

# 基于 SVR 的城市供水管网余氯预测分析\*

田一梅, 吴迷芳, 王 阳

(天津大学 环境科学与工程学院, 天津 300072)

**摘要:**支持向量机回归(Support Vector Regression SVR)算法是结构风险最小化原理在函数回归方面的应用。根据北方某城市供水管网余氯的人工采样数据,建立了基于SVR的余氯预测模型,并与人工神经网络、多元线性回归方法进行比较分析,结果表明:在有限样本情况下,SVR模型具有良好的泛化推广能力,各监测点模型预测平均相对误差为1.80%~8.73%,并可获得全局最优解,达到了实用要求,较好地解决了以往管网余氯小样本预测时,常常出现拟合精度高、预测效果较差的问题。

**关键词:**支持向量回归;供水管网;余氯;预测;模型

**中图分类号:**TU991.3 **文献标识码:**A **文章编号:**1006-7329(2006)02-0074-05

## Prediction and Analyses of Residual Chlorine Based on Support Vector Regression in Urban Water Distribution System

TIAN Yi - mei, WU Mi - fang, Wang Yang

(The College of Environmental Science and Engineering, Tianjin University, Tianjin 300072, P. R. China)

**Abstract:** Support vector regression (SVR) algorithm is an application of structural risk minimization principle in function regression. In this paper, a residual chlorine prediction model based on SVR is established by using the data of manual sampling residual chlorine of water distribution system in a certain city in the north of China. SVR model is compared with the artificial neural network and multivariate linear regression. The result shows that SVR model has better generalization ability for small samples, the predicted average relative error of all monitoring points is 1.80% ~ 8.73%, and can achieve unique and globally optimal solutions. It is practical and can solve the problem for small samples of residual chlorine when the fit precision of model is good but the predicted effect is worse.

**Keywords:** Support Vector Regression; water distribution system; residual chlorine; prediction; modeling

城市供水管网中的余氯是一种非稳定的物质,在管网输配中随时间而衰减,当余氯值衰减到一定程度,其杀菌能力降低,引起细菌滋生,水质恶化。对管网余氯进行预测是及时采取控制措施、防治管网水质二次污染的有效方法。

近几年,一些统计方法被应用于管网余氯预测,如人工神经网络、多元线性回归等,这些方法都需要大量的数据资料,但在实际问题中,由于大多城市的供水管网均未安装余氯在线监测装置或在线监测点数量很少,为建立余氯预测模型常常采用短期人工取样。由于人工取样获得的余氯样本数据很有限,因此,现有的预测方法难以取得理想的效果。

支持向量机(Support Vector Machine SVM)是20世纪90年代中期由Vapnik等人针对模式识别问题提

出的一种小样本通用学习算法<sup>[1,2]</sup>。该算法是建立在统计学习理论上,借助最优化方法解决机器学习问题的新工具。近年来,因其在理论研究和算法实现方面都取得了突破性进展,而表现出许多优于已有方法的性能。目前,该算法不仅成功地处理了模式识别、判别分析等分类问题,而且随着Vapnik对 $\epsilon$ -不敏感损失函数的引入,SVM已推广到解决非线性回归估计问题,并开始用于预测预报、建模与控制等领域<sup>[3,4]</sup>。

将SVM理论用于城市供水管网水质预测,根据管网余氯人工监测数据,建立了基于支持向量回归机(Support Vector Regression SVR)的管网余氯预测模型,并与人工神经网络模型、多元线性回归模型进行了比较分析。

\* 收稿日期:2005-12-30

基金项目:国家“十五”科技攻关项目(2002AA601120),天津市科委社发项目(033113111)

作者简介:田一梅(1959-),女,天津市人,副教授,硕士,主要从事给排水系统优化方面的研究。

## 1 支持向量回归机

支持向量回归机的理论基础可描述为:给定某训练样本,根据统计学习理论,要使回归函数的实际输出与理想输出之间的偏差尽可能小,应遵循结构风险最小化原则,而不是传统的经验风险(训练误差)最小化原则。目前,常用的支持向量回归算法有两种: $\varepsilon$ -支持向量回归机和 $\nu$ -支持向量回归机<sup>[5,6]</sup>。

### 1.1 支持向量回归机( $\varepsilon$ -SVR)

用线性回归函数 $f(x) = wx + b$ 拟合数据 $\{(x_i, y_i), i = 1, 2, \dots, n\}$ ,  $x_i \in R^d, y_i \in R, n$ 为样本数。

基于支持向量的最优回归函数是指满足结构风险最小化原理,即最小化

$$\min \frac{1}{2} \|w\|^2 + C \cdot R_{\text{emp}}[f] \quad (1)$$

其中 $C$ 是预先指定的常数, $R_{\text{emp}}[f]$ 是训练误差,可以采用 $\varepsilon$ -不敏感损失函数来度量:

$$|y - f(x)|_{\varepsilon} = \max\{0, |y - f(x)| - \varepsilon\} \quad (2)$$

这里 $\varepsilon$ 是事先给定的一个正数, $\varepsilon$ 规定了回归函数在样本数据上的误差要求。该 $\varepsilon$ -不敏感损失函数可看作在回归函数 $f(x)$ 周围构成的带子。则式(1)可表示为:

$$\min \frac{1}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n |y_i - f(x_i)|_{\varepsilon} \quad (3)$$

当 $|y_i - f(x_i)| \leq \varepsilon, i = 1, 2, \dots, n$ 时,式(3)等价于式(4),即训练误差作为优化问题的约束条件:

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 \\ & \text{s. t. } \begin{cases} y_i - f(x_i) \leq \varepsilon \\ f(x_i) - y_i \leq \varepsilon \end{cases} \quad i = 1, 2, \dots, n \end{aligned} \quad (4)$$

而当 $|y_i - f(x_i)| > \varepsilon, i = 1, 2, \dots, n$ 不能满足时,即约束条件不可实现时,引入松弛变量 $\xi_i, \xi_i^*$ 则式(4)变为:

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ & \text{s. t. } \begin{cases} y_i - f(x_i) \leq \varepsilon + \xi_i \\ f(x_i) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \quad (5)$$

式(5)中目标函数的第一项是为了控制回归函数的复杂程度,从而提高模型的推广能力,第二项则为减少训练误差,常数 $C$ 决定着两者之间的平衡,其值的大小控制着对超出误差 $\varepsilon$ 的样本点的惩罚程度。这是一个凸二次规划问题,利用Lagrange优化方法将上述问题转化为其对偶问题:

$$\max - \frac{1}{2} \sum_{i,j=1}^n (a_i - a_i^*)(a_j - a_j^*)(x_i \cdot x_j)$$

$$\begin{aligned} & + \sum_{i=1}^n (a_i - a_i^*)y_i - \varepsilon \sum_{i=1}^n (a_i + a_i^*) \\ & \text{s. t. } \begin{cases} \sum_{i=1}^n (a_i - a_i^*) = 0 \\ 0 \leq a_i, a_i^* \leq C \end{cases} \quad i = 1, 2, \dots, n \end{aligned} \quad (6)$$

解此最大值问题可以得到:

$$f(x) = \sum_{i=1}^n (a_i - a_i^*)(x_i \cdot x) + b \quad (7)$$

优化计算得到多数样本的 $a_i = a_i^* = 0$ ,对应的 $x_i$ 称为非支持向量,而 $a_i, a_i^*$ 不为零对应的样本就是支持向量,它们通常是位于回归函数的 $\varepsilon$ -带上或 $\varepsilon$ -带外的样本点。因此,SVR就是在样本集中选择在函数变化比较剧烈的位置上的样本数据(支持向量)进行回归估计,而且, $\varepsilon$ 的大小决定了支持向量的数目,并影响到函数估计的精度和模型的泛化能力。

由于上面式子中只涉及内积运算,因此,只要用核函数 $K(x_i, y_i)$ 替代式(6)、(7)中的内积运算,就可以实现非线性函数拟合。

### 1.2 $\nu$ -支持向量回归机( $\nu$ -SVR)

在 $\varepsilon$ -SVR中,需要事先确定 $\varepsilon$ -不敏感损失函数中的参数 $\varepsilon$ ,并通过 $\varepsilon$ 控制回归估计的精度。然而,在某些情况下选择合适的 $\varepsilon$ 以实现高精度估计却不是一件容易的事情。为此,人们提出了一种能够自动计算 $\varepsilon$ 的新方法: $\nu$ -SVR<sup>[7]</sup>作为 $\varepsilon$ -SVR的一种变形。在 $\nu$ -SVR中 $\varepsilon$ 不是作为参数而是作为优化问题的变量出现, $\nu$ 是新引入的反映错误样本点(即 $\varepsilon$ -带外的样本点)个数和支持向量个数的参数,且 $\nu$ 的取值范围限定在 $0 \leq \nu \leq 1$ ,从而使SVR的参数调节得以简化。

在 $\nu$ -SVR中,优化目标为:

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 + C \left( \nu \varepsilon + \frac{1}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \right) \\ & \text{s. t. } \begin{cases} y_i - f(x_i) \leq \varepsilon + \xi_i \\ f(x_i) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \\ \varepsilon > 0 \end{cases} \quad i = 1, 2, \dots, n \end{aligned} \quad (8)$$

得到对偶问题后,由优化充要条件(Karush-Kuhn-Tucker, KKT),在最优点满足:

$$\sum_{i=1}^n (a_i + a_i^*) = C\nu \quad (9)$$

设支持向量个数为 $p$ ,错误样本个数 $q$ ,由于错误样本点均满足 $a_i^* = \frac{C}{n}, a_i = 0$ 或 $a_i = \frac{C}{n}, a_i^* = 0$ ,故有 $\frac{q}{n} \leq \nu$ ,即 $\nu$ 是错误样本的个数所占总样本点数的份额的上界;而对于支持向量,因其满足 $a_i^* \leq \frac{C}{n}, a_i = 0$ 或

$a_i \leq \frac{C}{n}, a_i^* = 0$ , 故由式(9)有  $C\nu = \sum_{i=1}^n (a_i + a_i^*) \leq p \frac{C}{n}$ ,

因此有  $\nu \leq \frac{p}{n}$ , 即  $\nu$  是支持向量的个数所占总样本点数的份额的下界, 因此  $\nu$  取值范围为:

$$\frac{q}{n} \leq \nu \leq \frac{p}{n} \quad (10)$$

式(10)为  $\nu$  值的选取提供了一个依据, 且用  $\nu$  可控制支持向量个数或错误样本点个数。从这个意义上说,  $\nu$ -SVR 优于  $\varepsilon$ -SVR。

与  $\varepsilon$ -SVR 相同, 当引进适当的核函数  $K(x_i, x_j)$  就可以推广为一般的  $\nu$ -SVR。

## 2 基于支持向量回归机的管网余氯预测模型

管网余氯在管网输配过程中因各种因素而逐渐衰减, 特别是当水输配到下游或管网末梢时, 若水中余氯过低则易引发水中细菌学指标超标, 造成饮用水质的微生物风险。因此建立管网余氯预测模型, 其目的是根据管网上游监测点的水质监测数据, 对下游或末梢水中余氯进行准确预测, 如发现水质异常, 可及时采取相应措施, 保证供水卫生安全。

为研究管网余氯变化规律, 选择北方某城市供水管网中的局部管网作为实验区。该区占地面积 150 万  $m^2$ , 用水人口约 5 万人。区内管网中大多为灰口铸铁管, 管径 DN400 ~ DN100, 水源来自城市南部的一座地表水水厂, 水厂采用氯胺消毒。本课题对该实验区管网水质进行了大量的监测分析, 并建立了基于支持向量回归机的余氯预测模型。

### 2.1 余氯影响因素的选择与数据获取

管网水中余氯的变化受多种因素的影响, 主要包括出厂水余氯浓度、氯消毒方式、水在管网中的输配时间、管道内部腐蚀程度以及水质状况如浊度、pH、微生物等。在考虑这些因素对余氯预测模型的影响时应遵循两方面原则, 一是合理制定管网水取样方案, 如通过管网监测点布局、各监测点取样频率即可反映出上述部分因素对余氯的影响; 二要综合考虑项目检测分析的复杂程度、分析时间、所需仪器条件以及模型的实用性而选择有代表性的测试项目, 像水中细菌总数虽然是余氯的主要影响因素, 但由于细菌总数检测的延迟性(通常细菌总数需先培养 24 h 后测定), 而不能作为余氯实时预测模型的影响变量。

根据上述原则, 制定了实验区管网数据获取方案。首先结合实验区管网结构、水流方向及现场查勘, 选择布置了 8 个监测点进行人工取样。其中实验区管网入口处 3 个点, 管网中途 3 个点, 管网末梢 2 个点。再通

过对管网不同工况的水力分析, 计算出管网中任意上下游两个监测点间的水力停留时间, 并以此作为该两点的取样时间间隔, 按照每小时一次的频率连续进行 3 d 实测, 其测试项目均选择能用便携仪器现场即可测定的项目包括: 温度、pH 值、浊度、溶解氧、游离余氯和总余氯, 每个监测点获得 30 组实验数据。通过对实验数据各监测指标相关性分析, 最终选择各预测点上游的部分监测点的总氯为影响变量建立了单因子余氯预测模型, 选择各预测点上游的部分监测点的总氯、浊度、温度和 pH 值为影响变量, 建立了多因子余氯预测模型。

### 2.2 SVR 余氯预测模型的建立与预测结果

由于各监测指标量纲不同、数据的数量级不同, 直接用原始数据建模, 必将突出温度、pH 值的影响, 而减弱余氯、浊度的作用。因此, 需按式(11)对实验数据进行标准化处理, 以消除各影响变量因量纲和单位不同对模型的影响。

$$x'_{ij} = \frac{2(x_{ij} - \min\{x_{ij}\})}{\max\{x_{ij}\} - \min\{x_{ij}\}} - 1 \quad (11)$$

式中:  $x'_{ij}$  分别为第  $i$  监测点第  $j$  项指标标准化前后的变量。

建立了管网中途和末梢共 5 个监测点余氯预测的支持向量回归模型。每个预测点均以其上游监测点  $t$  时刻的水质作为输入变量(单因子模型为余氯, 多因子模型为余氯、浊度、温度和 pH 值四个参数), 以预测点  $t+1$  时刻( $t \sim t+1$  为两点间水力停留时间)的余氯为输出变量, 分别采用各点实验数据的 25 组数据作为训练集建立模型, 5 组数据作为测试集进行模型验证。

建立 SVR 模型时, 首先需要解决 SVR 选型、核函数及模型参数的选择问题。按前文所述为简化参数调节, SVR 选型以  $\nu$ -SVR 模型为主。而核函数的选择则需对各种核函数的性能进行分析对比。尽管理论上讲只要满足 Mercer 条件的函数都可选为核函数, 但对于具体的预测问题, 由不同核函数得到的预测结果将有很大不同。通过对 5 个点模型的反复调试及交互检验, 发现核函数中高阶多项式的性能最优, 其次是径向基核函数, 最后是 Sigmoid 核函数和线性核函数, 故最终选用多项式  $K(x_i, y_i) = (0.2x_i \cdot x_j)^3$ , 参数  $\nu$  为 0.4 ~ 0.6, 建立了 5 个预测点基于 SVR 的余氯预测模型。图 1 为模型建立与预测流程, 表 1 为各预测点余氯预测结果。

表 1 计算出各预测点单因子、多因子余氯预测模型的预测误差, 总体来讲多因子预测误差较之单因子要小, 而管网中输配时间长、从多个方向来水的预测误差要大。前者验证了管网水中余氯的变化受多种因素

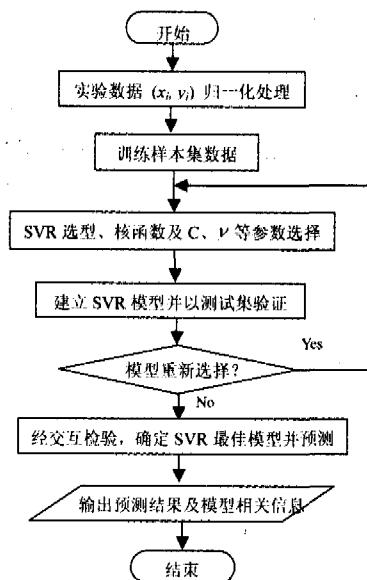


图1 SVR模型建立与预测流程

影响,后者则反映出管网末梢或多点供水管段中余氯的变化受随机因素的影响更大,特别是沿途用户用水量的随机变化,导致水在管网中停留时间不同,从而影响到余氯的衰减,并加大了预测误差。虽然从模型实用来讲,上述单因子、多因子余氯预测模型均已能满足日常管理对水质预报的要求,但从提高模型预测精度(特别是管网末梢)考虑,应选择采用多因子余氯预测模型。

表1 各预测点余氯预测结果

预测点序号	预测点位置	实测均值/ mg · L <sup>-1</sup>	预测均值 /mg · L <sup>-1</sup>		预测平均绝对 相对误差/%		备注
			单因子	多因子	单因子	多因子	
1	末梢	0.50	0.463	0.495	7.82	5.56	距实验区入口最近末梢点
2	中途	0.56	0.598	0.597	6.84	6.68	—
3	中途	0.32	0.339	0.334	13.92	5.16	从两个方向来水
4	中途	0.25	0.251	0.249	9.33	1.80	从两个方向来水
5	末梢	0.05	0.058	0.054	16.75	8.73	实验区最远端末梢点

2.3 SVR 预测与神经网络、多元回归预测对比分析

为了对比 SVR 与神经网络、多元回归方法的总体预测效果,采用相同数据样本建立了各点多因子余氯预测的人工神经网络和多元线性回归模型,三种方法预测结果见表2。

表2 不同预测方法预测结果比较

预测点序号	平均绝对相对误差/%		
	支持向量回归机	人工神经网络	多元线性回归
1	5.56	9.70	6.85
2	6.68	6.56	11.54
3	5.16	9.72	19.43
4	1.80	6.90	21.03
5	8.73	18.79	27.16

表2显示出了5个预测点模型中,SVR预测效果最好,而多元线性回归方法的预测结果较差,其误差较之SVR方法增加了23%~1068%。鉴于SVR方法与神经网络方法在模型构建上有更多相似之处,故以预测点1为例,就两种方法的拟合、预测结果进行对比分析,见表3、表4。

表3 预测点1拟合结果比较

相对误差的绝对值/%	误差分布/%	
	支持向量机	神经网络
0~5	40	64
5~10	12	16
10~33.43	48	20
绝对平均值/%	10.41	4.71

表4 预测点1预测结果比较

检验数据 序号	实测值/ mg · L <sup>-1</sup>	支持向量回归机		人工神经网络	
		预测值/ mg · L <sup>-1</sup>	相对误差/ %	预测值/ mg · L <sup>-1</sup>	相对误差/ %
1	0.54	0.491	-9.07	0.63	16.89
2	0.46	0.477	3.70	0.45	-2.27
3	0.47	0.503	7.02	0.46	-2.63
4	0.49	0.506	2.16	0.56	14.85
5	0.53	0.499	5.85	0.47	-11.84
绝对平均值			5.56		9.70

由表3、表4可以看出,神经网络的拟合精度较高,但预测精度却有所下降,特别是预测误差大于10%的比例较大;而SVR方法的拟合效果虽不及人工神经网络方法,但SVR的预测结果却好于神经网络的预测值,预测的相对误差都在10%以内,且SVR的预测精度高于其拟合精度,说明SVR方法具有较强的推广泛化能力,这正是SVR优越于神经网络方法的关键所在。

由于人工神经网络方法是建立在对客观事物进行大量试验和观测的基础上,当观测样本不够充分时,拟合样本难以将所有事件特征涵盖在内,而且神经网络方法存在着追求训练误差最小,对有限训练数据具有很强的学习记忆能力、构建的模型过于精细等问题,因此容易出现拟合精度高而预测效果不好的所谓过学习现象。而SVR在模型的复杂程度与有限样本的适应性之间反复协调,从训练样本中选择有代表性的、在函数变化比较剧烈的位置上的样本作为支持向量构建模型,显然此种方法构建的模型与传统的用所有训练数据构建模型比较,虽然其训练样本的拟合精度容易受损,但却提高了模型的预测外推能力。如本例中的预测点1,在25个样本数据中选择了10个大都位于波峰或波谷位置上的样本作为支持向量,见图2中较大的菱形点,构造ν-SVR模型。10个支持向量的拟合误差平均值为18.97%,而另15个非支持向量的拟合误差平均值为4.7%,总平均拟合误差为10.41%,拟

合精度可谓不高,但按照 10 个支持向量构造的  $\nu$ -SVR 模型获得的平均预测误差仅为 5.56%。由此看出,模型设计合理是模型推广应用的关键。

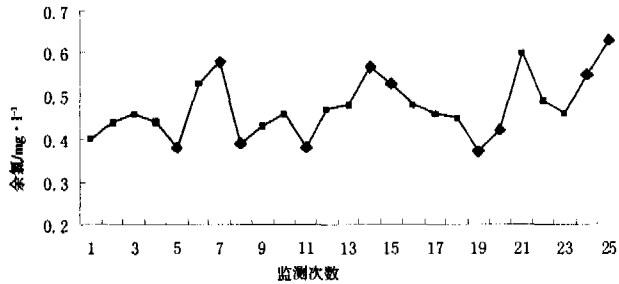


图2 拟合样本集中的支持向量

此外,在建模过程中对不同的初始参数,SVR 每次训练都可以得到相同的结果,而神经网络在神经元数等网络结构给定相同的情况下,每次的训练结果都会有些差异,这说明在 SVR 算法中由于对数据的训练就相当于解一个有线性约束的二次规划问题,因此能得到全局最优解,解决了在神经网络方法中无法避免的局部极值问题。

#### 4 结语

管网余氯是反映管网水质的重要指标,及时准确的余氯预测对管网日常运行管理和制定突发水质事故的应急措施具有重要意义。建立的基于支持向量回归机的余氯预测模型,经多个预测点验证,具有较强的函数表达能力、泛化能力和学习效率,并以较高预测精度达到了实用要求。较好地解决了管网余氯小样本预测时,传统算法中拟合精度高、预测效果仍较差的问题。

由于 SVR 算法能够利用有限的样本信息构建最

佳模型,非常适合当前管网余氯人工取样的小样本预测,即使将来管网安装了余氯在线装置,有足够多的样本数据,仍然可以用 SVR 进行余氯的建模预测,因为 SVR 所具有的严密的理论体系、全局最优解和良好的泛化能力等优越性能是其他算法难以比拟的,而且由于 SVR 方法只取支持向量作为训练样本,则在大样本数 SVR 余氯建模中,大大减少了有用的样本数据,同时可将最新监测数据纳入样本,采用增量学习法 (Incremental Learning) 通过优化计算<sup>[8]</sup>建立预测模型,从而实现供水管网余氯的在线预报。

#### 参考文献:

- [1] Vapnik V N. Statistical learning theory [M]. New York, 1998.
- [2] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
- [3] 马笑箭, 柴毅. 基于支持向量机的故障过程趋势预测研究[J]. 系统仿真学报, 2002, 14(11): 1548-1551.
- [4] 王定成, 方廷健. 支持向量机回归在线建模与应用[J]. 控制与决策, 2003, 18(1): 89-91.
- [5] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机 [M]. 北京: 科学出版社, 2004.
- [6] 杜树新, 吴铁军. 用于回归估计的支持向量机方法[J]. 系统仿真学报, 2003, 15(11): 1580-1585.
- [7] Schölkopf B, Smola A J, Bartlett P L. New Support Vector Algorithms [J]. Neural Computation, 2000, 12: 1207-1245.
- [8] Carozza M, Rampone S. Towards an incremental SVM for regression [A]. Proc of the IEEE - ENNS Int Joint Conf on Neural Networks [C]. 2000, 6: 405-410.

## 更正启事

《重庆建筑大学学报》2006年28卷第1期第54页图5后第1段至参考文献之前应更正为:

次弯矩有多大,是弯矩调幅设计中值得关心的问题。根据较多的工程实践和试验分析<sup>[5~9]</sup>,在符合相对受压区高度  $\xi \leq 0.3$  和总调幅量控制在 20% 以内的情况下,一般荷载弯矩调幅多为荷载弯矩的 10% 以上,等效荷载次弯矩大体为 5% 或更大一些,轴力次弯矩仅为 1% ~ 2%。……,可以取此类框架节点的总调幅量不超过 20%。……,总调幅量控制在 15% ~ 20% 之间;……,总调幅量控制在 10% ~ 15% 之间。

#### 4 小结

预应力混凝土框架弯矩调幅设计除遵循截面相对

受压区高度  $\xi \leq 0.3$  及总调幅量控制在 20% 以内两点以外,还需注意以下各点:

……

3) 次弯矩在预应力框架梁弯矩调幅设计中起着重要作用。框架梁弯矩调幅考虑次弯矩影响的总调幅量,当次弯矩全部起有利作用时不超过 20%;当等效荷载次弯矩起有利作用而轴力次弯矩起不利作用时控制在 10% ~ 20% 之间;当等效荷载次弯矩起不利作用时或等效荷载次弯矩和轴力次弯矩都起不利作用时控制在 10% ~ 15% 之间。