DOI:10.16652/j.issn.1004-373x.2013.22.028



# 基于熵权和层次分析法的数据质量评估研究

杨栋枢，杨德胜
(安徽南瑞继远软件有限公司，安徽 合肥 230088)

摘 要：数据质量评价是电力企业运营监测(控)中心数据质量管理的重要工作，其中关键的一项就是对运营监测(控)数据质量进行分析评估。针对电力企业运营监控中心数据质量现状，建立了运营监控中心数据质量评价指标体系。通过构建基于熵权与层次分析法的运营监控中心数据质量组合权重评价模型，应用电力企业运营监测(控)中心支撑信息系统数据验证了运营监控中心数据质量组合权重评价模型的有效性和适用性。

关键词：数据质量评价；指标体系；熵权；层次分析法
中图分类号：TN911-34 文献标识码：A 文章编号：1004-373X(2013)22-0039-04

## Data quality assessment based on entropy weight and AHP

YANG Dong-shu，YANG De-sheng
(Anhui Nari Software Co.，Ltd.，Hefei 230088，China)

Abstract：Data quality evaluation is an important work of the data quality management of the power business operations monitoring (control) center，in which one of the key tasks is to analyze and evaluate the operation monitoring (control) data quality. According to the status quo of data quality of the business operations mornitoring (control) center，an index system for data quality evaluation of the operation monitoring (control) center was established. By building data quality combination weight evaluation model based on entropy weight and AHP，the data from support information system of the monitoring (control) center was used to verify the validity and applicability of data quality combination weight evaluation model of the monitoring (control) center.

Keywords：data quality assessment；index system；entropy weight；AHP

## 0 引 言

随着电力企业运营监测(控)中心支撑信息系统建设逐步深入，运营监测(控)中心支撑信息系统也面临着复杂多变的数据环境。信息数据的大量产生，以及从各种渠道收集获取的不符合系统要求的数据造成了维度不完整、数据取值范围不一致、历史数据缺失、频度不一致和计量单位不一致等问题，从而对企业的数据分析、数据应用影响非常严重。

数据作为运营监测(控)中心支撑信息系统的基础和核心，对运营监测(控)中心支撑信息系统起着至关重要的作用，数据质量的高低直接对整个系统有直接的影响。好的数据质量是各种数据分析能够得到有意义结果的基础条件，而质量低劣的数据已经成为影响企业进行正确决策的重要因素。

运营监测(控)中心支撑信息系统在公司生产经营中的支撑作用越来越突出，企业级数据资源已成为公司重要的核心资源，为公司领导及各业务部门及时全面掌握生产经营情况以及科学分析决策提供了重要依据。因此，公司确定将对运营监测(控)中心支撑信息系统数据开展数据质量评价、通报、考核工作。

数据质量评估是为了保障电力企业两级运营监测(控)中心数据及时、完整、准确地接入，提升数据质量，为了准确而客观地评价数据质量，有效地指导电力企业运营监测(控)中心工作开展，必须在建立数据质量评价指标体系的基础上，科学地确定各项指标的权重。

本文基于当前电力企业运营监测(控)中心数据质量研究的成果，设置评价指标体系，利用熵权法法[1]和层次分析法[2]分别确定运营监测(控)中心数据质量评价指标的客观指标权重和主观指标权重，然后综合评价指标的主观权重和客观权重计算各评价指标的组合权重，而建立了电力企业运营监测(控)中心数据质量评价的熵组合权重评价模型。

收稿日期：2013-06-15

## 1 运营监测(控)中心数据质量评价指标体系

### 1.1 数据质量评价指标选择原则

作为电力企业运营监测(控)中心数据质量评价模型,需要通过对一组关键性指标的监测和分析,以此来反应电力企业运营监测(控)中心支撑信息系统数据质量水平情况,因此建立科学的公司数据质量水平评价指标是建立电力企业运营监测(控)中心数据质量水平评价模型的重要环节。本文根据电力企业运营监测(控)中心支撑信息系统数据现状,在进行了大量的分析研究的基础上,根据以下原则建构了电力企业运营监测(控)中心数据质量评价指标体系:

⑴ 全面覆盖、重点突出。数据质量管理工作范围覆盖运监中心所有业务数据,包括各源业务系统线上自动接入的系统数据和各业务部门以线下方式手工录入的各类数据;重点核查系统自动接入数据,并进行数据溯源、分析和数据评价,以逐步提高线上自动接入比例,减少线下手工录入数据,支撑运营监测(控)工作及时有效开展。

⑵ 统一规范、客观高效。制定统一规范的数据质量规则和评价工作流程,并将规则和流程固化到运营监测(控)信息支撑系统中,依托系统对数据质量进行在线监测,客观、真实、即时反映数据质量情况。

⑶ 循序渐进、持续优化。以运营监测(控)数据需求为基础,根据数据接入实际情况,不断丰富数据质量核查规则,完善和提升评价标准,动态调整评价指标,持续优化评价体系,实现以通报评价促进数据及时、完整、准确接入,逐步提升运营监测(控)数据质量。

⑷ 定性与定量相结合的原则。电力企业运营监测(控)中心数据质量评价中,根据不同评价内容的特点采用不同性质的评价指标,能够更准确地反应电力企业运营监测(控)中心数据质量的现状和趋势。

⑸ 实用性与可比性原则:电力企业运营监测(控)中心数据质量评价指标设计要具有可行性、可操作性、实用性以及能够进行纵向比较和横向比较,指标要简化以及数据易于获取。

### 1.2 运营监测(控)中心数据质量指标分析

根据以上原则和数据质量原理,结合电力企业运营监测(控)中心数据的实际情况,本文选取的指标维度包括以下四个维度:

⑴ 数据接入情况要求数据在规定的时间内接入全部接入系统。主要从数据应接入数量、实际接入数量、指标历史数据等方面进行核查。数据接入情况用指标数据自动采集率和指标历史数据接入率两个指标来度量,其定义为:指标数据自动采集率反映指标通过系统接入的自动化程度,是指实际由源业务系统自动接入的指标数据占指标体系中应接指标总数的比例;指标历史数据接入率反映历史数据接入情况的指标,指历史数据在规定的时间和频度周期内接入系统,由月指标历史数据接入率、周指标历史数据接入率、日指标历史数据接入率等构成。

⑵ 数据质量及时性规则要求数据在规定的时间和频度周期内接入系统。主要从各源业务系统数据接入及时性及各省(市)公司数据上报及时性等方面进行核查。数据质量及时性用指标数据接入及时率来度量,其定义为:指标数据接入及时率反映数据接入及时情况的指标,指数据在规定的时间和频度周期内接入系统的比例,由月指标数据及时率、周指标数据及时率、日指标数据及时率等构成。

⑶ 数据质量完整性规则要求数据记录内容完整。主要从数据业务维度组合完整、单位维度完整、指标值完整等方面进行核查。数据质量完整性用指标数据完整率来度量,其定义为:指标数据完整率反映数据接入完整情况的指标,指数据记录内容完整,包括数据业务维度组合完整、单位维度完整,指标值完整等,由月指标数据完整率、周指标数据完整率、日指标数据完整率等构成。

⑷ 数据质量准确性规则要求数据符合各业务规则和业务实际。主要从数据间业务逻辑准确和数据内各种维度、频度、字段之间业务逻辑准确,以及数据值、精度属性准确等方面进行核查。数据质量准确性用指标数据准确率来度量,其定义为:指标数据准确率反映接入数据准确情况的指标,指数据符合各业务规则和业务实际,包括各种数据间业务逻辑准确和数据内各种维度、频度、字段之间业务逻辑准确,以及数据值、数据精度等属性准确,由月指标数据准确率、周指标数据准确率、日指标数据准确率等构成。

本文依据上述原则和文献[3-5]的研究成果以及结合电力企业运营监测(控)中心数据实际情况,确定电力企业运营监测(控)中心数据质量评价指标体系如图1所示。

## 2 运营监测(控)中心数据质量评价模型

### 2.1 层次分析法及应用

层次分析法是美国匹茨堡大学教授T.L.Saaty提出的一种定性与定量相结合的决策分析方法,将决策者的决策思维过程与经验判断模型化、数量化的过程。层次分析法在信息系统数据质量综合评价领域得到广泛应

用[6]。采用层次分析法将目标问题分层逐步分解细化，将专家经验知识引入不同的层次中。问题分解的最低层元素是可以明确获取和度量的电力企业运营监测（控）中心数据质量评价的各个单项指标，以相对标度作为电力企业运营监测（控）中心数据质量评价度量的测度，从而回避了绝对量求解的困难。

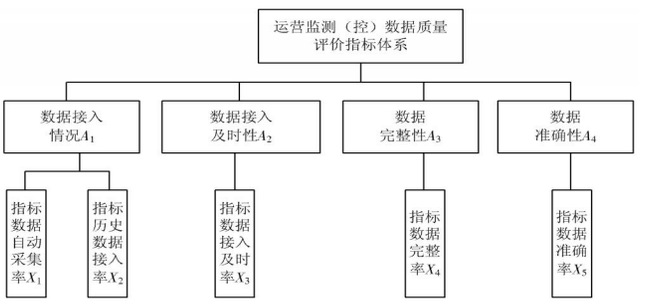


图1 电力企业运营监测（控）中心数据质量评价指标体系

### 2.2 熵权法及应用

运用层次分析法确定各层次评价指标的权重对专家经验水平要求很高，评价结果受人为主观因素影响较大，而信息熵可以有效地弥补这一不足。信息熵是用来度量随机变量不确定程度，可以用来解决信息量的度量问题，即对不确定性的了解所需的信息量，可以被用来消除不确定性的多少来表示。如果某评价指标的熵越小，说明该指标提供的信息量就越大，在综合评价中所起的作用就越大，权重就越高[1]。应用熵权法[7-13]可以尽可能消除人为因素对应用层次分析时计算各指标权重的影响，使评价结果更为实际。熵权计算方法如下：

设有 $m$ 个待评对象，$n$ 个评价指标，则原始数据矩阵为：

$$X=\left(x_{ij}\right)_{m\times n} \tag{1}$$

对于某个评价指标，信息熵为：

$$E_j=-\sum_{i=1}^{m}p_{ij}\frac{\ln p_{ij}}{\ln m} \tag{2}$$

式中 $p_{ij}=\dfrac{x_{ij}}{\sum_{i=1}^{m}x_{ij}}$

评价指标的信息效用价值为 $D_j=1-E_j$，则该评价指标的熵权为：

$$W_j=\frac{D_j}{\sum_{j=1}^{n}D_j} \tag{3}$$

### 2.3 组合评价模型

基于直接计算电力企业运营监测（控）中心数据质量评价的困难及采用层次分析法的不足，本文采用组合权重法对电力企业运营监测（控）中心数据质量进行度量，主要步骤如下：

步骤1：电力企业运营监测（控）中心数据质量识别，确定电力企业运营监测（控）中心数据质量评价指标 $x_i$，建立电力企业运营监测（控）中心数据质量评价指标体系。

步骤2：应用层次分析法获得运营监测（控）中心数据质量评价指标的主观权重 $W_j$，$j=1,2,\cdots,n$。

步骤3：根据式（1）~式（3）应用熵权法计算运营监测（控）中心数据质量评价指标的客观权重 $w_j$，$j=1,2,\cdots,n$。

步骤4：计算综合权重

$$\bar{\omega}=\left\{\frac{W_1w_1}{\sum_{j=1}^{n}W_jw_j},\frac{W_2w_2}{\sum_{j=1}^{n}W_jw_j},\cdots,\frac{W_nw_n}{\sum_{j=1}^{n}W_jw_j}\right\}=(\omega_1,\omega_2,\cdots,\omega_n)$$

$$\text{s.t.}\quad \sum_{j=1}^{n}\omega_j=1;\quad \omega_j>0 \tag{4}$$

步骤5：计算评价结果。利用式（4）可得第 $i$ 个电力企业运营监测（控）中心数据质量 $A_i$：

$$A_i=\sum_{j=1}^{n}\omega_jX^*\times 100$$
$$i=1,2,\cdots,m;j=1,2,\cdots,n \tag{5}$$

式中：$\omega_j$ 为综合权重；$X^*=(x_{i1},x_{i1},\cdots,x_{in})$ 为预处理后的电力企业运营监测（控）中心数据质量评价指标度量值。

## 3 算例分析

以电力企业8家下属单位运营监测（控）中心基本信息为依据，根据本文所建立的电力企业运营监测（控）中心数据质量评价指标体系对这8家单位进行综合数据质量评价。

（1）获取运营监测（控）中心数据质量评价指标数据。指标数据主要来源于电力企业运营监测（控）中心支撑信息系统，部分指标数据通过人工线下收集，经过预处理后的电力企业运营监测（控）中心数据质量评价指标度量值见表1。

（2）对图1所示的5个指标应用层次分析法计算运营监测（控）中心数据质量评价指标的主观权重向量为：

$$\boldsymbol{W}=(0.350\ 0,0.162\ 5,0.162\ 5,0.162\ 5,0.162\ 5)$$

（3）应用熵权法，运用式（1）~式（3）计算运营监测（控）中心数据质量评价指标的客观权重向量为：

$$\boldsymbol{w}=(0.234\ 0,0.190\ 0,0.192\ 0,0.192\ 0,0.192\ 0)$$

（4）应用式（4）计算得到运营监测（控）中心数据质量评价指标的综合权重为：

$$\bar{\omega}=(0.396\ 9,0.149\ 8,0.151\ 1,0.151\ 1,0.151\ 1)$$

⑸ 计算运营监测（控）中心数据质量综合评价结果，应用式⑸可得各单位运营监测（控）中心数据质量综合评价度量值：

$$A=(88.04,84.55,86.48,80.00,76.81,66.59,80.65,73.60)$$

根据上述结果可以得到公司的8家下属单位运营监测（控）中心数据质量综合排序为：

$$A_1>A_3>A_2>A_7>A_4>A_5>A_8>A_6$$

该评价结果和公司下属8家单位的运营监测（控）中心数据质量综合评价结果与实际情况一致，从而证明了该评价模型的有效性。

表1 公司数据综合治理水平评估的指标值

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ | $A_7$ | $A_8$ |
|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0.85 | 0.80 | 0.86 | 0.75 | 0.70 | 0.50 | 0.74 | 0.65 |
| $X_2$ | 0.75 | 0.70 | 0.65 | 0.68 | 0.60 | 0.75 | 0.74 | 0.72 |
| $X_3$ | 1.00 | 1.00 | 1.00 | 0.90 | 0.95 | 0.85 | 0.93 | 0.90 |
| $X_4$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 0.90 |
| $X_5$ | 0.85 | 0.80 | 0.82 | 0.75 | 0.70 | 0.50 | 0.78 | 0.65 |

## 4 结 语

本文将层次分析法和熵权法结合起来建立了电力企业公司运营监测（控）中心数据质量综合评价模型，充分发挥二者的优势，取长补短，使构成的模型具有两者的优点。应用电力企业运营监测（控）中心支撑信息系统的数据进行案例分析，结果表明该评价模型适用有效，能够较全面反映电力企业运营监测（控）中心数据质量总体水平。

## 参 考 文 献


[1] 文军.基于熵权法的航空公司绩效评价研究[J].科学技术与工程，2009，9（22）：6939-6941.

[2] 张永艳.应用层次分析法确定政府网站绩效评估指标权重的确定[J].现代商贸工业，2010（2）：239-241.

[3] 黄武锋，郑华.面向企业信息化的数据质量评估研究[J].计算机技术与发展，2011，21（1）：186-188.

[4] 王继民，赵运革，徐波.数据质量评估方法在水利普查中的应用[J].水利发展研究，2012（7）：11-14.

[5] 杨青云，赵培英，杨冬青，等.数据质量评估方法研究[J].计算机工程与应用，2004（9）：13-15.

[6] 高起蛟，严凤斌，池斌，等.层次分析法（AHP）在数据质量评估中的应用[J].信息技术，2011（3）：168-170.

[7] 冯义，李洪东，田廓，等.基于熵权和层次分析法的电力客户风险评估及其规避[J].继电器，2007，35（24）：67-73.

[8] 陈文斌，龚代圣.基于AHP熵权法的信息化厂商评价模型及应用[J].现代电子技术，2012，35（12）：102-106.

[9] 冯义，李洪东，田廓，等.熵权系数法在湖北某高速公路投标决策中的应用[J].工程造价管理，2009（3）：34-36.

[10] 刘宁，高飞燕.基于AHP-FCE的供应商选择问题研究与应用[J].计算机技术与发展，2009，19（11）：11-13.

[11] 朱强，阎子刚.基于AHP与TOPSIS算法的供应商选择决策方法[J].物流与信息，2008，17（2）：197-199.

[12] 王道平，王煦.基于AHP/熵值法的钢铁企业绿色供应商选择指标权重研究[J].软科学，2010，24（8）：117-122.

[13] 罗军刚，解建仓，阮本清.基于熵权的水资源短缺风险模糊综合评价模型及应用[J].水利学报，2008（9）：92-97.



作者简介 杨栋枢 男，1970年出生，安徽宿州人，高级工程师。主要从事电力业务建模、数据分析及电力行业信息化建设项目管理工作。

杨德胜 男，1982年出生，硕士，工程师。主要从事电力行业数据挖掘和人工智能方面的工作。