Multiple-Clustering ARMAX-Based Predictor and its Application to Freeway Traffic Flow Prediction

Cheng-Ju Wu, Thomas Schreiter, and Roberto Horowitz

Abstract—An adaptive predictor for a linear discrete time-varying stochastic system is proposed in this paper in order to predict freeway traffic flow at a specific location over a one-hour horizon. Historical sensor data is first clustered by the $K$-means method to obtain the representative data pattern of the sensor. For each $K$-means cluster and using the clusters centroid as the exogenous input, the time-varying output of the sensor is subsequently modeled as an ARMAX stochastic process, and identified in real time using a recursive least squares (RLS) with forgetting factor algorithm. Based on the identified ARMAX model, a $D$-step ahead optimal predictor is generated for each cluster and its associated estimated error prediction variance calculated. The cluster and its associated ARMAX estimate that produces the smallest estimated $D$-step ahead error prediction variance is selected at each sampling time instant to generate the optimal $D$-step ahead predictor of the sensor output. The proposed technique is applied to empirical vehicle detector station (VDS) data to forecast both freeway mainline and on-ramp traffic flow at specific locations over a horizon of one hour. Results indicate that the proposed traffic flow predictor often offers superior flexibility and overall forecast performance compared to using either only historical data or only real-time sensor data on both normal commute days and days when unusual incidents occur.

I. INTRODUCTION

Stochastic prediction and statistical learning techniques are gaining increased attention for managing large complex systems. One such system is a freeway traffic network that is equipped with inductive loop detectors. These loop detectors provide traffic flow data, which are gathered by traffic management centers (TMC) in order to control traffic. The main challenge of a TMC is to reduce congestion, since it leads to wasted time, air pollution, waste of gasoline, and reduced driver safety. Congestion can be either recurrent or non-recurrent. Recurrent congestion occurs regularly during rush hours, which can peak either in the morning or in the evening or in some instances both. Non-recurrent congestion is caused by special events, such as road accidents. To reduce congestion, TMCs rely on accurate real-time traffic information and near-future traffic predictions to evaluate potential traffic management strategies.

A Decision Support System (DSS) system is currently being developed by the Connected Corridors program at the University of California PATH program. With the use of real-time traffic information provided by loop detectors and Lagrangian (mobile) sensors, the DSS will be able to perform model-based near term (e.g., one-hour) forecasting of traffic flow in a corridor and to evaluate a suite of possible management strategies, such as incident response through re-routing and lane management, freeway ramp metering, and arterial signal timing. The current DSS development effort at PATH includes traffic macro-simulation and estimation modules: road network representation, traffic state estimation, fundamental diagram calibration, split ratio prediction, and boundary flow prediction. This paper focuses on the boundary flow prediction module, which predicts the incoming flows at the boundaries of the road network for the near future.

Several stochastic prediction and statistical learning methods have been proposed to forecast traffic flow. Time series models, such as autoregressive moving average (ARMA) and autoregressive moving average with exogenous inputs (ARMAX) models [1] have been widely used in the field of traffic flow prediction. For example [2] proposed using a seasonal autoregressive integrated moving average (ARIMA) model to forecast periodic traffic flow. More recently [3] proposed the use of Kalman filtering techniques to for traffic flow prediction. In the statistical learning approach, neural network models [4], clustering analysis [5], support vector machine [6] and probabilistic graphical models [7] have been used to predict traffic flow. However, statistical learning-based models need more computation time and more parameters than ARMA-like time series models. As the number of parameters in the model increases, it becomes more computationally complex and more difficult to calibrate. In the application of on-line traffic prediction, a model must be both computationally efficient and easy to calibrate. Therefore, an ARMA-like model with fast computation and with small number of parameters is suitable for on-line traffic prediction application.

This paper proposes a multi-clustering ARMAX-based traffic flow predictor, which combines the strength of using large amounts of historical data with using real-time measurement data. The rest of this paper is organized as follows. First, the concept of multi-clustering ARMAX-based prediction is introduced in Section II. The data clustering and model parameter estimation algorithms used in this paper are briefly reviewed in Section III. The multi-clustering ARMAX-based traffic flow predictor is described in Section IV. Section V sets up experiments, which apply the proposed method to empirical freeway traffic data in order to predict the flow one hour ahead. The empirical...
freeway traffic data were obtained through the Performance Measurement System (PeMS)[8]. Results for both recurrent and non-recurrent congestion are presented in Sections VI and VII, respectively. Finally, conclusions and future work are presented in Section VIII.

II. MULTIPLE-CLUSTERING ARMAX-BASED PREDICTION CONCEPT

![Diagram](image)

Fig. 1. Predictions based on one cluster

This section presents a methodology for implementing a multiple-clustering predictor, which can make $D$ steps ahead predictions, given clustered historical data and the most recently measured data of a sensor. Assume that a large amount of historical time series data has been classified into $K$ clusters, with each cluster being characterized by its most representative time series sequence. Let $m \in \{1 \cdots K\}$ be the index associated with $m$th cluster and time series profile $\{u_m(j)\}_{j \in [0,N]}$ for $j = 0, 1, \ldots, N$, be cluster $m$'s representative time series. Let $\{y(j)\}_{j \in [0,k]}$ be the sensor’s most recently obtained data from 0 to the current time step $k < N$. As illustrated in Fig. 1, given $\{u_m(j)\}_{j \in [0,N]}$ and $\{y(j)\}_{j \in [0,k]}$, we will determine a $D$-steps ahead prediction of the sensor output $y_m^p(k+D|k) = E\{y(k+D)|\{y(j)\}_{j \in [0,k]}, \{u_m(j)\}_{j \in [0,N]}\}$, based only on cluster $m$’s representative historical data and a recursively estimated ARMAX model. We will also recursively estimate the $D$-steps ahead error prediction variances

$$E\{(e_m^p(k+D))^2|k\} = E\left\{(y(k+D) - y_m^p(k+D|k))^2\right\}$$

for cluster $m$.

Assuming that optimal $D$-step ahead predictors have been obtained for all $K$ clusters, i.e., $y^p(k+D|k) = \{y_1^p(k+D|k), \ldots, y_K^p(k+D|k)\},$ and their respective estimated error prediction variances, i.e., $E^p(k+D|K) = \{E\{(e_1^p(k+D))^2|k\}, \ldots, E\{(e_K^p(k+D))^2|k\}\},$ the optimal $D$-steps ahead predictor for the sensor output will be selected as the predictor from the cluster that has the smallest estimated error prediction variance, i.e.,

$$y_k^p(k+D|k) = y_{m^*(k,D)}^p(k+D|k)$$

$$m^*(k,D) = \arg\min_m E^p(k+D|K).$$

ARMAX models are primarily used to represent stationary or slowly varying stochastic processes. However, freeway traffic flow is not necessarily a stationary process. For example, daily freeway traffic flow patterns may vary from day to day. In order to deal with daily traffic flow pattern variations, the multiple-clustering ARMAX-based prediction concept presented in this paper utilizes clustering algorithms, such as the $K$-means clustering method, to aggregate daily flow variations into $K$ representative flow patterns, based on large amounts of historical freeway traffic flow data. In each traffic flow cluster, the predominantly non-stationary behavior is captured by the cluster mean, which is used as an exogenous deterministic input to the ARMAX model.

III. DATA CLUSTERING AND ARMAX PARAMETER ESTIMATION

This section presents how typical traffic flow patterns are extracted from a large historical database, and how these are in turn combined with recent data to estimate ARMAX-based $D$-steps ahead predictors. First, the well-known $K$-means algorithm [9] is employed to classify the historical data into $K$ clusters of flow profiles. Each cluster is represented by its centroid data flow for an entire day. Subsequently, for each cluster, an ARMAX model is recursively identified, using the cluster’s centroid as the exogenous input and the actual flow data as the output, and its estimated parameters are used to construct the $D$-steps ahead predictor for the cluster.

A. Data Clustering

Let a vehicle detector station (VDS) measure the flow over the course of $S$ full days. These data are represented as a set

$$M = \{x_1, x_2, \cdots, x_S\}$$

of historical traffic flow data, where each element

$$x_i = [x_i(1), x_i(2), \cdots, x_i(N)]^T \in \mathbb{R}^N$$

is a traffic flow profile of day $i$, for $i = 1, \ldots, S$, with $x_i(k)$ as the flow measured at time step $k$.

Since historical traffic flow data can be used for the traffic flow prediction [5], the most representative traffic flow profiles were extracted from the historical flow data by the $K$-means algorithm. The goal of the $K$-means algorithm is to classify a set of data in (1) into $K$ clusters such that

$$J = \sum_{m=1}^{K} \sum_{i=1}^{S} r_{im}\|x_i - u_m\|^2$$

is minimized, where $\|\cdot\|^2$ denotes the squared Euclidean distance between $x_i$ and $u_m$. The indicator function $r_{im}$ is 1 if $x_i$ belongs to cluster $m$, and 0 otherwise.

$$u_m = [u_m(1), u_m(2), \cdots, u_m(N)]^T \in \mathbb{R}^N$$

is the centroid of cluster $m$, for $m = 1, \ldots, K$, which is seen as the most representative flow profile of cluster $m$. 

4398
B. ARMAX Model Parameter Estimation

In order to model the traffic flow over time for each cluster, an autoregressive moving average with exogenous inputs (ARMAX) model [10] is employed in this research. Consider a recursive ARMAX model for each cluster \( m \) in every time step \( k \) described by the following linear stochastic difference equation

\[
A_m(q^{-1})y(k) = B_m(q^{-1})u_m(k + p) + C_m(q^{-1})w(k)
\]

(5)

where \( y(k) \) is the measured flow of the VDS at time step \( k \), \( u_m(k + p) \) is the historical flow data in cluster \( m \) at time step \( k + p \), \( p \) is the forward shift steps or backward shift steps of \( u_m(k) \), and \( w(k) \) is an innovation sequence, for example a zero-mean white noise. \( A_m(q^{-1}) \), \( B_m(q^{-1}) \) and \( C_m(q^{-1}) \) are scalar polynomials in the backward shift operator \( q^{-1} \) \([q^{-1}y(k) = y(k-1)]\) of orders \( n_a, n_b \) and \( n_c \) in cluster \( m \), respectively, defined by

\[
\begin{align*}
A_m(q^{-1}) &= 1 + a_{1,m}q^{-1} + \ldots + a_{n_a,m}q^{-n_a} \\
B_m(q^{-1}) &= b_0,m + b_{1,m}q^{-1} + \ldots + b_{n_b,m}q^{-n_b} \\
C_m(q^{-1}) &= 1 + c_{1,m}q^{-1} + \ldots + c_{n_c,m}q^{-n_c}.
\end{align*}
\]

(6)

The coefficients of the polynomials in (6) are estimated using an extended recursive least squares identification algorithm [10] with forgetting factor and covariance resetting [11], which is given by

\[
\begin{align*}
\hat{y}_m(k) &= \phi_m^T(k)\hat{\theta}_m(k-1) \\
\hat{\theta}_m(k) &= \hat{\theta}_m(k-1) + P_m(k)\phi_m(k)[y(k) - \hat{y}_m(k)] \\
L_m(k) &= \frac{P_m(k-1)\phi_m(k)\phi_m^T(k)}{\lambda + \phi_m^T(k)P_m(k-1)\phi_m(k)} \\
P_m(k) &= \frac{1}{\lambda}[P_m(k-1) - L_m(k)P_m(k-1)] + \mu I,
\end{align*}
\]

(7) - (10)

where \( 0 < \lambda < 1 \) is the forgetting factor, \( \mu \) is the covariance resetting factor, \( \hat{\theta}_m(k) \) is the unknown parameter vector, \( \phi_m(k) \) is the regressor vector, \( I \) is the identity matrix and \( P_m(k) \) is the symmetric matrix.

Therefore, the estimated system parameters are given by

\[
\begin{align*}
\hat{A}_m(q^{-1}) &= 1 + \hat{a}_{1,m}(k)q^{-1} + \ldots + \hat{a}_{n_a,m}(k)q^{-n_a} \\
\hat{B}_m(q^{-1}) &= \hat{b}_0,m(k) + \hat{b}_{1,m}(k)q^{-1} + \ldots + \hat{b}_{n_b,m}(k)q^{-n_b} \\
\hat{C}_m(q^{-1}) &= 1 + \hat{c}_{1,m}(k)q^{-1} + \ldots + \hat{c}_{n_c,m}(k)q^{-n_c},
\end{align*}
\]

(11)

where the posteriori estimation output \( \hat{y}_m(k) \) and posteriori estimation error \( e_m(k) \) in each cluster \( m \) is given by

\[
\begin{align*}
\hat{y}_m(k) &= -\hat{A}_m^*(q^{-1})y(k) + \hat{B}_m(q^{-1})u_m(k + p) + \hat{C}_m(q^{-1})e_m(k), \\
e_m(k) &= y(k) - \hat{y}_m(k),
\end{align*}
\]

(12) - (13)

with \( \hat{A}_m^*(q^{-1}) = \hat{A}_m(q^{-1}) - 1 \) and \( \hat{C}_m^*(q^{-1}) = \hat{C}_m(q^{-1}) - 1 \). The parameters are updated by an extended recursive least squares (RLS) parameter adaptation algorithm (PAA) in such a way as to make the residual \( e_m(k) \) in (13) converge to an innovation sequence.

IV. Multi-Clustering ARMAX-Based Prediction

Using the polynomials of the ARMAX model estimated in the previous section, a Bezout equation is solved and traffic flow is predicted for each of the \( K \) clusters. The optimal predictor among these \( K \) predictions is selected based on the minimum prediction error variance criterion. This section first explains the predictor itself and then the minimum variance criterion.

A. Bootstrapping Predictor

Assume that the noise polynomial \( \hat{C}_m(q^{-1}) \) estimated by the ARMAX model of (6) is asymptotically stable (i.e., all roots of the polynomial \( \hat{C}_m(z) \) lie outside the unit circle). Given the historical centroid flow data \{\( u_m(j) \)\}_{j \in [0,N]} of cluster \( m \) and current measurement flow data \{\( y(0), \ldots, y(L) \)\}, where \( L \) is the current time step, the optimal predictor of \( y(L + D) \) based on the \( m^{th} \) cluster centroid data, denoted as \( y_m^*(L + D|L) \), satisfies the following difference equation [10]:

\[
\begin{align*}
\hat{C}_m(q^{-1})y_m^*(k + D|k) &= \hat{G}_m(q^{-1})y(k) + \hat{F}_m(q^{-1})\hat{B}_m(q^{-1})u_m(k + D + p) \\
\hat{F}_m(q^{-1}) &= 1 + \hat{f}_{1,m}(k)q^{-1} + \ldots + \hat{f}_{D-1,m}(k)q^{-(D-1)}, \\
\hat{G}_m(q^{-1}) &= \hat{g}_0,m(k) + \hat{g}_{1,m}(k)q^{-1} + \ldots + \hat{g}_{n_a-1,m}(k)q^{-(n_a-1)},
\end{align*}
\]

(14) - (15)

The polynomials \( \hat{G}_m(q^{-1}) \) and \( \hat{F}_m(q^{-1}) \) are uniquely defined by solving the Bezout equation

\[
\hat{C}_m(q^{-1}) = \hat{F}_m(q^{-1})\hat{A}_m(q^{-1}) + q^{-D}\hat{G}_m(q^{-1}).
\]

(17)

Notice that \( \hat{F}_m(q^{-1}) \) and \( \hat{G}_m(q^{-1}) \) depend on \( k \) and \( D \). However, we omitted the indexes for legibility.

Let \( \hat{H}_m(q^{-1}) \) be the product of the two polynomials \( \hat{F}_m(q^{-1}) \) and \( \hat{B}_m(q^{-1}) \):

\[
\hat{H}_m(q^{-1}) = \hat{F}_m(q^{-1})\hat{B}_m(q^{-1})
\]

(18)

\[
\begin{align*}
\hat{H}_m(q^{-1}) &= \hat{b}_0 + (\hat{b}_1 + \hat{f}_1\hat{b}_0)q^{-1} + \ldots + \hat{f}_{D-1}\hat{b}_{n_a}q^{-(D-1+n_a)} \\
&= \hat{h}_0 + \hat{h}_1q^{-1} + \ldots + \hat{h}_{D-1+n_a}q^{-(D-1+n_a)}.
\end{align*}
\]

Since the prediction made in each time step \( k \) can be used as the information for the prediction in the next time step, (14) can be written in the form of a matrix multiplication
called bootstrapping predictor:
\[ y_{m}^{p}(k + D|k) = \begin{bmatrix} y_{m}^{p}(k + D - 1|k) \\ y_{m}^{p}(k + D - 2|k) \\ \vdots \\ y_{m}^{p}(k + D - n_{c}|k) \end{bmatrix} \]
\[ + \begin{bmatrix} \hat{y}_{0,m} \\ \hat{y}_{1,m} \\ \vdots \\ \hat{y}_{n_{a} - 1,m} \end{bmatrix} \begin{bmatrix} y(k) \\ y(k - 1) \\ \vdots \\ y(k - n_{a} + 1) \end{bmatrix} \]
\[ + \begin{bmatrix} \hat{h}_{0,m} \\ \hat{h}_{1,m} \\ \vdots \\ \hat{h}_{D - 1 + n_{b}} \end{bmatrix} \begin{bmatrix} u_{m}(k + D + p) \\ u_{m}(k + D + p - 1) \\ \vdots \\ u_{m}(k + p - n_{b} + 1) \end{bmatrix}. \]

B. Prediction Select Criterion

The bootstrapping predictor in (19) produces one prediction for each of the $K$ clusters at time step $k$. Among these predictions, the optimal one is selected by the minimum estimated prediction error variance criterion, as is defined in the following. The estimation error of cluster $m$ in (13) is defined as
\[ e_{m}(k) = y(k) - \hat{y}_{m}(k). \]

The estimation error variance $E\{e_{m}(k)^{2}\}$ is approximated by
\[ E\{e_{m}(k)^{2}\} \approx W_{m}(k) = \frac{1}{N_{T} + 1} \sum_{i=0}^{N_{T}} (e_{m}(k - i))^{2}, \]
where the integer $N_{T} > 0$ is a parameter to be chosen. We also define the $D$-steps ahead prediction error for cluster $m$ as
\[ e_{m}^{p}(k + D|k) = y(k + D) - \hat{y}_{m}^{p}(k + D|k). \]

Since the $D$-steps ahead prediction error for cluster $m$ cannot be determined at time $k$, its estimated variance will be approximated using $W_{m}(k)$ as follows [10]
\[ E\{e_{m}^{p}(k + D)^{2}\} \approx W_{m}^{p}(k + D) = \sum_{i=1}^{D-1} \hat{f}_{i,m}(k)^{2}W_{m}(k), \]
where $\hat{f}_{i,m}(k)$'s are the coefficients of the polynomial $\hat{F}_{m}(q^{-1})$ in (15).

The optimal $D$-steps ahead predictor is determined as
\[ y_{m}^{p}(k + D|k) = y_{m}^{p_{o}(k,D)}(k + D|k), \]
where $m_{o}(k, D)$ is the index of the cluster with the smallest prediction error estimated variance approximation, i.e.,
\[ m_{o}(k, D) = \arg\min_{1 \leq m \leq K} \{W_{m}^{p}(k + D), \cdots, W_{K}^{p}(k + D)\}. \]

V. Experimental Setup

This section describes the experimental setup, where the proposed traffic flow prediction method is applied to empirical data to predict the near-future traffic flow.

A. Historical Traffic Flow Data

Vehicle detector station (VDS) data along the freeway corridor I-15 North, near San Diego, California, were obtained by the Performance Measurement System PeMS [8] from August 1 to December 31, 2012. The data were aggregated into 15-minute data, which results in 96 observations per day ($N = 96$). These 15-minute data were used to execute the $K$-means clustering for obtaining six historical centroid flow profiles $\{u_{m}(k)_{k \in [0, 96]}\}$, $m = 1, \cdots, 6$ of a VDS.

In order to demonstrate the adaptability of the proposed algorithm, we choose two days of flow measurement data of the mainline VDS 1108592. One day is January 30, 2013, a day when recurrent traffic flow occurred. The other day is January 27, 2013, the day a special sporting event occurred, namely the Super Bowl XLVII game, which led to a non-recurrent traffic flow pattern that differs severely from historical data. The measured flow data of the Super Bowl game day is shown Fig. 9 as the blue line. The flow drastically decreases during the time that the game took place (between 15:00 and 20:00), which deviates significantly from the typical weekend traffic flow pattern.

B. Simulation Parameters

The parameters used for this simulation are as follows. The order of ARMAX model $(n_{a}, n_{b}, n_{c})$ in (6) is $(1, 0, 1)$. The shift of centroid in each cluster $u_{m}(k)$ is $p = 0$ (no shift). The forgetting factor and the regularization factor of the extended least squares recursive parameter estimation algorithm used in this paper can be refer to [11]. As described previously, the number of clusters was selected as $K = 6$. The prediction step was set to $D = 4$, in order to make a one-hour ahead traffic flow prediction.

C. Prediction Performance Evaluation Criterion

To evaluate the prediction performance, the variance of the true prediction error
\[ e^{p}(k + D) = y(k + D) - \hat{y}_{m}^{p}(k + D|k), \]
was approximated by
\[ \text{Var}(k, N_{T}) = \frac{1}{N_{T} + 1} \sum_{i=0}^{N_{T}} (e^{p}(k - i + D))^{2}. \]

In this simulation example $N_{T}$ was set to 10, which means that the estimated prediction error variance is calculated by the data collected over the interval from the current time to two and half hours in the past.
D. Naive Predictor

For comparison purposes, we used a zero order hold predictor, which assumes that the traffic flow remains constant for the next $D$ steps ahead. The $D$ steps naive predictor is therefore given by

$$y_{ZOH}(k + D) = y(k). \quad (28)$$

VI. RESULTS AND ANALYSIS OF A DAY WITH A RECURRENT TRAFFIC PATTERN

Fig. 2 shows the prediction results for a normal day. The flow measurements $y(k)$ are indicated by the solid blue line. The optimal historical cluster centroid profile $u^o(k)$ is shown by the starred black line. Since the actual flow measurements profile of the current day, $y(k)$, are close to the cluster centroid profile $u^o(k)$, the day is considered a regular day. The one-hour flow (four steps) prediction made by the proposed predictor is indicated by the red solid line, which is very close to the actual measured flow data, so only a very small prediction error occurs. Because the regular day does not have any special events or incidents, the current data flow profile ends up closely matching one of the cluster’s historical flow centroid profiles (primarily clusters 3 and 1).

In contrast, the naive predictor shows a large delay with respect to the the current data and the proposed predictor. These results indicate that the use of historical flow data greatly enhances traffic flow prediction in a day with a recurrent traffic pattern.

The starred black line in the top sub-figure in Fig. 3 is the selected cluster at each time step $k$ during the normal day. The bottom sub-figure shows the selected cluster index at time step $k$. As shown in the figure, in this case cluster 3 was the most frequently used cluster for forecasting traffic flow one hour ahead.

VII. RESULTS AND ANALYSIS OF A DAY WITH A NON-RECURRENT TRAFFIC PATTERN

This section presents the results for a day with a non-recurrent traffic flow profile.

A. Parameter Estimation

Using the current flow measurement, $y(k)$, and the centroid of each cluster, $u_m(k)$, at each time step, the coefficients in (11) were estimated by an extended least squares identification algorithm with forgetting factor and covariance resetting. Fig. 4, Fig. 5, and Fig. 6 show the parameter estimates of $A_m(k)$, $B_m(k)$, and $C_m(k)$, respectively, for each of the six clusters.

As the parameter values of $A_m(k)$ and $B_m(k)$ show, clusters 2 and 6 have a different parameter adaptation profile than the other four clusters. In contrast, the parameter values of $C_m(k)$ do not differ much between clusters.

Fig. 7 shows the estimated prediction error variance for each cluster (six dashed color lines labeled as $m = 1, \ldots, 6$),
which are calculated using Eq. (23), and the real prediction error variance (red thick solid line). The optimal input profile $u^o(k)$ at time $k$, which was selected by the minimum prediction error variance criterion (25), is shown as the black starred line in Fig. 8 top. The index of optimal cluster of $u^o(k)$ is shown in the bottom sub-figure in Fig. 8.

The green dashed line in Fig. 7, labeled as $m = 2$, has the lowest estimated prediction error variance between 00:00 and 13:00. Therefore during this period of time, the optimal cluster is cluster 2, and the prediction is made by using the ARMAX parameters of cluster 2, which are also shown in Fig. 8.

However, between 18:00 and 20:00, the lowest estimated prediction error variance is associated with cluster 3, which is caused by the sudden decrease in traffic during the Super Bowl game. As shown in the bottom sub-figure of Fig. 8, the optimal cluster index switches at the beginning of that time period to adapt for the precipitous drop in flow; cluster 3 was the most often selected cluster during this period.

B. Prediction Results

The one hour ahead prediction results are shown in Fig. 9. The four-step zero order hold predictor (naive predictor) $\hat{y}_{ZOH}(k)$ and historical optimal centroid data profile $u^o(k)$ are also shown as a benchmark. The naive predictor (dotted pink line) has an obvious delay at the beginning of the prediction (before 12:00), while the bootstrapping predictor (solid red line) and selected historical centroid profile $u^o(k)$ (starred black line) match well with the current measurement data (solid blue line). However, between 15:00 to 20:00, the flow starts to drop because of the Super Bowl game. Since this special flow pattern differs strongly from the pattern in the historical database, it results in a large prediction error variance, which is shown in Fig. 10. Compared with historical data, the naive predictor and the proposed bootstrapping predictor both forecast the future flow with relatively low variance between 17:00 and 20:00, as shown...
in Fig. 10. Between 20:00 and 22:00, the measured flow increases sharply, and our proposed method predicts this behavior as a small sharp peak, as shown in Fig. 9.

Fig. 9. One hour traffic flow prediction during day with non-recurrent traffic pattern

Fig. 10. Comparison between the prediction error variances of different prediction methods during day with non-recurrent traffic pattern

VIII. CONCLUSIONS AND FUTURE WORK

We developed a traffic flow predictor that is based on empirical real-time flow data from freeways. By clustering the large amount of historical data and extracting the most representative flow pattern (the centroid of a cluster), each data cluster is used to estimate an ARMAX model that is used for making a one hour ahead traffic flow prediction. By applying the minimum prediction error variance criterion, the cluster that leads to the lowest prediction variance estimate was selected as the optimal predictor. Results using empirical traffic data showed that the proposed method has the ability to predict the flow well even when real-time measurements are very different from historical data. The results also show that the proposed predictor outperforms both the zero-order hold predictor and the historical data centroid profile in terms of adaptability and overall accuracy when both recurrent and nonrecurrent traffic conditions are forecasted.

Future research will investigate the correlation between sensors at different locations, i.e., taking the spatial correlation into consideration when making a prediction. Although this paper only used the freeway traffic flow data as an example to implement the proposed methodology, the multi-clustering ARMAX-based predictor can be applied to other large-scale systems that are instrumented with a large number of sensors.

ACKNOWLEDGMENT

This work is supported by California Department of Transportation under the Connected Corridors program and by the National Science Foundation (NSF) through grant CDI-0941326.

REFERENCES