

基于模糊层次分析法和熵权法的金融数据仓库数据质量评估

——以H公司为例

王 浩, 孟 飞

上海理工大学管理学院, 上海

收稿日期: 2025年2月20日; 录用日期: 2025年3月12日; 发布日期: 2025年3月24日

摘 要

本文以H公司为例, 探讨金融数据仓库的数据质量评估。首先介绍证券行业数据仓库数据内容及特点, 构建包含完整性、准确性等7个一级指标及相关二级指标的评价体系并量化。接着阐述模糊层次分析法和熵权法, 前者通过构建层次模型和模糊判断矩阵确定主观权重, 后者经数据标准化等步骤计算客观权重, 两者结合得出综合权重。通过对H公司主体、交易等四个主题域数据集的算例分析, 包括指标量化、权重计算及质量评估, 结果表明主体和渠道数据在准确性及一致性方面有不足, 研究为金融数据仓库数据质量管理提供了科学方法和改进方向。

关键词

金融数据仓库, 数据质量评估, 模糊层次分析法, 熵权法

Data Quality Evaluation of Financial Data Warehouse Based on Fuzzy Analytic Hierarchy Process and Entropy Weight Method

—Taking H Company as an Example

Hao Wang, Fei Meng

School of Management, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 20th, 2025; accepted: Mar. 12th, 2025; published: Mar. 24th, 2025

Abstract

This research focuses on the data quality evaluation of financial data warehouses, taking H Company as an example. Firstly, it introduced the data content and characteristics of the security industry data warehouse, constructed an evaluation system including 7 first-level indicators such as integrity and accuracy and related second-level indicators, and quantified them. Then, it elaborated on the fuzzy analytic hierarchy process and the entropy weight method. The former determines subjective weights by constructing a hierarchical model and a fuzzy judgment matrix, while the latter calculates objective weights through steps such as data standardization. The two methods are combined to obtain comprehensive weights. Through a case analysis of the data sets of four theme domains such as the main body and transactions of Company H, including index quantification, weight calculation, and quality assessment, the results show that the main body and channel data have deficiencies in terms of accuracy and consistency. This study provides a scientific method and an improvement direction for the data quality management of financial data warehouses.

Keywords

Financial Data Warehouse, Data Quality Assessment, FAHP, Entropy Method

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着我国证券市场的快速发展,金融机构在数据管理和质量控制方面面临诸多挑战。数据仓库作为证券公司决策支持的关键系统,在风险管理、市场分析、客户服务和合规报送等方面发挥着核心作用。然而,多源性、实时性和复杂性的数据特性使得数据质量问题频发,对准确性和可靠性提出更高要求。

数据仓库作为企业数据管理和分析的重要平台,其数据质量管理成为构建和运营的关键因素。国外研究主要集中在理论框架、评价指标设计、数据清洗技术和数据治理模型等方面。学者们提出了多维度的质量评价体系[1]-[3],并开发了如EM算法、DBSCAN等优化数据质量的技术[4],同时探索了基于元数据管理和数据可观测性的治理模式[5] [6],也针对金融这一特定领域的数据合规管理提供了新视角[7]。国内研究侧重于行业实践和本地化应用,涵盖元数据质量控制、多维评估方法和ETL集成技术[8]-[10]。通过数据治理框架和成熟度模型,推动数据质量在银行、监管等领域的落地应用,通过自动化质量监控平台满足监管合规的需求[11]。但当前金融数据质量管理的研究主要集中在银行、保险等领域,对证券行业的数据仓库质量评价的探讨较为有限。

本研究结合模糊层次分析法(FAHP)和熵权法(EWM),构建综合赋权模型,在消除主观偏差的同时引入数据驱动[12],为数据质量评估提供科学依据,并通过H公司的实证分析验证模型效果。本研究为证券行业数据仓库的数据质量管理提供了理论框架和实践指导。

2. 金融数据仓库数据质量评价指标体系

2.1. 证券行业数据仓库数据内容及特点

证券行业数据仓库涵盖多源、实时的丰富数据,全面记录证券市场的核心交易活动及相关信息。这

些数据详细包含了交易的具体时间、价格、成交量以及交易对手等关键信息,还囊括了上市公司的基础信息、财务报表和公告信息等多维度数据。此外,数据仓库还涵盖客户基本信息(如姓名、联系方式、身份证号)及交易记录、持仓状况等。证券行业数据仓库中具有以下特点:

(1) 主题导向性:以特定业务主题如(营销、合规管理)为核心,定义相关的指标,旨在为关键业务问题提供决策支持[13]。

(2) 数据集成性:整合包括交易系统、公告系统、会员管理系统等异构数据源,确保信息的统一性和一致性。

(3) 数据持久性:确保加载后的数据稳定可靠,输入与输出保持一致,为分析提供坚实基础。

(4) 高度动态性:数据随市场实时更新,涵盖交易价格、成交量及投资者行为等复杂信息,展现出显著的即时性特征[14]。

2.2. 数据质量评价指标选取

在构建数据质量评价体系的指标选取过程中,严格遵循“标准映射-行业适配-企业验证”原则。首先参考 ISO/IEC 25012 [15]数据质量模型的完整性、准确性等核心维度,从国际通用标准层面保障理论完备性,随后结合《证券期货业数据分类分级指引》[16]等监管要求,充分考虑证券行业高频交易、跨系统一致性等特有场景,聚焦数据的及时性、准确性以及跨系统一致性等关键点。最后,通过对 H 公司近三年数据质量事件库的回溯测试,对初步拟定的指标进行筛选。剔除数据存储压缩率这类与数据质量核心问题关联度低的冗余指标,确定了 7 个一级指标和 18 个二级指标,切实为评价体系注入了实用性与有效性。具体而言,各项指标特性如下。

完整性着重衡量数据记录与字段信息的完备程度,力求从源头上杜绝数据出现记录缺失或字段空值的情况,确保数据信息无死角覆盖。准确性聚焦于精准评估数据能否如实反映其所代表的实体或概念本质,保障字段值的精确无误以及业务逻辑的协调统一。唯一性将关注点投射于数据的独特属性,全力规避数据冗余与冲突乱象,保障每条记录或字段在既定范围内具备独一无二的特性。一致性致力于展现数据在不同时间节点、多元来源以及各异数据集之间的协调融合程度,确保数据生态的和谐统一。及时性着重衡量数据采集、处理与更新的时效水准,以高效的数据流转赋能实时决策支持。规范性聚焦于审视数据是否遵循预定的标准规范,力求达成数据格式的整齐划一与标准化呈现。效用性则专注于评估数据模型设计的优化成效以及对业务开展的支撑赋能水平,确保数据资产价值最大化释放。

上述评价指标体系全面覆盖了数据质量的关键维度,为证券行业数据仓库的质量管理提供了科学的框架。金融数据仓库数据质量评价指标层次结构如图 1 所示。

2.3. 数据质量评价指标量化

在明确各指标选取依据及意义后,需对其进行量化处理,将抽象的质量概念转化为具体可测的数据,为后续权重分配和综合质量评估提供坚实的数值基础。具体量化公式如下。

1. 完整性指标

数据记录完整率:用于评估数据表中记录的完备性。数据值完整率:反映表关键字段非空值的情况。数据关联率反映表间关联关系的完整程度。

$$\text{数据记录完整率} = \frac{\text{实际采集到的记录数}}{\text{应采集的记录数}}$$

$$\text{数据完整率} = \frac{\text{表关键字段非空值总数}}{\text{关键字段非空总数} \times \text{记录总数}}$$

$$\text{数据关联率} = \frac{\text{相关表间正确关联的记录数}}{\text{应关联的记录总数}}$$

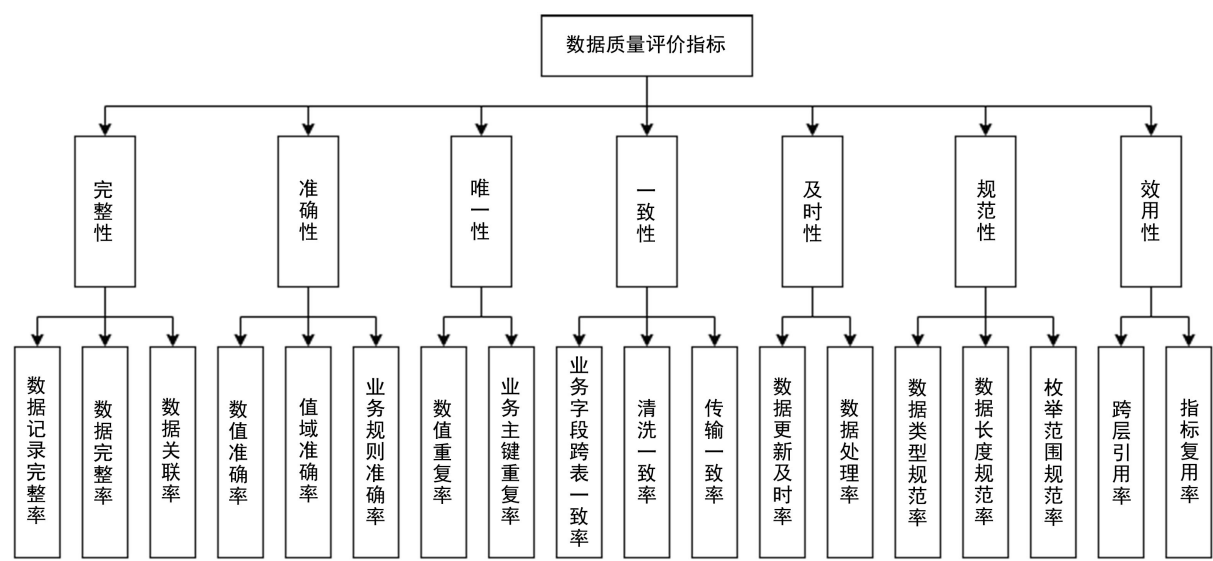


Figure 1. The hierarchical structure of financial data warehouse data quality evaluation indicators
图 1. 金融数据仓库数据质量评价指标层次结构

2. 准确性指标

数值准确率：评估字段值的精确性，计算需依据具体业务场景确定准确值范围，再统计实际准确值的数量占总数据量的比例。值域准确率：检查字段值是否在有效范围内。业务规则准确率：验证数据对业务逻辑的符合程度，比如校验当日规模是否与份额乘以当日净值相等。

$$\text{数值准确率} = \frac{\text{字段值准确的记录数}}{\text{字段值总记录数}}$$

$$\text{值域准确率} = \frac{\text{字段值处于值域范围内记录数}}{\text{字段值总数据量}}$$

$$\text{业务规则准确率} = \frac{\text{符合预设业务规则记录数}}{\text{表总记录数}}$$

3. 唯一性指标

数值重复率：衡量特定字段值的重复频率。业务主键重复率：检查具有唯一标识意义的字段是否存在重复。

$$\text{数值重复率} = \frac{\sum(\text{count}(\text{唯一值}) - 1)}{\text{总记录数}}$$

$$\text{业务主键重复率} = \frac{\text{总记录数} - \text{count}(\text{distinct主键})}{\text{业务主键总记录数}}$$

4. 一致性指标

传输一致率：衡量数据在跨系统传输后的保真性。清洗一致率：反映数据清洗前后的一致性。业务字段跨表一致率：用于衡量不同表间业务字段值的一致性。

$$\text{传输一致率} = \frac{\text{传输后数据匹配项数}}{\text{传输前数据项总数}}$$

$$\text{清洗一致率} = \frac{\text{清洗后与原始数据一致的数据项数}}{\text{原始数据项总数}}$$

$$\text{业务字段跨表一致率} = \frac{\text{跨表匹配的业务字段值的数量}}{\text{检查的业务字段值的总数}}$$

5. 及时性指标

数据更新及时率：数据是否按预定时间间隔更新，反映了数据在时间维度上的新鲜程度和时效性。

数据处理及时率：数据从采集到可用的时效性，体现了数据在处理流程中的流转速度和及时性。

$$\text{数据更新及时率} = \frac{\text{规定时间内更新数据记录数}}{\text{应更新的数据总记录总数}}$$

$$\text{数据处理率} = \frac{\text{规定时间内处理的数据项数}}{\text{总数据量}}$$

6. 规范性指标

通过数据类型规范率：确保字段值符合预定的数据类型定义。数据长度规范率：检查字段值的长度是否符合设定的限制。枚举范围规范率：验证字段值是否在规定的枚举范围内。

$$\text{数据类型规范率} = \frac{\text{符合类型规范的数据项数}}{\text{总数据项数}}$$

$$\text{数据长度规范率} = \frac{\text{符合长度规范的数据项数}}{\text{总数据项数}}$$

$$\text{枚举范围规范率} = \frac{\text{符合枚举规范的数据项数}}{\text{总数据项数}}$$

7. 效用性指标

跨层引用率衡量了数据在不同层级间的流动性和依赖性，对于评估数据仓库的整合能力和规范开发能力至关重要[17]。指标复用率反映了数据仓库中指标设计的通用性和效率。

$$\text{跨层引用率} = \frac{\text{跨层引用 ODS 表数量}}{\text{总 ODS 层数据表数}}$$

$$\text{指标复用率} = \frac{\text{DWD 层指标被 DWS 引用次数}}{\text{DWD 层总指标数}}$$

通过上述量化方法，研究能够将主观判断转化为客观数据，支持权重分配及综合质量评估，为企业优化数据质量管理提供清晰指导。

3. 金融数据质量评价模型

为全面评估金融数据仓库的质量，本研究构建了基于模糊层次分析法和熵权法的组合赋权模型。模糊层次分析法引入专家主观判断，熵权法则从数据特性出发，通过综合权重方法，兼顾主观经验与客观数据，实现科学有效的评价。

3.1. 模糊层次分析法

本节首先介绍模糊层次分析法的原理及应用方法。模糊层次分析法(Fuzzy AHP)是一种多准则决策工具，结合了模糊数学与层次分析的逻辑框架，适合应对不确定性因素和主观判断[12]。该方法将决策问题细致地划分为三个核心层次：目标层、准则层和方案层，采用 0.1~0.9 的标度方法对指标成对两两

比较, 形成模糊互补矩阵。最后通过模糊一致矩阵, 利用公式计算各指标权重。

基于数字标度建立起两层次之间的模糊判断矩阵 $R = (r_{ij})_{n \times n}$, 若模糊判断矩阵 $R = (r_{ij})_{n \times n}$ 满足 $0 \leq r_{ij} \leq 1, (i=1, 2, \dots, n; j=1, 2, \dots, n)$ 且满足 $r_{ij} + r_{ji} = 1$, 则矩阵 R 为模糊互补矩阵。根据下列公式计算模糊一致矩阵 $F = (f_{ij})_{n \times n} (i, j=1, 2, \dots, n)$

$$\begin{cases} f_i = \sum_{k=1}^n a_{ik} \\ f_{ij} = \frac{f_i - f_j}{2n} + 0.5 \end{cases} \quad (1)$$

利用如下模糊层次分析法权重计算公式对各指标权重进行计算:

$$\begin{cases} \beta = \frac{n-1}{2} \\ w_{f_i} = \frac{1}{n\beta} \sum_{j=1}^n f_{ij} + \frac{1}{n} - \frac{1}{2\beta} \end{cases} \quad (2)$$

w_{f_i} 是第 i 个指标所对应的权重, 则模糊层次分析法求得的权重向量 w_f 为:

$$w_f = [w_{f1}, w_{f2}, \dots, w_{fn}]$$

此外, 模糊层次分析法还包括对比较矩阵进行严格的一致性检验, 确保了评价结果的合理性和可信度。

3.2. 熵权法

为增强评价模型的客观性, 3.2 节将介绍熵权法的步骤和计算过程。熵权法(Entropy Weight Method, EWM)是一种广泛应用于多指标综合评价中的客观赋权方法。通过计算各指标的熵值, 该方法可以量化指标的有效性和重要性。通常包含下面几个步骤:

设第 i 个方案中的第 j 个指标的评估值为 $x_{ij} (i=1, 2, \dots, n; j=1, 2, \dots, m)$ 。首先, 对原始数据进行标准化处理, 避免因数据单位不同造成的计算偏差[18] [19]。正向指标标准化公式为:

$$x'_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \quad (3)$$

负向指标, 如数值重复率, 标准化公式为:

$$x'_{ij} = \frac{\min(x_{ij}) - x_{ij}}{\max(x_{ij}) - \min(x_{ij})} \quad (4)$$

接着计算每个指标在各个方案中的比重 p_{ij} , 公式为:

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^n x'_{ij}} \quad (i=1, 2, \dots, n; j=1, 2, \dots, m) \quad (5)$$

然后, 根据比重计算每个指标的熵值 e_j , 其中, $k = \frac{1}{\ln(n)}$ 为常数, 用于标准化熵值。

$$e_j = -k \sum_{i=1}^n p_{ij} \ln(p_{ij}) \quad (i=1, 2, \dots, n; j=1, 2, \dots, m) \quad (6)$$

计算第 j 个指标的差异系数 g_j ，反映指标的有效信息量，计算公式为：

$$g_j = 1 - e_j \quad (j = 1, 2, \dots, m) \quad (7)$$

差异系数越大，表示指标的区分度越高。最后，根据差异系数计算每个指标的权重 w_{ej} ：

$$w_{ej} = \frac{g_j}{\sum_{j=1}^m g_j} \quad (j = 1, 2, \dots, m) \quad (8)$$

可得到用熵权法计算得出的各评估指标的权重向量 w_e ：

$$w_e = [w_{e1}, w_{e2}, \dots, w_{en}]$$

这种方法的引入，显著减少了人为因素的干扰，为数据质量评价提供了更具客观性的权重分配方式。

3.3. 求组合赋权权重

在模糊层次分析法和熵权法结果的基础上，本研究采用几何平均法[12]将两种权重相结合，生成综合权重向量。公式如下：

$$W_i = \frac{\sqrt{w_{fi} w_{ei}}}{\sum_{i=1}^n \sqrt{w_{fi} w_{ei}}} \quad (9)$$

其中，模糊层次分析法求得的权重为 w_f ，由熵权法求得的权重为 w_e ，综合权重既保留了主观判断的经验性，又兼顾了客观数据的科学性。

3.4. 综合评价方法

利用综合权重与量化评分，计算数据质量综合得分。综合评价模型如下：

$$S = \sum_{i=1}^n w_i \times Q_i \quad (10)$$

其中， S 为数据仓库项目的数据质量的综合得分。通过预设等级标准，评估数据质量等级，并为企业识别薄弱环节与改进方向提供科学依据。

4. 算例分析

为验证评价模型的有效性，本章以 H 公司数据仓库的四个主题域为例，通过实际数据计算各指标量化值，并综合权重得出各主题域的质量评价结果，分析其数据质量状况。

4.1. 数据集与主题域选择

本研究采用分层抽样方法，从 H 公司金融数据仓库中抽取主体、交易、资产和渠道四个主题域数据作为分析对象，覆盖 2023 年 10 月至 2024 年 10 月期间累计产生的 1.2 亿条记录，其中核心业务数据的时间跨度为 2024 年 3 月至 2024 年 10 月。抽样过程按主题域进行分层，各层样本量占比分别为：交易域 42%、资产域 28%、渠道域 19%、主体域 11%，与各域日均数据增量比例保持严格对应。这些主题域构成了金融数据仓库的重要支柱，涵盖了 H 公司业务的主要数据流，具有关键支撑作用。交易域涵盖证券交易、基金申购、债券买卖等各类金融交易记录，主体域包含客户、机构等参与主体的详细信息，资产域涉及 H 公司和客户持有的金融资产信息，包括股票、债券、基金、衍生品等多种类别的资产数据，渠道域涵盖 H 公司通过线上(如移动端 APP、网上交易平台)和线下(如柜台、自助终端)提供服务的各类

数据。

4.2. 指标量化与权重计算

根据 2 章构建的指标体系, 利用生产环境真实数据计算四个主题域的量化值(见表 1)。随后, 采用模糊层次分析法与熵权法相结合的综合赋权模型, 计算每个指标的综合权重。此过程有效整合了专家判断与数据客观特性, 为综合质量评价奠定了基础。

Table 1. Quantification of indicators by subject area
表 1. 主题域指标量化表

评价指标	指标量化值			
	交易	主体	资产	渠道
数据记录完整率	1	0.64	0.75	0.72
数据完整率	1	0.66	0.82	0.73
数据关联率	1	0.75	0.82	0.73
数值准确率	0.99	0.64	0.93	0.76
值域准确率	0.99	0.66	0.88	0.77
业务规则准确率	0.99	0.65	0.9	0.77
数值重复率	0.03	0.23	0.20	0.13
业务主键重复率	0.02	0.24	0.15	0.14
业务字段跨表一致率	0.99	0.74	0.85	0.88
清洗一致率	0.88	0.66	0.76	0.72
传输一致率	0.99	0.75	0.85	0.88
数据更新及时率	0.88	0.80	0.76	0.75
数据处理率	0.82	0.80	0.76	0.75
数据类型规范率	0.93	0.79	0.79	0.73
数据长度规范率	0.89	0.76	0.76	0.87
枚举范围规范率	0.72	0.77	0.77	0.72
跨层引用率	0.18	0.26	0.28	0.39
指标复用率	0.79	0.86	0.83	0.67

4.2.1. 基于模糊层次分析法的评价指标权重分析

首先, 我们需要根据式(1)构建基于 7 个一级评价指标的模糊互补评价矩阵。一级指标模糊一致矩阵 F 如下:

$$F = \begin{pmatrix} 0.500 & 0.329 & 0.404 & 0.442 & 0.421 & 0.396 & 0.367 \\ 0.671 & 0.500 & 0.575 & 0.613 & 0.592 & 0.567 & 0.538 \\ 0.596 & 0.425 & 0.500 & 0.538 & 0.517 & 0.492 & 0.463 \\ 0.558 & 0.388 & 0.463 & 0.500 & 0.479 & 0.454 & 0.425 \\ 0.579 & 0.408 & 0.483 & 0.521 & 0.500 & 0.475 & 0.446 \\ 0.604 & 0.433 & 0.508 & 0.546 & 0.525 & 0.500 & 0.471 \\ 0.633 & 0.463 & 0.538 & 0.575 & 0.554 & 0.529 & 0.500 \end{pmatrix}$$

根据式(2), 利用模糊层次分析法计算的一级评价指标权重向量 w_{fa} 为:

$$w_{fa} = [0.112, 0.169, 0.144, 0.132, 0.139, 0.147, 0.157]$$

同样利用式(2)对二级评价指标权重向量进行求解, 求得权重如下:

$$\begin{aligned} w_{fb1} &= [0.296, 0.358, 0.346] & w_{fb2} &= [0.371, 0.333, 0.296] & w_{fb3} &= [0.45, 0.55] \\ w_{fb4} &= [0.384, 0.308, 0.308] & w_{fb5} &= [0.70, 0.30] & w_{fb6} &= [0.459, 0.333, 0.208] \\ w_{fb7} &= [0.20, 0.80] \end{aligned}$$

最后将一级指标的权重值和它包含的二级指标的权重值相乘, 从而得到模糊层次分析法计算的主观权重 W_f :

$$W_f = \begin{pmatrix} 0.0332 & 0.0401 & 0.0387 & 0.0627 & 0.0563 & 0.0500 \\ 0.0648 & 0.0792 & 0.0506 & 0.0407 & 0.0407 & 0.0973 \\ 0.0417 & 0.0675 & 0.0489 & 0.0306 & 0.0314 & 0.1256 \end{pmatrix}$$

4.2.2. 基于熵权法的评价指标权重分析

利用表 1 中的数据值, 根据熵权法的步骤(3.2 节), 得出各评价指标的客观权重向量。具体计算步骤结合式(3)~式(8)求得各评价指标基于熵权法的客观权重 W_e :

$$W_e = \begin{pmatrix} 0.0607 & 0.0556 & 0.0776 & 0.0443 & 0.0450 & 0.0439 \\ 0.0603 & 0.0451 & 0.0420 & 0.0512 & 0.0430 & 0.0744 \\ 0.0585 & 0.0558 & 0.0836 & 0.0831 & 0.0393 & 0.0366 \end{pmatrix}$$

4.2.3. 综合权重的确定

根据 3.3 节式(9), 将模糊层次分析法和熵权法得到的主客观权重进行几何平均, 以得出综合权重向量 W 。

$$W = \begin{pmatrix} 0.0464 & 0.0489 & 0.0567 & 0.0545 & 0.0521 & 0.0485 \\ 0.0647 & 0.0619 & 0.0477 & 0.0473 & 0.0433 & 0.0881 \\ 0.0511 & 0.0635 & 0.0662 & 0.0552 & 0.0364 & 0.0702 \end{pmatrix}$$

4.3. 综合质量评价

根据 4.2 节中的表 1 各自指标的量化结果, 运用综合评价方法中的式(10)将数据质量管理中每个影响因素的定量评分与综合权重结合, 分别计算出四个主题域的综合评价价值。

$$S = (0.7802, 0.6553, 0.7131, 0.6669)$$

从综合评价价值来看, 各主题域的数据质量状况存在明显差异。交易域综合评价价值最高, 表明其数据质量在多个方面表现出色。数据记录完整率、数据完整率、数值准确率等多项指标接近或达到较高水平。资产域综合评价价值处于较上水平, 数据质量有一定保障, 但仍存在提升空间。例如数据记录完整率、数据关联率等指标表现较好, 但在准确性和规范性方面还有改进余地。

主体域和渠道域综合评价价值较低, 数据质量问题较为突出。主体域在准确性方面, 3 个二级指标准确率均处于较低水平, 如数值准确率仅为 0.64, 这意味着数据在反映真实业务情况时存在较大偏差。一致性方面, 业务字段跨表一致率等指标虽有一定表现, 但整体仍有待提高。主体域数据来源复杂多样, 不同开户方式下数据格式与标准不统一, 分支机构填报风格各异, 使得数据整合极易产生错误。此外, 客户信息变动频繁, 而机构获取和更新信息不及时也加剧了数据的不准确和不一致。渠道域同样面临准

确性和一致性问题, 其数值准确率为 0.76 反映出数据存在一定程度的错误和冗余。渠道域涵盖线上线下多种渠道, 各渠道厂商交互接口标准不一致, 以及网络波动和人工失误等原因都容易导致数据传输和交互过程中出现数据不一致等问题。

针对上述评价结果, 为提升 H 公司数据仓库的数据质量, 提出以下对策:

建立统一的数据标准和接口规范: 制定适用于所有开户方式和分支机构的数据标准, 统一规定客户信息的必填字段、数据类型、长度限制等, 避免因格式和标准不统一导致的数据错误。对各渠道的接口进行全面梳理和统一规划, 建立接口标准文档, 明确接口的数据格式、传输方式、加密要求等, 要求各渠道和系统严格按照标准进行开发和对接。

强化人员培训与自动化工具应用: 加强操作人员培训, 规范操作流程, 以减少人为操作风格差异导致的错误。引入自动化的数据整合工具, 减少人工干预, 提升数据处理的准确性和效率。

网络架构优化与数据采集机制增强: 优化网络架构, 增加数据采集的重试机制和数据校验机制, 确保在网络不稳定情况下采集到的数据完整且准确。建立与客户信息变动和机构信息更新相关的实时监测系统, 一旦有信息变更, 能够及时同步到数据仓库中, 确保数据的及时性和准确性。

5. 结论

本研究基于模糊层次分析法和熵权法的组合模型, 构建了一套适用于金融数据仓库的数据质量评价体系, 并以 H 公司数据仓库为例进行了实证分析。通过对数据完整性、准确性、一致性、及时性等关键质量指标的量化评估, 识别了不同主题域在数据质量方面的优势与不足。研究结果表明, 模糊层次分析法能够有效反映专家对各项数据质量指标的重要性判断, 熵权法则进一步校正了评价过程中的主观偏差。通过两种方法的组合赋权, 不仅提升了数据质量评价的科学性和准确性, 还为数据仓库的数据管理提供了明确的改进方向。

研究中发现, H 公司数据仓库的关键领域在于主体和渠道数据的准确性及一致性相对较低, 这表明公司在数据清洗、格式标准化和实时性处理上存在改进空间。结合本研究提出的指标体系和权重计算模型, 公司可以进一步优化数据质量管理策略, 增强在风险管理、客户服务、合规报送等方面的数据支持能力, 从而提升业务效率和决策准确性。

参考文献

- [1] Liaw, S.-T., Guo, J.G.M., et al. (2021) Quality Assessment of Real-World Data Repositories across the Data Lifecycle: A Literature Review. *Journal of the American Medical Informatics Association*, **28**, 1591-1599. <https://doi.org/10.1093/jamia/ocaa340>
- [2] Ya, L., Li, Y., Song, H., et al. (2020) Method for Calculating the Weights of Internet+Government Service Data Quality Assessment Indexes Based on Analytic Hierarchy Process. *Journal of Physics: Conference Series*, **1584**, Article 012043. <https://doi.org/10.1088/1742-6596/1584/1/012043>
- [3] Ya, L., Li, Y., Song, H., et al. (2020) Studies on Data Quality Evaluation Index System for Internet Plus Government Services in Big Data Era. *Journal of Physics: Conference Series*, **1584**, Article 012014. <https://doi.org/10.1088/1742-6596/1584/1/012014>
- [4] Gupta, N. (2023) Optimising Data Quality of a Data Warehouse Using Data Purgation Process. *International Journal of Data Mining, Modelling and Management*, **15**, 1-13. <https://doi.org/10.1504/IJDMMM.2023.10052818>
- [5] Chu, Y.C., Yang, S.S. and Yang, C.C. (2001) Enhancing Data Quality through Attribute-Based Metadata and Cost Evaluation in Data Warehouse Environments. *Journal of the Chinese Institute of Engineers*, **24**, 497-507. <https://doi.org/10.1080/02533839.2001.9670646>
- [6] Travkin, P. (2023) Data Observability: Is It Data Quality Monitoring or More? *Database Trends and Applications*, **37**, 1-5.
- [7] Soňa, K. (2023) Data Governance Model to Enhance Data Quality in Financial Institutions. *Information Systems Management*, **40**, 1-10. <https://doi.org/10.1080/10580530.2022.2042628>

-
- [8] 周艳会, 曾荣仁. 基于元数据的数据质量管理研究[J]. 信息技术与信息化, 2020(7): 26-29.
 - [9] 刘智锋, 王继民, 李倩. 元数据质量评价研究综述[J]. 情报理论与实践, 2022, 45(7): 42-48.
 - [10] 赵恩毅. 大数据中的数据清洗与预处理技术研究[J]. 信息记录材料, 2024, 25(3): 195-197.
 - [11] 宋俊典, 刘丰源. 一种支持数据质量评价的方法与应用研究[J]. 计算机应用与软件, 2018, 35(5): 328-333.
 - [12] 计蓉, 侯慧娟, 盛戈皞, 等. 基于组合赋权法和模糊综合评价的电力设备状态数据质量评估[J]. 高电压技术, 2024, 50(1): 274-281.
 - [13] 赵毅. 基于大数据平台构建数据仓库的研究与实践[J]. 中国金融电脑, 2017(5): 37-42.
 - [14] 刘明吉, 张晓京, 刘洪杰, 等. 数据仓库在证券交易中的研究与应用[J]. 计算机工程, 2000(2): 47-49+94.
 - [15] ISO/IEC JTC 1/SC 7. ISO/IEC25012-2008 软件工程, 软件产品质量要求和评估(SQuaRE)数据质量模型[S]. 日内瓦: 国际标准化组织, 2008.
 - [16] 姚前, 蒋东兴, 刘铁斌, 周云晖, 王东明, 毛嘉伟, 等. GB/T42775-2023 证券期货业数据安全风险防控数据分类分级指引[S]. 北京: 中国标准出版社, 2023.
 - [17] 刘益江, 毛宁, 陈庆新. 一种评估数据仓库设计质量的方法[J]. 计算机技术与发展, 2012, 22(9): 161-165.
 - [18] 申梁, 陈立强, 黄勇. 基于模糊层次分析法和熵权法的岩溶发育强度评价方法探讨[J]. 勘察科学技术, 2024(2): 30-35.
 - [19] 杨栋枢, 杨德胜. 基于熵权和层次分析法的数据质量评估研究[J]. 现代电子技术, 2013, 36(22): 39-42.