智能交通系统数据抽样方法研究

朱小锋

摘要：智能交通系统在现代交通领域具有美好的应用前景，是应该大力发展的方向。而数据管理特别是数据的抽样技术，在 ITS 研究中起着举足轻重的作用。本文结合国外抽样技术的应用实践以及数理统计的相关理论，提出了智能交通中数据管理的思想及数据抽样模型。

关键词：智能交通，数据管理，抽样模型

一、抽样的目的以及质量控制

按照传统的方法，有两种方式可供对 ITS 数据进行存储：一个是只存储集成后的 ITS 数据，但是假如日后我们有了新的需求，那么这些数据有可能就无法充分满足需求；另外就是存储尽可能多的原始 ITS 数据，但由于数据量庞大，对于数据管理工作以及日后的科学研究都将带来很大的不便。

为了解决这一对矛盾，我们最好选择一些最具有代表意义的 ITS 数据，譬如特定的路段数据、高峰小时的数据等等。这就是我们所说的数据抽样，它不仅能满足用户的需求，方便科学研究，还能节省相当的存储空间。

另外一方面，数据样本的质量控制问题也不容忽视，包括两个方面的内容：一方面是对采集的原始数据进行质量控制，保证从数据采集源点传来的数据的完整、正确；另一方面是对经过抽样、压缩和存储处理后的输出数据进行质量控制，保证输出数据的真实有效。本论文主要涉及的是针对原始数据的质量控制问题。

二、数据抽样的基本思路

根据本文的设想，数据抽样应该按照这样几个步骤：首先我们将获得的原始数据进行分组，每组代表每星期中特定的一天。然后，我们将这些数据按照本文所建立的数学模型进行处理。当新的数据得到后，我们再对它们进行滚动迭代，直到满足用户的需求为止。

三、抽样模型的建立

原始数据包括时间、采集器编号和数量、每个采集器采集到的交通流量、占有率、速度，根据这些数据来推断一定时间跨度（如：一个月，一个季度，甚至一年）内的有规律的某一天或某几天的数据性质（例如：周一、每个月的第一天等等），或者说确定最佳的抽样日。

模型中变量定义如下：

t -- 样本；

n_t -- 经过修补后的每天的总样本数；

w -- 一星期中的一天（$w = 1,2,...,n_w$）；

n_w -- 一周中的天数 (一般为 7)；

k -- 周（第几周）；

n_k -- 取样的总周数；

v_t^{wk} -- 在第 k 周中的第 w 天的交通流量；

o_t^{wk} --在第 k 周中的第 w 天的占有率；

s_t^{wk} --在第 k 周中的第 w 天的速度；

d -- 采集器序号；

n_d -- 采集器总数；

v_{td}^{wk} -- d 采集器在第 k 周中的第 w 天的交通流量；

o_{td}^{wk} -- d 采集器在第 k 周中的第 w 天的占有率；

s_{td}^{wk} -- d 采集器在第 k 周中的第 w 天的速度；

$v_{t_r d}^{wk}$ -- d 采集器在第 k 周中的第 w 天的交通流量原始数据；

$o_{t_r d}^{wk}$ -- d 采集器在第 k 周中的第 w 天的占有率原始数据；

$s_{t_r d}^{wk}$ -- d 采集器在第 k 周中的第 w 天的速度原始数据；

t_r -- 原始数据样本；

第一步： 数据拟合或内插

原始数据: $v_{t_r d}^{wk}$， $o_{t_r d}^{wk}$， $s_{t_r d}^{wk}$

$$v_{td}^{wk} = F_v^{wk}(v_{t_r d}^{wk}) \qquad (1\text{-}1)$$

$$o_{td}^{wk} = F_o^{wk}(o_{t_r d}^{wk}) \qquad (1\text{-}2)$$

$$s_{td}^{wk} = F_s^{wk}(s_{t_r d}^{wk}) \qquad (1\text{-}3)$$

$F_v^{wk}, F_o^{wk}, F_s^{wk}$ 是内插函数

第二步：对采集器求和

$$v_t^{wk} = \sum_{d=1}^{n_d} v_{td}^{wk}$$

$$\forall\ t = 1,2,...,n_t;\ w = 1,2,...,n_w;\ k = 1,2,...,n_k \qquad (1\text{-}4)$$

$$o_t^{wk} = \sum_{d=1}^{n_d} o_{td}^{wk}$$

$$\forall\ t = 1,2,...,n_t;\ w = 1,2,...,n_w;\ k = 1,2,...,n_k \qquad (1\text{-}5)$$

$$s_t^{wk} = \sum_{d=1}^{n_d} s_{td}^{wk}$$

$$\forall\ t = 1,2,...,n_t;\ w = 1,2,...,n_w;\ k = 1,2,...,n_k \qquad (1\text{-}6)$$

第三步：对 k 取均值

$$\bar{v}_t^w = \frac{1}{n_k}\sum_{k=1}^{n_k} v_t^{wk}$$

$$\forall\ t = 1,2,...,n_t;\ w = 1,2,...,n_w \qquad (1\text{-}7)$$

$$\bar{o}_t^w = \frac{1}{n_k}\sum_{k=1}^{n_k} o_t^{wk}$$

$$\forall\ t = 1,2,...,n_t;\ w = 1,2,...,n_w \qquad (1\text{-}8)$$

$$\bar{s}_t^w = \frac{1}{n_k}\sum_{k=1}^{n_k} s_t^{wk}$$

$$\forall\ t = 1,2,...,n_t;\ w = 1,2,...,n_w \qquad (1\text{-}9)$$

第四步： 误差平方和 SSE

$$SSE^{wk} = \alpha_v \sum_{t=1}^{n_t}(v_t^{wk} - \bar{v}_t^w)^2 + \alpha_o \sum_{t=1}^{n_t}(o_t^{wk} - \bar{o}_t^w)^2 + \alpha_s \sum_{t=1}^{n_t}(s_t^{wk} - \bar{s}_t^w)^2$$

$$\forall\ w = 1,2,...,n_w;\ k = 1,2,...,n_k$$

其中 $\alpha_v + \alpha_o + \alpha_s = 1$

$$0 \le \alpha_v \le 1 \quad 0 \le \alpha_o \le 1 \quad 0 \le \alpha_s \le 1 \qquad (1\text{-}10)$$

第五步：优化 k_w^{opt}：

$$k_w^{opt} = \arg\min SSE^{wk},$$

$$= \arg\min_{k} \quad [\quad \alpha_v \sum_{t=1}^{n_t} (v_t^{wk} - \bar{v}_t^{w})^2 + \alpha_o \sum_{t=1}^{n_t} (o_t^{wk} - \bar{o}_t^{w})^2 + \alpha_s \sum_{t=1}^{n_t} (s_t^{wk} - \bar{s}_t^{w})^2 \quad]$$

$$\forall \ w = 1,2,...,n_w \qquad (1\text{-}11)$$

四、抽样模型的分析

为了进一步细化和充实本文所建立的数学模型，按照模型的五个步骤具体阐述如下：

第一步： 数据拟合

在取样过程中（即得到原始数据的过程）难免会产生误差，因此，必须对原始数据拟合。本文采用 $F_v^{wk}, F_o^{wk}, F_s^{wk}$ 表示拟合函数，而没有直接采用具体的函数，这是因为数据拟合有多种方法可供选择，譬如分布拟合检验、一元线性回归、多元线性回归等，这就要求我们根据需要选择适当的函数。具体的拟合方法是：对原始数据通过预处理检索，定位出奇异数据和所缺少的数据，然后根据历史同期数据统计分布规律，将差值补齐。如果不能得到历史同期数据，则可以用数学方法，例如求平均值、插值拟合等获得新的数据。

第二步：对采集器求和

由第一步我们得到了比较合理的原始数据，接着对这些原始数据进行处理，即求出一天内的交通量、速度、车道占有率的总和，进行数据的集成，集成度为一天。(注：原始数据是每隔固定时间探测器测得的数据。)

由这一步，我们可以得到一天的交通量、速度、占有率的数据。

第三步：对 k 取均值

对 k 取均值，即对交通量、速度、车道占有率求均值，相当于求数学期望。当面对庞大的数据，并且要找到一个比较能说明实际情况的数据的时候，求数学期望往往是一种很好的方法，而且这种方法在实际中得到了广泛的应用。

第四步： 误差平方和 **SSE**

尽管均值已经比较符合要求了，但是，有必要研究一下这些随机变量与其均值的偏离程度，以便找到更好的方法来优化均值。在这里，用每天的数据值与均值的差的平方来表示，这种方法类似求方差的过程。由于涉及到交通量、速度、占有率三个变量，所以，给出的公式的数学内涵是求三维随机变量协方差。系数可以理解为权重，系数值的确定应该结合实际，由专家分析的经验值来确定。当然，根据实际需要，也可以由研究者自己决定。

分析误差平方和 SSE 的目的是为了优化样本均值，因此，接下来研究的是怎样对数据进行优化的问题。

第五步：优化

与数据拟合、抽样相似，优化也存在多种方法。例如：平分法、黄金分割法、牛顿法、二次规划法、线性加权和法、极大极小值法等。

本文采用的是极小值法，即算出误差平方和 SSE 的最小值，把这个最小值作为最优化条件，并加到均值上去，得到的即是最佳抽样日的数据。

到此，整个抽样过程的模型设计结束。根据这种抽样方法，我们可以做整个路网的数据抽样优化，也可以做某个路段或区域的数据抽样优化。

五、结束语

智能交通系统是 21 世纪现代化交通运输体系的发展方向，其数据管理的好坏，直接关系到 ITS 的发展前景。论文从 ITS 数据管理出发，对其抽样方法进行了初步的探讨。分析了抽样基本思路，并建立了抽样模型，实践证明，该模型切实有效。当然，在实际的运用中，这个模型需要不断的完善。

（作者工作单位：北方交通大学）

参考文献：

[1] 陆化普，史其信 ITS-新一代道路交通系统.公路交通科技，1998：33-35

[2] 史其信 21 世纪的智能交通系统.交通世界，2001，3：62-65

[3] 骤盛，谢式千 概率论与数理统计.高等教育出版社，1999：141-145

[4] 陆化普 智能运输系统 人民交通出版社 2002.1

[5] 于雷 智能数据抽样系统讲义 2002.6

诱发交通量产生机理的研究

关宏志　邵洁　池洪波

摘　要：本文列举分析了有关诱发交通量的定义，从交通需求的角度给出了诱发交通的定义。详细讨论了诱发交通量的形成的机理。结合四阶段预测法和交通需求预测理论对诱发交通量预测方法进行了初步探讨。

关键词：诱发交通量，四阶段预测方法

1．问题的提出

交通需求量的预测是交通规划的基础。在预测交通需求的四阶段预测法中，人们考虑了交通的发生、吸引，交通量的分布，交通方式划分以及交通量的分配。作为当今最为完善的交通需求预测理论体系，四阶段预测法得到了广泛的应用。

随着人们对交通需求发生机理认识的加深，人们意识到：交通基础设施（例如：道路、桥梁等）的建设、交通服务水平的改变，不仅会改变路网内交通量的分布格局，而且会改变对象地区交通需求的总量。于是，人们提出了诱发交通(induced traffic)[1]的概念。

时至今日，人们已经根据已有的认识提出了多种描述诱发交通的机理，开发了多种模型。然而，人们对诱发交通的认识并未达成一致，人们对于诱发交通的定义及预测方法存在着诸多的不同的理解。这些都直接影响到了诱发交通量的预测结果。

[1] induced traffic 的中文叫法并不统一，有些研究报告称其为新增交通量、新生交通量、诱增交通量等。