Journal of System Simulation

Volume 29 | Issue 9 Article 34

6-2-2020

Research of State Vector in Short-Term Passengers Flow Forecasting Based on Nonparametric Regression

Guo Han
School of Civil and Transportation Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;

Pengpeng Jiao
School of Civil and Transportation Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;

Research of State Vector in Short-Term Passengers Flow Forecasting Based on Nonparametric Regression

Abstract

Abstract: KNNR (K Nearest Neighbor Based Nonparametric Regression) Method was used for short-term traffic forecast and *the choice of state vector was studied*. The result shows that taking the data of some historical periods as the state vector has a good prediction. Although the correlation of the historical passenger flow between different Rail transit sites is significant, it neglects the fact that the passengers enter each station is independent. So taking the historical passenger flow of adjacent sites as the state vector is not appropriate.

Keywords

short-term passengers flow forecasting, nonparametric regression, state vector, k nearest neighbors

Recommended Citation

Guo Han, Jiao Pengpeng. Research of State Vector in Short-Term Passengers Flow Forecasting Based on Nonparametric Regression[J]. Journal of System Simulation, 2017, 29(9): 2128-2134.

第 29 卷第 9 期 系统仿真学报© Vol. 29 No. 9
2017 年 9 月 Journal of System Simulation Sep., 2017

非参数回归短时客流预测中状态向量研究

郭晗，焦朋朋
(北京建筑大学土木与交通工程学院，北京 100044)

摘要：采用 K 近邻非参数回归的方法对轨道交通站点短时进站客流量进行了预测，*并对状态向量的选择进行了研究*。预测结果表明：以预测时段前 m 个时段的历史数据作为状态向量具有较好的预测效果，而相邻站点历史客流数据虽然在数值上与预测站点的客流数据具有较大相关性，但由于其忽视了各站进站客流是相对独立的，因此不宜作为状态向量。

关键词：短时客流预测；非参数回归；状态向量；K 近邻
中图分类号：U491 文献标识码：A 文章编号：1004-731X (2017) 09-2128-07
DOI: 10.16182/j.issn1004731x.joss.201709034

Research of State Vector in Short-Term Passengers Flow Forecasting Based on Nonparametric Regression

Guo Han, Jiao Pengpeng
(School of Civil and Transportation Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: KNNR (K Nearest Neighbor Based Nonparametric Regression) Method was used for short-term traffic forecast and *the choice of state vector was studied*. The result shows that taking the data of some historical periods as the state vector has a good prediction. Although the correlation of the historical passenger flow between different Rail transit sites is significant, it neglects the fact that the passengers enter each station is independent. So taking the historical passenger flow of adjacent sites as the state vector is not appropriate.

Keywords: short-term passengers flow forecasting; nonparametric regression; state vector; k nearest neighbors

引言

城市轨道交通建设正朝着越来越复杂的路网结构和满足不断增长的客流需求的方向发展，因此畅通、便捷的交通路网是市民们普遍渴望的。短时客流预测是对短时间(5~15 min)内客流量进行预测，对城市轨道交通建设起着至关重要的作用。从运营方面来说，如果客流预测结果偏大，将导致运营费用和维修费用增加，使运营公司长期处于亏损状态，政府将给予大量财政补贴；如果客流预测结果偏小，则会造成轨道交通的服务质量下降，运力不足。所以客流预测为城市轨道交通的建设规模、经济效益等提供理论基础和依据。

短时客流预测模型有很多，如线性模型、非线性模型、混合模型、其他模型等[1]。非参数回归(Nonparametric Regression，NPR)是一种利用历史数据与系统当前数据进行近似程度分析的预测方法，它不需对原始数据进行训练，也不需要对模型进行严格的参数标定，只需将历史数据与现状数据

收稿日期：2017-05-19 修回日期：2017-07-10;
基金项目：国家自然科学基金(51578040)，北京市科技新星计划(Z151100000315050)，北京市自然科学基金(8162013)；
作者简介：郭晗(1993-)，女，北京，硕士生，研究方向为交通运输规划；焦朋朋(通讯作者 1980-)，男，安徽，博士，教授，研究方向为交通运输规划。

进行匹配就可完成预测。由于非参数回归没有对原始数据进行平滑处理，保持了数据的原始特性，因此在预测特殊环境的客流参数时，非参数回归的预测效果更为准确。

非参数回归这一概念最早应用于国外的研究中。Davis G. A. & Nihan N. L.[2]将非参数回归应用于短时交通流预测领域，并提出基于非参数回归 K 最近邻(KNN)的方法。Smith B. L. & Demetsky M. J.(1997)[3]将历史平均法、时间序列法、BP 神经网络和非参数回归四种方法的优缺点进行了比较。Smith B.L.等[4]指出非参数回归的基本方法受模式识别的影响，且认为所有关系都包含在数据内，而不是包含在模型里。Williams B.M.[5]提出将匹配距离的倒数作为相似状态权重，并使用最相似状态下一时段交通状态进行预测。Clark S.[6]将非参数回归进行了多变量参数拓展，在短时交通流预测中同时考虑流量、车速、占有率 3 个交通流参数作为状态向量。Yuan Z.W. & Wang Y.H.[7]提出基于 KD-Tree 和聚类分析的非参数回归算法，利用空间自相关分析确定状态向量。Rasyidi M.A.等[8]考虑临近路段特征对目标路段车速的影响，提出目标路段速度的预测不仅与历史数据有关还与相邻路段有关。

近些年，国内学者也对非参数回归理论展开了深入的研究。宫晓燕[9]提出基于密集度的变 K 值搜索算法与基于动态聚类和散列函数的历史数据组织方式。张晓利等[10-11]提出基于平衡二叉树的 K-邻域非参数回归(KNN-NPR)的短时交通流预测方法，并将模式分析引入非参数回归预测中。张涛等[12]提出对于特殊日期的交通流预测应着重考虑 K=1 时的最近邻。于滨等[13]分析了 K 近邻算法的时间和空间参数，提出基于时空参数的 K 近邻模型。Hou X 等[14]基于传统 K 最近邻算法提出双层 K 最近邻算法。

本文参考短时交通流预测中状态向量的确定方法，将其应用于短时进站客流量预测中，并对状态向量的选取进行了研究，为该领域的研究理清思路，并为进一步的研究奠定坚实的基础。

1 非参数回归预测主要步骤

1.1 建立历史数据库

历史数据库由包含全部系统状态的历史观测数据组成，K 近邻非参数回归的预测效果取决于历史数据库的容量和质量，历史数据越多、越准确，越接近真实的系统状态，有利于得到较为准确的预测量。因此历史数据库要具有足够的数据储备，以保证在近邻搜索时能够得到尽可能多的近似状态数据。另外，历史数据库也要持续不断更新。当检测到新的数据时，应将其添加到历史数据库中，不断拓展现有数据库容量和样本，以满足数据更新和实时预测的需求。

1.2 确定状态向量

状态向量是指与预测量相关的因素组成的向量，为了提高非参数回归的运算速度以及准确性，通常选择对预测量影响较大的因素。目前，对于状态向量的确定方法没有一个统一的方法，大多数研究根据经验法或试探法选取状态向量[12]，部分是通过计算各影响因素与预测量之间的相关系数来确定状态向量[9]。之后，还有学者提出利用主成分分析法[15]和趋势法[16]选取状态向量。

1.3 计算距离度量准则

距离度量准则表示历史状态与当前状态的近似程度，通常用欧几里德距离(简称欧式距离)表示，即计算数据的离差平方和。由于不同状态向量对预测量的影响程度不同，因此用相关系数表示影响程度，并参与距离度量准则计算。以下介绍空间和时间两种维度下状态向量的距离度量准则：

(1) 空间状态向量的距离度量准则：表示同一时段，不同位置的状态向量的近似程度。

对于邻近站点与预测站点之间的距离度量准则，采用式(1)：

第 29 卷第 9 期 系统仿真学报 Vol. 29 No. 9
2017 年 9 月 Journal of System Simulation Sep., 2017

$$d_{it}=\sqrt{\frac{\sum_{J=j}^{j+k}\gamma_J\left[S_*(t)-S_{hJ}(t)\right]^2}{\sum_{J=j}^{j+k}\gamma_J}} \tag{1}$$

式中：d_{it} 表示邻近站点与预测站点在第 i 天第 t 时段的欧式距离；γ_J 表示 J 号站点与预测站点间的相关系数；$S_*(t)$ 表示预测站点 t 时段的现状进站客流量(人次)；$S_{hJ}(t)$ 表示 J 号站点 t 时段的历史进站客流量(人次)。

(2) 时间状态向量的距离度量准则：表示同一位置，不同时段的状态向量的近似程度。

对于历史时段与预测时段进站客流量的距离度量准则，采用式(2)：

$$D_{it}=\sqrt{\frac{\sum_{T=t}^{t-m+1}\left[S_*(T)-S_{h*}(T)\right]^2}{m}} \tag{2}$$

式中：D_{it} 表示预测站点在第 i 天第 t 时段的欧式距离；$S_*(T)$ 表示预测站点第 T 时段的现状进站客流量(人次)；$S_{h*}(T)$ 表示预测站点第 T 时段的历史进站客流量(人次)；m 表示与预测时段相邻时段的个数。

1.4 确定 K 值

近邻搜索的过程是以事先定义的度量距离为依据，从历史数据库中寻找与当前状态距离最短的历史数据标记为搜索近邻，并将这些近邻组成一个近邻子集。其中，对预测结果有较大贡献的近邻所组成的子集被称为最优近邻子集，用 K 表示最优近邻子集中近邻的个数。研究表明，K 值过大或过小均会影响预测效果，因此 K 值的选取十分重要。

1.5 预测算法

预测算法是指利用历史状态预测未来状态的算法，通常采用预测精度较高的带权重预测算法，即以匹配距离的倒数作为权重，以匹配距离所对应的下一时刻流量作为自变量，加权求和得到预测值。带权重的预测算法采用如下形式：

$$v(t+1)=\sum_{i=1}^{k}\beta_{it}v_{hi}(t+1)\text{，其中 }\beta_{it}=\frac{d_{it}^{-1}}{\sum_{i=1}^{K}d_{it}^{-1}}$$

式中：$v(t+1)$ 表示预测时段进站客流预测值(人次)；$v_{hi}(t+1)$ 表示历史数据库中与预测时段相同时段的进站客流量(人次)；β_{it} 表示第 t 个时段的权重.

由于匹配距离表示预测值与实际值的近似程度，欧式距离越小，表明数据匹配程度越高，继而对预测值的贡献就越大。

2 状态向量的确定

2.1 状态向量的组成

状态向量指与预测量相关的因素。对于道路交通系统来说，影响交通流预测的因素有很多，比如车辆的速度，天气情况，道路状况以及上下游路段的交通流量等。对于客运交通系统来说，影响客流预测的因素主要包括进出站客流量、不同交通方式间换乘客流量、各站点前 m 个时段历史客流量等。为了提高算法的速度，满足预测实时性的要求，通常选择对流量影响最大的因素。

本文的数据由第三方提供，数据种类仅包括各站进出站刷卡记录，因此，本文的预测对象为进站客流量，备选状态向量为相邻站点进站客流量、预测站点历史时段进站客流。

2.2 状态向量确定方法

现有研究中状态向量的确定方法有很多，如基于统计学理论提出的相关系数法，基于数学降维思想提出的主成分分析法，考虑交通流动态变化趋势的趋势法，通过数据建模简化数据的聚类分析等。这些确定方法的核心思想就是从种类丰富的交通流数据中提取出与预测量最相关的几种数据，精简预测输入量，提高模型的计算效率。相关系数法因为计算简单，通过计算不同数据间的相关性参数，判别数据的相似程度，因而广泛应用于状态向量的确定中。

本文应用相关系数法，分别从时间和空间两个维度确定状态向量，因此应用相关系数和自相关系数确定用于进站客流预测的状态向量。

3 实例分析

3.1 数据采集及处理

本文采用北京地铁 13 号线 2013 年 11 月各站的进站刷卡记录建立历史数据库，以 5 min 为统计单元，统计了一个月内各站点每天 4:45–24:00 的进站客流量。由于各站点全天进站客流的分布差异较大，考虑数据的有效性和质量，最终选取了 37 号站点作为预测站点，以 29 日的进站客流数据作为测试数据进行预测，预测时段为 5:30–22:59。

首先计算样本数据库中相邻站点间进站客流相关系数，以及预测站点的历史进站客流自相关系数，计算结果如表 1、表 2 所示。

表 1 各站点间相关系数
Tab. 1 Correlation coefficient between different sites

站名	S21	S23	S25	S27	S29	S33	S35
S37	0.006	0.086	0.036	0.060	0.610	0.905	0.977
站名	S39	S41	S43	S45	S47	S49	S51
S37	0.957	0.968	0.400	0.716	0.711	0.098	0.191

表 2 预测站点自相关系数
Tab. 2 Autocorrelation coefficient of predictive site

m 值	*m*=1	*m*=2	*m*=3	*m*=4	*m*=5
自相关系数	0.195	0.309	0.480	0.658	0.713
m 值	*m*=6	*m*=7	*m*=8	*m*=9	*m*=10
自相关系数	0.784	0.835	0.865	0.892	0.916

根据统计学知识，相关系数表示不同事件间相互影响程度，自相关系数表示的是同一事件在两个不同时期的相关程度。相关系数等于 0 表示两个事件完全不相关，相关系数接近于 1 表示两个事件具有较强相关性。从表 1、表 2 可以看出，各站进站客流和预测站点的历史进站客流都与预测站点进站客流具有不同程度的相关性。因此为了提高预测精度，本文选择相关系数大于 0.95，自相关系数大于 0.8 的影响因素作为状态向量，即以与 35 号站、39 号站、41 号站的历史数据，以及 37 号站预测时段前 7 个、8 个、9 个时段的历史数据作为状态向量。因为数据预测的起始时段为 5:30，样本数据统计的起始时段为 4:45，不能满足预测时段前 10 个时段的要求，因此 *m*=10 的历史数据不选为状态向量。

将以上确定的状态向量按照时间和空间分为两组，以预测站点前 *m* 个时段历史进站客流表示时间组，以不同站点的历史进站客流表示空间组，分别进行短时进站客流预测以比较状态向量选取的优劣。

3.2 客流预测实验设计

为了验证状态向量的适用性，设置了 4 组预测方案，如表 3 所示。

表 3 预测方案
Tab. 3 Prediction scheme

方案	*m*=7	*m*=8	*m*=9	35 号站	39 号站	41 号站
1	√					
2		√				
3			√			
4				√	√	√

高峰时段的进站客流对运营部门分配运力、调动相关人员疏导站内秩序起着关键的作用，因此需对高峰时期状态向量的预测效果进行详细分析。通过分析 37 号站点日进站客流的时间分布(如图 1)，可知 7:25–8:35 为该进站客流量的高峰时段。

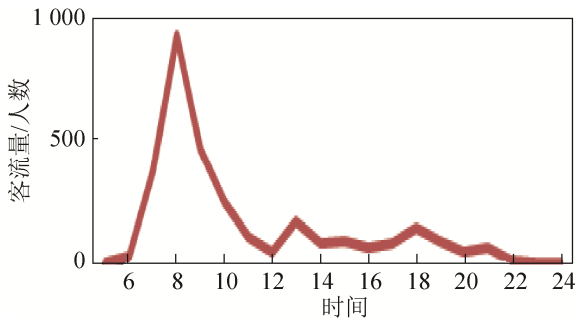

图 1 37 号站点日进站客流趋势图
Fig. 1 Trend of daily entering passenger flows on NO.37 site

3.3 实验结果

本文分别用平均绝对误差(MAE)、均方误差(MSE)、均方根误差(RMSE)、相对误差(RE)、修正的相对误差(MRE)评价预测结果。表 4 是 4 种预测方案的各项误差指标及对应的 K 值。

第 29 卷第 9 期 系统仿真学报 Vol. 29 No. 9
2017 年 9 月 Journal of System Simulation Sep., 2017

从表 4 中可以看出，方案 1，2，3 在 K 值分别取 12，18，17 时得到了最小的误差，且随着 m 值的增大，K 值的取值也逐渐增大。从各误差指标来看，方案 1 的预测效果最好，而方案 2 和方案 3 的预测效果差不多，这表明一方面虽然状态向量的组成元素越多，在近邻匹配的过程中越能体现真实的交通状况，但是另一方面，过多组成元素可能包含一些干扰因素，导致误差增大，影响预测效果。由于本文的数据是根据各站进站刷卡记录统计而成，而刷卡记录可能会由于设备传输、故障问题，导致记录中的进站时间存在偏差，因此这些误差可能会对近邻匹配过程造成干扰。

图 2 表示 4 种实验方案的修正相对误差曲线，从图中可以看出，以相邻站点进站客流量作为状态向量的方案(方案 4)，其修正相对误差呈现逐渐上升的趋势，这表明相邻站点的进站客流量与预测站点的进站客流量虽然在数值上有较强的相关性，但在实际运营中由于各站进站客流量是相互独立的，各站进站客流量不存在相互影响，因此选取相邻站点进站客流量作为状态向量是不合理的，应把这个状态向量剔除。

表 4 全天预测效果误差比较

Tab. 4 Comparison of full-day prediction effect errors

方案	MAE	K 值	MSE	K 值	MRE	K 值	RMSE	K 值
1	8.97	12	281.54	12	13.72%	12	16.78	12
2	9.01	15	289.95	18	13.78%	15	17.03	18
3	9.02	16	290.86	17	13.80%	16	17.05	17
4	28.83	1	2 344.81	1	44.11%	1	48.42	1

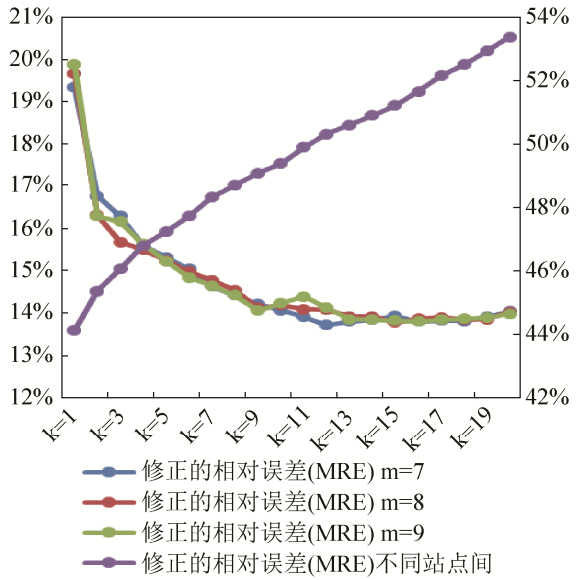

图 2 修正相对误差曲线

Fig. 2 Corrected relative error curve

而以预测站点历史数据作为状态向量的方案(方案 1，2，3)，其修正相对误差则呈现先下降后上升的趋势，这表明在相同 m 值不同 K 值时，随着 K 值的增大，误差曲线近似为 U 型，当误差值位于曲线底端时，预测效果最佳，对应 K 值为最优近邻个数。因此，预测站点前 m 个时段的历史数据与预测时段进站客流量有较强相关性，选择此因素作为状态向量可以得到较好的预测结果。

表 5 是方案一的高峰时段和全天时段预测结果的误差指标对比。从表中结果可以看出，高峰时段各项误差均比全天时段各项误差大，因为高峰时段进站客流量大，数据存在取整误差、数据收集误差等误差累积，对客流预测的影响较大。而对于全天时段来说，非高峰时段进站客流量少，不同时段数据变化量较小，对于客流预测的影响较小，因此全天时段的进站客流预测值与实测值偏差不大。

表 5 不同时段预测的误差对比

Tab. 5 Comparison of errors in different periods

误差指标	高峰时段	全天时段
平均绝对误差(MAE)	33.43	8.97
均方误差(MSE)	1 964.29	281.54
均方根误差(RMSE)	44.32	16.78

4 结论

非参数回归是一种无固定模型参数、移植性强、预测精度高的方法，它仅仅借助大量历史数据，无需繁琐的数学计算，就可对实时、非线性的数据进行预测，因此非参数回归的优势较为突出。本文基于非参数回归原理，利用相关系数法确定状态向量，对短时进站客流进行预测，得到了如下结论：

(1) 状态向量的确定要符合客观规律，虽然状态向量与预测对象在数值上显示出强相关性，但如果不符合客观规律也不适合作为状态向量。

(2) 状态向量组成元素的个数对预测效果有一定的影响，组成元素越多模型的计算量越大，虽然能体现数据的真实性，但会造成数据冗余，降低预测精度。

进一步研究可从以下几方面展开：

(1) 轨道交通站点周边土地利用性质、人口规模、多种交通方式接驳等因素都会对站点的客流量造成一定的影响，因此研究这些因素对进站客流的影响有助于提高模型的预测精度。

(2) 随着轨道交通的日益发展，周末或节假日的客流量也日趋增多，应对这些特殊日期的进站客流进行预测，并对相关影响该因素进行分析。

参考文献:

[1] 陆海亭, 张宁, 黄卫, 等. 短时交通流预测方法研究进展 [J]. 交通运输工程与信息学报, 2009, 7(4): 84-91. (Lu H T, Zhang N, Huang W, et al. Research Progress of Short Term Traffic Flow Prediction Methods [J]. Journal of Transportation Engineering and Information, 2009, 7(4): 84-91.)

[2] Davis G A, Nihan N L. Nonparametric Regression and Short Term Freeway Traffic Forecasting [J]. Journal of Transportation Engineering (S0733-947X), 1991, 117(2): 178-188.

[3] Smith B L, Demetsky M J. Traffic Flow Forecasting: Comparison of Modeling Approaches [J]. Journal of Transportation Engineering (S0733-947X), 1997, 123(4): 261-266.

[4] Smith B L, Williams B M, Oswald R K. Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting [J]. Transportation Research Part C Emerging Technologies (S0968-090X), 2002, 10(4): 303-321.

[5] Williams B M, Hoel L A. Modeling and Forecasting Vehicular Traffic Flow as A Seasonal Stochastic Time Series Process [D]. Charlottesville, USA: University of Virginia, 1999.

[6] Clark S. Traffic Prediction Using Multivariate Nonparametric Regression [J]. Journal of Transportation Engineering (S0733-947X), 2003, 129(2): 161-168.

[7] Yuan Z W, Wang Y H. Research on K Nearest Neighbor Non-parametric Regression Algorithm Based on KD-Tree and Clustering Analysis [C]// Yuan Z W. Research on K Nearest Neighbor Non-parametric Regression Algorithm Based on KD-Tree and Clustering Analysis. New York, USA: IEEE, 2012: 298-301.

[8] Rasyidi M A, Kim J, Ryu K R. Short-Term Prediction of Vehicle Speed on Main City Roads using the k-Nearest Neighbor Algorithm [J]. Journal of Intelligence and Information Systems (S0925-9902), 2014, 20(1): 121-131.

[9] 宫晓燕, 汤淑明. 基于非参数回归的短时交通流量预测与事件监测综合算法 [J]. 中国公路学报, 2003, 16(1): 82-86. (Gong X Y, Tang S M. Integrated Traffic Flow Forecasting and Traffic Incident Detection Algorithm Based on Non-Parametric Regression [J]. China Journal of Highway and Transport, 2003, 16(1): 82-86.)

[10] 张晓利, 贺国光, 陆化普. 基于 K-邻域非参数回归短时交通流预测方法 [J]. 系统工程学报, 2009, 24(2): 178-183. (Zhang X L, He G G, Lu H P. Short-Term Traffic Flow Forecasting Based on K-nearest Neighbors Non-Parametric Regression [J]. Journal of Systems Engineering, 2009, 24(2): 178-183.)

[11] Zhang X, Lu H. The Simulation Research of Nonparametric Regression for Short-Term Traffic Flow Forecasting [C]// Zhang X. The Simulation Research of Nonparametric Regression for Short-Term Traffic Flow Forecasting. New York, USA: IEEE, 2009: 626-629.

[12] 张涛, 陈先, 谢美萍, 等. 基于 K 近邻非参数回归的短时交通流预测方法 [J]. 系统工程理论与实践, 2010, 30(2): 376-384. (Zhang T, Chen X, Xie M P, et al. K-NN Based Nonparametric Regression Method for Short-Term Traffic Flow Forecasting [J]. Systems Engineering-Theory & Practice, 2010, 30(2): 376-384.)

（下转第 2139 页）

DOI: 10.16182/j.issn1004731x.joss.201709034