2013 American Control Conference (ACC)
Washington, DC, USA, June 17-19, 2013

An Air Traffic Prediction Model based on Kernel Density Estimation

Yi Cao,[1] Lingsong Zhang,[2] and Dengfeng Sun[3]

Abstract— **This paper revisits a link transmission model that is designed for nationwide air traffic prediction. The prediction accuracy relies on the estimate of traversal time of each link, which is obtained through statistical analysis of historical trajectories. As the most straightforward approach, the average traversal time is often used in the model implementation. But the outliers inherent in the data samples can easily distort the estimate. To address this issue, this paper proposes to use the mode of the traversal times which corresponds to the value reaching the peak of the probability density function of data samples. The continuous probability density function is estimated using a non-parametric approach, kernel density estimation. As the mode is resistant to the outliers, using the mode to parameterize the link transmission model is a more robust approach. Simulations based on historical traffic data of three months show that, in comparison with the conventional mean approach, use of the kernel density estimation in the sector count prediction leads to a 6% reduction in modeling errors.**

I. INTRODUCTION

Air traffic forecasting plays a more and more important role in air traffic management in the face of ever-increasing traffic demand. One way to predict the traffic is by propagating forward each aircraft trajectory through sophisticated flight dynamics [1]. However, for the Air Traffic Control System Command Center who concern more about a high level picture of traffic, detailed modeling of individual aircraft is not necessary and computationally inefficient. Aggregating trajectories into flows and taking advantage of flow properties of the traffic provide an alternative in predicting the traffic.

Traffic can be aggregated at different levels. The Linear Dynamic System Model (LDSM) proposed by Sridhar et al. formulates traffic at an Air Route Traffic Control Center (simply denoted as *Center* hereafter) level [2]. It forecasts aircraft count in each Center of the National Airspace System (NAS) with a time interval of 10 minutes. It is scalable to focus on a smaller airspace volume, such as a Center that consists of several *Sectors*. A flow-based model was proposed in [3], which was later improved to a more efficient one called Link Transmission Model (LTM) [4]. In the LTM, a flight path is defined as a sequence of directed links passing through sectors. An aircraft is assumed to traverse each of the links within an estimated traversal time such that the state of each link can be easily tracked. As a result, aggregation of links in each sector yields a traffic forecast for that sector.

Aggregate models must be parameterized before starting the time evolution of their system dynamics. For instance, in the implementation of LDSM, in order to estimate the inflows and outflows of each Center, the transition probabilities at the Center boundaries must be known [2]. Similarly, the nominal traversal times of links must be specified in the LTM to initialize the model [4]. A straightforward approach to obtain these parameters is to calculate the mean of observed values in history. However, the mean may not be the most representative value when the data samples are not in conformance of symmetric distributions, e.g. Gaussian distribution. An alternative measure may be more appropriate for the most representative value.

We propose to replace the mean with the mode of the underlying probabilistic density function (pdf), which is estimated by Kernel Density Estimation (KDE). KDE is a non-parametric method for estimating the distribution based on a finite set of data samples without any presumptive distributional properties [5]. KDE is widely used in computer vision to identify target object [6], [12]. Application of KDE in transportation can be found in [8] and [9]. Tabibiazar et al. used KDE to extract the congestion spot in road network based on collected car data. Laxhammar et al. used KDE to detect anomalies in the sea traffic. In both applications, the target's position was in conformance with unknown distributions, and KDE was used to approximate the pdf of the position variables so that the target position can be estimated with maximum likelihood. On close inspection of the traversal times of different links in the LTM, the distributional patterns are not always consistent with a single statistical model, thus which model best describes the distribution of traversal times of a particular link is never known a priori. In such a case, KDE is an appropriate tool.

This study is the first attempt to examine the use of KDE in an aggregate air traffic model. The LTM will be first reviewed In Section II. Section III introduces the kernel density estimator in the context of LTM. Numerical results are presented in Section IV, and concluding remarks are given in Section V.

II. THE LINK TRANSMISSION MODEL

A. *Link Representation of the Traffic Network*

The Link Transmission Model is an Eulerian Model [4], which focuses on the flow properties of the traffic rather than the flight dynamics of individual aircraft. The NAS is a highly hierarchical system, which is comprised of three layers, i.e. low altitude, high altitude, and superhigh altitude. In the horizontal direction, each layer are divided into a collection of small control volumes, called *sectors*, as

[1] Yi Cao is with School of Aeronautics and Astronautics, Purdue University. West Lafayette, IN 47907, USA cao20@purdue.edu

[2] Lingsong Zhang is with the Department of Statistics, Purdue University. West Lafayette, IN 47907, USA lingsong@purdue.edu

[3] Dengfeng Sun is with School of Aeronautics and Astronautics, Purdue University. West Lafayette, IN 47907, USA dsun@purdue.edu

978-1-4799-0178-4/$31.00 ©2013 AACC

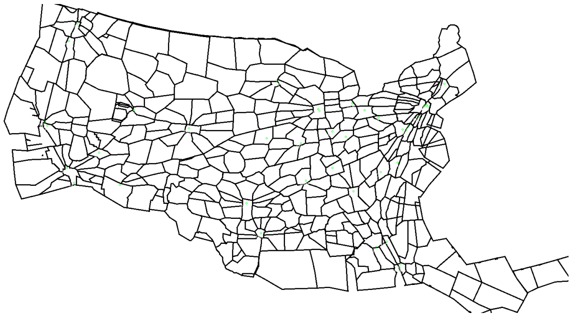
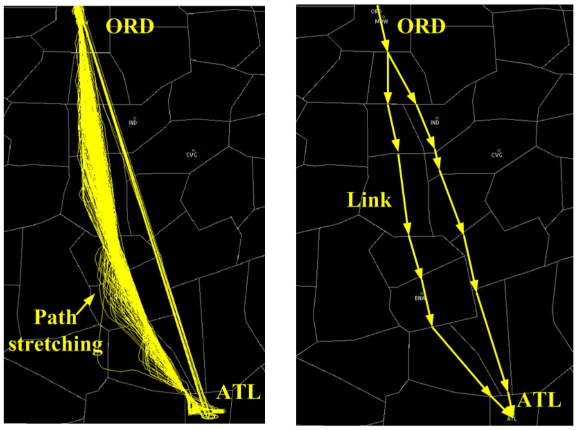


Fig. 1. Sectorization of the National Airspace System in the high altitude layer.



(a) Recorded flight trajectories (b) Link representation of the flows

Fig. 2. Link representation of the observed trajectories.

shown in Fig. 1. The LTM is designed to predict en route traffic rather than terminal operations, all flight operations are projected onto the high altitude layer to simplify the modeling, resulting in a planar representation of traffic flows is obtained, as illustrated in Fig. 2. Extracted from the historical data, flight trajectories from the Chicago O'Hare International Airport to the Atlanta International Airport are clustered into two flows, corresponding to two routings. Each flight path is further divided into a sequence of directed *links* that are abstracted as boundary-to-boundary arcs regardless of the actual flight trajectories inside the sectors. As a result, a link l_j is uniquely identified by a boundary pair. A sector may contain multiple links depending on its geometry. Similarly, a link can be a part of multiple paths that pass through the same boundary pair. A link network can be constructed by mining historical trajectory data, creating a link representation of the traffic network in the NAS.

The LTM is purposely designed as a fast-time NAS-wide traffic simulation tool, so it must avoid sophisticated flight dynamics computation. On a flight path, the traversal of flights is abstracted as one-dimensional movements. Suppose a flight path consisting of a link sequence $P = [l^p_0, l^p_1, \cdots, l^p_j \cdots, l^p_m]$, where p is the path index and j is the link index. The airspeed at which a flight pass through a link l^p_j is a complicated function of multiple parameters, such as aircraft type, gross weight, flight altitude, and so on. But LTM assumes a universal traversal time $t_{l^p_j}$ for all flights traversing the same link. Such simplification creates a fast-time, flight-independent, and flow-level traffic model. As a result, the link sequence corresponds to a traversal time series $t = [t_{l^p_0}, t_{l^p_1}, \cdots, t_{l^p_m}]$. Once taking off, an aircraft k traverses sequentially these links. Since a commercial carrier usually files its flight plan with the air traffic authority three hours before departure, its scheduled departure time t^k_{dep} is known. As a result, its arrival time at the j^{th} link can be predicted by accumulating the traversal times of the links:

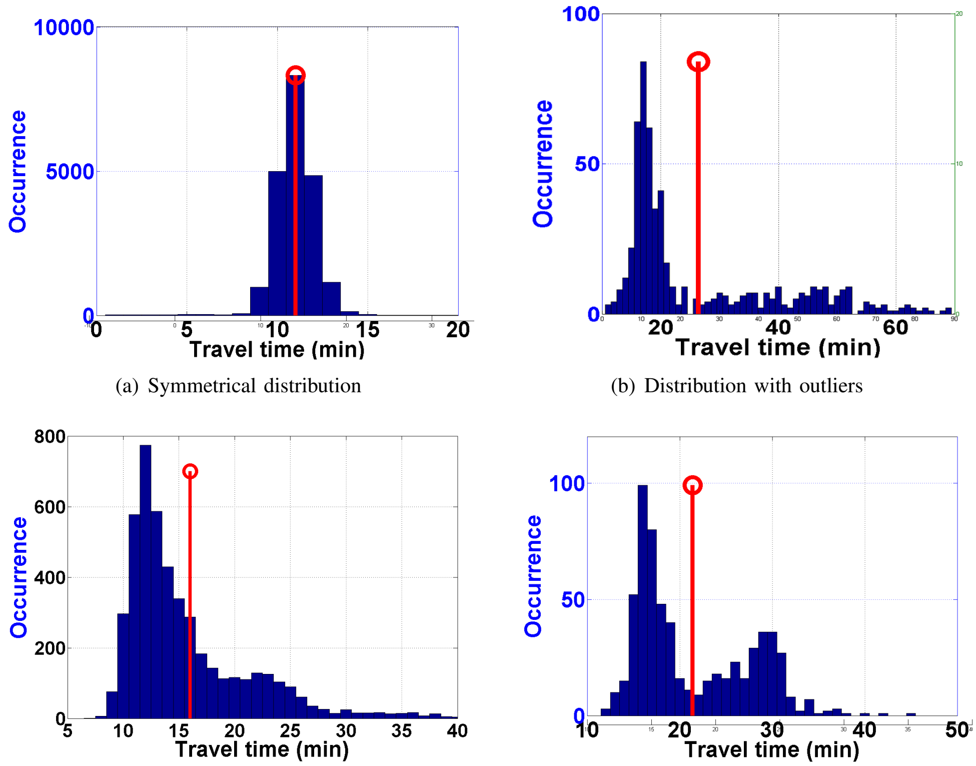
$$t^k_{arr}(j) = t^k_{dep} + \sum_{u=0}^{j} t_{l^p_u} \tag{1}$$

All scheduled flights are predicted in the same way. The traffic moves forward in a deterministic manner. The aircraft count in each sector at time step t is estimated by the following recursion:

$$S(t) = S(t-1) + \sum_{t^k_{arr}(j)=t, l^p_j \in S} 1 - \sum_{t^k_{arr}(j)=t, l^p_j \notin S \&\& l^p_{j-1} \in S} 1 \tag{2}$$

where $l^p_j \in S$ means the j^{th} links on flight path P is in sector S. The second term is the inflow into sector S, and the third term is the outflow. We refer to $S(t)$ as the sector count hereafter.

In the aggregate model, the traversal time is crucial to the prediction accuracy, which is calculated based on historical trajectory data. In an earlier implementation of LTM, the mean was used due to easy calculations [4]. But it is an accurate estimate of the mode only if the distribution of the data set is symmetrical and unimodal. Due to the highly aggregated property, the traversal times present a nature of wide dispersion, where the mean can hardly capture the distributional properties. Fig. 3 shows four typical distribution instances caught in historical traffic data. Fig. 3 (a) is a Gaussian distribution. The mean well represents the mode of the distribution in this scenario. Fig. 3 (b) shows a Gaussian-alike distribution with scattered outliers stretching out to far away from the main cluster. Those outliers cause the mean diverge from the mode. Outliers in this pattern are possibly due to tactic maneuvers like path stretching and air holding, which are used to delay flights for flow management purpose. The length of a delay is a random variable depending on various factors, it thereby results in wide spread of data. In addition, erroneous measurements by radar are also a source of anomalies. Fig. 3 (c) shows an asymmetrical distribution. This pattern is possibly due to the link abstraction. Different trajectories passing through from one boundary to the other are considered the same passage. The irregular geometry of sectors leads to a wide range of possible entrance and exit combinations. Fig. 3 (d) shows a bimodal distribution. It is clear that there are a major cluster and a minor cluster in the histogram. The mean is right between the two clusters, capturing none of the modes. Multimodal distribution can be caused by either a wide range of aircraft types or variations in flight speed at different flight levels.



(a) Symmetrical distribution

(b) Distribution with outliers

(c) Asymmetrical distribution

(d) Multimodal distribution

Fig. 3. **Mean leads to inconsistent performances.**

Given the highly dynamic nature of the air traffic system, asymmetric distributions are very common in the data set. As a result, use of the mean is not a reliable method to estimate the mode of underlying distributions. A robust method should be used to capture the mode of the distribution.

III. Kernel Density Estimation

The Kernel Density Estimation is a non-parametric way to estimate the probability density function $f(x)$ of a random variable x without assuming any distributional property a priori [5]. Given a set of data samples $\{x_0, x_1, \cdots, x_n\}$, one can always use histogram to generate a discrete probability mass function with a predefined resolution. A continuous probability density function by the estimator is expressed in the following form:

$$\hat{f}_{KDE}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) \tag{3}$$

where x_i refers to any of the data samples, and $K(\cdot)$ is the *kernel* function, $K_h(x) = \frac{1}{h}K(x/h)$, usually a symmetric unimodal probability density function, such as Gaussian. Coefficient n is the sample size. Coefficient h is a smoothing parameter that determines the width of the kernel function. (3) sums up the envelope of the kernel function centered at the data samples. The shape of the distribution to be estimated is approximated by the sum. The selection of h is crucial as an inappropriate value will either oversmooth the density function or make it spiky [10]. A comprehensive survey of bandwidth selection can be found in [11]. In this paper, we used the standard Gaussian kernel function $N(0,1)$ for easy implementation and nice theoretical properties:

$$K(\frac{x-x_i}{h}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-x_i}{h})^2} \tag{4}$$

We used the Direct Plug-In (DPI) bandwidth selector proposed in [12] because of its good performance (see in [11]).

Let K be the kernel function, and $f(x)$ be the underlying function. Function $R(K) = \int_{-\infty}^{+\infty} K(z)^2 dz$ measures the "roughness" for the kernel function, and $\sigma^2(K) = \int_{-\infty}^{+\infty} z^2 K(z) dz$ is the kernel variance. An asymptotic optimal bandwidth selector (given f is known) is

$$h_A = \left[\frac{R(K)}{\sigma^2(K)^2 R(f'')n}\right]^{\frac{1}{5}} \tag{5}$$

Note that $R(f'')$ is unknown. If we replace $R(f'')$ by an estimate, then it is called the Direct Plugin (DPI) bandwidth selector:

$$h_{DPI} = \left[\frac{R(K)}{\sigma^2(K)^2 \widehat{R(f'')}n}\right]^{\frac{1}{5}} \tag{6}$$

Note that $\widehat{R(f'')}$ may also depend on the unknown density function $f(x)$. A typical solution is to use several iterations to estimate $f(x)$. See more discussion in [12] and [10].

Once the continuous probability density function is calculated, it is easy to find the mode at which the pdf reaches

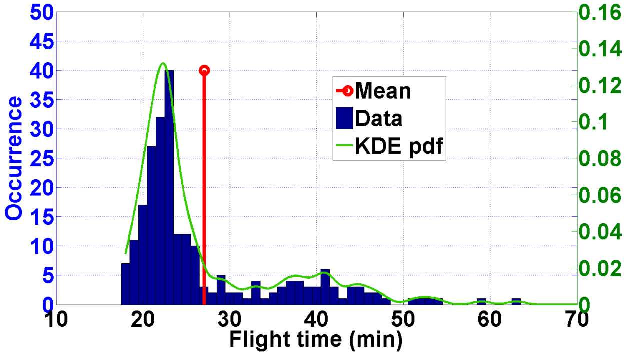


Fig. 4. Distribution of traversal time of a link. The mode of KDE pdf is 22 minutes, and the mean is 27 minutes.

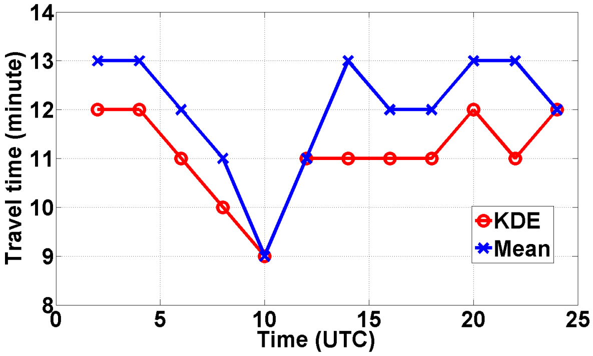


Fig. 6. The peak of KDE and the mean in every two hours prediction timeframe vary throughout the day.

its maximum value:

$$t_{l_j^p} = arg\ max_t \hat{f}_{KDE}(t) \tag{7}$$

An example is shown in Fig 4. The time resolution (time bin) used to generate the histogram and the evolution interval used in the LTM is one minute, mostly equal to the mean of radar update interval which typically ranges from 45 seconds to 70 seconds found in the radar track data. In Fig 4, in total, 226 samples of traversal time extracted from the radar data are fed into the estimator to generate a continuous pdf. The envelop of the continuous pdf matches the histogram well. Its mode (continuous) is close to the mode (discrete) of the histogram. In contrast, the mean is 5 minutes higher than the mode of KDE pdf due to outliers.

IV. Numerical Results

To validate the use of KDE in LTM, a NAS-wide air traffic simulation was conducted and numerical results were statistically examined. Three months (July, August and September in 2005) Aircraft Situation Display to Industry (ASDI) data were used [13], amongst which data of 81 days were used for parameter training, and data of the rest 10 days were used for model validation. There are on average 65,000 flight records found in each day's data. The large population pool is sufficient for constructing a link representation of flight routes between any airport pairs in the NAS. Flights that are destined for or originate from airports outside of the continental United States are considered to fly into/from a fictional "international" sector once they cross the NAS boundaries. This simplification results in a closed airspace system. Furthermore, the transition between the airport space and the sector boundaries are also considered as links. There are 479 sectors defined in the United States high altitude airspace. A network of 18286 links was constructed from the training data set. For each link, the traversal time was obtained by analyzing historical trajectories. Then the mode and the mean of travel times were used to parameterize the LTM respectively, enabling reproduction of the 10 days' traffic based on filed departures. Finally, the reproduced traffic was compared with the actual traffic to evaluate the model performance.

The traffic demand changes throughout the day. To account for the variation, the traversal time of each link is calculated based on samples in every two-hour period, it thus becomes a time variant parameter. Fig. 5 shows the traversal time distributions of a link in different time periods. The actual distribution (histogram) looks Gaussian alike. The KDE modes and the histogram mode agree for most of the time, whereas the mean is bigger than the KDE mode due to asymmetric distributions.

Fig. 6 shows the time evolution of the KDE mode and the mean of a particular link. The time histories of both values are in accordance with the change of traffic demand throughout the day. The traversal time reaches the minimum at time 10:00 AM. From Fig. 5, it is observed that there are less data samples between 4:00 AM through 10:00 AM, indicating low traffic in the early morning. Flights are less likely to be subject to air traffic control during this period. As a result, flights pass the sector quickly, resulting in shorter traversal times. The traversal time increases later on as the airspace gets busy, and maintains at a high level. Fig. 7 compares the KDE modes with the means of all links in the NAS network. The distribution shows that, for the majority of links, the mean is higher than the KDE mode, suggesting asymmetrical distribution with data samples prone to high values. Actually, a few outliers of high value can easily affect the mean if there are not sufficiently large size of data

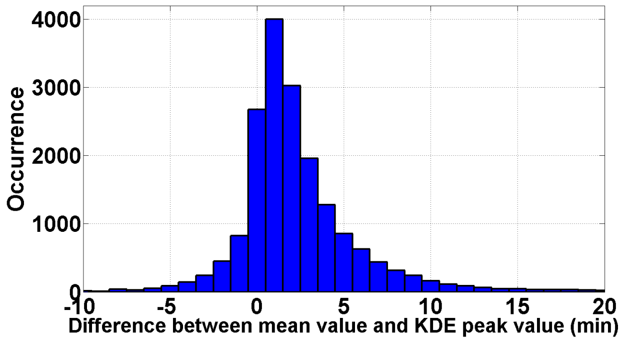


Fig. 7. The distribution of difference between the means and the KDE peaks of all links.

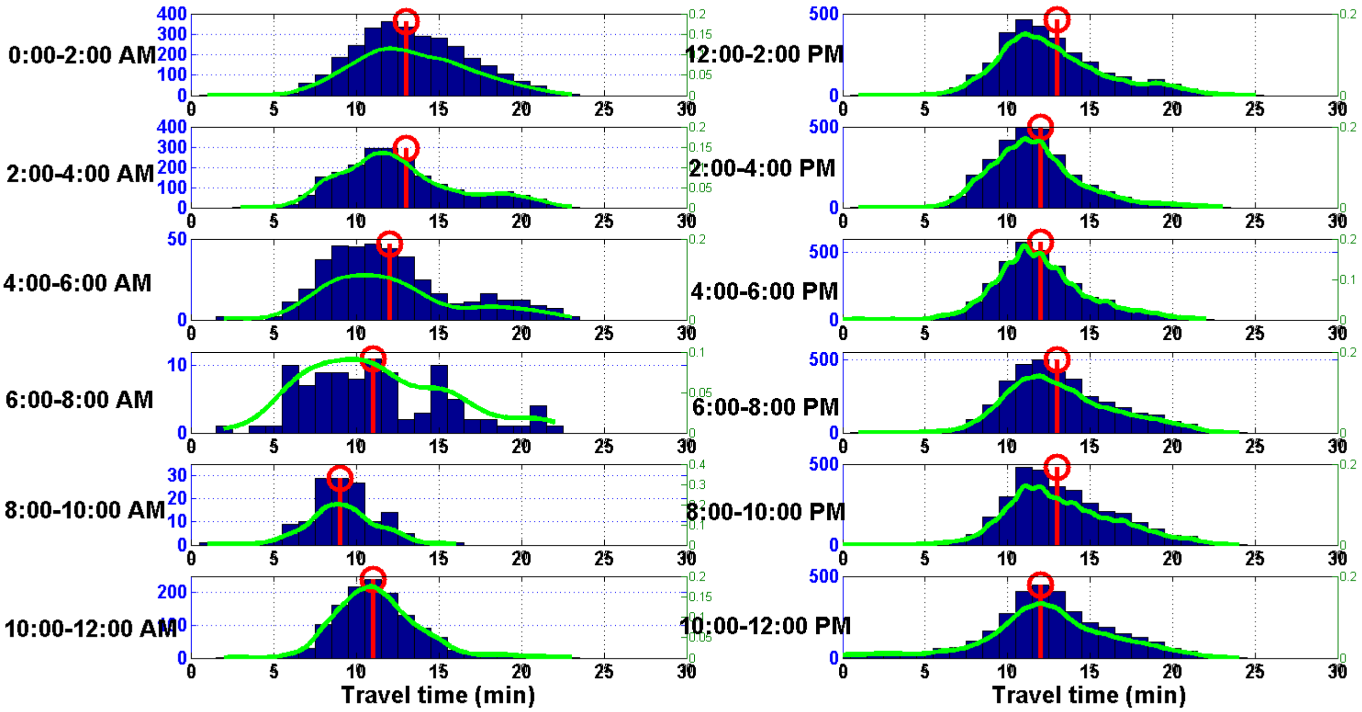


Fig. 5. The traversal time distribution of a particular link varies throughout 24 hours. The mean and the KDE mode change accordingly.

samples.

Fig. 8(a) shows the predicted traffic against the actual traffic in the sector ZAU83 of the Chicago Center during high traffic period. This sector has a large volume of traffic consisting of departures from Chicago Metroplex airports and overflights from neighboring sectors, thus representing a benchmark scenario. L^2 distances are summarized in Fig. 8(b). Overall, the traffic predicted by using KDE is closer to the actual traffic than the mean. To measure the model performance throughout the day, the Euclidean norm, also known as L^2 distance, was used:

$$L^2(\hat{S}) = ||\hat{S}|| = \sqrt{\sum_{t=1}^{T}(S(t)-\hat{S}(t))^2} \tag{8}$$

where $S(t)$ is the observed sector count at time step t, and T is the prediction timeframe. Statistics are shown in Table I. From the last column, It is clear that KDE achieves a lower L^2 distance value than the mean does. The second and third column present the total time (in hour) when the model overestimates and underestimates the traffic. For most of the time, the model associated with mean overestimates the traffic while the model associated with KDE does the opposite.

TABLE I

PREDICTION PERFORMANCE FOR ZAU83 ON SEPTEMBER 21st, 2005

Approach	Overest. (hr)	Underest. (hr)	Precise (hr)	L^2
Mean	13.25	10.25	0.5	231.6
KDE	10	12.75	1.25	196.5

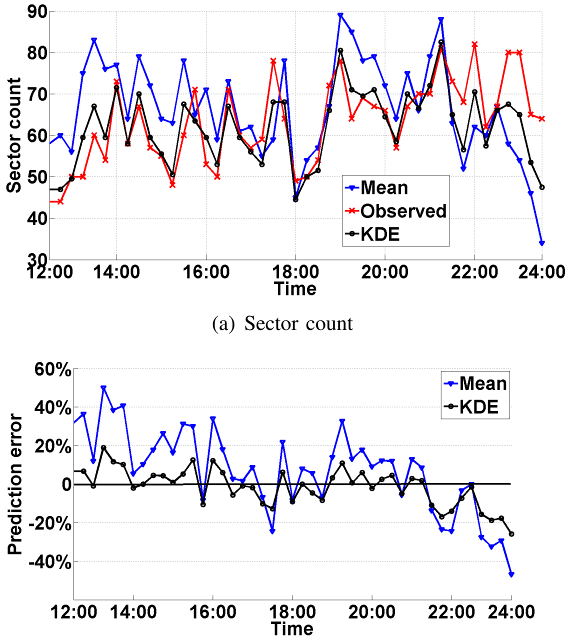
To examine the performance consistency, predictions over all sectors are presented in Fig. 9. To make the comparison intuitive to read, we define a metric, ratio of L^2, for each sector, calculated as follows:

$$Ratio(S) = \frac{L^2_{KDE}(\hat{S})}{L^2_{Mean}(\hat{S})} \tag{9}$$

The sectors are indexed and ordered in terms of $Ration(S)$. Only sectors where traffic is observed are shown. It can be seen that KDE achieves lower prediction errors than the mean in 79% of the sectors, and performs as the same as the mean in 6% of the sectors, and yields higher prediction errors in the rest sectors. The comparison of performance is summarized in Table II. The performance of KDE is quite consistent over the 10 days. In nearly 78% of the sectors KDE outperforms the mean. On average, prediction error associated with KDE is 94% of the prediction error associated with the mean, equivalent to 6% increases in prediction accuracy.

V. CONCLUSIONS

This paper investigates an application of the Kernel Density Estimation in the context of an air traffic prediction model. Numerical results indicates that use of KDE in estimating the probability density function of the traversal



(a) Sector count

(b) Prediction error

Fig. 8. Sector count prediction for ZAU83 on September 21^{st},2005.

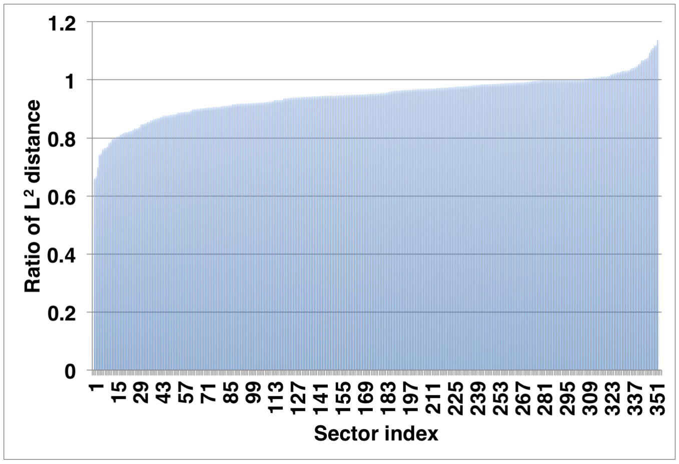


Fig. 9. The ratios of L^2 distance of the 351 sectors, September 21^{st}, 2005.

time of links can precisely capture the mode of the traversal time, resulting in better estimation than the conventional approach where the mean is used. The new approach reduces 6% of the prediction errors. Although the improvement is not very significant, this preliminary study has demonstrated the advantage of use of this statistical method in aggregate air traffic model.

We envision a better performance if a more complicated clustering techniques is employed. In the current implementation, we only used the highest local peak if the distribution is multimodual. As a direction of future work, Gaussian mix model can be used to label the modes, which may have direct relationship to aircraft weight categories or flight configurations. Moreover, the airspace can be divided into different layers, i.e. low altitude, high altitude, superhigh altitude, and the traversal times are further grouped by flight levels. As such, we can incorporate more flight information into the model and choose an appropriate traversal time for a particular flight. In doing so, the prediction errors are expected to further decrease.

TABLE II

STATISTICS OF $Ratio(S)$, SEPTEMBER 21-30, 2005

Date	Sector No.	< 1	$= 1$	> 1	Min	Max	Mean
21	354	79%	6%	15%	0.66	1.14	0.94
22	348	78%	4%	18%	0.64	1.24	0.94
23	350	79%	4%	17%	0.65	1.22	0.95
24	353	78%	5%	17%	0.57	1.24	0.95
25	351	79%	4%	17%	0.60	1.22	0.94
26	355	71%	6%	23%	0.67	1.41	0.96
27	350	80%	4%	16%	0.58	1.21	0.94
28	349	79%	3%	18%	0.66	1.14	0.94
29	355	77%	6%	17%	0.59	1.28	0.95
30	354	79%	6%	15%	0.57	1.19	0.95

REFERENCES

[1] A. M. Bayen, P. Grieder., G. Meyer, C. J. Tomlin., "Lagrangian Delay Predictive Model for Sector-Based Air Traffic Flow," Journal of Guidance, Control, and Dynamics, Vol. 28, No. 5, Sep.-Oct. 2005.

[2] B. Sridhar, T. Soni, K. Sheth, and G.B. Chatterji, "An Aggregate Flow Model for Air Traffic Management," Journal. Journal of Guidance, Navigation, Control. Dynamic, pp. 992-997, Jul-Aug. 2006.

[3] D. Sun, and A.M. Bayen, "Multicommodity Eulerian-Lagrangian Large-Capacity Cell Transmission Model for En Route Traffic," Journal of Guidance, Control and Dynamics. Vol. 31, No.3, May-Jun 2008.

[4] Y. Cao, D. Sun, A link transmission model for air traffic management, AIAA Journal of Guidance, Control and Dynamics. Vol.34, No. 5, 2011.

[5] B.W. Silverman, Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC, 1998.

[6] A. Elgammal, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, Proceedings of the IEEE, Vol. 90, pp. 1151-1163, Jul. 2002.

[7] A. Elgammal, R. Duraiswami, L. S. Davis, Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking, IEEE Transaction on Pattern Analysis and Machine Intelligence. Vol. 25, pp. 1499-1504, Nov. 2003.

[8] A. Tabibiazar, O. Basir, Kernel-based optimization for traffic density estimation in ITS, IEEE Vehicular Technology Conference, pp. 1-5, Dec. 2011.

[9] R. Laxhammar, G. Falkman, E. Sviestins, Anomaly detection in sea traffic - a comparison of the Gaussian Mixture Model and the Kernel Density Estimator, 12^{th} International Conference on Information Fussion, pp. 756-763, July, 2009.

[10] M. P. Wand, M. C. Jones. Kernel Smoothing. London: Chapman & Hall/CRC, 1995.

[11] M. C. Jones, J. S. Marron, S. J. Sheather, A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91 (433): 401407, 1996.

[12] S. J. Sheather, and M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, Journal of the royal Statistical Society, Series B, Vol. 53, pp. 683-690, 1991.

[13] Enhanced Traffic Management System (ETMS), Volpe National Transportation Center, U.S. Department of Transportation, Cambridge, MA, Tech. Rep VNTSC-DTS56-TMS-002, October 2005.