文章编号：1000-8608(2020)03-0293-07

基于 Gamma 分布的交通流时间序列分割模型

王本超[1,2]， 李 丹[1]， 秦 攀[1]， 顾 宏*[1]

（1.大连理工大学 控制科学与工程学院，辽宁 大连 116024；
2.辽宁警察学院，辽宁 大连 116036）

摘要： 准确获取交通流量变化点，对后续的交通流预测、分类及多时段控制具有重要意义. 鉴于交通流时间序列的非负性及异方差性，采用 Gamma 分布拟合交通流时间序列，并对其进行有效分割. 针对多元交通流时间序列，首先利用非负主成分分析方法实现降维并提取特征序列，之后利用最大似然估计得到 Gamma 分布参数，通过不同参数的 Gamma 分布拟合特征序列的不同片段，并由赤池信息准则（AIC）确定最优分割边界及分割阶数. 实验结果表明，所建立的分割模型能够反映不同时段的交通流变化，与现有分割方法相比，取得了更好的分割结果.

关键词： 交通流时间序列；Gamma 分布；时间序列分割；非负主成分分析

中图分类号： TP183 **文献标识码：** A **doi：** 10.7511/dllgxb202003010

0 引 言

交通作为城市发展的主要驱动力，对城市生产要素流动和日常生活有着显著影响. 随着车辆的日益增多，提高道路利用效率、缩短出行时间、缓解交通拥堵一直是管理者及科研人员的研究目标. 智慧交通控制是减小交叉口冲突、提高交通运行效率的一种有效途径，其中，多时段控制[1-2]是一种行之有效的常用方法. 依据交叉口流量的变化把一天 24 h 划分为若干时段，针对不同的交通时段采用不同的信号控制方案，交通信号机根据预先设置的时段划分方案自动进行控制方案切换，具有实现方法简单、可靠性高的优点.

对交通流量的有效分段可以提高短时交通流预测精度，优化交通诱导方案，合理化控制信号交替时间等，目前的研究已经取得了一定的进展. 徐琛等[2]用经典自回归滑动平均（autoregressive moving average，ARMA）模型描述交通流时间序列，在模型预测的基础上，通过预测值与真实值的差值是否达到既定阈值来判断是否为分割点. Salamanis 等[3]对交通流时间序列建立时空自回归差值滑动平均模型，使用滑动窗口方法搜索分割点. 然而基于回归的时间序列分割模型计算量较大，并且倾向于对交通流异常检测. Chang 等[4]提出用正态分布来近似表示时间序列从而获取时间序列变化趋势，并通过相邻点与前一段序列均值的最小差值来确定是否为分割点.

在实际应用中，由于邻近路口交通相互影响，具有密切相关性[5]，而反映区域交通状态的变量有很多，为了减少计算量，消除冗余及噪声信息影响，已有研究利用主成分分析（principal component analysis，PCA）方法[6]进行降维. Wang 等[7]与 Wagner-Muns 等[8]通过 PCA 方法对区域交通流量数据降维并进行短时交通流预测. 汤旻安等[9]通过 PCA 方法对兰州市城市交通流特征提取，并对比机动车流量、非机动车流量、引道延误的累积贡献率，总结出地形和车流量是造成拥堵的深层次原因. 李慧等[10]融合 PCA 与回声状态网络（echo state network）算法对交通流序列进行周期预测，并有效避免预测结果延迟.

鉴于交通流序列本身的非负性，取值范围是全体实数的正态分布并不适合描述交通流数据.

收稿日期： 2019-08-08； **修回日期：** 2020-03-31.
基金项目： 国家自然科学基金资助项目(61502074,61633006)；中国博士后科学基金资助项目(2016M591430)；大连理工大学基本科研业务费专项资金资助项目(DUT17RC(4)09).
作者简介： 王本超(1985-)，男，博士生，E-mail：wangbc@mail.dlut.edu.cn；李 丹(1977-)，女，博士，副教授，硕士生导师；秦 攀(1978-)，男，副教授，硕士生导师；顾 宏*(1961-)，男，教授，博士生导师，E-mail：guhong@dlut.edu.cn.

Gamma分布在拟合非负数据中有着广泛的应用，更适于表达交通自由流、拥挤流及间歇流等不同时段的状态[11]. 因此本文采用适合的Gamma分布来拟合交通流量，对系统把握各时段交通流的运动机理，制定信号灯控制方案具有重要意义.

为保留交通流数据非负性、异方差性以及流量浮动不对称性并且达到降维目的，本文引入非负主成分分析(nonnegative principal component analysis，NPCA)方法[12]，并在此基础上提出基于Gamma分布函数交通流时间序列分割方法. 该方法既考虑区域交通流整体分布差异，又将交通不同时段流量分布分割开，可为交通管理系统分时间管理提供理论依据.

1 模型建立

为了建立基于Gamma分布交通流时间序列分割模型，首先考虑原始多元交通流时间序列的非负性和相关性引入NPCA，然后根据不同时段流量差异，通过自顶向下方法，提出不同参数Gamma分布拟合特征序列片段，进而建立基于Gamma分布交通流时间序列分割模型，该模型能反映出不同时段交通流运动机制.

1.1 NPCA交通流时间序列降维

令$\boldsymbol{Y}=(\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_n)$表示多元交通流时间序列，其中$n$表示检测点数量，$\boldsymbol{y}_i=(y_{1i} \quad y_{2i} \quad \cdots \quad y_{Ti})^{\mathrm{T}}$表示第$i$个检测点的一系列观测值，$T$表示采集样本的时间长度. 这里$\boldsymbol{Y}$也可表示为一个$T\times n$矩阵，即

$$\boldsymbol{Y}_{T\times n}=\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{T1} & y_{T2} & \cdots & y_{Tn} \end{pmatrix}$$

考虑到各检测点的交通流量有较强的相关性，对多元交通流时间序列进行降维，能够降低序列的复杂性，进而在低维空间展开分析，即利用$m(m<n)$维时间序列$\boldsymbol{Z}$描述$\boldsymbol{Y}$.

主成分分析方法是一种经典降维方法，其基本思想是，通过线性变换矩阵将高维、相关性较高的数据映射到新的空间，使得变换后的各分量线性无关，提取出主要的分量从而达到降维的目的.

尽管主成分分析方法为目前一种成熟的特征提取算法，但是在对非负数据特征提取过程中，得到的主成分会出现负值，与交通流数据的非负特性不符. NPCA可以克服主成分分析的无方向性[13]，将原序列定向投影到非负空间，被看作是经典PCA的扩展. 通过对每个主成分序列(PC)的所有项施加非负性约束，以得到非负的特征. 这等价于求解如下二次规划问题：

$$\begin{aligned} &\max\ J(\boldsymbol{U})=\frac{1}{2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}\|_F^2 \\ &\text{s.t. } \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I},\ u_{ij}\geqslant 0 \end{aligned} \tag{1}$$

其中$\boldsymbol{U}=(\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \cdots \quad \boldsymbol{u}_n)\in\boldsymbol{R}^{n\times n}$，为非负的系数矩阵，$\boldsymbol{u}_i=(u_{1i} \quad u_{2i} \quad \cdots \quad u_{mi})^{\mathrm{T}}$. 在实际应用中，往往放宽条件，只要求在$\boldsymbol{U}$的每一列中有一个非负的项[14]，即每个PC只包含一个非负值条目，原非负正交条件的二次规划问题可以转化为

$$\max_{\boldsymbol{U}\geqslant\boldsymbol{0}} J(\boldsymbol{U})=\frac{1}{2}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}\|_F^2-\alpha\|\boldsymbol{I}-\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\|_F^2 \tag{2}$$

其中$\alpha\geqslant 0$为惩罚参数，用于控制$\boldsymbol{U}$的每一列标准正交度，以得到非负的特征. 放宽条件后，矩阵$\boldsymbol{U}$为非负近似正交矩阵$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\approx\boldsymbol{I}$. 非线性优化问题可以通过以下梯度下降法求解：

$$\boldsymbol{U}(l+1)=\boldsymbol{U}(l)-\eta(l)\frac{\nabla_{\boldsymbol{U}}J(l)}{\|\nabla_{\boldsymbol{U}}J(l)\|} \tag{3}$$

其中参数$\eta(l)$是第l次迭代步长. 对$\boldsymbol{U}$的目标函数的梯度计算为

$$\nabla_{\boldsymbol{U}}J(l)=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}+4\alpha(\boldsymbol{I}-\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U})\boldsymbol{U}^{\mathrm{T}} \tag{4}$$

本文在迭代中选择步长为1，以避免信任区域搜索的耗时. 通过非负主成分分析以进行有效的交通流模式特征提取，使得在对多元交通流数据降维的情况下，保持原数据的非负特性.

1.2 基于Gamma分布函数的序列分割

考虑交通信号管理需要，将提取的交通流特征序列按不同参数Gamma分布分段，为实现多时段控制提供分割基础. 实际问题中，包括交通流时间序列在内的复杂数据通常难以找到完全准确的函数进行描述. 对于阶数为k的分割问题，通常采用k个简单函数$f_i(\cdot)$，$i=1,2,\cdots,k$(常数、线性函数或概率密度函数)进行近似拟合，并依据拟合程度与模型复杂度得到分段结果，其中概率密度函数为常数及线性函数的一般形式.

通过不同参数Gamma分布与不同分割片段的交通流拟合，拟合函数$f(\cdot)$为Gamma分布概率密度函数. 假设第i段时间序列服从参数为μ_i和σ_i^2的Gamma分布，有

$$f_i(z_t\,|\,\mu_i,\sigma_i^2)=\frac{1}{(\sigma_i^2\mu_i)^{1/\sigma_i^2}}\frac{z_t^{1/\sigma_i^2-1}\mathrm{e}^{-z_t/\sigma_i^2\mu_i}}{\Gamma\left(\frac{1}{\sigma_i^2}\right)}$$

其中$E(z_t)=\mu_i$，$Var(z_t)=\mu_i^2\sigma_i^2$，$t\in[\tau_i,\tau_{i+1})$，进

而刻画不同时段的 Gamma 分布，其对数似然函数为

$$\ln L(\mu_i,\sigma_i^2\mid z_t)=\sum_{t=1}^{n}\ln f(z_t\mid\mu_i,\sigma_i^2)=\sum_{t=1}^{n}\left\{-\frac{1}{\sigma_i^2}\ln\mu_i-\frac{z_t}{\sigma_i^2\mu_i}-\ln\left[\Gamma\left(\frac{1}{\sigma_i^2}\right)\right]+\frac{1}{\sigma_i^2}\ln z_t-\frac{1}{\sigma_i^2}\ln\sigma_i^2-\ln z_t\right\}\tag{5}$$

令 $\theta_i=\{\mu_i,\sigma_i^2\}$ 为模型参数，可通过最大似然估计方法进行求解：

$$\hat{\theta}_i=\arg\max\ln L(\theta_i)\tag{6}$$

为了最大程度上表征交通流内部变化差异，将特征序列分割成若干片段，其最优划分使得各片段内的特征序列具有最高相似性，同时片段间的特征序列具有最大差异性. 拟合程度衡量及分割阶数确定方法可视为模型选择问题，通过定义代价函数 $cost(S)$ 找到同质部分[15]来确定.

设 k 为分割阶数，分割结果为 $S=\{S_1,S_2,\cdots,S_k\}$，$S_i=\{z_{\tau_i+1},z_{\tau_i+2},\cdots,z_{\tau_{i+1}}\}$ 表示第 i 个分割片段；$\boldsymbol{\tau}_{0:k}=(\tau_0\quad\tau_1\quad\cdots\quad\tau_k)$ 为分割点，也称为分割边界或者变异点，满足 $0=\tau_0<\tau_1<\cdots<\tau_k=T$. 本文用不同参数 Gamma 分布概率密度函数 $f_i(\cdot)$ 计算拟合程度，将统计分析中应用广泛的模型选择赤池信息准则(AIC)[16]作为代价函数，即

$$cost(S)=-2\sum_{i=0}^{k-1}\sum_{t=\tau_i+1}^{\tau_{i+1}}\ln(f_i(z_t,\hat{\theta}_i))+2\left|\hat{\Theta}\right|\tag{7}$$

其中 $\hat{\theta}_i$ 为第 i 段 Gamma 分布的参数估计值，$|\hat{\Theta}|$ 为序列分段后整个模型的参数个数.

在实际应用中，交通流分割阶数 k 的选择直接影响 Gamma 分布拟合交通流序列片段的优良性与复杂度. 由于交通流序列具有异方差性，若分割片段较少，将波动范围差异较大的交通流序列作为同分布划分成一个片段，会导致交通管理系统不能准确识别路面交通状态，管理效率偏低；若分割片段过多，Gamma 分布对片段的过拟合，会造成区域交通管理繁杂混乱[5]. AIC 是寻找可以最好地解释数据但包含最少自由参数的模型，从而平衡交通管理效率与复杂度. 因此，本文以最小 AIC 值确定分割阶数 k 以及分割点 $\boldsymbol{\tau}_{0:k}$，使得每个分割片段的数据符合最优的 Gamma 分布.

搜寻分割点的方式主要有 3 种. 滑动窗口(sliding windows)[17-18]方法简单直接，能支持在线分段，但是在很多真实的数据集上分割结果不佳，且不能分割成预定段数. 自顶向下与自底向上[19-21]属于逐阶寻找最优分割点的方法，支持预定分割阶数，实际使用时可根据分割段数及细致程度进行选择. 交通流时间序列分割问题中目标分割阶数较少，自底向上方法对于较长交通流序列的计算量过大，且初始阶段对小样本片段拟合分布函数估计参数精度不高，影响分割准确度，因此，本文选择自顶向下的分割方式.

1.3　分割算法流程

结合基于 Gamma 分布函数序列分割与非负主成分分析方法，实验步骤如下：

输入　多元交通流时间序列 $\boldsymbol{Y}=(\boldsymbol{y}_1\quad\boldsymbol{y}_2\quad\cdots\quad\boldsymbol{y}_n)$，最大分割阶数 $k_{\max}$.

输出　最优分割点 $\boldsymbol{\tau}_{0:k}=(\tau_0\quad\tau_1\quad\cdots\quad\tau_k)$.

步骤 1　对 $\boldsymbol{Y}=(\boldsymbol{y}_1\quad\boldsymbol{y}_2\quad\cdots\quad\boldsymbol{y}_n)$ 非负降维，得到主成分分量 $\boldsymbol{Z}$，各主成分分量贡献率从大到小为 $z_1,z_2,\cdots,z_n$.

步骤 2　对步骤 1 中确定的第一主成分 z_1 计算贡献率，若贡献率不小于 85%，则转步骤 3；否则，终止.

步骤 3　设定初始值 $k=1$ 及最大分割阶数 $k_{\max}$.

步骤 4　对 k 阶主成分序列片段增加 1 个分割点，并通过式(6)计算不同分割点下 $k+1$ 个分割片段的分布函数参数.

步骤 5　对步骤 4 得到的各分布函数计算式(7)并比较代价函数 $cost(S)$，取最小值的分割点为最优分割点 $\boldsymbol{\tau}_{0:k+1}$.

步骤 6　$k=k+1$. 若 $k\leqslant k_{\max}$，重复计算步骤 4 与步骤 5；否则，转步骤 7.

步骤 7　计算并比较不同阶数最优分割的 AIC，得到最优分割阶数及最优分割点.

2　实验与结果分析

本文所有实验均基于 R 语言和 GAMLSS 工具包[22]来实现.

2.1　实验数据

本文研究对象东联路是大连重要的交通枢纽道路，纵向穿越华东路、松江路、西南路、华北路、疏港路等大连市区的多条城市主干道以及哈大高铁专线，全长 11.3 km，路面宽度 24 m .

使用的数据为入市区方向 7 个检测点，采样

间隔为 15 min. 从 2016-01-12T00:15 到 2016-02-09T00:00 共 28 d 的流量(N)数据,将每天同一时间点的 28 个流量做箱线图,如图 1 所示,发现具有明显的白天方差较大而夜晚方差较小的异方差特点,即非平稳特性.

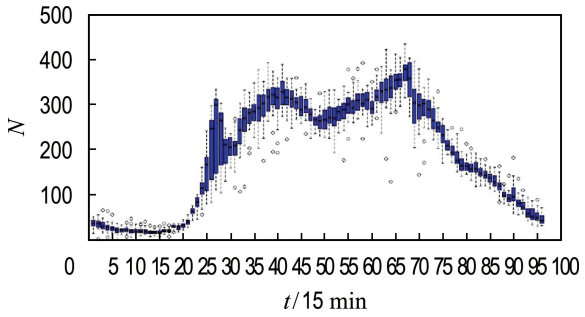

图 1 华北路口 28 d 流量箱线图

Fig. 1 Boxplot of 28-day flow at Huabei junction

选取第 1 d 从 2016-01-12T00:15 到 2016-01-13T00:00 用作序列分割,即序列长度 $T=96$. 各路口的时间序列图如图 2 所示.

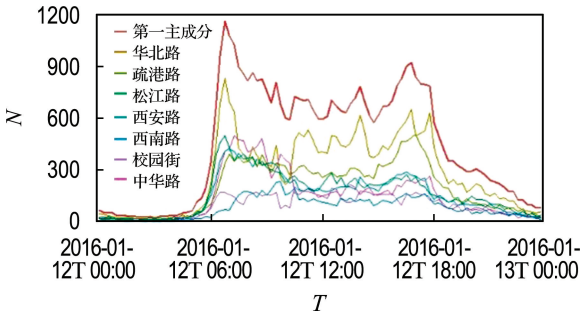

图 2 时间序列图

Fig. 2 Time series figure

2.2 非负主成分提取结果

采用非负主成分分析方法对东联路 7 个路口 1 d 的数据进行非负特征提取,计算各主成分贡献率(R),得到主成分贡献率如图 3 所示. 第一个主成分贡献率达到 87.17%,大于 85%,因此本文选择第一个主成分进行分析. 与多元交通流时间序列对比如图 2 所示.

2.3 最优分割点与分割阶数确定

本文研究分割意义在于将交通流序列按分布分段,指导道路诱导与控制对不同时段的交通状态做出应对,因此设定每段最少 4 个时间点,即 1 h,在分割点选取过程中舍掉间隔小于 4 的情形. 需要指出的是,由于每天 24 h 周而复始,对于将应用到每天循环的多时段控制,并没有真正意义上的起点与终点,因此即使将 24 h 的交通流序列分成两段,也要选取两个分割点.

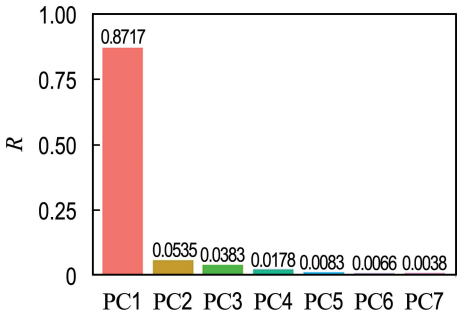

图 3 各非负主成分贡献率

Fig. 3 Contributing rate of each nonnegative principal component

分别计算分割阶数 $k=1,\cdots,6$ 时的代价函数式(7),依据 AIC 取最小值确定最优分割点,并将期望值以虚线形式标记入各分段,如图 4～9 所示.

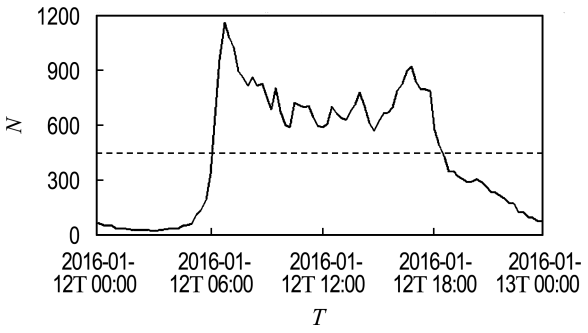

图 4 1 阶分割结果

Fig. 4 Segmentation result in order 1

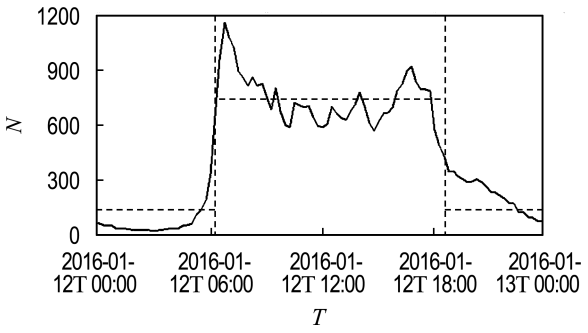

图 5 2 阶分割结果

Fig. 5 Segmentation result in order 2

计算不同阶数最优分割的 AIC,如表 1 所示.

从表 1 得出,当分割阶数 $k=4$ 时,模型 AIC 取最小值,因此对交通流特征序列的最优分割阶数为 4. 从分割片段中可以看出,与大连实际道路

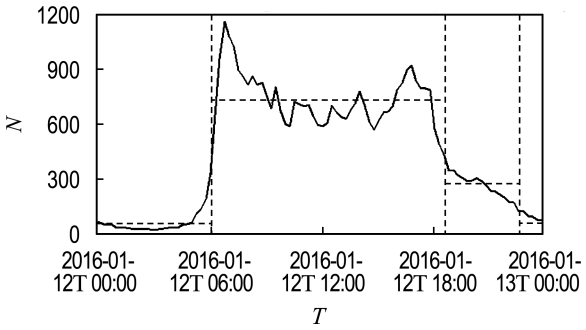

图 6　3 阶分割结果

Fig. 6　Segmentation result in order 3

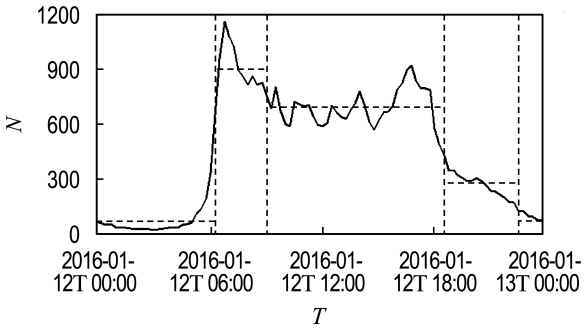

图 7　4 阶分割结果

Fig. 7　Segmentation result in order 4

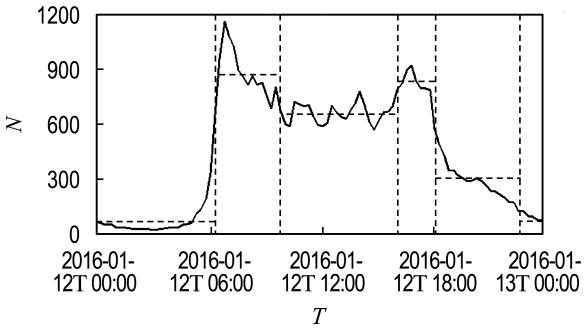

图 8　5 阶分割结果

Fig. 8　Segmentation result in order 5

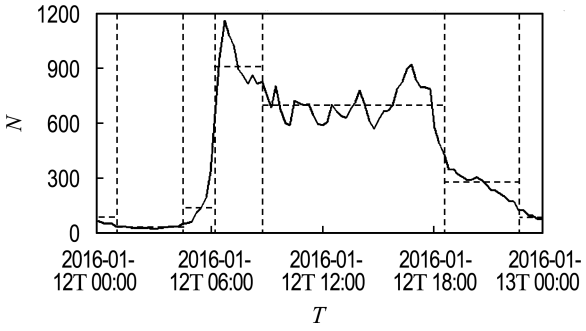

图 9　6 阶分割结果

Fig. 9　Segmentation result in order 6

表 1　不同分割阶数的 AIC

Tab. 1　AIC for different segmented orders

阶数	AIC	阶数	AIC
1 阶	1 006.56	4 阶	894.99
2 阶	973.25	5 阶	906.61
3 阶	916.57	6 阶	934.24

交通的早高峰、晚高峰、白天平峰及夜晚交通流趋势变化相吻合，具有一定实际解释意义.

为了验证模型的有效性，将本文方法与同样 $k=4$ 的基于正态分布[4]以及直线拟合[21]分割作对比，如图 10、11 所示，AIC 数值比较结果如表 2 所示.

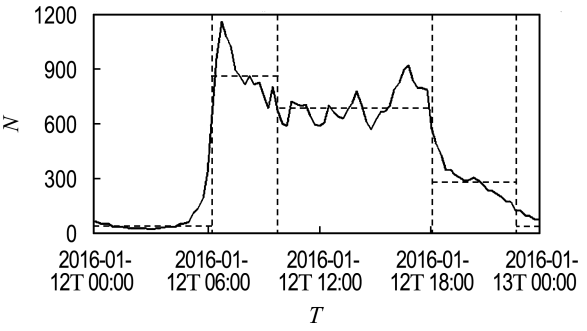

图 10　正态分布 4 阶分割结果

Fig. 10　Segmentation result in order 4 by normal distribution

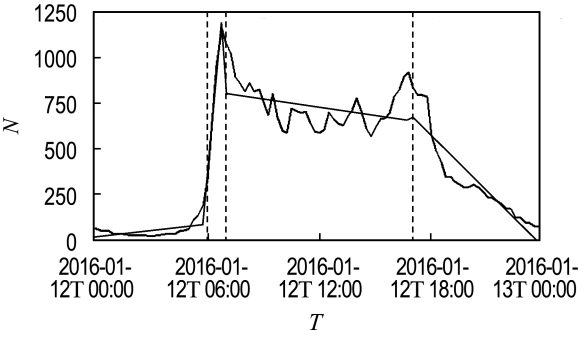

图 11　直线拟合 4 阶分割结果

Fig. 11　Segmentation result in order 4 by linear fitting

表 2　不同分割方法 4 阶分割的 AIC

Tab. 2　AIC of four-order segmentation by different methods

模型	AIC
Gamma 分布	894.99
正态分布	913.24
直线拟合	938.07

从图 10 与图 7 对比可以看出，基于 Gamma 分布与正态分布分割结果相似，仅在一些分割点上有些偏差，但从表 2 可以得出，本文研究的分割方法得到了更优的 AIC 值，表明本文方法对交通流特征序列具有更好的拟合结果；而基于直线拟合得到的分割结果更像是一种趋势，不能准确表达道路交通状态. 通过计算不同分割方法的 AIC 值，得到基于 Gamma 分布时间序列分割方法为最优分割方法.

3 结 语

随着人们生活水平的不断提高，机动车辆快速增长，有效提取道路通行状态并做出应对能够在一定程度上缓解交通压力. 准确掌握不同时段交通状态，为进一步优化信号灯控制、交通诱导以及平衡道路交通占有量提供了依据，具有便利人们出行的重要意义.

实验结果表明，用基于 Gamma 分布的时间序列分割方法能够更好地拟合不同时段交通流量分布. 尽管实验只是针对大连市东联路入市方向，但是本文方法也适用于其他交通流的分割，分段结果可以帮助了解不同路段交通流运动机制，为智能交通管理系统分时段管理交通提供依据.

本文选取的实验数据为一工作日交通流序列，由于节假日与工作日分别对应不同的流量模式，在实际应用中需加以区分，并且本文设定最小分割片段为 1 h 来表征某时段交通流量分布规律，后续工作将对交通流序列异常进行检测，以提高分段及预测模型的鲁棒性.

参考文献：

[1] ZHONG Zhizhen, HUA Nan, TORNATORE M, *et al*. Energy efficiency and blocking reduction for tidal traffic via stateful grooming in IP-over-optical networks [J]. **Journal of Optical Communications and Networking**, 2016, **8**(3): 175-189.

[2] 徐 琛，董德存，欧冬秀. 传感网中数据驱动的多时段控制方法优化研究 [J/OL]. 计算机工程与应用 [2019-07-18]. http://kns.cnki.net/kcms/detail/11.2127.tp.20190716.1537.014.html.
XU Chen, DONG Decun, OU Dongxiu. Time-of-day control optimization of data-driven urban road constant-peak-type intersections in sensor networks [J/OL]. **Computer Engineering and Applications** [2019-07-18]. http://kns.cnki.net/kcms/detail/11.2127.tp.20190716.1537.014.html. (in Chinese)

[3] SALAMANIS A, MELADIANOS P, KEHAGIAS D, *et al*. Evaluating the effect of time series segmentation on STARIMA-based traffic prediction model [C] // **Proceedings - 2015 IEEE 18th International Conference on Intelligent Transportation Systems: Smart Mobility for Safety and Sustainability, ITSC 2015**. Piscataway: IEEE, 2015: 2225-2230.

[4] CHANG Huijun, SHAN Hong, MA Tao. Segmentation, clustering and timing relationship analysis of MANET traffic flow [J]. **Telkomnika**, 2013, **11**(8): 4817-4823.

[5] PUN Lilian, ZHAO Pengxiang, LIU Xintao. A multiple regression approach for traffic flow estimation [J]. **IEEE Access**, 2019, **7**: 35998-36009.

[6] WOLD S, ESBENSEN K, GELADI P. Principal component analysis [J]. **Chemometrics and Intelligent Laboratory Systems**, 1987, **2**(1/2/3): 37-52.

[7] WANG Yitian, JAJA J. Analysis and forecasting for traffic flow data [J]. **Sensors and Materials**, 2019, **31**(6): 2143-2154.

[8] WAGNER-MUNS I M, GUARDIOLA I G, SAMARANAYKE V A, *et al*. A functional data analysis approach to traffic volume forecasting [J]. **IEEE Transactions on Intelligent Transportation Systems**, 2018, **19**(3): 878-888.

[9] 汤旻安，王攀琦. 兰州市交通拥堵研究 [J]. 西北大学学报(自然科学版)，2019，**49**(1)：71-77.
TANG Min′an, WANG Panqi. Research on traffic congestion in Lanzhou City [J]. **Journal of Northwest University (Natural Science Edition)**, 2019, **49**(1): 71-77. (in Chinese)

[10] 李 慧，奚园园，马宇鑫，等. 融合 PCA 和 ESN 的交通流周期预测模型 [J]. 西安电子科技大学学报(自然科学版)，2019，**46**(1)：20-26.
LI Hui, XI Yuanyuan, MA Yuxin, *et al*. Traffic flow cycle prediction based on the PCA-ESN model [J]. **Journal of Xidian University (Natural Science)**, 2019, **46**(1): 20-26. (in Chinese)

[11] 王晓原，张敬磊，马立云. 适应交通流演化的伽马分布形状参数估计 [J]. 计算机工程与应用，2014，**50**(5)：247-251.
WANG Xiaoyuan, ZHANG Jinglei, MA Liyun. Estimation of gamma distribution shape parameter

adapting to traffic flow evolvement [J]. **Computer Engineering and Applications**, 2014, **50**(5): 247-251. (in Chinese)

[12] HAN Xiaoxu. Nonnegative principal component analysis for cancer molecular pattern discovery [J]. **IEEE-ACM Transactions on Computational Biology and Bioinformatics**, 2010, **7**(3): 537-549.

[13] PLUMBLEY M D, OJA E. A "nonnegative PCA" algorithm for independent component analysis [J]. **IEEE Transactions on Neural Networks**, 2004, **15**(1): 66-76.

[14] HAN Xiaoxu, SCAZZERO J A. Protein expression molecular pattern discovery by nonnegative principal component analysis [M] // CHETTY M, NGOM A, AHMAD S, eds. **Pattern Recognition in Bioinformatics**. Berlin: Springer, 2008: 388-399.

[15] HUBERT P. The segmentation procedure as a tool for discrete modeling of hydrometeorological regimes [J]. **Stochastic Environmental Research and Risk Assessment**, 2000, **14**(4/5): 297-304.

[16] AKAIKE H. A new look at the statistical model identification [J]. **IEEE Transactions on Automatic Control**, 1974, **19**(6): 716-723.

[17] DAS G, LIN K I, MANNILA H, *et al*. Rule discovery from time series [C] // **Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, KDD'98**. Palo Alto: American Association for Artificial Intelligence, 1998: 16-22.

[18] KEOGH E, CHU S, HART D, *et al*. An online algorithm for segmenting time series [C] // **Proceedings - 2001 IEEE International Conference on Data Mining, ICDM'01**. Piscataway: IEEE, 2001: 289-296.

[19] KEOGH E, KASETTY S. On the need for time series data mining benchmarks: A survey and empirical demonstration [J]. **Data Mining and Knowledge Discovery**, 2003, **7**(4): 349-371.

[20] BORENSTEIN E, ULLMAN S. Combined top-down/bottom-up segmentation [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2008, **30**(12): 2109-2125.

[21] FIDLER S, MOTTAGHI R, YUILLE A, *et al*. Bottom-up segmentation for top-down detection [C] // **Proceedings - 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013**. Piscataway: IEEE Computer Society, 2013: 3294-3301.

[22] STASINOPOULOS D M, RIGBY R A. Generalized additive models for location scale and shape (GAMLSS) in R [J]. **Journal of Statistical Software**, 2007, **23**(7): 45907.

Gamma distribution based traffic flow time series segmenting model

WANG Benchao[1,2], LI Dan[1], QIN Pan[1], GU Hong[*1]

(1. School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China;
2. Liaoning Police College, Dalian 116036, China)

Abstract: It is significant to accurately obtain the change points of traffic flow for the subsequent traffic flow prediction, classification and multi-time control. Considering the nonnegative and heteroscedasticity, the traffic flow time series are fitted by Gamma distribution and segmented effectively. For multiple traffic flow time series, dimension reduction is carried out by the nonnegative principal component analysis (NPCA) for feature extraction. Then, the likelihood of the principal component is constructed to obtain the parameters of the Gamma distribution. Consequently, the change points are determined from degree of fitting using the different parameters of the Gamma distribution by maximizing the likelihood. The Akaike information criterion (AIC) is used to select the optimal segmentation order and boundary. The experimental results indicate that the proposed segmenting model can reflect the change of traffic flow at different time and has better segmentation results than other existing methods.

Key words: traffic flow time series; Gamma distribution; time series segmentation; nonnegative principal component analysis