致使局部数据出现一定程度偏差，不过利用均值修正方式依旧能体现整体发展趋向。从模拟实验情况看，不同区间修正数值虽有上下起伏波动情况，但均值能有效使极端数值带来的影响变得平滑，且随着迭代次数不断增多偏差会逐渐减小。与此同时只要存在一定比例真实反馈信息价格就会朝期望值方向汇聚，如实填写数据可推动价格往实际数值靠近，随机填写所产生误差仅会使修正过程中的震荡有所增加而不会改变价格汇聚方向。

4 模型构建与验证

4.1 模型准备

在进行模型构建之前需完成环境模拟、参数定义、逻辑流程设计以及参数确认等一系列基础性工作。在环境模拟环节把客户划分成价格敏感型、迟钝型，以及消极合作型（以贴合实际市场中不同商户对于价格的感知特性，价格模拟以“上帝价格”为基准通过多轮对价格的判断来观察修正后产生的效果。参数定义包含核心变量如上帝价格（GP）、判断价格（JP）、修正次数（i）等并明确各指标计算规则及边界条件，逻辑流程遵照“参数输入-客户模拟值生成-多轮修正-精度判断-结果输出”这样的闭环式设计，以保证每一步运算过程具备可追溯性。参数确认挑选 110、300、850 作为具有典型代表意义的价格，并设定±40、±20、1 等初始价差，再结合 300 个样本数量以及 10000 次模拟次数为模型验证奠定可靠的数据基础。

4.2 模型验证与优化

通过一次次版本更新来验证模型，使其精准度一步步得以提高。在 V1.0 阶段运用固定范围分段且依据比例精度判断，在价格较高区间修正效果能达标准，可在中低价位区间因对误差限定过严，致使许多修正未成功（LUj=16），例如 110 元价位在精度要求为 0.005 时经 16 轮未达收敛状态，此情况不符合实际业务所需标准。到 V2.0 时期引入“固定比例+固定数值”精度判断方法，将误差范围扩大到 99%+1，在中低价位区间可靠程度明显提升，110 元价位在精度达 99%时 4 步以内修正完成比例达 97.09%，但判断节点与修正数值固定不变问题仍未解决。发展至 V3.0 采用根据区间宽度动态调整比例方式，设定中间区间为±1%并依次往两边拓宽 2%，让修正幅度与价格基础数值相互匹配，最终达成中高价位区间 4 步以内收敛、低价位区间最大修正次数降至 9 次成果，彻彻底底解决极端情形下精度不够难题，为模型实际应用打下稳固基础。

5 模型应用与效果分析

5.1 问卷前期准备

对于问卷前期准备工作需兼顾有效性与实际操作情况，从理论层面要通过科学合理设计保证所获数据真实可信，可利用诸如重复出现问题、异常问题等检验手段筛选排查无效问卷。同时简化问卷内容以降低填写难度，防止因设计过复杂引起商户抵触心理，在问卷投放时依据选点原则挑选具代表性商户，对其展开培训使标准统一，且把控回收周期减少时间因素干扰，这些前期准备工作是确保后续数据质量根本，能为模型应用提供符合标准输入信息。以黄金叶价格采集问卷设计为例，问卷包含重复出现问题与异常问题。第 8 题和第 15 题皆问“黄金叶价格相较于 132 元是高还是低”，选项设为 A（低于 128.04）、B（128.04-130.68）等五个档次。同时设置第 23 题“黄金叶（硬帝豪）价格相较于 666 元是高还是低”，此次问卷投放对象是 754 户国家局选定商户，投放前通过在线培训让填写规则清晰明了，着重强调“我不卖不了解”选项要慎重选择。问卷借助问卷星平台发放且每天回收以避免价格波动产生影响，在 6 天周期里第一天发放问卷后发现 32%商户对重复出现问题回答存在矛盾，经再次培训后矛盾率降至 8%，异常问题错误率也从 15%下降到 3%，有力保证了数据可靠性，为之后模型运算提供高质量样本。

5.2 数据处理与问卷有效性检验

数据处理以及有效性查验乃挑选出可靠数据之重中之重，从道理上讲，需先清理掉如重复递交、许可证号有异常情况、选大量“不了解”选项等无效信息的问卷。再通过对相关问题查验以证实数据真实性，确保进入模型的数据达精度要求，此过程可去除随机误差，为模型校正提供信赖之输入信息，是问卷收集与模型运用间关键环节。就针对 754 户商户展开的六轮问卷而言，第一轮回收 745 份问卷，经清理后留 532 份有效问卷，具体清理情况为剔除 12 份重复递交、47 份许可证号存在异常及 154 份选“不了解”超 20 个产品规格的问卷。进行有效性检验时，剔除 89 份复现问题及 36 份反常问题，最终六轮问卷有效率均超 50%，其中第三轮回收 754 份问卷，有效 613 份，复现问题一致率达 92%，反常问题正确率达 97%，为模型 V3.0 校正提供符合标准数据，保证后续校正结果具科学性。

5.3 数据修正与最终结论

数据校正及结论验证于模型运用而言乃核心部分，理论上模型依商户评价对价格作动态调整，评价若为“过低”或“较低”则向上校正价格，评价若为“过高”或“较高”则向下校正价格，经数轮反复操作后借最终验证确定结果是否稳定，此过程可消除初始数据偏差、使价格趋近真实市场水平且为结论提供实际证明支持，直接体现模型有效性。以钻石（荷花）产

品为例，最初判断价格 350 元，第一轮回收 532 份有效问卷，其中 28%的人认为价格“过高”，35%的人认为“较高”，经均值计算后价格修正为 340 元，第二轮以 340 元为判断价格。41%的人认为“较高”，29%的人认为“合理”，价格又修正为 337 元，第三轮到第五轮价格依次修正为 335 元、334 元、335 元，第六轮验证以 335 元为判断价格，62%的商户觉得“合理”，21%的商户认为“较低”，17%的商户认为“较高”，修正后价格为 335.43 元，与第五轮结果偏差仅 0.13%。在此过程中每轮修正皆基于前一轮均值，如从 350 元到 335 元的五次反复调整般逐渐趋近稳定数值，第六轮验证进一步确认结果有效，证明模型可准确校正价格为市场监管提供可靠数据。

6 模型延伸：信息采集点质量评价

6.1 真实性判断

判断信息的真实性这一筛选有效信息采集点必不可少的基础工作，需通过一些特定方式识别那些无效的数据。如一方面设计专门问题，像设置特定价格判断问题且设定时长阈值，对于复现问题通过对同一品规价格判断重复提问以验证回答一致性。举例来说在问卷里两次提出“黄金叶相较于 132 元的价格判断”这样的问题，给出如“低于 128.04 元”“128.04-130.68 元”等五个档次选项，若两次回答不一样则判定该数据无效，对于反常问题依靠常识性答案筛查随意填写情况。例如“黄金叶相较于 666 元的价格判断”这一问题，因该品规实际售价远不到 666 元，所以只有“低于 646.02 元”才是正确有效答案，选其他选项则标记为不真实，指示性问题通过带有指令性选择检验填写问卷认真程度。比如“本题请选择 A”或者“黄金叶不管价格高低都要选择 B”，未按指令回答则当作无效处理，时长判断方面设定合理时间界限，比如 30 道题目的问卷以 90 秒为界限，答卷用时短说明答题者无法完成真实思考，这样的答卷会被剔除。

6.2 可靠度计算

计算信息采集点数据的可靠度，是通过将扫码价格和问卷选项进行逻辑匹配实现把数据一致性用具体数值量化出来这一目的。就拿黄金叶来讲，若某商户扫码价格是 130 元，对应问卷选项是 B（128.04-130.68 元），说明逻辑一致记为 T，若选择的是 E（高于 135.96 元），则表示存在矛盾记为 F，对所有品规都按此方式比对后按“可靠度=$\Sigma T/(\Sigma T+\Sigma F)$”公式计算，结果在 0%-100%范围之间。实际操作中把可靠度达到 60%作为合格标准，低于此数值说明采集点数据矛盾率高会被判定为质量不好，此计算方式考虑到价格可能短期波动情况，采用对多个品规累计评分，比如某商户 30 个品规里有 25 个逻辑一致则可靠度是 83.3%判定为合格，若只有 15 个逻辑一致则说明稳定性不足会被标记出来。

6.3 偏离度标记

借助方差运算得出的偏离度标记，主要用于衡量采集点扫码价格与市场平均价格在总体上的偏差状况，属具有参考性质的指标。以市场价格 JP 为参照标准，某个品规扫码价 Pi 的偏离数值按$(Pi-JP)^2 \div (JP^2)$计算，若涉及多个品规则取平均值，计算公式为$DE=\Sigma(Pi-JP)^2 \div (n\times JP^2)$。举例而言，市场价格 100 元时，扫码价 101 元（偏差 1%）对应 DE 值是 1，扫码价 120 元（偏差 20%）对应 DE 值则是 400，表明该数值越大偏离情况越明显。实际运用中，DE 值主要用于标注，不会直接判定质量，因价格偏差可能出于店铺自身定位或区域市场行情等合理理由。例如 A 商户有 30 个品规，DE 值是 25，意味着整体偏离程度处于中等水平。B 商户 DE 值为 400，说明其价格出现系统性偏高情况，需结合具体经营场景分析，不能直接否定其数据价值。此指标能为了解采集点价格特点提供额外观察角度，与真实性、可靠度共同构成多维度评价系统，兼顾数据质量与市场多样性。

7 结语

针对卷烟价格数据采集过程中存在的复杂性及质量参差不齐等问题，构建的基于区间统计分析的迭代式价格信息采集模型，通过多轮迭代修正机制有效提高了数据精准程度，克服了传统采集方式的局限性。在此基础上延伸形成的信息采集点质量评价体系，从真实性、可靠度、偏离度三个维度实现了对采集点的科学评估。经实际验证，该模型及评价体系与烟草行业实际情况适配，为价格监管工作、政策制定提供可靠支撑，也为烟草行业在数字化转型进程中管控数据质量提供可行路径，未来还有望进一步推广到更多商品领域的价格采集及评价场景中。

参考文献：

[1] 李蕊.浅谈几个著名的大数定律及应用[J].科学咨询(科技管理),2010(12):64-65.

[2] 滕泰,羿伟强,赵虹,等.全球大宗商品供求价格弹性分析[J].世界经济研究,2006,(06):59-64.

[3] 许宏锋.高等教育管理审计[M].河海大学出版社,2007.11.