**文章编号**：1673-5196(2018)01-0095-05

# 基于四分位数对路段速度的预测优化

朱昶胜，李　硕，王永贤，刘敬帅

（兰州理工大学 计算机与通信学院，甘肃 兰州　730050）

**摘要**：使用出租车GPS数据作为基础，采用更加合理有效的路段速度作为交通状态参数，分析路网划分后的路段速度时间序列，利用四分位数特性优化算法，提高预测模型的合理性和准确性，并通过真实历史数据验证方法的可靠性.从带有随机性和不确定性的交通流变化中，通过分析找出其中的规律性，以预测未来几个时段的交通流变化.结果表明四分位法既体现出了路段速度的变化趋势，同时削弱了极端值和异常值的影响，能够展现出合理的交通状态变化过程，并且其计算简便，为大规模数据处理有效节省了计算资源.对计算结果的曲线拟合证明了四分位法处理路段速度的可靠性，对交通状态预测具有重要意义.

**关键词**：四分位数；路段速度；优化；交通状态预测

**中图分类号**：TP391　　**文献标志码**：A

# Optimization of road-segment speed prediction based on tetra-quantile

ZHU Chang-sheng，LI Shuo，WANG Yong-xian，LIU Jing-shuai

（College of Computer and Communication，Lanzhou Univ. of Tech.，Lanzhou　730050，China）

**Abstract**：Taking taxi GPS data as basis and logical and valid road-segment speed as traffic state parameter，the time series of road-segment speed on meshed road net is analyzed and the tetra-quantile characteristic optimization algorithm is used to improve the rationality and accuracy of the prediction model and verify the reliability of this method with real historical data. By means of analysis，the regularity of traffic flow variation is found from its randomness and uncertainty to predict it in oncoming couple of time-intervals. The result shows that the tetra-quantile method will embody the trend of road-segment speed as well as weaken the influence of its extreme and abnormal values，showing a reasonable traffic state variation process. Besides，its calculation will be simple，saving computing resources effectively for large-scale data processing. The fitted road-segment speed curves will successfully verify that the road-segment processing with tetra-quantile method is reliable and significant for traffic state prediction.

**Key words**：tetra-quartile；road-section speed；optimization；traffic state prediction

随着经济和社会的发展，城市规模日益扩大，城镇居民的出行要求逐步增加，使得困扰世界各大城市的交通拥挤问题变得更加严峻.交通拥挤不仅使道路通行能力降低、行车速度下降、交通延误增大，还造成巨大的经济损失.解决城市交通拥挤问题的传统方法，如增加城市道路、修建高架桥等交通基础设施，显露出难以克服的局限.随着对交通问题研究的深入，交通对策设计必须逐步从设施供给为主的硬对策，转向设施供给与需求管理相结合的软硬协同[1-4].

智能交通系统被认为是缓解道路交通拥堵、减少汽车尾气排放污染和交通事故等交通问题的有效方法之一.采用有效的道路交通状态分析技术对前端采集的历史数据及当前交通状况进行快速分析和实时、准确的预测与判别，可以提高控制与诱导的效果，为道路交通系统的交通信息服务、交通诱导、交通管制以及交通拥堵问题的缓解提供强有力的技术支持.

交通流预测是智能交通系统的重要内容，同时也是交通信息服务、交通控制与诱导的重要基础，要解决的问题就是如何从带有随机性和不确定性的交通流变化中，根据来自各种交通流信息采集设备的

**收稿日期**：2016-08-30

**作者简介**：朱昶胜(1974-)，男，甘肃秦安人，博士，教授.

交通流参数数据，结合其他影响因素，进行数据的系统分析，找出其中的规律性，建立相应的预测方法和模型，以预测未来几个时段的交通流变化[5-7].

近年来，对交通流预测的研究更倾向于对交通量进行预测[8-11].但不同的道路，以及同一条道路的畅通与拥挤的不同情况等多种因素，使其可能对应相同的流量水平，所以仅以交通量来描述和预测道路的交通状况是不合适的.相比而言，采用路段速度来描述路段上的交通状况要合理得多.

出租车由于其出行率高，运行时间长，行驶范围广，占交通流比例较高，且路网覆盖率高等特点，其GPS数据成为优质的FCD数据来源，依靠企业调度管理系统直接产生的数据也能够一定程度上满足采用浮动车方法检测道路交通流行驶车速的要求.因此相比于公交车数据的单调和滞后，利用出租车GPS数据对交通流预测能有效、准确、实时地反映道路交通状态[12-13].

而在描述路段速度的过程中，通常采用统计时段的平均值，或计量车辆通过路段的时间等方法[14-15].但由于交通流变化不仅与本路段过去几个时段的交通流有关，还受上下游的交通流及天气变化、交通事故和交通环境等因素的影响，交通流变化过程体现出实时、非线性、高维、非平稳随机等特点.随着统计时段的缩短，交通流变化的随机性和不确定性越来越强，统计时段中路段速度值的合理性将会直接影响最后的预测结果.

本文采用出租车GPS数据作为基础，使用路段速度作为交通状态参数，分析路网划分后的路段速度时间序列，利用四分位数特性优化算法进而提高预测模型的合理性和准确性，并通过真实历史数据验证方法的可靠性.

# 1 探索性数据分析

## 1.1 数据预处理

本文采用的出租车FCD数据集由甘肃省天水市交通局提供，数据真实，包括车量编号、追踪时间、GPS(包括经度、维度)、行驶速度、载客状态、海拔和里程数等众多变量.根据需要，提取部分变量数据作为在速度预测过程中的影响因子作为新的数据集，示例如图1所示.

首先，原始数据通过GPS定位与地图匹配，按照空间划分为各个路段的数据子集.在此过程中，由于存在GPS的误差和其他变量的异常值，需要对数据进行清理与矫正.数据预处理过程是大数据处理的关键步骤，直接关系到后续计算的效率与结果.之

| CAR_ID | TRACK_TIME | LONGITUDE | LATITUDE | SPEED | … |
|---|---|---|---|---|---|
| 290500 | 25-08-2014 12:45:43 | 105.728357 | 34.585251 | 22.5 | … |
| … | … | … | … | … | … |

图1 数据示例

Fig. 1 Illustration of data

后将数据子集根据时间排序则得到本条路段的原始速度时间序列.如何在数据清洗后通过计算利用不同车辆的速度值时间排序代表路段的速度，进而发现与预测道路的交通状态则是另一个关键问题.

## 1.2 速度值分析

在处理数据的过程中，速度变量具有其特殊性.原因在于实际情况中汽车行驶速度状态具有巨大的不确定性和随机性.例如，在极短的时间间隔内，速度可能变化特别大，不具有平稳变化的特性.另外，加速超车、临时停车、突发事件、出租车上下乘客等各种情况均会轻易且明显地影响车辆速度值，进而对评价路段速度以及后续的挖掘性数据分析过程造成不利影响.由于数据量大，数据采集时间间隔小，相邻数据并不一定来自同一数据来源车辆，直接使用原始数据对路段速度进行评价是不合理的.必须设定一个统计时段，对时间段内的速度值合理分析，得到一个经过计算的最终值来代表路段速度.因此，选取适当的方式处理统计时段中的速度值来表示路段速度是尤为重要的.

图2为数据集中2014年6月18日一天内某市某路段的速度频率图.通常情况下，平均数、众数、中位数是最常用到的数学统计量.

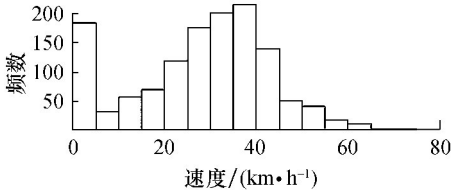


图2 某路段6月18日速度频数图

Fig. 2 Velocity frequency diagram on a road in Jun 18th

平均数是统计学中最基本、最常用的一种平均指标.算术平均数是一个良好的集中量数，具有反应灵敏、确定严密、简明易解、计算简单、适合进一步演算和较小受抽样变化的影响等特点，也是计算路段速度时通常使用的方法.但每个数据的大小变化都会影响到算术平均数的最终结果，极易受极端值的影响.但仔细观察数据集速度值分布情况可知，速度的极端值较多，明显不能使用简单的算术平均数来代表一个统计时段内的速度值.

众数是一组数据中出现次数最多的数值.众数不受极端数据的影响,并且求法简便,选择众数表示数据的"集中趋势"比较适合.但在本文采用的数据集中,速度值一般在区间[0,60]内(单位:km/h).由于速度值精确到小数点后 1 位,且确定统计时段后,每个速度值的频数都非常低且接近,甚至可能都为 1,所以很难得到一个合适众数来代表的速度.若采用分组分类统计的方式,计算代价过大,并且很难针对不同道路选定合适的速度值范围分类.

中位数是处于一串数的中间位置的数,以它在所有值中所处的位置确定的全体值集的代表值.中位数不受分布数列的极大或极小值影响,从而在一定程度上提高了中位数对分布数列的代表性.但有些离散型变量频数分布偏态时,中位数的代表性会受到影响.例如本文数据集中的速度分布,简单地利用中位数表示难以达到理想效果.

图 3 是 6 月 18 日某路段每小时速度分布的箱形图.例如在下午 13 时、15 时,在中位数上下两端的速度分布密度明显不均衡,这表示整段道路中,车辆更稳定于在以较密集分布区的速度下行驶.因此,单纯利用中位数不足以代表统计时段的路段速度特征.同时通过观察图像可知,中位数分布位置变化不够明显,利用中位数也不能得到一天中此路段明显的速度变化趋势.

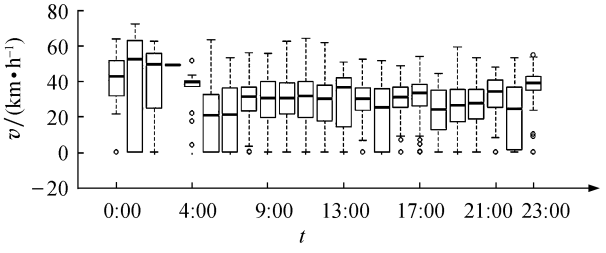


**图 3　某路段 6 月 18 日速度分布箱形图**

**Fig. 3　Velocity box-plots diagram on a road in Jun 18th**

## 2　数学模型

### 2.1　四分位数

在统计学中,把所有数值由小到大排列并分成四等份,处于三个分割点位置的数值就是四分位数(Quartile).其中,处于数列排序后第 25%位置的数字称为下四分位数,记为 $Q_1$;处于第 50%位置的数字即中位数,记为 $Q_2$;处于第 75%位置的数字称为上四分位数,记为 $Q_3$.

四分位数包含中位数.同时,可以使用四分位数之间的差值即间距获得平均值和众数的优点.例如,若中位数与下四分位数间距较小,则说明处于此区间的车辆速度值较其他区间更为稳定,代表整体速度的权重更大.四分位数的分布特征能反映出数据的分布特征,因此利用四分位数对速度值处理不仅能够更准确地代表在一定统计时间段内路段上所有速度值,还能较好地处理极端值对整体数据的影响.

### 2.2　函数公式

要构建的路段速度表达函数需要体现出速度的分布,同时又不受极端值影响,并且应该增强不同时间段的速度变化趋势,以便于后续的时间序列分析和预测.同时计算得到的速度值必须有合理的上下界限,并取在数据分布最密集的区间内.

待定系数的反比例函数 $y=\frac{1}{x+k}+b$ 可以满足上述条件.以上下四分位数的均值与中位数的差值为自变量,以路段速度为因变量,通过边界值与所需条件确定函数中的系数.

记 $Q_1$ 与 $Q_3$ 的均值为 $P$,即 $P=\frac{1}{2}(Q_1+Q_3)$,则有:

1) 当 $Q_2=P$ 时,说明速度的分布较为平衡,此时,利用 $Q_2$ 代表路段速度是合适的:

$$v=Q_2 \tag{1}$$

2) 当 $Q_2>P$ 时,中位数更加接近下四分位数,此种分布情况下代表值函数表达式为

$$v=-\left(Q_2-P+\frac{2}{Q_3-Q_2}\right)^{-1}+\frac{1}{2}(Q_2+Q_3) \tag{2}$$

3) 当 $Q_2<P$ 时,中位数更加接近上四分位数,此种分布情况下代表值函数表达式为

$$v=\left(\frac{2}{Q_2-Q_1}-Q_2+P\right)^{-1}+\frac{1}{2}(Q_1+Q_2) \tag{3}$$

综上,合并式(1~3),暂取 1 h 为统计时段,则某小时内的路段速度可表示为

$$v_h=\begin{cases}-\left(Q_{2h}-P+\dfrac{2}{Q_{3h}-Q_{2h}}\right)^{-1}+\dfrac{1}{2}(Q_{2h}+Q_{3h}) & (Q_{2h}\geqslant P)\\ -\left(\dfrac{2}{Q_{2h}-Q_{1h}}-Q_{2h}+P\right)^{-1}+\dfrac{1}{2}(Q_{1h}+Q_{2h}) & (Q_{2h}<P)\end{cases} \tag{4}$$

## 3　数据统计分析

### 3.1　路段速度

经过对数据的处理,可以得到任意一天中每小时的统计时段内的路段速度.通过编程实现进而统

计分析，可以将初步结果进行可视化展现.

图 4 是 2014 年 6 月 17 日某市某路段每小时内速度均值与基于四分位数处理路段速度值的对比情况. 从图中可以明显地观察到，使用四分位法对速度值处理之后，路段速度的状态变化更加明显，并且能呈现出均值所不能反映出的变化状况，与均值法相较更能如实地反应交通状态. 例如在早上 6:00～7:00、午间 12:00 左右、下午 17:00 左右以及晚间 20:00 左右均会出现交通高峰期导致路段速度明显降低，而均值法曲线在中午 12:00 左右并没有体现出路段速度的明显变化.

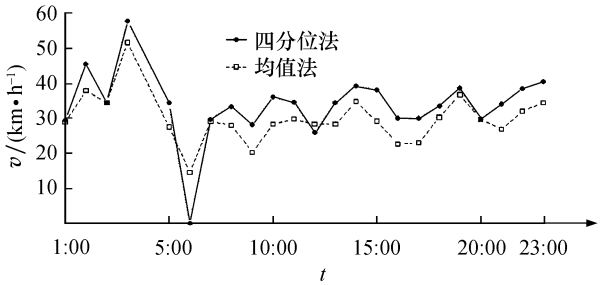


**图 4 某路段一天内四分位法和均值法计算得到速度值的对比**

**Fig. 4 Comparison of velocities evaluated with tetra-quantile and mean methods**

### 3.2 曲线拟合

经过对多种曲线拟合方法的对比，本文最终选用自然样条曲线拟合的方式对四分位法速度值进行曲线拟合. 样条拟合定义如下式：

$$S_i(x) = a_{i0} + a_{i1}(x - x_i) + a_{i2}(x - x_i)^2 + a_{i3}(x - x_i)^3 \quad (i = 0,1,\cdots,n-1) \tag{5}$$

其中：$S_i(x)$是在$[x_i, x_{i+1}]$分段区间的拟合曲线；$n$为数据区间个数；$a_{i0}$、$a_{i1}$、$a_{i2}$、$a_{i3}$为待定系数.

鉴于凌晨 0:00～5:00 的行人稀少，数据量较少，路段速度的变化随机性巨大且参考价值不高. 因此舍弃 0:00～5:00 的数据后，选取合适的自由度，对数据统计分析.

图 5 是某日某路段日间速度值的拟合曲线，结果表明，自然样条曲线拟合方式可以表达出速度变化趋势，在不同时段呈现出合理的变化趋势，同时适当地削弱了个别异常值的影响.

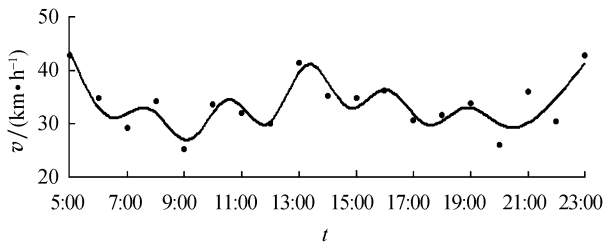


**图 5 某日某路段日间速度值的拟合曲线**

**Fig. 5 Daily velocity fitting curve on a road section in one day**

## 4 结果验证与可视化对比

根据数学模型，使用四分位法对真实历史数据处理，得到统计时段的路段速度数据集. 从结果集中随机选取 4 天的数据，使用四分位法对数据进行曲线拟合，观察实际效果以验证其可行性.

图 6 是某路段 2014 年 6 月 17 日、19 日、20 日、27 日的日间路段速度变化曲线汇总. 经过观察可知，使用四分位法处理后的路段速度变化曲线呈现出了明显的路段速度变化的趋势.

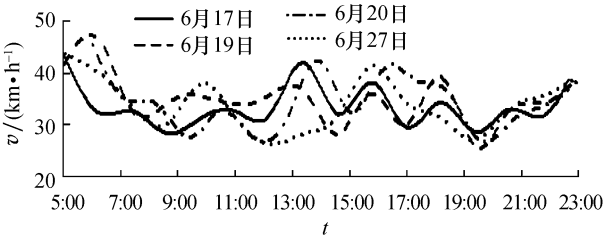


**图 6 某路段多天的四分位法日间路段速度变化曲线**

**Fig. 6 Multi-day variation curve of daily road-section velocity on a road-section evaluated with tetra-quantile method**

图 6 中，凌晨时间段由于行人稀少而普遍车速较快；在 7:00～9:00 时、12:00 左右的居民出行高峰期，路段速度明显低下；而在下午时段路段速度呈现出规律的波动；随着夜晚的到来车速再次逐步升高. 同时，曲线的变化趋势在多日对比情况下呈现出大致相同的规律性，能为时间序列分析提供基础的样本支持，这对今后的交通状态预测具有重大意义.

图 7 是使用均值法对 2014 年 6 月 17 日、19 日、20 日、27 日的相同数据处理后的路段速度进行曲线拟合汇总. 从处理过后的路段速度曲线拟合图可以明显地观察到，均值法速度变化曲线杂乱不堪，速度变化不合情理且受异常值影响严重. 更为重要的是，速度变化曲线在多日数据间无法呈现出规律性，几乎对预测交通状态没有帮助.

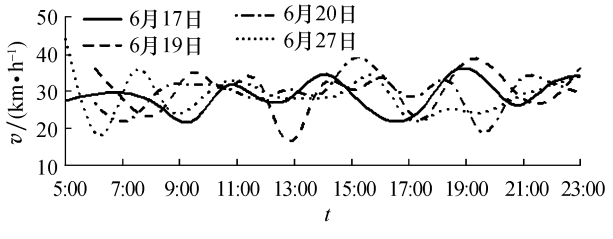


**图 7 某路段多天的均值法日间路段速度变化曲线**

**Fig. 7 Multi-day variation curve of daily road-section velocity on a road section**

## 5　结论

1) 交通数据因其特殊性,直接使用原始数据进行统计分析会伴随着诸多问题.传统均值方法处理路段速度难以寻找其规律性.本文提出一种基于四分位数的路段速度计算方法,在体现出路段速度的变化趋势的同时削弱了极端值和异常值的影响,展现出了合理的交通状态变化过程,并且计算简便,能有效地为大规模数据实时处理节省计算资源.

2) 采用自然样条曲线拟合方法,选用合适的自由度,实现了统计时段的路段速度曲线拟合,并成功呈现出不同日期同一路段的交通状态规律性,对交通状态预测具有重要意义.

3) 将基于四分位法的路段速度变化曲线和基于均值法的速度变化曲线对比,在使用同样数据的情况下明显体现出四分位法的优势,证明了使用四分位法处理路段速度的可靠性.

致谢:本文得到兰州理工大学红柳杰出人才基金(J201304)的资助,在此表示感谢.

**参考文献:**

[1] AGBOLOSU-AMISON S J, PARK B, YUN I. Comparative evaluation of heuristic optimization methods in urban arterial network optimization [C]//12th International IEEE Conference on Intelligent Transportation Systems. [S. l.]: IEEE, 2009: 1-6.

[2] BAN X, HERRING R, HAO P, *et al*. Delay pattern estimation for signalized intersections using sampled travel times [C]// Proceedings of the 88th Annual Meeting of the Transportation Research Board. Washington: [s. n.], 2009.

[3] FABRITIIS C, RAGONA R, VALENTI G. Traffic estimation and prediction based on real time floating car data [C]//11th International IEEE Conference on Intelligent Transportation systems. [S. l.]: IEEE, 2008: 197-203.

[4] HERRING R, HOFIEITNER A, AMIN S, *et al*. Using mobile phones to forecast arterial traffic through statistical learning [C]//89th Annual Meeting Transportation Research Board. Washington: [s. n.], 2010.

[5] HUNTER T, HERRING R, ABBEEL P, *et al*. Path and travel time inference from GPS probe vehicle data [J/OL]. [2016-06-20]. https://www.researchgate.net/publication/229034951_Path_and_Travel_Time_Inference_from_GPS_Probe_Vehicle_Data.

[6] WORK D, BLANDIN S, TOSSAVAINEN O P, *et al*. A distributed highway velocity model for traffic state reconstruction [J/OL]. [2016-06-20]. http://piccoli.camden.rutgers.edu/files/IEEE_TAC27.pdf.

[7] YIM Y B Y, CAYFORD R. Investigation of vehicles as probes using global positioning system and cellular phone tracking: field operational test [R]. California: Institute of Transportation Studies, University of California, 2001.

[8] 张　红,朱昶胜.基于大数据的智能交通体系架构[J].兰州理工大学学报,2015(2):112-115.

[9] 杨　涛.基于浮动车技术的路段交通流量推算研究[D].北京:北京交通大学,2006.

[10] 张周强,杜豫川.国外浮动车技术发展及其数据利用综述[C]//第七届华人交通运输学术大会论文集.上海:上海人民交通出版社,2007.

[11] 姚智胜.基于实时数据的道路网短时交通流预测理论与方法研究[D].北京:北京交通大学,2007.

[12] 朱丽云,温慧敏,孙建平.北京市浮动车交通状况信息实时计算系统[J].城市交通,2008,6(1):77-80.

[13] 王晋生,周伟刚,王　渤.GPS浮动车数据在城市交通管理和出行服务中的应用[J].公安交通科技窗,2009(1):31-33.

[14] 董均宇.基于GPS浮动车的城市路段平均速度估计技术研究[D].重庆:重庆大学,2006.

[15] 翁剑成,荣　建,于　泉.基于浮动车数据的行程速度估计算法及优化[J].北京工业大学学报,2007,33(5):459-464.