Journal of Southeast University (English Edition) Vol. 21, No. 3, pp. 359 – 362 Sept. 2005 ISSN 1003—7985

Application of cluster analysis and stepwise regression in predicting the traffic volume of lanes

Zhang He[1] Wang Wei[1] Gu Huaizhong[2]

([1] College of Transportation, Southeast University, Nanjing 210096, China)

([2] Research Institute of Nanjing Public Security Traffic Science and Technology, Nanjing 210001, China)

Abstract: Because of the difficulty to obtain the traffic flow information of lanes at non-detector intersections in most metropolises of the world, based on the relationships between the lanes of signal-controlled intersections, cluster analysis and stepwise regression are integrated to predict the traffic volume of lanes at non-detector isolated controlled intersections. First cluster analysis is used to cluster the lanes of non-detector isolated signal-controlled intersections and the lanes of all signal-controlled intersections with detectors. Then, by the results of cluster analysis, the traffic volume samples are selected randomly and stepwise regression is used to predict the traffic volume of lanes at non-detector isolated signal-controlled intersections. The method is tested by the traffic volume data of lanes of the road network of Nanjing city. The problem of predicting the traffic volume of lanes at non-detector isolated signal-controlled intersections was resolved and can be widely used in urban traffic flow guidance and urban traffic control in cities without enough intersections equipped with detectors.

Key words: intelligent transportation systems (ITS); cluster analysis; stepwise regression

With ever fast progress of the world's economy, the traffic volume of the world's metropolises is increasing quickly, and the consequence of this is severe traffic jams. To some degree it restricts people from going out. In the USA, the loss in monetary terms to traffic jam and delay is estimated to be more than 100 billion US dollars[1]. If these problems are not solved, they will continue to constitute an obstacle to social progress. To effectively improve the management of a city's traffic, every country applies modern science and technology, such as traffic control systems, traffic flow guidance systems, intelligent transportation systems etc. As road networks develop in most metropolises around the world there are also increased numbers of dots within the roads network (Here dots refer to signal-controlled intersections.) The modern management of road networks in a city must be based on the traffic flow information. So it is important to study the traffic flow information of signal-controlled intersections. But at present the data of the traffic volume is mainly obtained from the detectors at signal-controlled intersections. In most metropolises not all signal-controlled intersections are equipped with detectors. It should be observed that within the respective metropolis, the number of the signal-controlled intersections with detectors is less than ten percent of all signal-controlled intersections within the city. And there are lots of non-detector signal-controlled intersections. So the traffic flow information for non-detector signal-controlled intersections cannot be obtained easily, especially the traffic volume of lanes at non-detector isolated signal-controlled intersections (because the isolated signal-controlled system is the basis of urban traffic control system). It is very important to predict the traffic volume of lanes at non-detector isolated signal-controlled intersections. In order to obtain the traffic volume of lanes at non-detector isolated signal-controlled intersections, we must study the relative relationships between the non-detector isolated signal-controlled intersections and the signal-controlled intersections with detectors, and predict the traffic volume of lanes at non-detector isolated signal-controlled intersections by using the traffic flow volumes of lanes at the signal-controlled intersections with detectors. In this paper we will introduce the integration of cluster analysis and stepwise regression to predict the traffic volume of lanes at non-detector isolated signal-controlled intersections.

1 Basic Idea

We will predict the traffic volume of lanes at non-detector isolated signal-controlled intersections. Although not all signal-controlled intersections of metropolises have been equipped with detectors and the number of these intersections is limited, there are numerous lanes at these intersections. If these lanes are equipped with detectors, the number of the detectors will become very large. So in this paper we will utilize cluster anal-

Received 2004-11-02.

Foundation item: The National Natural Science Foundation of China (No. 50378016).

Biographies: Zhang He(1971—), male, doctor; Wang Wei(corresponding author), male, doctor, professor, wangwei@seu.edu.cn.

ysis to cluster the lanes of non-detector isolated signal-controlled intersections and the lanes of isolated signal-controlled intersections with detectors, and then we will utilize stepwise regression to predict the traffic volume of lanes at non-detector isolated signal-controlled intersections[2-6]. The purpose is to simplify the variables of the equation and keep the equation stable.

2 Computational Procedures

2.1 Clustering procedure

In this paper the variable is an isolated signal-controlled intersection, and the sample is traffic volume. We will use the similarity coefficient method to cluster the variables. The basic idea[7-9] is that if the characteristics of two variables are close to each other, their similarity coefficient will be close to 1 (or -1), otherwise their similarity coefficient will be close to 0. So the more relevant variables will be clustered to one category and the less relevant ones will be clustered to different categories.

The similarity coefficient is defined as follows: If C_{ij} represents the similarity coefficient between the variable y_i and the variable y_j, then C_{ij} must be satisfied with the following relations:

$$C_{ij} = \pm 1 \Leftrightarrow y_i = a y_j \qquad a \neq 0,\ a \text{ is constant}$$
$$|C_{ij}| \leqslant 1 \qquad \text{for all } i \text{ and } j$$
$$C_{ij} = C_{ji} \qquad \text{for all } i \text{ and } j$$

There are many kinds of similarity coefficient. In this paper we choose the relativity coefficient as the similarity coefficient.

In this paper we adopt the maximum similarity coefficient as the cluster criterion. That is, we must cluster two kinds of intersections with the maximum similarity coefficient to one category. The similarity coefficient $\boldsymbol{R}_{\mathrm{pu}}$ between category G_{p} and category G_{u} is

$$\boldsymbol{R}_{\mathrm{pu}} = \max_{q_i \in G_{\mathrm{p}},\, q_j \in G_{\mathrm{u}}} r_{ij} \tag{1}$$

The concrete procedure is as follows:

① Calculate the similarity coefficient among all variables and get a similarity coefficient matrix $\boldsymbol{R}_{(0)}$; therefore each variable becomes its own kind.

② Find the maximum similarity coefficient in the matrix $\boldsymbol{R}_{(0)}$ except for the main diagonal line and define it as $\boldsymbol{R}_{\mathrm{pu}}$. Then cluster category G_{p} and category G_{u} to one category. We define it as G_{s}, then $G_{\mathrm{s}} = \{G_{\mathrm{p}}, G_{\mathrm{u}}\}$.

③ Calculate the similarity coefficient between the new category and the other category.

$$\boldsymbol{R}_{\mathrm{sk}} = \max_{q_i \in G_{\mathrm{s}},\, q_j \in G_{\mathrm{k}}} r_{ij} = \max\{\max_{q_i \in G_{\mathrm{p}},\, q_j \in G_{\mathrm{k}}} r_{ij},\ \max_{q_i \in G_{\mathrm{u}},\, q_j \in G_{\mathrm{k}}} r_{ij}\} = \max\{\boldsymbol{R}_{\mathrm{pk}}, \boldsymbol{R}_{\mathrm{uk}}\} \tag{2}$$

④ We get the new similarity coefficient matrix and define it as $\boldsymbol{R}_{(1)}$.

⑤ For $\boldsymbol{R}_{(1)}$, we repeat the same step for $\boldsymbol{R}_{(0)}$, then we get the new similarity matrix $\boldsymbol{R}_{(2)}$. Repeat the same for $\boldsymbol{R}_{(2)}$, then we get the new similarity matrix $\boldsymbol{R}_{(3)}$, and so on until all the kinds of variables will be clustered to one kind.

In this paper we will use the data gotten from thirty-three lanes detectors of the different isolated signal-controlled intersections in Nanjing city on August 6th, 2004. The detectors in Tab. 1 and Fig. 1 are shown as follows: 1—ccm1, 2ccm2, 3—ccm3, 4—ccm4, 5—ccm5, 6—ccm6, 7—ccm7, 8—ccm8 (ccm1 to ccm8 represent the names of the detectors which are located in Caochangmen Avenue.); 9—jy1, 10—jy2, 11—jy3, 12—jy4, 13—jy5 (jy1 to jy5 represent the names of the detectors which are located in Jianye Avenue.); 14—lxy1, 15—lxy2, 16—lxy3, 17—lxy4, 18—lxy5, 19—lxy6, 20—lxy7, 21—lxy8 (lxy1 to lxy8 represent the names of the detectors which are located in Luxiying Avenue.); 22—ezs1, 23—ezs2, 24—ezs3, 25—ezs4, 26—ezs5, 27—ezs6 (ezs1 to ezs6 represent the names of the detectors which are located in East Zhongshan Road.); 28—zy1, 29—zy2, 30—zy3, 31—zy4, 32—zy5, 33—zy6 (zy1 to zy8 represent the names of the detectors which are located in Zhongyang Road.).

The cluster analysis results are shown in Tab. 1 and Fig. 1.

Tab. 1 Parameters of system cluster

No.	Similarity coefficient	Detector number		No.	Similarity coefficient	Detector number		No.	Similarity coefficient	Detector number	
		Row	Column			Row	Column			Row	Column
1	0.992 6	4	3	12	0.930 3	5	2	23	0.685 3	33	27
2	0.990 4	17	16	13	0.929 7	13	12	24	0.676 3	31	29
3	0.987 9	19	18	14	0.924 5	6	2	25	0.675 4	8	1
4	0.987 3	20	18	15	0.910 8	31	28	26	0.641 1	31	11
5	0.983 7	21	18	16	0.898 4	8	2	27	0.597 8	33	14
6	0.981 6	17	15	17	0.889 0	8	7	28	0.597 4	22	11
7	0.971 8	26	24	18	0.8774 4	30	23	29	0.499 5	10	1
8	0.970 3	21	15	19	0.849 8	30	29	30	0.344 1	10	9
9	0.967 7	4	2	20	0.812 5	32	29	31	0.342 3	22	9
10	0.954 1	25	23	21	0.742 7	21	14	32	0.295 4	22	14
11	0.934 3	26	23	22	0.708 7	13	11				

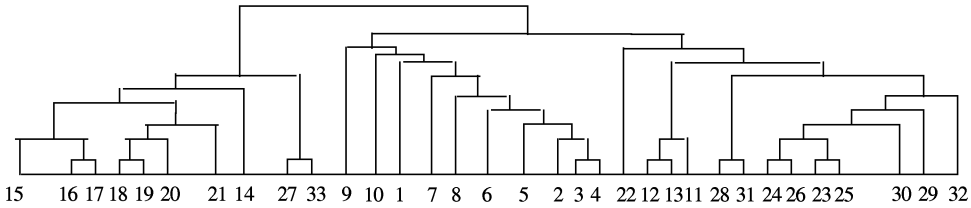

Fig. 1 Pedigree chart of system cluster

2.2 Stepwise regression procedure

On the basis of cluster analysis, we utilize stepwise regression[7, 9, 10] to predict the traffic volume of lanes at non-detector isolated signal-controlled intersections. Here we select No. 30 detector at Jianye Avenue as the lane of the non-detector isolated signal-controlled intersection at which the traffic volume will be predicted. From Fig. 1, we know that No. 30, No. 29, No. 32, No. 23, No. 24, No. 25 and No. 26 detectors have been clustered into one category. So we should only utilize the traffic volume of the lanes of these detectors to predict the traffic volume of the No. 30 detector. The stepwise regression procedure is as follows:

① Calculate the average of the original matrix

$$\left.\begin{aligned} \bar{x}_i &= \frac{1}{n}\sum_{K=1}^{n} x_{Ki} \\ \bar{y} &= \frac{1}{n}\sum_{K=1}^{n} y_K \end{aligned}\right\} \quad i = 1, 2, \ldots, p \tag{3}$$

② Calculate the deviation-from-average matrix

$$\left.\begin{aligned} S_{ij} &= S_{ji} = \sum_{K=1}^{n} (x_{Ki} - \bar{x}_i)(x_{Kj} - \bar{x}_j) \\ S_{iy} &= S_{yi} = \sum_{K=1}^{n} (x_{Kj} - \bar{x}_i)(y_K - \bar{y}) \\ S_{yy} &= \sum_{K=1}^{n} (y_K - \bar{y})^2 \end{aligned}\right\}$$
$$i, j = 1, 2, \ldots, p \tag{4}$$

③ Calculate relative coefficient matrix

$$r_{ij} = r_{ji} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} \quad i, j = 1, 2, \ldots, p, y \tag{5}$$

The results are shown in Tab. 2.

④ Adhibit (or eliminate) regression variables

We can prove that in the $m+1$ step and the $m+2$ step of stepwise regression the variables that have been adhibited into linear regression equation are not eliminated in the $m+3$ step. So in the first three steps we will repeat the adhiditing procedure. Utilize $\boldsymbol{R}^{(k)}(k=0, 1, 2)$ to calculate the contribution of every variable in the first step.

$$V_i^{(m)} = \frac{(r_{iy}^{(k)})^2}{r_{ij}^{(k)}} \quad i = 1, 2, \ldots, p; k = 0, 1, 2; m = 1, 2, 3 \tag{6}$$

Find the sequence k_1 which makes $V_i^{(1)}$ the maximum, then we have

$$V_{k1}^{(m)} = \max_{1 \leqslant i \leqslant p} \{V_i^{(m)}\} \qquad m = 1, 2, 3 \tag{7}$$

Then calculate

$$F_y = \frac{V_{k1}^{(m)}(n-2)}{(r_{yy}^{(k)} - V_{k1}^{(m)})} \qquad m = 1, 2, 3; k = 0, 1, 2 \tag{8}$$

If $F_y > F_{\alpha 1}$, x_{K1} will be adhibited into linear regression equation and then the relative coefficient matrix $\boldsymbol{R}^{(1)} = (r_{ij}^{(1)})$ is calculated; if $F_y \leqslant F_{\alpha 1}$, stop calculating. In this way, the results are shown as follows: all the adhibited variables are ezs3, ezs6, zyl2; all the eliminated variables are ezs3, zyl2; the variables which are not adhibited into the regression equation are ezs4, ezs5, zyl5. From all of above, only one variable is adhibited into regression. It can make the regression stable and increase the accuracy of the prediction.

⑤ The linear regression equation is as follows:

$$y = (4.134 - 105.43x') \times 10^{-7} \tag{9}$$

The F test value is 14.131 and its theoretical value is 0.5, and the multiple relative coefficient is 0.927 8. Here the value of y is standard value, because the value of x' is standard value. The predicted value of y is shown in Tab. 3.

From Tab. 3, we find there exists error between the original value and the predicted value. The reason is that the variable which has been adhibited into Eq. (9) cannot include all the information for all variables.

Tab. 2 Relative coefficients

Detector	zy3	ezs3	ezs4	ezs5	ezs6	zy2	zy5
zy3	1.000 0	0.927 6	0.971 8	0.145 7	0.790 3	0.805 2	0.899 7
ezs3	0.927 6	1.000 0	0.921 5	0.220 0	0.810 0	0.830 4	0.841 7
ezs4	0.971 8	0.921 5	1.000 0	0.108 3	0.777 4	0.819 0	0.884 4
ezs5	0.145 7	0.220 0	0.108 3	1.000 0	−0.041 3	0.138 4	−0.062 1
ezs6	0.790 3	0.810 0	0.777 4	−0.041 3	1.000 0	0.812 5	0.849 8
zy2	0.805 2	0.830 4	0.819 0	0.138 4	0.812 5	1.000 0	0.697 5
zy5	0.899 7	0.841 7	0.884 4	−0.062 1	0.849 8	0.697 5	1.000 0

Tab. 3 Prediction results

No.	Original value	Predicted value	No.	Original value	Predicted value	No.	Original value	Predicted value
1	28	53. 567 5	9	251	251. 338 2	17	266	232. 511 8
2	107	100. 513 2	10	235	255. 713 4	18	275	206. 850 9
3	68	79. 864 2	11	247	258. 126 5	19	246	231. 931 6
4	219	240. 313 2	12	245	248. 878 9	20	236	219. 128 5
5	208	243. 619 7	13	239	268. 064 6	21	229	212. 045 9
6	235	244. 669 6	14	266	248. 611 9	22	193	186. 458 6
7	225	253. 327 7	15	249	238. 710 5	23	244	229. 859 1
8	259	255. 860 6	16	228	240. 976 4	24	225	222. 057 6

3 Conclusion

These methods based on the relative relationships among the traffic volumes obtained from the detectors of the signal-controlled intersections is adopted to analyze traffic volumes. So the problem of predicting the traffic volume of lanes at the non-detector isolated signal-controlled intersections is solved, which makes the traffic flow guidance system and urban traffic control system in the cities that do not have enough signal-controlled intersections equipped with detectors become possible. Meanwhile it provides the theoretical basis for deciding which are strategic intersection(s) and the macro management of the intersections. With these methods, when there are lots of signal-controlled intersections with detectors, the stepwise regression and the cluster analysis method are chosen to predict the traffic volume of lanes at non-detector isolated signal-controlled intersections. It can reduce the workload, but it has some disadvantages on prediction accuracy.

References

[1] Zhao Yilin. *Vehicle reckoning and position system* [M]. Beijing: Publishing House of Electronics Industry, 1999. (in Chinese)

[2] Wang W, Guo X C. *Traffic engineering* [M]. Nanjing: Southeast University Press, 2000. (in Chinese)

[3] Wang W, Xu J Q, Yang T, et al. *Urban transportation planning theory and application* [M]. Nanjing: Southeast University Press, 1998. (in Chinese)

[4] Ceylan H, Bell M G H. Traffic signal timing optimization based on genetic algorithm approach, including drivers' routing [J]. *Transportation Research*, *Part B*, 2004, **38**(4): 329 - 343.

[5] Luo J Y, Xing Y. *Analysis method of economic statistic and prediction* [M]. Beijing: Tsinghua University Press, 1987. (in Chinese)

[6] Loo Hong K, Chow Andy H F. Control strategies for oversaturated traffic [J]. *Transportation Engineering*, 2004, **130**(4): 466 - 479.

[7] Zhang He, Wang Wei. Research on the method used in predicting the traffic volume of non-detector single-signal-controlled intersections [A]. In: *Proceedings of Urban Transport* 2005 [C]. Algarve, Portuga, 2005. 12 - 14.

[8] Wang D H, Yang Z S, Zhang H. Cluster analysis used in dealing with traffic flow information [A]. In: *Proceedings of IEEE IVEC'*99 [C]. Changchun, 1999, **1**: 114 - 116.

[9] Jiang Rui, Wu Qingsong, Zhu Zuojin. A new continuum model for traffic flow and numerical tests [J]. *Transportation Research*, *Part B*, 2002, **36**(5): 405 - 419.

[10] Yang Zhaosheng, Zhang He, Li Juan. Step-by-step regression used in traffic flow prediction of non-detector road intersections [J]. *Journal of Jilin University* (*Engineering and Technology Edition*), 2002, **32**(4): 19 - 25. (in Chinese)

聚类分析和逐步回归法在车道流量预测中的综合应用

张 赫[1] 王 炜[1] 顾怀中[2]

([1] 东南大学交通学院,南京 210096)

([2] 南京市公安交通管理局交通科学研究所,南京 210001)

摘要:针对目前国内外大中城市中普遍存在的无检测器信号交叉口车道交通流信息难于获取的情况,基于信号控制交叉口车道之间的相关性,综合应用聚类分析和逐步回归法预测单点无检测器信号控制交叉口车道流量. 首先应用聚类分析将单点无检测器信号控制交叉口的车道与有检测器信号控制交叉口的车道交通流量进行聚类,然后在聚类分析结果的基础上随机选取车道交通流量样本运用逐步回归法预测单点无检测器信号控制交叉口的车道流量,此方法经过南京市的具体车道流量数据验证. 此类问题的解决,可广泛应用于城市交通流诱导系统以及交通控制系统.

关键词:智能运输系统;聚类分析;逐步回归

中图分类号:U491. 1^{+}4