2010 17th International Symposium on Temporal Representation and Reasoning

Local Polynomial Regression Models for Average Traffic Speed Estimation and Forecasting in Linear Constraint Databases

Hang Yue, Elizabeth Jones*
Nebraska Transportation Center
Civil Engineering Department
University of Nebraska-Lincoln
Email: yuehang366@gmail.com
*ejones1@unl.edu

Peter Revesz
Computer Science & Engineering Department
University of Nebraska-Lincoln
Email: revesz@unl.cse.edu

Abstract

Constraint databases have the specific advantage of being able to represent infinite temporal relations by linear equations, linear inequalities, polynomial equations, and so on. This advantage can store a continuous time-line that naturally connects with other traffic attributes, such as traffic speed. In most cases, vehicle speed varies over time, that is, the speed is often nonlinear. However, the infinite representations allowed in current constraint database systems are only linear. Our article presents a new approach to estimate and forecast continuous average speed using linear constraint database systems. Our new approach to represent and query the nonlinear average traffic speed is based on a combination of local polynomial regression and piecewise-linear approximation algorithm. Experiments using the MLPQ constraint database system and queries show that our method has a high accuracy in predicting the average traffic speed. The actual accuracy is controllable by a parameter. We compared the local linear regression model with the local cubic model by using a field experiment. It was found that the local cubic model follows more closely the raw data than the linear model follows.

1. Introduction

Today relational databases are popularly applied to many traffic systems [1, 2], such as traffic management systems, public transit systems and Advanced Traveler Information Systems (ATISs). However, when the relational model is used to handle spatial data, the points, lines and polygons in space are discretely saved in tables and often lose some of the important spatial relationships [3]. Having only a finite set of tuples in the relational tables may make it difficult to see the intuitive relationship among the important traffic attributes [4].

Time discontinuity is another significant deficiency in relational databases. Users expect that, regardless of the magnitude of change, the complete snapshot produced at each time slice could duplicate all the unchanged data in the database. However, relational databases with time discontinuity [3] cannot store the complete information of moving objects, such as vehicles.

Constraint databases are viewed as a special kind of post-relational databases, although they share with relational databases some important features, such as, formal model-theoretic semantics, and various high-level query languages including SQL and Datalog [5]. Constraint databases have some additional useful features like the ability to represent infinite relations by various types of constraints, to describe continuous temporal and spatiotemporal data in arbitrarily high-dimension [6].

Linear equations over rational numbers form a type of constraint in constraint databases (other constraints include linear inequalities over rational numbers or polynomial equations over real numbers). Once linear equations support the storage of continuous curve, constraint databases can serve as a more useful tool for traffic data archiving and query operations. As a nonparametric method, local polynomial regression is a real-time curve model without data pre-classification. It follows the curved tendency of the raw data over the entire estimating region and learns the functions from the dataset [7]. Further, a complicated global regression model can be easily approximated by a local polynomial regression model using the band width and weight.

The aim of our study is to develop the local polynomial regression models to estimate and predict nonlinear average traffic speed with a continuous time-line in linear constraint databases. The development of these new models means that

1530-1311/10 $26.00 © 2010 IEEE
DOI 10.1109/TIME.2010.24

constraint databases have the capability to model and store continuous nonlinear data. In addition, constraint databases have far-reaching potentials to evaluate and analyze the information of moving traffic objects (vehicles and pedestrians) on the basis of statistical nonparametric methods.

Section 2 reviews the related literature. Section 3 discusses the local polynomial regression models and piecewise-linear approximations. Section 4 presents the experiments that test the accuracy of using constraint databases to predict the average traffic speed. Finally, Section 5 gives a brief discussion and some concluding remarks.

2. Literature review

Local polynomial regression without data pre-classification has several advantages. It can avoid the drawbacks of the traditional kernel regression methodologies, such as the Nadaraya-Watson estimator [8, 9] and the Gasser-Müller estimator [10]. The Nadaraya-Watson estimator produces an undesirable bias, and the Gasser-Müller estimator must pay a price in variance to manipulate a random design model. Moreover, local polynomial regression with high curvature adapts well to the bias problems at boundaries [11, 12].

As opposed to the local model, the global model deviates from the data pattern and requires the offline training [7], such as neural networks and time series models. The local modeling is not the approximate function with more accuracy from it, and this feature avoids negative interference exhibited by the global models. Moreover, the local linear method is preferable to the local constant regression in traffic data analysis [7, 13]. About local constant regression, Smith et al. [14] and Faouzi [15] respectively implement the k-nearest neighbor method and kernel estimator in transportation.

Travel time estimation has been an interesting research area related to traffic short-term prediction. [16] describes how speed-based travel time estimation has two application contexts: on-line real-time prediction and off-line historic data analysis. [17] emphasizes the combination of real-time and historic data for dynamic travel time prediction. The method of travel time estimation in [18] relies solely on average speeds.

3. Method

3.1. Local polynomial regression

3.1.1. Definition

There is a basic difference between a parametric approach and a nonparametric approach. The former assumes a pre-specified functional form for the density estimator, while the latter does not. The density estimation in nonparametric regression can effectively describe the overall pattern in a set of data. Suppose that in a sample of random pairs $(x_1, y_1),\ldots, (x_n, y_n)$, the response variable y_i is assumed to satisfy [13]:

$$y_i = m(x_i) + \upsilon^{1/2}(x_i)\varepsilon_i \qquad (1)$$

where m is a function to be estimated; υ is a variance function; ε_i is an independent random variable with zero mean and unit variance; x_i is a random variable having common density f; i =1,…, n.

A local polynomial estimator $\hat{m}(x;p,h)$ [19, 20, and 21] can be developed via "locally" fitting a p^{th} degree polynomial $\sum_{j=0}^{p} \beta_j (x_i - x)^j$ to (x_i, y_i) using weighted least squares. Bandwidth h is assumed to approach zero at a rate slower than n –1, that is:

$$\lim_{n\to\infty} h = 0 \qquad \lim_{n\to\infty} nh = \infty$$

The function of local polynomial estimator for the true function Y is shown below:

$$\hat{Y} = \hat{m}(x;p,h)$$

$$= e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x y = e_1^T \hat{\beta} = \hat{\beta}_0 \qquad (2)$$

where e_1 is a $(p+1)\times 1$ vector having 1 in the first entry and zero elsewhere;

$y = (y_1,\cdots,y_n)^T$ is a vector of responses;

$$W_x = diag\{K_h(x_1 - x),\cdots,K_h(x_n - x)\}$$

is an $n\times n$ diagonal matrix of weights;

$$X_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{bmatrix}$$

is an $n\times(p+1)$ design matrix, n is the number of observations;

$\hat{\beta} = (\hat{\beta}_0,\cdots,\hat{\beta}_p)^T$

is able to minimize the locally weighted polynomial regression $\sum_{i=1}^{n}\{y_i - \sum_{j=0}^{p}\beta_j (x_i - x)^j\}^2 K_h(x_i - x)$;

$K_h(\cdot) = K(\cdot/h)/h$ is a kernel function scaled by h (kernel function is usually a unimodal symmetric probability with $\int K(x)dx = 1$).

Figure 1 illustrates the important aspects of local polynomial regression theory. Y (green curve) is the true model, and $\hat{Y}$ (red curve) is the result of the local polynomial regression. The bandwidth h is a nonnegative number controlling the size of the local neighborhood. $K_h(x_i - x)$ is the weight assigned to y_i , and this weight depends on the height of the kernel function centered about the particular point x. The data closer to x carry more influence in the value of $m(x)$, not assuming a specific form of the regression function m(x). There are some shape choices about kernel function [22], such as Epanechnikov, Biweight, Triweight, Normal, Uniform, Triangular, etc. The shape choice about kernel function is not that important for data estimation and analysis as the bandwidth selection.

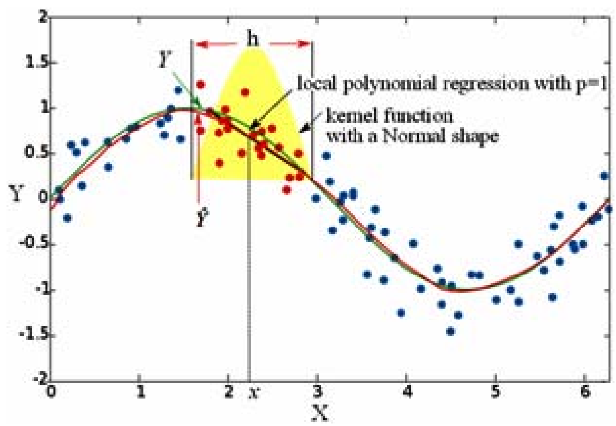

Figure 1. Local polynomial regression model.

3.1.2. Bandwidth selection

The bandwidth choice is particularly important to highlight the significant structure in a set of data. [23] executes a survey of several bandwidth selections for the density estimation, and these selectors are Biased Cross-validation (BCV) [24], Least Squares Cross-validation (LSCV) [25], Rule-of-Thumb (ROT), Solve-the-equation (STE) [19, 26, 27, 28, and 29], and Smoothed Bootstrap [30], and [23] summarizes that ROT has a small variance yet an unacceptable large mean; LSCV has a good mean yet a too large variance; BCV suffers from unstable performance; both STE and smoother bootstrap have a correctly centered distribution in mean and an acceptable variance.

[31] compares three plug-in bandwidth selection strategies [13], such as ROT, STE, and Direct Plug-in (DPI) via data simulation and analysis, and the result is that DPI and STE have the same appealing performance. Moreover, DPI does not need the extra complication of requiring a root-finding procedure and minimization. Hence, DPI is selected as the approach of the bandwidth calculation in our paper. On the basis of several important assumptions, [31] clarifies the rules and calculation steps about DPI bandwidth.

3.1.3. Order choice

In terms of the order of polynomial fit for the asymptotic performance of $\widehat{m}(\cdot;p,h)$, [13] shows that fitting the polynomials of higher order leads to a possible bias reduction and a variance increase, and odd order fits are preferable to even order fits in the problem of the variability increase. Further, even order fits achieve a lower efficiency in a bias reduction, especially in boundary regions and highly clustered design regions. According to the practical performance in many cases, the order of polynomial fits, which are beyond cubic fit, need a very large sample to actualize a significant improvement. Therefore, our study proposes to use p=1 or p=3. A local cubic fit (when p=3) has more degrees of freedom for estimating a high curve region in a set of data than a linear fit (when p=1), although a cubic fit has a higher requirement concerning its calculation and sample variability than a local fit does [22].

3.2. Piecewise-linear approximation

In a piecewise-linear approximation [32] data points (x_i, y_i) with i=1,2,…,n, the relation between the piecewise-linear function $f(x_i)$ and y_i satisfies:

$$|f(x_i) - y_i| \le \Psi \text{ for each } (x_i, y_i) \qquad (3)$$

The maximum error threshold Ψ controls the maximum difference between the original data points and the piecewise-linear function. It means that the original data points are always within a narrow band with width Ψ around the piecewise-linear function, as shown in Figure 2.

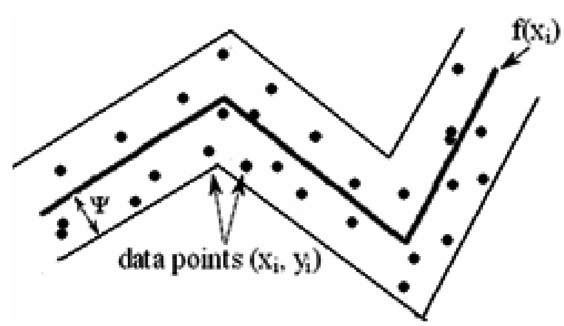

Figure 2. Piecewise-linear approximation.

The piecewise-linear approximation compresses the discrete data points into a piecewise-linear function for data interpolation and faster query. Our study innovatively uses this algorithm to approximately transform a curve into a piecewise-linear form.

4. Experimental results

4.1. Data collection

The values of vehicle speeds used in this research were collected by Cambridge Systematics Inc. at the detector station (717490) of U.S. Highway 101 in Los Angeles, California on June 8, 2005 [33]. Figure 3 shows a schematic of the detector placement for the five lane highway section of this detector station. Each station records the time that each vehicle occupies the detector as it travels over it. The detector also counts the number of vehicles passing over it. The occupancy and the flow rate of vehicles are then used along with the mean vehicle length to determine speed (speed = flow / occupancy / mean vehicle length). These data for occupancy, flow and speed are aggregated and computed over a specified period of time and only the aggregated data are stored. The data at this station are averaged over 5 minute periods. The data points (the open circles in Figure 4) represent the sequential five-minute average speed values in miles per hour (mph).

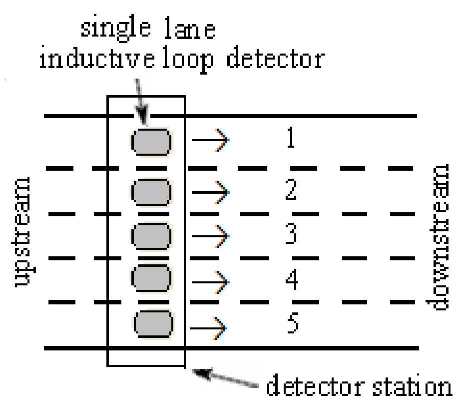

Figure 3. Loop detector station

4.2. Model implementation

The local polynomial models with the bandwidth h=11.5 and *Normal shape* estimate the average speed of vehicles in all lanes. Moreover, the models alter the discrete vehicle speed points into the continuous speed curves. Figure 4 shows the local linear model (black curve) and the local cubic model (red curve). The solid circles in this figure represent the average speed of five lanes reported at the detector station.

Figures 5 and 6 respectively display the speed-time pairs in the local linear and cubic models with those average speed values from the detector station (the continuous curve about the local linear or cubic regression model is divided into 361 data points, and these data points are called *speed-time pairs*). Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are computed for an estimation of accuracy. Their definitions are shown in the following equations, where n is the number of the average speed reported by the detector station from 4:30 pm to 6:30 pm (n=25), Y_i is the average speed from the detector station, and $\hat{Y}_i$ is the speed of the speed-time pairs in the local linear or cubic model:

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \qquad MAE = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n}$$

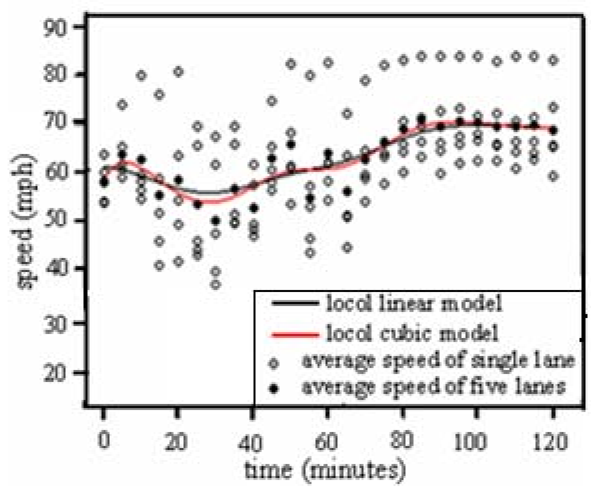

Figure 4. Local linear and cubic models for traffic speed data.

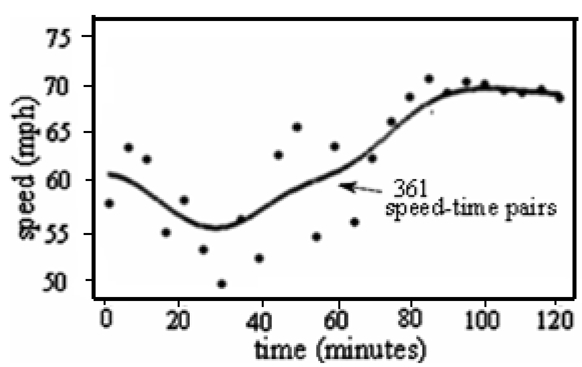

Figure 5. Speed-time pairs in local linear model and average speed from detector.

The accuracy estimations about the linear and cubic models are given in Table 1. The results show that the cubic model is closer to the speed values recorded by the detector station than the linear model. The data patterns in Figures 4, 5, and 6 display that the local cubic fit has more degrees of freedom for estimating a high curve region, which is consistent with the description in [22]. Due to the traffic congestion at the rush hours, the speed is the lowest around 5:00 pm, i.e. x=30, and then the speed begins to rise and has a tendency to level off after 6:00 pm, i.e., x=90. From 4:30 pm (x=0) to 4:40 pm (x=10) and from 5:20 pm (x=50) to 5:40 pm (x=70), the local linear and cubic models show significantly different results (see Figures 5 and 6), and the local cubic model is more accurate in following the raw data.

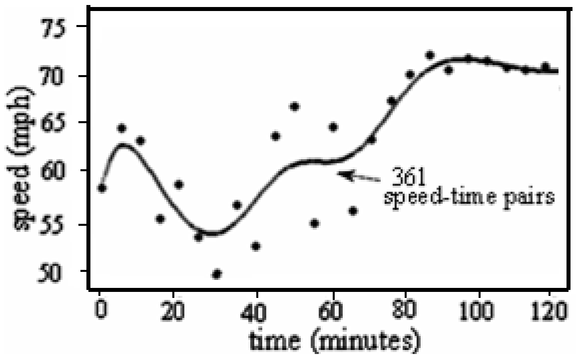

Figure 6. Speed-time pairs in local cubic model and average speed from detector.

Table 1. Model estimation

Model	MSE	RMSE	MAE
Linear	15.014	3.875	0.046
Cubic	7.561	2.75	0.0331

4.3. Database implementation and queries

After using the speed-time pairs in the local polynomial models, the piecewise-linear approximation can transform the continuous speed curves in Figure 4 into the corresponding linear functions / constraints with high accuracy for the storage and query of the speed curves. More importantly, the accuracy can be adjusted and controlled via the error threshold Ψ, and the accuracy is higher with a smaller error threshold.

The query design and results are displayed using the MLPQ system [6]. MLPQ allows Datalog queries, minimum and maximum aggregation operators over linear objective functions, and some other operators. MSE, RMSE, and MAE are also applied for an estimation of accuracy about each sub-function of the piecewise function, where n is the number of the speed-time pairs related to a certain sub-function, Y_i is the speed of the pairs, and $\hat{Y}_i$ is the speed calculated by the sub-function.

Based on the piecewise-linear algorithm with the error threshold Ψ=0.05, the speed-time pairs' data points in Figures 5 and 6 are compressed into some linear functions, which are respectively shown in Tables 2 and 3. The MSE, RMSE, and MAE columns summarize the accuracy analysis of every piecewise-linear function calculated by the piecewise-linear approximation algorithm.

The piecewise-linear segments listed in Tables 2 and 3 are actualized to exert the data analysis in constraint databases. In Figure 7, the MLPQ database system shows the curves similar to those in Figure 4. Table 4 lists the model speed values evaluated by the local linear and cubic regressions and the query results from constraint databases. The query results are very close to the velocity values in the local linear and cubic models. Table 4 also shows that the cubic regression has a better result than the linear regression.

Table 2. Local linear model

Function	Piecewise-linear function		MSE	RMSE	MAE
1	Y = −0.11278*X + 61.20767	X∈[0.0, 4.2]	0.001	0.0317	0.00045
2	Y = −0.24021*X + 61.74287	X∈[4.2, 11.7]	0.00106	0.0325	0.00048
3	Y = −0.27471*X + 62.14656	X∈[11.7, 18.6]	0.0015	0.0387	0.00060
4	Y = −0.16853*X + 60.17166	X∈[18.6, 24.6]	0.00114	0.0337	0.00053
5	Y = −0.02836*X + 56.72327	X∈[24.6, 29.7]	0.00099	0.0314	0.0005
6	Y = 0.10796*X + 52.67477	X∈[29.7, 35.1]	0.001	0.0316	0.0005
7	Y = 0.22043*X + 48.72692	X∈[35.1, 51.0]	0.00093	0.0306	0.00047
8	Y = 0.16662*X + 51.47116	X∈[51.0, 62.4]	0.0014	0.0374	0.00059
9	Y = 0.26155*X + 45.54756	X∈[62.4, 69.9]	0.00106	0.0325	0.00046
10	Y = 0.32257*X + 41.28253	X∈[69.9, 81.3]	0.00142	0.0377	0.00055
11	Y = 0.21982*X + 49.63603	X∈[81.3, 88.2]	0.00116	0.0341	0.00045
12	Y = 0.10476*X + 59.78426	X∈[88.2, 95.4]	0.00111	0.0333	0.00043
13	Y = 0.00948*X + 68.87408	X∈[95.4, 105.3]	0.0011	0.0331	0.00042
14	Y = −0.03706*X + 73.77512	X∈[105.3, 120.0]	0.00031	0.0177	0.00021

Table 3. Local cubic model

Function	Piecewise-linear function		MSE	RMSE	MAE
1	$Y = 1.2905*X + 58.61258$	$X \in [0.0, 1.2]$	0.00095	0.03083	0.00043
2	$Y = 0.82242*X + 59.17428$	$X \in [1.2, 2.7]$	0.00109	0.03296	0.0005
3	$Y = 0.40767*X + 60.29409$	$X \in [2.7, 4.2]$	0.00123	0.03505	0.00052
4	$Y = 0.06462*X + 61.73487$	$X \in [4.2, 6.3]$	0.00117	0.03426	0.00046
5	$Y = -0.23598*X + 63.62872$	$X \in [6.3, 8.7]$	0.00084	0.02895	0.0004
6	$Y = -0.4859*X + 65.803$	$X \in [8.7, 13.8]$	0.00107	0.0327	0.00048
7	$Y = -0.62621*X + 67.73928$	$X \in [13.8, 15.3]$	0.00122	0.03499	0.00051
8	$Y = -0.48699*X + 65.60911$	$X \in [15.3, 20.7]$	0.00107	0.03275	0.00051
9	$Y = -0.31532*X + 62.05559$	$X \in [20.7, 24.6]$	0.00109	0.03308	0.00054
10	$Y = -0.11443*X + 57.11376$	$X \in [24.6, 28.2]$	0.00108	0.03288	0.00054
11	$Y = 0.0943*X + 51.22758$	$X \in [28.2, 31.8]$	0.00104	0.03225	0.00053
12	$Y = 0.28838*X + 45.05575$	$X \in [31.8, 36.0]$	0.00109	0.033	0.00054
13	$Y = 0.41237*X + 40.59212$	$X \in [36.0, 44.7]$	0.00092	0.03026	0.00048
14	$Y = 0.26747*X + 47.06927$	$X \in [44.7, 48.9]$	0.00092	0.0304	0.00045
15	$Y = 0.11142*X + 54.69993$	$X \in [48.9, 53.4]$	0.00086	0.02937	0.00043
16	$Y = 0.02482*X + 59.32448$	$X \in [53.4, 60.9]$	0.00126	0.03553	0.00055
17	$Y = 0.17216*X + 50.35145$	$X \in [60.9, 65.1]$	0.00099	0.03151	0.00045
18	$Y = 0.33456*X + 39.77924$	$X \in [65.1, 69.3]$	0.00096	0.03101	0.00044
19	$Y = 0.46527*X + 30.72121$	$X \in [69.3, 80.1]$	0.00075	0.02731	0.00036
20	$Y = 0.32499*X + 41.95741$	$X \in [80.1, 84.9]$	0.0011	0.0331	0.00043
21	$Y = 0.16552*X + 55.49663$	$X \in [84.9, 89.7]$	0.00103	0.03216	0.00041
22	$Y = 0.02861*X + 67.777$	$X \in [89.7, 95.7]$	0.00108	0.0329	0.00042
23	$Y = -0.06537*X + 76.77111$	$X \in [95.7, 113.7]$	0.00081	0.02847	0.00035
24	$Y = 0.01361*X + 67.79179$	$X \in [113.7, 120.0]$	0.00037	0.01912	0.00023

Table 4. Result comparison

Time	Local linear model			Local cubic model		
	Model speed	Query result	Error	Model speed	Query result	Error
0	61.14	61.21	−0.07	58.61	58.61	0.00
5	60.56	60.54	0.02	62.10	62.06	0.04
10	59.38	59.34	0.04	60.97	60.94	0.03
15	57.97	58.03	−0.06	58.37	58.35	0.02
20	56.71	56.8	−0.09	55.86	55.87	−0.01
25	55.91	56.01	−0.1	54.26	54.25	0.01
30	55.8	55.91	−0.11	54.01	54.06	−0.05
35	56.39	56.45	−0.06	55.13	55.15	−0.02
40	57.49	57.54	−0.05	57.11	57.09	0.02
45	58.72	58.65	0.07	59.08	59.11	−0.03
50	59.78	59.75	0.03	60.29	60.27	0.02
55	60.6	60.64	−0.04	60.66	60.69	−0.03
60	61.39	61.47	−0.08	60.81	60.81	0.00
65	62.43	62.55	−0.12	61.58	61.54	0.04
70	63.86	63.86	0.0	63.30	63.29	0.01
75	65.53	65.48	0.05	65.64	65.62	0.02
80	67.16	67.09	0.07	67.91	67.96	−0.05
85	68.47	68.32	0.15	69.53	69.58	−0.05
90	69.34	69.21	0.13	70.33	70.35	−0.02
95	69.79	69.74	0.05	70.48	70.5	−0.02
100	69.94	69.82	0.12	70.26	70.23	0.03
105	69.87	69.87	0.0	69.89	69.91	−0.02
110	69.69	69.7	−0.01	69.54	69.58	−0.04
115	69.48	69.51	−0.03	69.36	69.36	0.00
120	69.3	69.33	−0.03	69.43	69.42	0.01

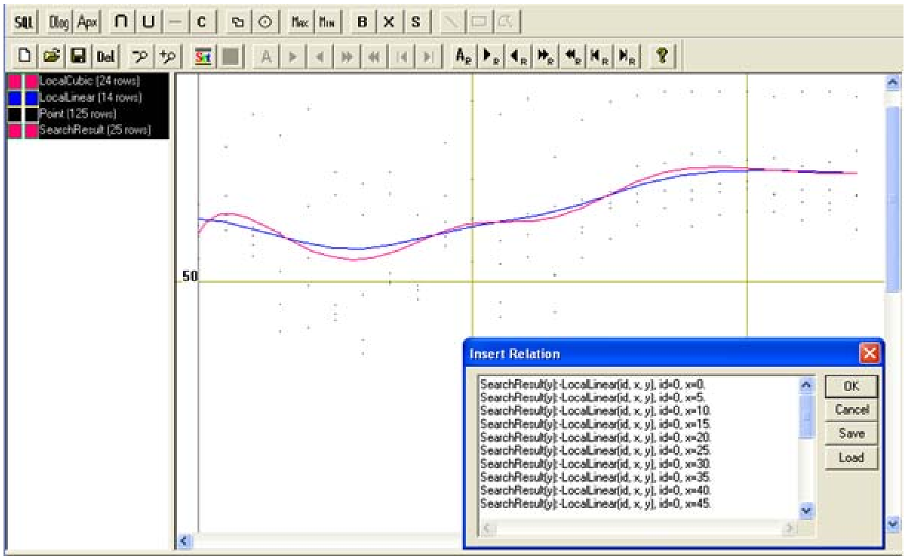

Figure 7. Piecewise model speed in MLPQ.

5. Conclusions

Linear constraint databases are able to store and query discrete data points or linear data. However, research related to the use of constraint databases for nonlinear data storage and advanced statistical modeling is not well developed. Our article proposes a novel and flexible approach to use the local polynomial regression and piecewise-linear approximation algorithm for the estimation and prediction of nonlinear traffic average speed in constraint databases. The experiment results show that this approach has a high accuracy in the storage of continuous nonlinear data in linear constraint databases for a transportation application.

Our study details the definition of local polynomial regression models, their bandwidth selection, and their order choice. The local cubic fit has more degrees of freedom for estimating a high curve region in traffic speed data than the linear fit, and yet the local cubic fit has a higher requirement concerning its calculation. Fortunately, the development of the software package can execute the complex calculations concerning the local cubic model, and overcome this demerit to satisfy traffic data operation in a large data size. More importantly, our approach can make a piecewise-linear function with a high accuracy for any nonparametric regression model.

Future research would concentrate on the micro-simulation of traffic moving objects by storing traffic geographical information and vehicular trajectory data extracted from traffic video. The development of vehicular trajectory data models need to implement spatiotemporal data analysis in constraint databases. Moreover, the analysis of other important traffic parameters, such as volume, flow rate, and queue length, would be also necessary in databases.

6. Acknowledgements

This research was supported by Nebraska Department of Roads Project No. ITS-STWD (66). The authors are solely responsible for the contents of this paper.

References

[1] Bob McQueen, Kan Chen and Rick Schuman. *Advanced Traveler Information Systems*. Artech House Publishers, 2002.

[2] Schofer, J.L., F.S. Koppelman, and W.A. Charlton. Perspectives on Driver Preferences for Dynamic Route Guidance Systems. *Transportation Research Record* 1588, pp. 26-31, 1997.

[3] Linxin Li, Xingyou Zhang, Reinhard Piltner. A Spatiotemporal Database for Ozone in the Conterminous U.S. Proceeding of the *Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)*. Budapest, Hungary, June 15-17, 2006.

[4] Fleming CC, Von Halle B. *Handbook of Relational Database Design*. Addison-Wesley, Reading, pp.605, 1989.

[5] Peter Revesz. Constraint Databases: A Survey. *Semantics in Databases*, Springer Berlin, Vol. 1358, pp. 209-246, 1998.

[6] Peter Revesz. *Introduction to Database: From Biological to Spatiotemporal*. Springer, New York, 2010.

[7] Hongyu Sun, Henry X. Liu, Heng Xiao, Rachel R. He, and Bin Ran. Use of local linear regression model for short-Term traffic forecasting. *Transportation Research Board*, No. 1836, Vol. 18, pp.59-71, 2003.

[8] Nadaraya, E.A. On estimating regression. Theory Probab. Appl. 10, pp.186-190, 1964.

[9] Wathson, G. W. and Leadbetter, M.R. Hazard analysis I. *Biometrika* 51, pp.175-84, 1964.

[10] Gasser, T. and Müller, H.-G. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (eds. T. Gasser and M. Rosenblatt). Springer-Verlag, Heidelberg, pp. 23-68, 1979.

[11] Silverman, B.W. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[12] Fan, J. and Marron, J.S. Fast implementations of nonparametric curve estimators. J. *Comput. Graphical Statist*. 3, pp.35-56, 1994.

[13] Fan J. and Gijbels I. *Local polynomial modeling and its applications*. Chapman & Hall, London, New York, 1996.

[14] Smith, B., Williams, B. & Oswald, K. Parametric and nonparametric traffic volume forecasting. *Transportation Research Record.* CDROM, 1999.

[15] Faouzi, E. Nonparametric Traffic Flow Prediction Using Kernel Estimator. Transportation and Traffic Theory: Proceedings of the *13th International Symposium on Transportation and Traffic Theory*, Lyon, France, pp.24-26, July, 19, 1996.

[16] Ruimin Li, Geoffrey Rose, and Majid Sarvi. Evaluation of speed-based travel time estimation models. *Journal of Transportation Engineering*, Vol. 132, ASCE, 2006.

[17] Steven I-Jy Chien and Chandra Mouly Kuchipudi. Dynamic travel time prediction with real-time and historic data. *Journal of Transportation Engineering*, Vol. 129, ASCE, 2003.

[18] Lu Sun, Jun Yang, and Hani Mahmassani. Travel time estimation based on piecewise truncated quadratic speed trajectory. Transportation Research Part A, Vol. 42, Page 173-186, 2008.

[19] Scott, D.W., Tapia, R.A. and Thompson, J.R. Kernel density estimation revisited. *Nonlinear Anal. Theory Meth. Applic.* 1, pp.339-372, 1977.

[20] Cleveland, W. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.*74, pp.829-836, 1979.

[21] Fan, J. Design-adaptive nonparametric regression. *Journal of the American Statistical Association.* 87, pp.998-1004, 1992.

[22] Wand, M. P. and Jones, M. C. *Kernel Smoothing*. Chapman and Hall, London, 1995.

[23] Jones, M.C. Marron J.S. and Sheather S. J. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association.* 91, No. 433, 1996.

[24] Scott, D.W. and Terrell, G.R.. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association.* 82, pp.1131-1146, 1987.

[25] Bowman, A. W. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71, pp.353-360, 1984.

[26] Sheather, S.J. An improved data-based algorithm for choosing the window width when estimating the density at a point. Comp. *Statist. Data Anal*. 4, pp.61-65, 1986.

[27] Park, B. U. and Marron, J.S. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association.* 85, 66-72, 1990.

[28] Sheather, S.J. and Jones, M.C. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* 53, pp.683-690, 1991.

[29] Engel, J., Herrmann, E. and Gasser, T. An iterative bandwidth selector for kernel estimation of densities and their derivative. *J. Nonparametric Statist*. to appear, 1995.

[30] Faraway, J.J., and Jhun, M. Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association.* 85, pp.1119-1122, 1990.

[31] Ruppert, D., Sheather, S. J. and Wand, M. P. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association,* 90, pp.1257–1270, 1995.

[32] Peter Revesz, Rui Chen, Min Ouyang. Approximate Query Evaluation Using Linear Constraint Databases. Proceeding of the *Eighth International Symposium on Temporal Representation and Reasoning (TIME'01)*. Cividale Del Friuli, Italy, June 14-16, 2001.

[33] Cambridge Systematics, Inc. Prepared for Federal Highway Administration. NGSIM U.S. 101 Data Analysis (4:30 p.m. to 6:30 p.m.), June 8, 2005.