DOI:10.11705/j.issn.1672-643X.2014.01.47

基于交互验证的数据质量评估方法的研究

凌 云, 吕王勇, 张里静
(四川师范大学 数学与软件科学学院, 四川 成都 610068)

摘 要: 数据是决策者分析问题的重要依据,其质量的高低直接影响统计分析的科学性和决策的正确性。文章提出基于交互验证的数据质量评估方法,在最小均方误差准则下,建立最优的交互验证模型,对数据质量作评估。最后,通过对成都市的生活用水量的实证分析表明,运用交互验证方法,可使数据质量的评估更加合理,符合实际情况。

关键词: 数据质量评估; 最小均方误差; 交互验证; 生活用水量; 成都市
中图分类号:TV213 **文献标识码**: A **文章编号**: 1672-643X(2014)01-0214-03

Study on assessment method of data quality based on cross-validation

LING Yun, LÜ Wangyong, ZHANG Lijing
(*School of Mathematics and Software Science, Sichuan Normal University, Chengdu* 610068, *China*)

Abstract: Data is an important basis for decision-maker to analyze problems. The data quality straightly affects the scientificalness of statistical analysis and the correctness of policy-making. The article proposed a method of data quality assessment based on cross-validation and established the optimal model of cross-validation to assess data quality in criterion of the minimum mean square error. Finally, through the empirical analysis of Chengdu domestic water consumption, the results indicated that the use of cross-validation method made the assessment of data quality more reasonable and conformed the actual situation.

Key words: assessment of data quality; minimum mean square error; cross-validation; demestic water consumption; Chengdu

数据质量评估是对调查、汇总、整理完毕的数据的质量进行科学的、实事求是的分析和评价,从准确性、适用性、完整性、可靠性以及可获得性[1] 5个方面对数据进行评估。数据质量的高低直接影响统计分析的科学性和决策的正确性,所以,数据质量评估的目的是为了掌握数据的可靠程度,以便正确使用数据,并有针对性的采取措施,不断提高数据的质量[2]。目前关于数据质量评估的方法非常多。许涤龙等[3]详细讨论了逻辑关系检验法、计量模型分析法、核算数据重估法、统计分布检验法、调查误差评估法以及多维评估法在数据质量的评估中的运用;卢二坡等[4]运用稳健MM估计方法评估GDP数据质量;高起蛟等[5]运用层次分析法评估报表数据质量;李庭辉等[6]运用时间序列匹配性方法评估中国GDP数据质量。交互验证方法至今没有用于数据质量评估中。因此,本文提出基于交互验证的数据质量评估方法,选取成都市1978-2011年的生活用水量数据[7],在最小均方误差准则下,建立最优的交互验证模型,对数据质量作评估。

1 交互验证的数据质量评估方法

交互验证是一种基于拟合数据模型的验证方法,其基本思想是:给定样本数据 $X=(x_1,x_2,\cdots,x_n)$,将其随机分为 A_1、B_1 两部分,$A_1=(x_{A_1},x_{A_2},\cdots,x_{A_p})$, $B_1=(x_{B_1},x_{B_2},\cdots,x_{B_q})$,其中 $p<q<n$ 且 $p+q=n$(图1),以 B_1 中的数据作为样本,根据样本数据的变化特点建立有效模型 $f_1(x_{B_1})$,拟合 A_1 中的数据,将拟合值 $\hat{A}_1=(\hat{x}_{A_1},\hat{x}_{A_2},\cdots,\hat{x}_{A_{AP}})$ 与真实值做比较,计算均方误差:

$$MSE_1=\frac{\sum_i^p(x_{A_i}-\hat{x}_{A_i})^2}{p} \tag{1}$$

收稿日期:2013-06-20; **修回日期**:2013-07-29
基金项目:四川省教育厅重点项目(12ZA137); 四川省重点实验室(PJ201107)
作者简介:凌 云(1989-),女,四川资阳人,在读研究生,主要从事数据质量评估研究。
通讯作者:吕王勇(1979-),女,四川邛崃人,博士,副教授,主要从事数据质量评估研究。

再将样本数据随机分为 A_2、B_2 两部分(图 1),以 B_2 中的数据作为样本,根据样本数据的变化特点建立有效模型 $f_2(x_{B_i})$,拟合 A_2 中数据的值,将拟合值与真实值做比较,计算均方误差:

$$MSE_2 = \frac{\sum_{i}^{p} (x_{A_i} - \hat{x}_{A_i})^2}{p} \tag{2}$$

如此重复 m 次,可得到 $MSE_1, MSE_2, \cdots, MSE_m$,当 MSE 最小时,建立的有效模型 f 能够比较准确的刻画样本数据 $X = (x_1, x_2, \cdots, n)$,即在最小均方误差准则下,建立了最优的交互验证模型,再通过模型拟合 $\hat{X} = (\hat{x}_1, \hat{x}_2 \cdots, \hat{x}_n)$,分析拟合值与实际值之间的相对误差:

$$W_i = \left| \frac{x_i - \hat{x}_i}{x_i} \right| \tag{3}$$

在置信水平的条件下找出异常值。对于找出的异常值,还需要从统计意义上判断其显著性。本文采用 T - G 检验法[8]对判断的异常值进行统计意义上的显著性检验,对数据质量做出评估。

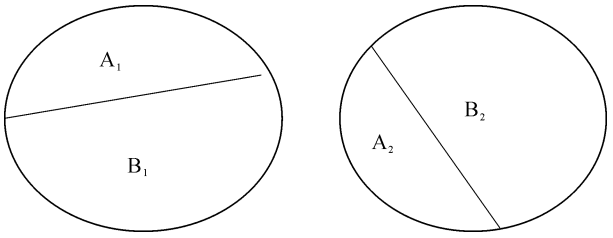

图 1　样本数据划分示意图

2　基于交互验证的成都市 1978 - 2011 年的生活用水量数据质量评估

水是人类生活必不可少的自然资源,随着经济的快速发展,成都市人口日益集聚,城市不断扩建,生活用水不断增加,准确统计生活用水量数据,评估其数据质量,对基于该数据来分析影响因素[9]和对未来做预测[10-11]有重要意义。目前,还没有人对成都市生活用水量作数据质量评估,因此,以该数据为样本的分析和预测,其可信度不完全令人信服。

2.1　交互验证

对成都市 1978 - 2011 年的生活用水量数据 $y_t (t = 1, 2, \cdots, 34)$,由于时间序列数据的特殊性以及样本数据容量较小,选择忽略末端的几个数据,建立交互验证模型。

根据数据 y_t 的时序图(图2),可以看出,随着时间得推移,该序列呈现上升的趋势 $f(t)$,是非平稳的。因此,对于该序列考虑用组合模型来刻画:

$$y_t = f(t) + u_t \tag{4}$$

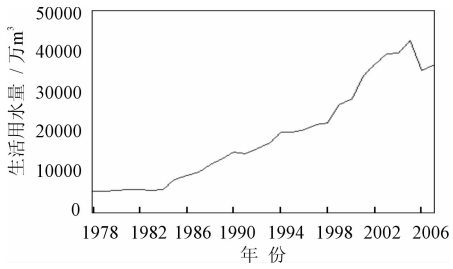

图 2　生活用水量时序图

又根据水资源的有限性,生活用水量不可能永远随着时间的推移呈上升趋势,而是在某个时间以后,呈现出平稳的趋势。故对于确定性趋势部分 $f(t)$,考虑用 S 型曲线模型来拟合。

利用 SPSS 软件[12-14]建模如下:

表 1　交互验证模型　　万 m^3

数据划分		交互验证模型	MSE
A	B		
2011	1978 - 2010	$f(t) = \frac{52400}{1 + 16.757e^{-0.134t}}$, $u_t = 0.5946u_{t-1} + \varepsilon_t$	739583.43
2010、2011	1978 - 2009	$f(t) = \frac{51950}{1 + 16.745e^{-0.135t}}$, $u_t = 0.5946u_{t-1} + \varepsilon_t$	388109.00
2009 - 2011	1978 - 2008	$f(t) = \frac{52490}{1 + 16.781e^{-0.134t}}$, $u_t = 0.6137u_{t-1} + \varepsilon_t$	720501.53
2008 - 2011	1978 - 2007	$f(t) = \frac{52490}{1 + 16.781e^{-0.134t}}$, $u_t = 0.6021u_{t-1} + \varepsilon_t$	2476077.00

由最小均方误差原则得到最优的交互验证模型是以 1978 - 2009 年数据建立的模型。

2.2　生活用水量数据质量评估

由于生活用水数据的收集与统计可能会因为中间多次的数据处理,出现较大的误差,故本文给出置信水平 $\alpha = 0.10$,即相对误差 $w_i > 0.1$,则认为对应的数据是异常的,再对该数据进行统计上的显著性检验。

通过最优的交互验证模型,计算 1978 - 2011 年的拟合值,并与真实值做比较,得到相对误差 $w_1, w_2, \cdots, w_{34}$,拟合值的相对误差见表 2。

表 2 生活用水量拟合相对误差表

年份	相对误差	年份	相对误差	年份	相对误差	年份	相对误差	年份	相对误差
1978	$w_1 = 0.34$	1985	$w_8 = 0.08$	1992	$w_{15} = 0.09$	1999	$w_{22} = 0.00$	2006	$w_{29} = 0.12$
1979	$w_2 = 0.27$	1986	$w_9 = 0.12$	1993	$w_{16} = 0.00$	2000	$w_{23} = 0.01$	2007	$w_{30} = 0.10$
1980	$w_3 = 0.02$	1987	$w_{10} = 0.02$	1994	$w_{17} = 0.06$	2001	$w_{24} = 0.07$	2008	$w_{31} = 0.00$
1981	$w_4 = 0.04$	1988	$w_{11} = 0.02$	1995	$w_{18} = 0.06$	2002	$w_{25} = 0.10$	2009	$w_{32} = 0.06$
1982	$w_5 = 0.01$	1989	$w_{12} = 0.10$	1996	$w_{19} = 0.15$	2003	$w_{26} = 0.08$	2010	$w_{33} = 0.01$
1983	$w_6 = 0.12$	1990	$w_{13} = 0.09$	1997	$w_{20} = 0.10$	2004	$w_{27} = 0.03$	2011	$w_{34} = 0.02$
1984	$w_7 = 0.27$	1991	$w_{14} = 0.07$	1998	$w_{21} = 0.11$	2005	$w_{28} = 0.08$		

在置信水平下,1978、1979、1983、1984、1986 和 2006 年的数据判定为异常值。对 1978 年、1979 年的数据,是由于对 u_t 部分的估计时,取初始值 $u_0 = 0$,使得估计偏差较大,可以不用考虑它的异常性。

对剩余 4 个判定的异常值做统计意义上的显著性检验,得出 1984 和 2006 年的数据在统计意义上具有显著性。出现异常的原因有很多,比如自然灾害,成都市居民的收入,人口的流动,或者生活用水价格的调整等等。2006 年生活用水量出现异常就受到当年发生百年不遇特大旱灾的影响。

3 结 语

本文运用于交互验证的数据质量评估方法,对成都市生活用水量数据做出评估,结果表明:1978 – 2011 年的生活用水量数据中有异常值,直接用该数据对未来几年的用水量作预测,其可信度不能完全令人信服。因此,对该数据中的异常值可以作一定的处理,用处理后的数据再作预测,能够提高生活用水的预测精度。

基于交互验证的数据质量评估方法是辨识数据异常值的有效方法,该方法充分利用数据所包含的信息,在最小均方误差准则下,建立最优的交互验证模型,对数据质量作评估,使得到的结果更加可靠,更加合理,也更符合实际情况。

参考文献:

[1] 常 宁. IMF 的数据质量评估框架及启示[J]. 统计研究,2004(1):27 – 30.

[2] 姜忠波. 浅谈统计数据质量评估[J]. 科技资讯,2006(18):240.

[3] 许涤龙,叶少波. 统计数据质量评估方法研究述评[J]. 统计与信息论坛,2011,26(7):3 – 14.

[4] 卢二坡,黄炳艺. 基于稳健 MM 估计的统计数据质量评估方法[J]. 统计研究,2010,27(12):16 – 22.

[5] 高起蛟,严凤斌,池 斌. 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 – 173.

[6] 李庭辉,薛丽娜. 基于时间序列匹配性的 GDP 数据质量评估研究[J]. 财经理论与实践,2012,33(1):119 – 123.

[7] 成都市统计局编. 成都统计年鉴 2007[M]. 北京:中国统计出版社,2007.

[8] 刘叶玲,翟建国. 一种改进的检验离群值的方法[J]. 统计与决策,2007(7):139 – 140.

[9] 于 兰. 基于模糊理论的成都市水资源价值评判[J]. 水资源与水工程学报,2007,18(4):79 – 81.

[10] 唐 亭,袁 鹏,凌 旋,等. 生活用水量组合预测模型及其应用[J]. 水电能源科学,2013,31(1):26 – 28.

[11] 刘治学,张 鑫,王颖华. 包头市市区居民生活用水量预测分析[J]. 水资源与水工程学报,2012,23(5):67 – 70.

[12] 章元明,盖钧镒. Logistic 模型的参数估计[J]. 四川畜牧兽医学院学报,1994,8(2):47 – 52.

[13] 戴国俊,王金玉,杨建生,等. 应用统计软件 SPSS 拟合生长曲线方程[J]. 畜牧与兽医,2006,38(9):28 – 30.

[14] 薛连民,邢 浩. 基于 S 型成长曲线模型的软基沉降预测[J]. 科技资讯,2010(14):87.

[15] 刘 洪,黄 燕. 我国统计数据质量的评估方法研究——趋势模拟评估法及其应用[J]. 统计研究,2007,24(8):17 – 21.