

基于 Shapley 值的西安市生活需水量组合预测研究

熊光红^{1,2}

(1. 长安大学 水与发展研究院, 陕西 西安 710054; 2. 西安科技大学 建筑与土木工程学院, 陕西 西安 710054)

摘要: 利用等维灰数递补法、非线性回归、广义回归神经网络等 3 种方法进行了西安市生活需水量预测, 比较了不同方法的预测误差, 然后根据各种方法对总预测值的信息贡献能力形成 Shapley 值组合需水量预测方法, 计算了不同预测方法的 Shapley 值及其组合权重, 形成组合预测需水量模型。预测结果表明: 组合方法误差曲线平缓、平均误差值较小, 具有一定的预测精度, 适用于需水量的中短期预测。

关键词: 生活需水量预测; Shapley 值法; 组合预测模型

中图分类号: TU991.31

文献标识码: A

文章编号: 1672-643X(2013)01-0010-04

Research on combination forecast method in predication of domestic water demand in Xi'an based on shapley value

XIONG Guanghong^{1,2}

(1. Reserch Institute for Watrer and Develpoment, Chang'an University, Xi'an 710054, China;

2. Department of Architecture and Civil Engineering, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: The same dimension gray recurrence prediction, multiple nonlinear regression prediction, GRNN neural network prediction methods were done to get domestic water demand of Xi'an, and then prediction error of different methods was compared. The paper proposect the combination forecast method based on shapely value which equal to the information contribtive ability of predictive value calculated the shapley value and combination weights of different forecast methods, formed combination forecast model for water demand. The result showed that the combination method had a smoothing error curve and a smaller average error value. It has a certain prediction accuracy and is applicable to water demand in the short-term water demand forecast.

Key words: domestic water demand prediction; Shapley value method; combination prediction model

1 概述

我国北方大部分地区均在不同程度上存在水资源短缺问题。作为西部大开发的战略核心城市,西安近年来工农业经济发展迅猛,城市规模不断扩大,种种原因造成城市需水量不断增加,同时由于水资源总量的限制,城市生活需水量短缺的压力愈来愈大。为了缓解“水荒”,政府已先后实施或正在实施“引乾济石”、“引红济石”、“引汉济渭”等一系列大型调水工程,但要从根本上解决这一问题,合理的水资源配置和利用至关重要,而准确的需水量预测是水资源合理规划的重要依据。

本文根据西安市近 20 年来的生活需水量及人口、经济、自然条件等因素的历史数据,运用 3 种方法进行

需水量预测,并进一步分析各种方法对需水量预测值的信息贡献能力,采用 Shapley 值法进行组合预测,获得预测精度较高的未来 10 年西安市生活需水量。

2 西安市生活需水量预测

生活需水量影响因素颇多,各因素之间交互作用关系复杂。每种预测方法从不同侧面考虑了影响需水量的主要因素,但也存在预测信息的缺失。本文首先采用时间序列等维灰色预测、非线性多元回归预测以及人工神经网络预测 3 种方法进行 2011-2020 年生活需水量预测。

2.1 需水量预测结构体系

城市生活需水量不可避免的受到自然环境、社会、经济及工程技术等因素的影响,各影响因素交互

收稿日期:2012-11-01; 修回日期:2012-11-27

基金项目:“111”干旱半干旱地区水文生态及水安全学科创新引智基地项目(B08039); 中国工程院咨询项目:防旱抗旱确保粮食及农村供水安全战略研究

作者简介:熊光红(1976-),女,陕西西安人,博士研究生,研究方向水资源管理。

作用形成了生活需水量的灰色增长型数列。本文对需水量影响因素数据进行了统计分析^[1]并借鉴相关研究成果^[2-3],选取城市人口总量 X_1 、城镇人口数量 X_2 、城镇化率 X_3 、城镇人口年可支配收入 X_4 、城镇人口恩格尔系数 X_5 、年均降雨量 X_6 、年均气温 X_7 等 7 个指标形成需水量预测的系统结构。1990 - 2010 年西安市生活需水量数据见表 1,其他数据统计值由西安市 2011 年统计年鉴及相关资料获得。

表 1 1990 - 2010 年西安市生活需水量					万 m ³
年份	需水量	年份	需水量	年份	需水量
1990	10356	1997	21345.0	2004	27068.2
1991	10364	1998	22530.5	2005	28440.2
1992	10874	1999	23716.0	2006	31214.8
1993	11516	2000	25195.9	2007	35473.3
1994	12108	2001	25445.1	2008	37077.4
1995	16791	2002	25877.3	2009	38094.2
1996	20562	2003	26375.7	2010	39745.1

2.2 灰色 GM(1,1) 预测

需水量历史数据形成增长型灰色数列,忽略数列增长的影响因素,采用灰色理论进行未来发展趋势的预测,对处于增长期的短期需水量预测便捷可行^[4]。该方法通过分析时间序列与用水量历史数据之间关系,视这一关系为不透明的“灰箱”,建立灰色系统模型(1),进而获得未来发展趋势的预测值。

$$X^0(t+1) = -a\left[X^0(1) - \frac{\mu}{a}\right]e^{-at} \tag{1}$$

式中: $X^0(t+1)$ 为预测时间数据; $X^0(1)$ 为已知第一时间数据; μ 、 a 分别为预测待定参数。表 1 中需水量的历史数据满足光滑级比的要求,可用于灰色预测。本文据此建立等维灰色 GM(1,1) 预测模型。灰色预测模型算子的处理和选择对预测精度有较大影响,为比较分析,笔者分别选取原始数据、滑动平均数据、累减数据代入模型。模型数据向量的维度亦影响预测结果,改进地等维灰数递补模型可降低初始数据和未来随机因素对模型的影响,因此在这里采用这一模型。预测结果误差见图 1。

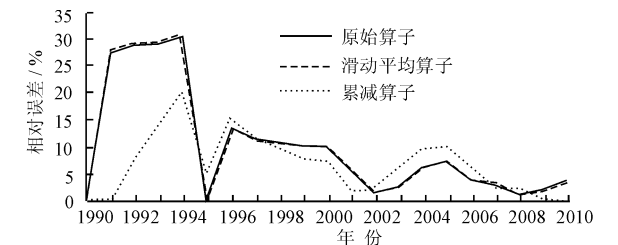


图 1 三种不同算子预测误差图

3 种不同算子的级比光滑因子最大值分别为 1.3868、1.2613、1.2032,预测平均误差分别为 9.949%、9.973%、6.636%,可见光滑因子较小的累减算子作为预测数据具有较高精度,因此采用这一算子作为未来 10 年需水量预测算子。西安市 2011 - 2020 年生活需水量等维灰数递补预测结果见表 6。

2.3 多元非线性回归预测

时间序列需水量预测方法未考虑区域需水量的影响因素及其随机变化,而事实上城市需水量与城市社会经济、工程技术和自然环境形成一个复杂的相互作用的系统。因此研究生活需水系统的组成结构及各因素之间的依赖关系,建立回归预测模型具有一定的合理性。

表 2 主成分分析结果		
主成分	特征值	贡献率
1	4.824	68.91
2	1.090	15.57
3	0.608	8.69

非线性回归分析中,综合考虑各因子与需水量之间的非线性关系,将变量逐个引入,选择合适的回归曲线比较其偏与显著性因子,然后建立各因子曲线与需水量之间的非线性回归模型^[5]。城市生活需水量的影响因素很多,各因素与需水量之间呈现复杂的非线性关系,因子之间也具有一定的相关性,因子的选取对预测精度的影响较大。本文首先将上述城市需水量系统结构中全部因子均纳入回归模型进行预测。其次考虑因子之间的相关性,取其主成分建立非线性回归模型。主成分分析结果见表 2,分析结果显示前两个主成分的贡献率达到 84.5%,因此回归模型取前两个主成分即可。最后根据主成分中最大载荷因子的分析结果表 3,建立主要因子的非线性回归模型,表 2 中主成分 1 对 X_1 (人口总量)有很大载荷,主成分 2 对 X_4 (人均可支配收入)有很大载荷,因此选取这两个主要因子。3 种回归模型的均方根误差依次 0.948、0.982、0.991,预测值平均误差分别为: 9.324%、5.651%、3.444%。由此可见,主因素非线性回归模型预测精度较高,其非线性回归方程见式(2)。

$$X = 103338.65 + 5.153 \times 10^{-5} X_1^3 + 76378974.98 \frac{1}{X_1} + 8.743 \times 10^{-9} X_2^3 - 41231.2 \ln X_2 + 1052.27 X_2^{0.58} \tag{2}$$

根据西安市“十二五”发展规划确立的人口和经济发展目标,将相关数据带入回归方程可得到未

来十年西安市生活需水量的预测数据见表 6。

表 3 主成分载荷

因子	主成分		因子	主成分	
	1	2		1	2
X_1	0.988	0.053	X_5	-0.776	0.170
X_2	0.896	0.084	X_6	0.219	0.504
X_3	0.838	0.089	X_7	0.679	-0.473
X_4	0.772	0.922			

2.4 人工神经网络预测

人工神经网络是目前进行需水量预测较为主流的方法,该方法摒弃了人为因素的影响,根据需水量及其影响因素历史数据的特点进行学习,从而产生权值矩阵和阈值矩阵,适应了需水量的复杂性、非线性和不确定性^[6-7]。笔者选用收敛速度较快的广义回归神经网络(GRNN) 建立预测模型^[8-9],在选择预测影响因素时,参考主成分分析结果,以第 1 和第 2 主成分数据作为网络的输入变量。在进行网络训练中,把样本分为两部分,第一部分 1990 - 2007 年数据作为学习样本,确定网络参数,以 2008 - 2010 年数据作为训练样本,对调整好的样本进行训练。

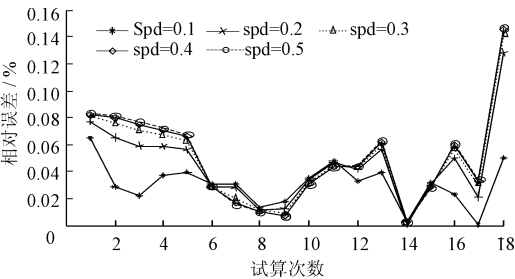


图 2 网络逼近误差

图 2 为采用不同光滑因子(spd) 学习的误差比较图。当 spd=0.1 时预测误差最小,因此在进行预测中光滑因子值取 0.1。利用训练好的网络进行未来 10 年西安市生活需水量预测的结果见表 6。

3 生活需水量的 Shapley 值组合预测

3.1 Shapley 方法

Shapley 值法是一种解决多人合作利益分配的数学方法。借鉴这一理论,不同预测方法获得的需水量结果均对实际需水量及其预测值有一定贡献,而贡献的程度又与这一方法的预测误差有关,因此根据各方法预测结果的误差,求解 Shapley 值,以其作为组合预测权值,建立组合需水量预测模型,能够综合多方信息获得较为准确的预测结果^[10-11]。

3.2 各种预测方法 Shapley 值的求解

各种方法的平均预测误差计算式:

$$E_i = \frac{1}{n} \sum_{j=1}^n E_{ij} \quad (j = 1, 2, \cdots, n) \quad (3)$$

计算结果见表 4,组合模型子集的误差计算结果见表 5。

表 4 不同模型需水量预测误差表 %

年份	模型 1 误差	模型 2 误差	模型 3 误差	组合预 测误差
1990	0	6.854231	4.707493	5.026973
1991	0	2.357898	4.178683	1.076965
1992	7.92	10.78724	0.787027	3.034515
1993	13.96	2.510297	2.916632	0.813112
1994	20.01	19.24600	5.458463	7.376511
1995	4.97	0.160992	2.668981	0.655345
1996	15.38	0.184793	2.329806	2.969960
1997	11.62	0.030858	2.813009	2.710661
1998	9.69	1.108757	1.655655	0.757150
1999	7.87	2.255089	2.593810	3.065109
2000	7.24	6.476419	0.483220	3.663651
2001	2.09	3.943320	0.412258	1.622329
2002	2.31	0.449459	1.130252	0.46185
2003	6.38	1.812850	0.346071	1.628716
2004	9.59	7.110357	0.952118	4.414041
2005	10.01	1.512011	0.367229	1.579739
2006	5.51	1.300226	2.675993	2.450358
2007	2.45	2.784059	2.096897	0.384077
2008	2.12	0.500728	6.069746	2.493244
2009	0.23	0.705062	4.426894	2.398388
2010	0.01	0.150836	3.535418	1.649283
平均	6.636	3.444000	2.505000	2.392000

表 5 组合模型子集的误差

序号	子集	误差	序号	子集	误差
1	E(1)	6.636	5	E(1,3)	4.571
2	E(2)	3.444	6	E(2,3)	2.975
3	E(3)	2.505	7	E(1,2,3)	4.195
4	E(1,2)	5.040			

模型 i 的 Shapley 值由式(4)计算。

$$E'_i = \sum_{i \in s(i)} w(|s|) [E(s) - E(s \setminus i)] \quad (i = 1, 2, \cdots, n) \quad (4)$$

式中: E'_i 为第 i 个预测模型的 Shapley 值; s 为包含预测模型的 i 个组合; $|s|$ 为组合中预测模型的个数; $(s \setminus i)$ 表示在组合中去除模型 i 。 $w(|s|)$ 为加权因子:

$$w(|s|) = \frac{(n - |s|)! \cdot (|s| - 1)!}{n!} \quad (5)$$

3 种方法的 Shapley 值分别为 3.229、0.815、0.1306,3 个模型 Shapley 值相加 $E'_1 + E'_2 + E'_3 = 4.196$, 等于 3 种模型组合误差,这说明各种预测方法分摊的

误差之和等于总误差。将各预测模型的 Shapley 值代入权重计算公式:

$$\lambda_i = \frac{1}{n-1} \cdot \frac{\bar{E} - E'_i}{\bar{E}} \quad (i = 1, 2, \cdots, n) \quad (6)$$

求得 3 种预测方法的权重分别为: $\lambda_1 = 0.115$, $\lambda_2 = 0.401$, $\lambda_3 = 0.484$ 。基于 Shapley 值的生活需水量组合预测模型为:

$$X = 0.115X_1 + 0.401X_2 + 0.484X_3 \quad (7)$$

式中: X_1 、 X_2 、 X_3 分别为灰色预测模型、主因素非线性回归模型、主成分 GRNN 预测模型所获得的生活需水量值,组合预测模型预测需水量误差见表 4 第 5 列,平均误差为 2.392。各种预测方法的需水量误差绘制成曲线,如图 3 所示。

模型 1 与模型 2 前期预测的误差较大,后期预

测误差趋缓,模型 3 总体预测误差较小,但在后期预测误差有不断增大的趋势。组合预测误差绝对值较小,误差曲线趋势平缓,因此采用组合预测模型进行未来需水量预测,一方面可以降低预测模型的总体误差,还能改善神经网络预测模型在未来预测误差加大的趋势。用组合需水量预测模型,式(8)获得需水量数据,如表 6 所示。

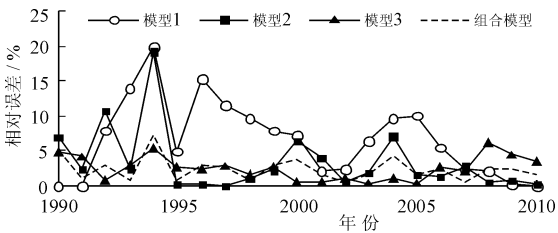


图 3 不同预测模型的误差比较图

表 6 组合预测结果 万 m³

模 型	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
需水量	1	41499	43280	45085	46913	48766	50643	52545	54471	56424
	2	45325	59839	68062	73479	79743	84059	86396	89556	91308
	3	43460	46049	48417	49345	52015	53895	54673	54905	55323
	组合	43982	51260	55911	58743	62761	65617	67149	68750	71688

说明:表中模型 1:等维灰色 GM(1,1), 预测模型、模型 2:非线性回归模型, 模型 3 为 GRNN 网络预测模型。

4 结 语

本文运用几种方法对西安市生活需水量进行预测,并就其预测误差进行了比较分析。同时考虑各种方法对需水量的信息贡献能力,运用 Shapley 值法建立了组合预测模型,预测了西安市 2011 - 2020 年生活需水量,分析结果表明:①各种需水量预测方法均存在其优点和不足之处,因此综合考虑各种方法对需水量的信息贡献程度,以 Shapley 值作为组合权重的组合预测方法,能够有效克服需水量预测过程中随机因素的影响及预测病态方程的出现。② Shapley 值法获得需水量总体误差变化曲线平稳,平均误差较小,具有一定的预测精度,对于中短期需水量预测具有一定的优势。③本文在计算组合权重时仅考虑了各种预测方法对预测总误差的分担能力,以其作为不同方法对需水量的信息贡献能力,若能考虑预测不同方法的研究偏好,及所用数据与需水量之间的相关性,对组合权重进行修正,将能进一步提高预测精度。

参考文献:

[1] 李安贵,张志安. 模糊数学及其应用[M]. 北京:冶金工

业出版社, 2005.
[2] 沈大军,陈雯,罗健萍. 城镇居民生活用水的计量经济学分析与应用实例[J]. 水利学报,2006,37(5):593 - 597.
[3] 张 灵,陈晓宏,刘青娥. 基于 IEA 的需水预测投影寻踪模型研究[J]. 灌溉排水学报,2008,27(1):73 - 76.
[4] 王 涛,张宏伟,牛志广. GM 组合模型用于城市生活用水量预测[J]. 水资源保护,2007,23(5):28 - 30.
[5] 杜发兴,徐 刚. 偏最小二乘回归模型在城市需水预测中的应用[J]. 水力发电,2008,34(6):20 - 23.
[6] 乔维德. 基于遗传神经网络的城市用水量预测研究[J]. 水科学与工程,2007(3):1 - 3.
[7] 俞亭超,张土乔,毛根海,等. 预测城市用水量的人工神经网络模型研究[J]. 浙江大学学报(工学版),2004,38(9):1156 - 1161.
[8] 张娅莉,喇果彦. GRNN 神经网络在信息分析预测中的应用[J]. 数据采集与处理,2009,24(S1):100 - 103.
[9] 崔纪委,鄧志红,刘力鹏. 基于 PSO - GRNN 方法的城市用水量短期预测[J]. 人民长江,2011,42(21):56 - 59.
[10] 朱晓曦,张 潜. 基于 Shapley 值的组合预测方法在福建省农业总产值预测中的应用[J]. 安徽农业科学,2010,38(9):4419 - 4421.
[11] 戴建华,薛恒新. 基于 Shapley 值法的动态联盟伙伴企业利益分配策略[J]. 中国管理科学,2004,12(4):33 - 36.