Title

Improving the Resolution of Gridded-Hourly Mobile Emissions: Incorporating Spatial Variability and Hangling Missing Data

Authors
Hicks, J.
Niemeier, Debbie

Publication Date
2001-05-01

PERGAMON

Transportation Research Part D 6 (2001) 153–177

TRANSPORTATION
RESEARCH
PART D

www.elsevier.com/locate/trd

Improving the resolution of gridded-hourly mobile emissions: incorporating spatial variability and handling missing data

J. Hicks [a], D.A. Niemeier [b,*]

[a] *Fehr and Peers Associates, Inc., 3685 Mt. Diablo Blvd., Suite 301, Lafayette, CA 94549, USA*
[b] *Department of Civil and Environmental Engineering, University of California, One Shields Avenue, Davis, CA 95616, USA*

Abstract

To more accurately predict hourly running stabilized link volumes for emissions modeling, a new method was recently developed that disaggregates the period-based model link volumes into hourly volumes using observed traffic count data and multivariate multiple regression (MMR). This paper extends the MMR methodology with clustering and classification analyses to account for spatial variability and to accommodate model links that do not have matching observed traffic count data. The methodology was applied to data collected in the South Air Basin. The spatial analysis resulted in identifying five clusters (or 24-h profiles) for San Diego and two clusters for Los Angeles. The MMR models were then estimated with and without clustering. For San Diego, the disaggregated model volumes with clustering were much closer to the observed volumes than those without clustering, with the exception of the a.m. period. For most hours in Los Angeles, the predicted volumes with clustering were only slightly closer to the observed volumes than those predicted without clustering, suggesting that spatial effects are minimal in Los Angeles (i.e., that 24-h volume profiles are fairly similar throughout the region) and clustering is not necessary. Finally, two classification models, one for San Diego and one for Los Angeles were developed and tested for network link data that does not have matching observed count data. The results indicate the procedure is relatively good at predicting a cluster assignment for the unmatched location for Los Angeles but less accurate for San Diego. © 2001 Published by Elsevier Science Ltd.

1. Introduction

Travel demand forecast models estimate network link traffic volumes by modeling period (e.g., a.m.-peak or p.m.-peak), where periods can include any number of hours. Alternatively,

* Corresponding author. +1-530-752-8918; fax: +1-530-752-7872.
E-mail address: dniemeier@ucdavis.edu (D.A. Niemeier).

1361-9209/01/$ - see front matter © 2001 Published by Elsevier Science Ltd.
PII: S1361-9209(00)00020-1

photochemical air quality models, such as EPA's urban airshed model (UAM), require estimated hourly volumes as input for computing emissions. Traditionally, these hourly emissions are estimated within a post-processor, such as the direct travel impact model (DTIM) developed by the California department of transportation (Caltrans). Here we specifically focus on the running stabilized emissions, where in DTIM, or most currently DTIM3, the disaggregation of period to hourly volumes is typically performed using trip end proportions by hour estimated from travel surveys (California Department of Transportation, 1998).

More recent air quality modeling [1] is taking advantage of a new method to disaggregate the period-based demand model link volumes into hourly volumes (Lin and Niemeier, 1998). This method uses observed traffic count data to stochastically estimate hourly allocation factors representing the expected value of traffic occurring during each hour within a modeling period. These allocation factors can be used to disaggregate the period-based model link volumes into hourly profiles, which can then be directly input into models such as DTIM.

In an early exploratory application of the method comparing emissions totals estimated using the traditional methodology (based on travel survey data) and the stochastic methodology (based on observed count data) in the Sacramento region, it was shown that the traditional methodology could be highly inaccurate in some cases (Niemeier et al., 1999a). For example, the traditional methodology produced an hourly estimate of carbon monoxide emissions 15% higher than estimated by the new methodology. This can obviously have profound implications for transportation conformity analysis.

While the new methodology significantly improves the estimation of hourly running stabilized volumes, to date, its application has been limited in two respects. First, it does not take spatial patterns into account. That is, the estimated allocation factors are assumed to be constant across space. The second limitation is that the method can only be applied to model links with matching count data. Since the vast majority of travel demand network links do not have matching count data (e.g., a permanent automated counter), an extension to the method is needed to account for these locations in the stochastic modeling. This paper outlines a theoretical modeling framework, with an exploratory application, addressing both of these limitations.

We begin the paper with a brief review of the theoretical model. We then describe extensions to the method that allow spatial variability to be incorporated and provide a means for accounting for unmatched network locations. The result of an empirical application to the South Coast Air Quality Basin is discussed in Section 4. Finally, we conclude with a review of the study and discussion of future research efforts.

2. The multivariate multiple regression theoretical model

The new method is based on a multivariate multiple regression (MMR) model that assumes correlation across the observed hourly counts within each modeling period (the reader should see Lin and Niemeier, 1998 and Niemeier et al., 1999a for details). Assume that there are J modeling

[1] For instance, the method is currently being used to develop the new South Coast Basin mobile emissions inventory and data are being collected for use of the method in the Central California Ozone Study.

periods, and T_j is the number of hours within modeling period j. Let i represent any hour within a 24-h period; so $i = 1, \ldots, 24$, and t_j represent the subset of hours i in modeling period j. For example, $j = 1$ might be the a.m.-peak, which might include hours $i = 7, 8$, and 9; so $t_1 = \{7, 8, 9\}$, and $T_1 = 3$. Inherently, $\sum_{j=1}^{J} T_j$ will equal 24. If there are M model links, then the model form for determining the allocation factors would be

$$y_{i,m} = \beta_i x_m^j + \varepsilon_{i,m}, \quad i \in t_j, \; j = 1, \ldots, J, \tag{1}$$

where $y_{i,m}$ is the observed volume at link m for hour i contained in the subset of t_j, x_m the estimated travel demand volume for link m during period j, β_i the proportion of the demand volume occurring during hour i, and $\varepsilon_{i,m}$ represents the model error term. There is no intercept (β_0) included in the equation since the hourly volumes proportions by definition sum to one.

The error terms are distributed with a mean of zero and a variance of σ_i^2, $\varepsilon_{i,m} \sim (o, \sigma_i^2)$; that is, the variance in error terms occurs across hours but not across links. Also, the error terms are independent and identically distributed (i.i.d.). If inferences about the value of β were necessary, then an assumption of normality would also be required.

The β's are unknown parameters and represent the expected value of the proportion of traffic occurring in each hour within a modeling period. The parameters are estimated using an ordinary least squares estimator (OLS) defined by

$$\hat{\beta}_i = \frac{\sum_{m=1}^{M} x_m y_{i,m}}{\sum_{m=1}^{M} x_m^2}. \tag{2}$$

3. Extending the MMR methodology

As noted in the introduction, two extensions to the theoretical framework are necessary: first to account for spatial variability and second to handle missing data. In the first extension, we use clustering analysis to account for spatial variability in the estimated hourly profiles. In the second extension, we draw upon classification analysis as a mechanism for estimating the hourly profiles for locations that do not have matching observed counts.

3.1. Clustering analysis

There are two reasons to add a clustering analysis to the MMR modeling methodology. First, the hourly traffic patterns on any given network link can substantially vary by time of day across the region. Fig. 1 shows two count locations within the San Diego region with distinct 1997 average weekday, hourly profiles.

From Fig. 1, we see that if the a.m.-period model volume for location ID 1 was disaggregated into three hourly volumes (7, 8, and 9 a.m.), then the largest factor would be assigned to 8 a.m., with the next largest to 9 a.m., and then 7 a.m. That is to say that the highest proportion of traffic during the 3-h period occurred during the second hour, followed by the third, then the first. This allocation would be different for location ID 2; where the highest factor would be assigned to 9, 8, and then 7 a.m. As these proportions vary across the hours within a modeling period, it would be

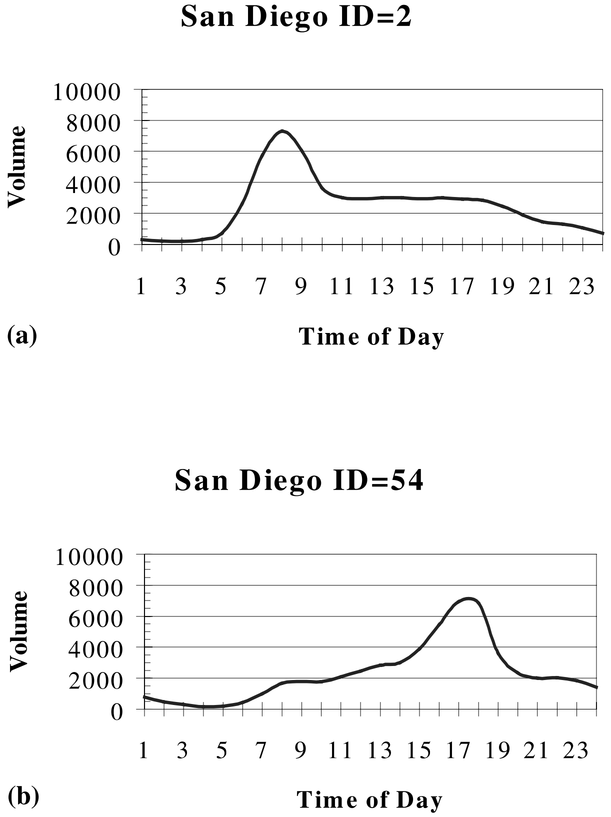

Fig. 1. Temporal traffic patterns of sample counter locations.

more accurate to estimate separate allocation factors for locations of markedly differing temporal patterns. One means for accounting for this spatial variability is by grouping, or clustering, locations based on hourly profiles, with one set of allocation factors estimated for each cluster.

The second advantage of introducing clustering into the MMR procedure is that it provides a basis for assigning allocation factors to unmatched model links. Since unmatched links do not have observed counts available, estimated model period-based volumes for these links cannot be disaggregated into hourly profiles. However, after clusters have been determined for all matched links, the estimated model volumes for the unmatched links can be used to classify the unmatched links into one of the predetermined clusters, and the estimated allocation factors for that cluster can be used to estimate the hourly profiles for the unmatched links.

For the MMR extension, we use the average linkage agglomerative hierarchical clustering method, which can be summarized in three steps:

1. Each of N count locations are assigned to a unique cluster so that each of the N clusters contains a single entity.
2. The statistical (Euclidean) distance between each pair of clusters is calculated.
3. The pair of clusters having the minimum distance between them is merged. There are then $N-1$ clusters remaining after merging.

Steps 2 and 3 are repeated until an appropriate number of clusters are determined based on a suitable criterion. One method that can be used for determining the termination point is the comparison of the pseudo *F*-ratio values, the ratio of the mean square error between clusters to the mean square error within clusters, where a large *F*-ratio is preferred. When there is more than one location within a cluster the Euclidean distance is calculated for each pair of locations, where one member of the pair comes from one cluster and the other member of the pair comes from the other cluster. The average distance between clusters is then used to determine which clusters should be merged (Johnson and Wichern, 1992). In this analysis, observations within clusters are assumed to be uncorrelated. This assumption is likely to be violated on an individual link-by-link analysis, however, if the clusters are sufficiently large, the procedure is robust.

Before clustering locations, the daily temporal traffic pattern at each location is mathematically described in order to calculate the Euclidean distance between each location, or cluster. The 24-hourly traffic counts could be used directly to describe each count location, however, the analysis is simplified by describing the hourly variables in terms of fewer (less than 24) uncorrelated linear combinations or *principal components* (Mardia et al., 1979). Once the principal components are established, the Euclidean distance between the jth and kth location for an analysis containing P principal components is defined as

$$d(j,k) = \left[\sum_{p} \left(Y_{jp} - Y_{kp} \right)^2 \right]^{1/2}, \tag{3}$$

where Y_{jp} denotes the pth principal component for the jth location. The two clusters (initially consisting of one location each) with the minimum Euclidean distance between them are grouped together to form a new cluster and new principal components are calculated for all clusters (or locations). The distances between all remaining $N-1$ clusters are then compared, after which two more clusters are merged. The closest clusters are continually merged until an appropriate number of clusters is determined. The MMR procedure can then be applied to the matched links.

3.2. Modified MMR procedure

Before the MMR model can be estimated, the general form, presented in Eq. (1), must be modified slightly to accommodate the clustering. From Eq. (1), recall that there are J modeling periods, T_j is the number of hours within modeling period j, and i represents any hour within a 24-h period; so $i = 1, \ldots, 24$. Also recall that t_j represents the subset of hours i in modeling period j. Let $c = 1, 2, \ldots, C$ denote the cluster ID's and $k_c = 1, 2, \ldots, K_c$ the index for locations in cluster c, with K_c representing the total number of locations in cluster c. The counter volume observed on the kth location in cluster c at hour i is formulated as

$$y_{i,k_c} = \beta_i^c x_{k_c}^j + \varepsilon_{i,k_c}, \quad i \in t_j, \;\; j = 1, \ldots, J. \tag{4}$$

For example, suppose we wanted to disaggregate the a.m.-peak period from 6 to 9 a.m. ($i = 7, 8, 9$) into hourly volumes. The β's must be estimated separately for each cluster. For cluster c, the equations would be

$$\begin{aligned} y_{7,k_c} &= \beta_7^c x_{k_c}^{\text{a.m.}} + \varepsilon_{7,k_c},\\ y_{8,k_c} &= \beta_8^c x_{k_c}^{\text{a.m.}} + \varepsilon_{8,k_c},\\ y_{9,k_c} &= \beta_9^c x_{k_c}^{\text{a.m.}} + \varepsilon_{9,k_c}. \end{aligned} \tag{5}$$

The resulting estimated allocation factors (for the a.m.-peak case β_7^c, β_8^c, β_9^c) are then used to disaggregate the a.m. period travel demand model volume into three separate hours within the period. For example, with estimated factors of (0.3, 0.2, 0.5), 30%, 20%, and 50% of the a.m. period model volume would be allocated to 7, 8, and 9, respectively. Factors are estimated for remaining a.m.-peak clusters and for all clusters within any remaining modeling periods (e.g. p.m.-peak, off-peak).

3.3. Classification analysis

The final step in the procedure requires consideration of the remaining unmatched links, those travel demand model links having no matching observed counts. This is a common problem since automatic counts and/or manual counts are likely to be far fewer in scope than the roadway system represented by the travel demand network. Without observed counts, the next best method for handling these locations would ideally be to classify the links into one of the predetermined clusters. Once a cluster is identified, the allocation factors for that cluster can then be used to disaggregate the estimated period-based model volumes into hourly volumes.

The procedure for classifying an unmatched link into a cluster is relatively simple. Suppose the period-based model volumes for the unmatched link and the period-based average model volumes for each cluster were plotted on a three-dimensional graph (assuming there are three modeling periods). For instance, if there are five clusters, then six points would be plotted on the graph, one for the unmatched link and one for each of the five clusters. The statistical distance from the unmatched link point to each of the five cluster points is computed, and the unmatched link is classified into minimum distance cluster.

The squared statistical distance from the unmatched link point to the point for cluster i, which is standardized to account for differences in variation among modeling periods, is calculated as follows:

$$D_i^2(x_0) = (\mathbf{x}_0 - \bar{x}_i)^{\mathrm{t}} S_{\text{pooled}}^{-1} (\mathbf{x}_0 - \bar{x}_i), \tag{6}$$

where x_0 represents the vector of model volumes for the unmatched link, $\bar{x}_i$ the vector of average model volumes for cluster i, S_{pooled}^{-1} the inverse of the pooled covariance matrix of the cluster model volumes, and $(\mathbf{x}_0 - \bar{x}_i)^{\mathrm{t}}$ is the transpose of the column vector $(\mathbf{x}_0 - \bar{x}_i)$, the difference between the new model volumes and the average model volumes for links in cluster i.

This statistical distance, based on Fisher's discriminant function, assumes that the volumes within each modeling period and cluster are normally distributed with equal covariance among clusters; hence the use of a pooled covariance matrix. The covariance matrix represents the linear association between each of the variables, or in this case the modeling periods (Johnson and Wichern, 1992). The squared distance represents the actual distance from the model volumes of the new link to the average model volumes for each of the other links (represented by $(\mathbf{x}_0 - \bar{x}_i)$)

standardized to account for differences in variation among modeling periods (represented by $(\mathbf{x}_0 - \bar{x}_i)^{\mathrm{t}} S_{\mathrm{pooled}}^{-1}$).

After determining the distance between the new link and each predetermined cluster, the new link is classified into the cluster that is closest (i.e., the cluster at a minimum distance away). After all of the model links, both matched and unmatched, are assigned to a cluster, the allocation factors already determined from the MMR procedure are used to disaggregate the estimated period-based model volumes into hourly volumes. For example, if the allocation factors for cluster 1 during the a.m. period (7, 8, and 9 h) were 0.2, 0.5, and 0.3, respectively, then the estimated hourly model volumes for link k in cluster 1 would be

$$\begin{aligned} \hat{y}_{7,k_1} &= 0.2x_{k_1}^{\text{a.m.}}, \\ \hat{y}_{8,k_1} &= 0.5x_{k_1}^{\text{a.m.}}, \\ \hat{y}_{9,k_1} &= 0.3x_{k_1}^{\text{a.m.}}, \end{aligned} \tag{7}$$

where $\hat{y}_{i,k_1}$ is the estimated hourly volume for hour i that occurs during the a.m. modeling period and $x_{k_1}^{\text{a.m.}}$ is the a.m. period model volume. This simple multiplication is applied to all links falling into cluster 1 for all 24 hours.

4. Empirical application

The extended MMR methodology was recently applied to data collected as part of a 1997 Southern California Ozone Study (SCOS97) sponsored by the california air resources board (CARB). The purpose of the study was to improve understanding of the formation and movement of ozone within the region. The long-term intent of the study is to help identify future control measures leading to ozone reduction in the South Coast region. Past efforts to reduce ozone levels have been relatively successful, reducing exposure by 80% between 1981 and 1995, but levels are still considered too high. In fact, the South Coast region currently experiences some of the highest levels of ozone in the country (California Environmental Protection Agency, 1997).

The SCOS97 study began on 15 June 1997, culminated on 15 October 1997 and included a 55,000 square-mile region, with boundaries of San Luis Obispo to the north, Mexico to the south, the Pacific Ocean to the west, and the Nevada/Arizona border to the east. The data included ozone readings collected from air monitoring stations, ozone concentrations, and movement information gathered from a lidar system, wind information using radar wind profilers, and weather data gathered by weather balloons and airplanes.

Also as part of this effort, the University of California, Davis was charged with collecting observed traffic data with the intent of implementing the MMR model to improve hourly resolution of running stabilized emissions (Niemeier et al., 1999b). Realtime traffic counts were collected from 15 June to 15 October for 1609 automatic count locations on the MODCOMP system in Los Angeles, Ventura, and Orange counties. Data for San Diego and Imperial counties were collected from 162 automatic count locations beginning on 25 June and ending 15 October.

The Southern California Association of Governments (SCAG) provided travel demand model information for the Los Angeles region and the 1994 model volumes for four modeling periods across 2775 freeway model links. The four modeling periods included the a.m.-peak (6–9 a.m.), midday (9 a.m.–3 p.m.), p.m.-peak (3–7 p.m.), and night time (7 p.m.–6 a.m.). In addition, they provided a list of state-plane coordinates for each of the 1836 freeway model nodes that defined the links. The San Diego Association of Governments (SANDAG) provided similar data for the 1995 San Diego travel demand model. Specifically, model volumes were provided for three modeling periods: a.m.-peak (6–9 a.m.), p.m.-peak (3–6 p.m.), and off-peak (9 a.m.–3 p.m. and 6 p.m.–6 a.m.) for 1196 freeway model links.

Arcview was used to create maps displaying all of the count locations and model nodes, as well as to store pertinent information related to each location, such as an identification number, the nearest cross street, or a freeway name and post-mile description. The map also allowed the matching of the real-time count locations to modeling links. In Los Angeles, there were 2775 freeway travel demand model links, with 1244 matching realtime active count locations. For San Diego, only 140 active count locations could be matched to 1196 freeway model links. In many cases, unmatched links abutted or were very near matched links, which clearly gave rise to the need for a method to assign unmatched links to clusters.

4.1. Clustering results

To begin the analysis each set of 24-hourly volumes, or variables, were reduced using principal components. For both the San Diego and Los Angeles data, the variables reduced to two principal components. The two San Diego components described 91% of the total variance, while the Los Angeles components explained 57% of the total variance (Table 1). For San Diego, each of the remaining PC's contributed less than 4% each towards the total temporal variability. For Los Angeles, additional PC's helped explain a little more of the variability, but it was difficult to interpret these additional PC's in terms of temporal traffic patterns. So for both regions, two PC's replaced the original 24-hourly variables.

The coefficients (eigenvectors) of the two principal components for both regions are given in Table 2. For San Diego, the first two PC's are given by

$$\begin{aligned} Y_1 &= -0.036\,X_1 - 0.017\,X_2 + \cdots - 0.061\,X_{24}, \\ Y_2 &= -0.012\,X_1 - 0.009\,X_2 + \cdots - 0.026\,X_{24}, \end{aligned} \tag{8}$$

where $X_1, X_2, \ldots, X_{24}$ represent the 24-hourly variables. The new variables Y_1 and Y_2 are uncorrelated. Similarly, the PC's for Los Angeles were formulated as

$$\begin{aligned} Y_1 &= -0.068\,X_1 - 0.032\,X_2 + \cdots - 0.137\,X_{24}, \\ Y_2 &= -0.103\,X_1 - 0.076\,X_2 + \cdots - 0.142\,X_{24}. \end{aligned} \tag{9}$$

Looking at the coefficients of the first PC for San Diego, large positive coefficients were found between hours 5 and 9 a.m. (β_6–β_9), large negative coefficients were found between 3 and 6 p.m. (β_{16}–β_{18}), and small magnitudes were found for the remaining coefficients. This PC reflects a strong morning, weak evening peaking characteristic. The second San Diego PC had large positive coefficients between 5 and 7 a.m. (β_6–β_7) and between 3 and 6 p.m. (β_{16}–β_{18}) and large negative coefficients between 9 a.m. and 1 p.m. (β_{10}–β_{13}).

Table 1
Total variance explained by principal components

San Diego region		Los Angeles region	
PC	Variance explained	PC	Variance explained
1	0.84	1	0.36
2	***0.91***	2	***0.57***
3	0.95	3	0.68
–	–	–	–
–	–	–	–
–	–	–	–
24	1.00	24	1.00

For Los Angeles, relatively large positive coefficients were computed between 5 and 9 a.m. (β_6–β_8) and suggest that the first PC has a strong a.m.-peak component. The large positive coefficients between 3 and 7 p.m. (β_{16}–β_{19}) suggest that the second PC has a strong p.m.-peak component.

Table 2
Eigenvectors (coefficients)

Coefficient	San Diego		Los Angeles	
	PC1	PC2	PC1	PC2
β_1	−0.036	−0.012	−0.068	−0.103
β_2	−0.017	−0.009	−0.032	−0.076
β_3	−0.005	0.003	−0.017	−0.062
β_4	0.017	0.014	0.022	−0.082
β_5	0.058	0.059	0.109	−0.189
β_6	0.269	0.286	0.259	−0.309
β_7	0.518	0.420	0.472	0.043
β_8	0.429	0.107	0.534	0.228
β_9	0.278	−0.133	0.336	0.120
β_{10}	0.121	−0.275	0.131	−0.083
β_{11}	0.053	−0.274	0.021	−0.172
β_{12}	0.008	−0.259	−0.038	−0.116
β_{13}	−0.017	−0.227	−0.078	−0.059
β_{14}	−0.034	−0.181	−0.104	−0.056
β_{15}	−0.108	−0.041	−0.093	0.144
β_{16}	−0.245	0.216	−0.107	0.339
β_{17}	−0.345	0.392	−0.137	0.423
β_{18}	−0.333	0.399	−0.114	0.481
β_{19}	−0.189	−0.053	−0.197	0.227
β_{20}	−0.102	−0.125	−0.238	−0.024
β_{21}	−0.087	−0.107	−0.198	−0.127
β_{22}	−0.092	−0.097	−0.164	−0.212
β_{23}	−0.082	−0.078	−0.161	−0.194
β_{24}	−0.061	−0.026	−0.137	−0.142

Table 3
Pseudo F-ratio values by total number of clusters

Region	Total number of clusters									
	1	2	3	4	5	6	7	8	9	10
San Diego	–	274	228	280	298	264	242	227	204	259
Los Angeles	–	64	56	51	57	46	39	35	168	213

Clustering analysis was then performed using the set of principle components. The appropriate number of clusters was determined using the pseudo F-ratio; the pseudo F-ratio is the ratio of the mean square error between clusters to the mean square error within clusters, where a large F-ratio is preferred. Table 3 shows the pseudo F-ratios for a variety of total cluster numbers for San Diego and Los Angeles.

For San Diego, the ratio peaks at five clusters ($F = 298$); for Los Angeles count locations were divided into nine clusters. While the pseudo F-ratio is higher for 10 and 11 clusters (not shown), the sample sizes within these additional clusters were extremely small. Of the nine clusters, seven clusters had very small sample sizes and were later merged into the nearest cluster. This resulted in two clusters for Los Angeles.

The average pattern for each of the five San Diego clusters is shown in Fig. 2. The first cluster (16 locations) exhibits a sharp a.m.-peak, while cluster 5 (10 locations) exhibits a high p.m.-peak. Clusters 2 (62 locations) and 3 (36 locations) had moderate a.m. and p.m.-peaks; however, cluster 2 was higher in the a.m. and cluster 3 was higher in the p.m. Cluster 4 (15 locations) exhibited a mild a.m.-peak with a moderate p.m.-peak.

Of the two clusters formed for the 1238 Los Angeles count locations, 683 counters fell into cluster 1 and 555 into cluster 2. Average 24-hourly traffic patterns for both clusters are shown in Fig. 3. The first cluster exhibits a slightly larger a.m.-peak than p.m.-peak, while the second cluster exhibits a large p.m.-peak and a small a.m.-peak.

4.2. Allocation factor estimation

After clustering the matching count locations, allocation factors were estimated for each cluster and each modeling period. For San Diego, there were a total of five regression models fit for each of the three modeling periods. For Los Angeles, there were two regression models fit for each of the four modeling periods. Table 4 shows the estimated allocation factors for all hours of all modeling periods. Note that for each cluster, the sum of factors within a modeling period must equal one. So for San Diego cluster 1, the factors from 6 to 9 a.m. (β_7–β_9), during the a.m.-peak, sum to one ($0.351+0.368+0.281=1$). In Table 4, the modeling periods for both regions are shaded differently.

Examination of the San Diego coefficients revealed that β_7 estimates for clusters 1 and 2 were larger than the estimates for clusters 3, 4, and 5. Conversely, the β_9 estimates were smaller for clusters 1 and 2. However, the differences between clusters during the p.m.-peak period were not

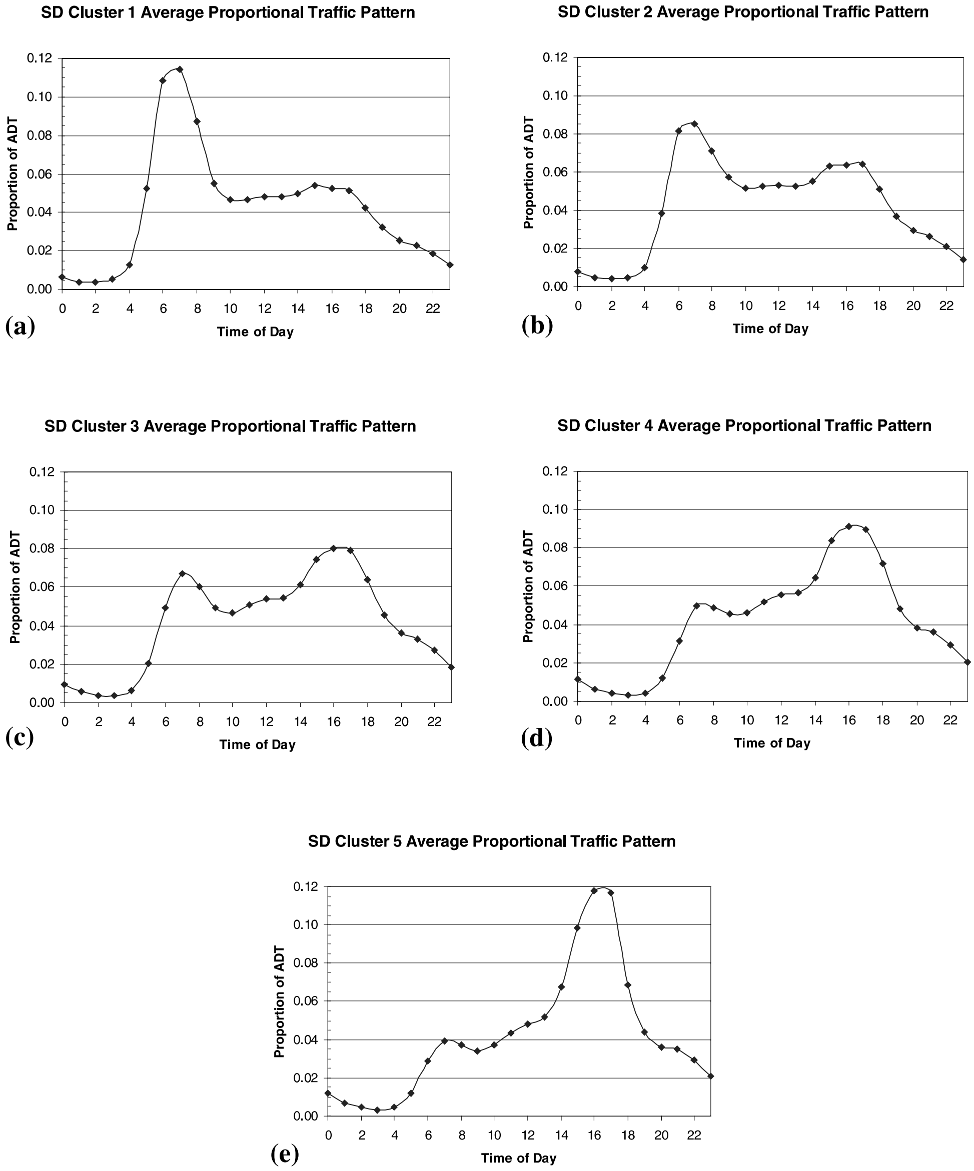

Fig. 2. Average temporal traffic pattern for SD clusters.

so evident. Nor were differences between the three hours detected. The most obvious difference was that cluster 5 had a smaller proportion of traffic from 3 to 4 p.m. (β_{16}) compared to the other clusters. For the 18-h off-peak period, the highest proportion of volume was given to hours from 9 a.m. to 3 p.m. and from 6 to 8 p.m., regardless of the cluster. In addition, cluster 1 had a high proportion for β_6 (5–6 a.m.) leading into the a.m.-peak.

For the Los Angeles a.m.-peak, the allocation factors were higher for cluster 1 during the first two hours (6–8 a.m.) and higher for cluster 2 during the third hour (8–9 a.m.). In addition, the maximum allocation for cluster 1 was from 7 to 8 a.m. (β_8) at 36%. The maximum allocation for

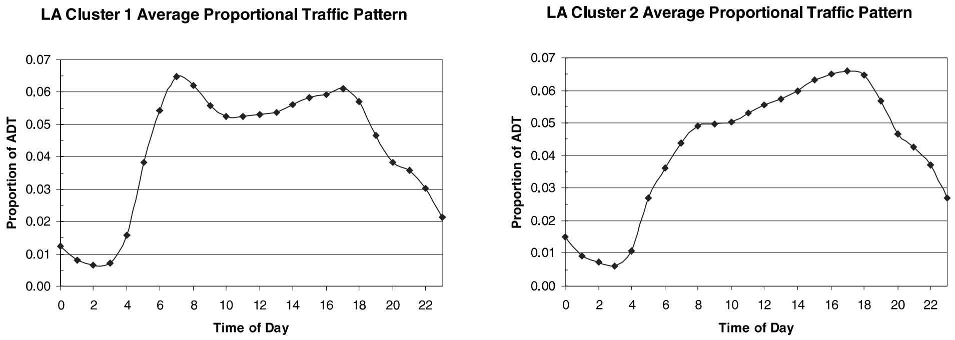

Fig. 3. Average temporal traffic pattern for LA clusters.

cluster 2 was from 8 to 9 a.m. (β_9) at 39%. For the mid-day period, the coefficients were stable over time for cluster 1 with slight increases during the first and last hours (9–10 a.m. and 2–3 p.m.) and increasing over time for cluster 2.

The coefficients for the p.m.-peak period were nearly the same for both Los Angeles clusters and all hours; about 25%. There was, however, a slight decrease in magnitude of the last allocation factor (6–7 p.m.) as the p.m.-peak commuting decreases. Finally, for the night-time period in Los Angeles, the coefficients were highest leading into the a.m.-peak period and out of the p.m.-peak period for both clusters. Otherwise, the factors remained low during all other hours of the night.

Fig. 4 plots predicted hourly volumes with and without clustering and the observed count volumes for location 2 in San Diego. In other words, in the without-clustering case, only one set of hourly allocation factors was estimated for all links. In the with-clustering scenario, different sets were estimated for each cluster. From the figure, it can be seen that the disaggregated model volumes with clustering were much closer to the observed volumes than those without clustering for most of the day, excluding the a.m. period. Similar figures can be drawn for all 138 other locations.

A similar plot for location 21 570 in Los Angeles reveals that the clustering does not seem to substantially improve the accuracy of the allocation factor estimates. Fig. 5 shows that the predicted disaggregated hourly traffic volumes at a randomly selected location were very similar for both with and without clustering, while both were slightly different from the observed volumes. But for most hours, the predicted volumes with clustering were slightly closer to the observed volumes. This could also imply that the clustering was not necessary for the Los Angeles area; that is, hourly profiles do not seem to vary much across the region.

4.3. Classifying unmatched links

Recall that there were large number of unmatched links in both San Diego and Los Angeles. To complete the analysis, these unmatched links were assigned to clusters. To assign the unmatched links to a cluster, the average model volumes for each cluster, the inverted covariance matrices of

Table 4
Allocation factor estimates

Region	Cluster	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
San Diego	1	0.012	0.007	0.007	0.010	0.024	0.098	0.351	0.368	0.281	0.103	0.087	0.087
	2	0.014	0.008	0.007	0.008	0.018	0.065	0.341	0.355	0.304	0.102	0.092	0.093
	3	0.016	0.009	0.006	0.005	0.010	0.032	0.271	0.381	0.348	0.083	0.079	0.086
	4	0.019	0.010	0.007	0.005	0.006	0.018	0.228	0.386	0.386	0.074	0.075	0.084
	5	0.021	0.012	0.008	0.006	0.009	0.023	0.284	0.375	0.341	0.062	0.066	0.077
Los Angeles	1	0.047	0.031	0.025	0.029	0.064	0.153	0.299	0.358	0.343	0.173	0.163	0.163
	2	0.054	0.033	0.026	0.022	0.036	0.090	0.269	0.342	0.389	0.154	0.155	0.163

Region	Cluster	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}	β_{21}	β_{22}	β_{23}	β_{24}
San Diego	1	0.091	0.091	0.093	0.341	0.333	0.326	0.080	0.061	0.047	0.043	0.035	0.024
	2	0.094	0.093	0.097	0.329	0.333	0.338	0.090	0.064	0.051	0.046	0.036	0.024
	3	0.092	0.093	0.104	0.321	0.343	0.336	0.109	0.078	0.062	0.056	0.047	0.031
	4	0.091	0.094	0.108	0.324	0.342	0.334	0.119	0.081	0.064	0.062	0.050	0.034
	5	0.086	0.092	0.120	0.297	0.353	0.350	0.120	0.078	0.065	0.062	0.052	0.037
Los Angeles	1	0.164	0.166	0.171	0.248	0.251	0.257	0.244	0.174	0.144	0.135	0.116	0.081
	2	0.171	0.175	0.182	0.243	0.251	0.256	0.251	0.198	0.164	0.149	0.131	0.096

model volumes, and the unmatched links' model volumes were required to compute Eq. (6) and are shown in Tables 5–8.

With the values given in the above tables, Eq. (6) was formulated for each cluster in each region. For San Diego, there were five equations, each one calculating the distance from an unmatched link to a San Diego cluster. These equations were

$$
\begin{aligned}
D_1^2(x_0) = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 16370 \\ 14312 \\ 35879 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 2.24\text{E}-07 & 4.68\text{E}-08 & -9.45\text{E}-08 \\ 4.68\text{E}-08 & 2.35\text{E}-07 & -1.13\text{E}-07 \\ -9.45\text{E}-08 & -1.13\text{E}-07 & 8.40\text{E}-08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 16370 \\ 14312 \\ 35879 \end{pmatrix}\right], \\
D_2^2(x_0) = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 18011 \\ 17850 \\ 44163 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 2.24\text{E}-07 & 4.68\text{E}-08 & -9.45\text{E}-08 \\ 4.68\text{E}-08 & 2.35\text{E}-07 & -1.13\text{E}-07 \\ 9.45\text{E}-08 & -1.13\text{E}-07 & 8.40\text{E}-08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 18011 \\ 17850 \\ 44163 \end{pmatrix}\right], \\
D_3^2(x_0) = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 13932 \\ 20794 \\ 43742 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 2.24\text{E}-07 & 4.68\text{E}-08 & -9.45\text{E}-08 \\ 4.68\text{E}-08 & 2.35\text{E}-07 & -1.13\text{E}-07 \\ -9.45\text{E}-08 & -1.13\text{E}-07 & 8.40\text{E}-08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 13932 \\ 20794 \\ 43742 \end{pmatrix}\right], \\
D_4^2(x_0) = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 14570 \\ 24594 \\ 50901 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 2.24\text{E}-07 & 4.68\text{E}-08 & -9.45\text{E}-08 \\ 4.68\text{E}-08 & 2.35\text{E}-07 & -1.13\text{E}-07 \\ -9.45\text{E}-08 & -1.13\text{E}-07 & 8.40\text{E}-08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 14570 \\ 24594 \\ 50901 \end{pmatrix}\right], \\
D_5^2(x_0) = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 8785 \\ 18870 \\ 34608 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 2.24\text{E}-07 & 4.68\text{E}-08 & -9.45\text{E}-08 \\ 4.68\text{E}-08 & 2.35\text{E}-07 & -1.13\text{E}-07 \\ -9.45\text{E}-08 & -1.13\text{E}-07 & 8.40E08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{off}} \end{pmatrix} - \begin{pmatrix} 8785 \\ 18870 \\ 34608 \end{pmatrix}\right].
\end{aligned}
\tag{10}
$$

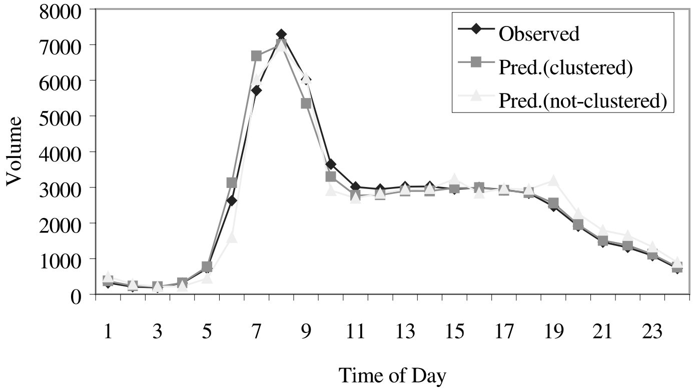

Fig. 4. Comparison of predicted hourly volumes for SD (location 2).

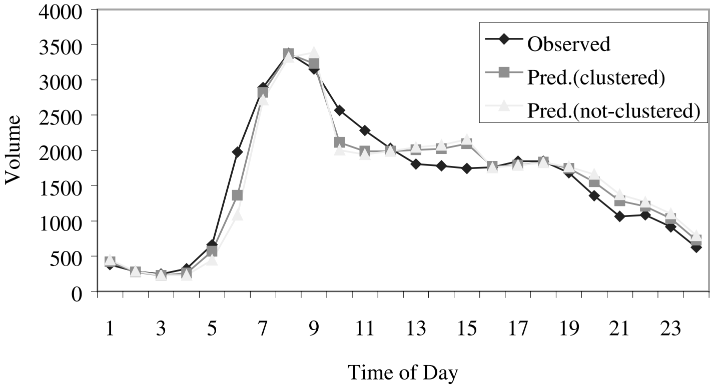

Fig. 5. Comparison of predicted hourly volumes for LA (location 21570).

Table 5
San Diego average model volumes for each cluster

Modeling period	Cluster ID				
	1	2	3	4	5
A.M.-peak	16 370	18 011	13 932	14 570	8785
P.M.-peak	14 312	17 850	20 794	24 594	18 870
Off-peak	35 879	44 163	43 742	50 901	34 608

Table 6
Los Angeles average model volumes for each cluster

Modeling period	Cluster ID	
	1	2
A.M.-peak	19 263	16 170
Mid-day	26 466	27 387
P.M.-peak	24 916	31 028
Night-time	7664	8629

Table 7
Inverted covariance matrix for San Diego

	A.M.	P.M.	Off
A.M.	2.243E−07	4.679E−08	−9.453E−08
P.M.	4.679E−08	2.350E−07	−1.126E−07
Off	−9.453E−08	−1.126E−07	8.404E−08

Table 8
Inverted covariance matrix for Los Angeles

	A.M.	Mid-day	P.M.	Nite
A.M.	1.184E−07	−1.164E−07	5.390E−08	1.190E−08
Mid-day	−1.164E−07	1.586E−07	−8.620E−08	−2.540E−08
P.M.	5.390E−08	−8.620E−08	7.050E−08	4.900E−09
Nite	1.190E−08	−2.540E−08	4.900E−09	5.880E−08

For Los Angles, only two equations were formulated, one for each cluster. They were

$$
\begin{aligned}
D_1^2 = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{mid}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{nite}} \end{pmatrix} - \begin{pmatrix} 19263 \\ 26466 \\ 24916 \\ 7664 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 1.184\text{E} - 07 & -\,1.16E - 07 & 5.39\text{E} - 08 & 1.19\text{E} - 08 \\ -1.164\text{E} - 07 & 1.59\text{E} - 07 & -\,8.62\text{E} - 08 & -\,2.54\text{E} - 08 \\ 5.39\text{E} - 08 & -\,8.62\text{E} - 08 & 7.05\text{E} - 08 & 4.90\text{E} - 09 \\ 1.19\text{E} - 08 & -\,2.54\text{E} - 08 & 4.90\text{E} - 09 & 5.88\text{E} - 08 \end{bmatrix} \\
& \times \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{mid}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{nite}} \end{pmatrix} - \begin{pmatrix} 19263 \\ 26466 \\ 24916 \\ 7664 \end{pmatrix}\right],
\end{aligned}
$$

and

$$
\begin{aligned}
D_2^2 = {} & \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{mid}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{nite}} \end{pmatrix} - \begin{pmatrix} 16170 \\ 27387 \\ 31028 \\ 8629 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 1.184\text{E} - 07 & -\,1.16\text{E} - 07 & 5.39\text{E} - 08 & 1.19\text{E} - 08 \\ -1.164\text{E} - 07 & 1.59\text{E} - 07 & 8.62\text{E} - 08 & -\,2.54\text{E} - 08 \\ 5.39\text{E} - 08 & -\,8.62\text{E} - 08 & 7.05\text{E} - 08 & 4.90\text{E} - 09 \\ 1.19\text{E} - 08 & -\,2.54\text{E} - 08 & 4.90\text{E} - 09 & 5.88E - 08 \end{bmatrix} \\
& \times \left[\begin{pmatrix} x_0^{\text{a.m.}} \\ x_0^{\text{mid}} \\ x_0^{\text{p.m.}} \\ x_0^{\text{nite}} \end{pmatrix} - \begin{pmatrix} 16170 \\ 27387 \\ 31028 \\ 8629 \end{pmatrix}\right].
\end{aligned}
\tag{11}
$$

Before classifying the unmatched links, the equations were tested in two ways using the matched links' model volumes. First, every matched link was classified into a cluster using Eqs. (10) and (11), and the results were compared to the link's cluster determined from the observed counts. A misclassification rate was calculated based on the percent of links not properly classified. Second, modified versions of the above equations were formulated using only 75% of the matched link data. The remaining 25% of the links were then classified, and the results were compared to the clustering results determined from the observed counts. Again, misclassification rates were calculated. Using the first testing procedure, the San Diego model had a misclassification rate of 0.449 (Table 9). For Los Angeles, the misclassification rate was 0.263 (Table 10).

For San Diego, the model generally misclassified a link into a cluster with the next similar pattern. For example, the misclassifications for links in cluster 3 were mostly into clusters 4 and 5. Fig. 4 indicates that clusters 4 and 5 were more like cluster 3 than were clusters 1 and 2. While these links were misclassified, using the wrong allocation factors in these cases would not be as inaccurate as using the factors from clusters 1 or 2.

Comparison of the properly classified and misclassified links in San Diego showed that certain directions were proportionally misclassified more often. The percent of northbound links misclassified was 58%; for westbound, eastbound and southbound, the percentages were 31%, 37%,

Table 9
Classification table for San Diego (test procedure 1)

Correct cluster	Model predicted cluster					Total
	1	2	3	4	5	
1	14 87.5%	2 12.5%	0 0.0%	0 0.0%	0 0.0%	16 100%
2	22 36.1%	33 54.1%	5 8.2%	1 1.6%	0 0.0%	61 100%
3	3 8.3%	3 8.3%	14 38.9%	9 25.0%	7 19.4%	36 100%
4	0 0.0%	0 0.0%	3 20.0%	7 46.7%	5 33.3%	15 100%
5	0 0.0%	0 0.0%	2 20.0%	0 0.0%	8 80.0%	10 100%

Table 10
Classification table for Los Angeles (test procedure 1)

Correct cluster	Model predicted cluster		Total
	1	2	
1	331 73.6%	119 26.4%	450 100%
2	101 26.0%	287 74.0%	388 100%

and 49%, respectively. So northbound links were more likely to be misclassified and westbound links were least likely in proportion to the number of count locations in each direction. Examining differences by freeway revealed that interstate (I) 15 was proportionally misclassified the most (excluding state route (SR) 78 with only one count location) with a rate of 74%. On the other hand, SR-94 had the smallest misclassification rate at 22%. I-5 had the next smallest rate of 31%. Table 11 summarizes the rates for all freeways in the region as well as average total model volumes (sum of three modeling periods).

Fig. 6 shows the relationship between total model volume and misclassification rate. The fitted line demonstrates that as the misclassification rate increases, the total model volume decreases. We observe this same relationship when comparing the average total model volumes for the properly classified and misclassified links. The average total model volume for misclassified links was 73,262 vehicles, with a standard deviation of 24,074. The mean for properly classified links was 82,808 with a standard deviation of 22,889. Clearly the San Diego links with smaller daily volumes had a much higher chance of being misclassified.

Inspection of the Los Angeles classifications showed few differences between the properly and improperly classified links. Comparing flow direction revealed that 17% of the westbound links were misclassified. Eastbound links, on the other hand, had a misclassification rate of 24%. North and southbound links had similar rates of misclassification (~30%). Many differences were detected in the misclassification rates by freeway. Table 12 displays all of the rates for each freeway. Ignoring the extremes, which are influenced by very small samples sizes, the highest misclassification rates (>35) were found for interstate I-5, SR-118, SR-133, SR-134, SR-170, and I-605.

Table 11
Misclassification rates for SD freeways

Freeway	Misclassification rate (%)	Total no. count locations	Average total model volume	Standard deviation of total model volumes
I-5	31	29	88 949	11 599
I-8	42	26	73 306	35 384
I-15	74	31	87 740	23 482
SR-78	100	1	49 728	–
SR-94	22	27	66 661	11 064
SR-125	67	3	34 426	18 316
SR-163	63	8	64 435	12 077
I-805	38	13	89 380	9991

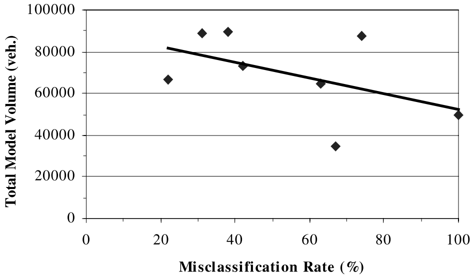

Fig. 6. Misclassification rate by total model volume for San Diego.

Table 12
Misclassification rates for LA freeways

Freeway	Misclassification rate (%)	Total no. count locations	Average total model volume	Standard deviation of total model volumes
I-5	41	126	84 612	26 385
I-10	18	111	92 186	17 408
SR-55	32	19	80 290	12 575
SR-57	28	32	78 436	8743
SR-60	4	46	77 557	6468
SR-71	0	1	25 909	–
SR-91	22	69	78 874	10 287
US-101	26	89	89 621	23 209
I-105	30	30	55 135	14 350
I-110	32	41	86 377	22 707
SR-118	45	31	43 670	12 574
SR-133	50	2	9398	1309
SR-134	36	25	74 401	6323
SR-170	40	5	67 662	4271
I-210	9	43	68 376	13 116
I-405	14	93	90 217	16 649
I-605	40	45	78 143	14 517
I-710	33	30	76 446	12 638

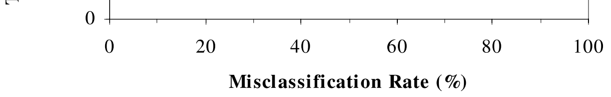

Fig. 7. Misclassification rate by total model volume for Los Angeles.

Inspection of Fig. 2 reveals that these freeways are distributed fairly evenly over the network. Interestingly, I-5 had the second smallest misclassification rate in the San Diego region and one of the highest in the Los Angeles region.

Fig. 7 demonstrates the relationship between the total model volumes and the misclassification rates. Like San Diego, a negative correlation is suggested between total model volume and misclassification rate; however, of a lesser magnitude. The average total model volume for all misclassified links was 78,178 with a standard deviation of 23,378. The mean and standard deviation for all properly classified links were 81,956, and 20,438, respectively.

For the second test procedure, Eqs. (10) and (11) were reformulated using 75% of the data, randomly chosen. The modified equations for San Diego were

$$D_1^2(x_0)=\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}18014\\ 15207\\ 39511\end{pmatrix}\right]'\begin{bmatrix}1.88\mathrm{E}-07 & 4.25\mathrm{E}-08 & -8.40E-08\\ 4.25\mathrm{E}-08 & 2.15\mathrm{E}-07 & -1.01\mathrm{E}-07\\ 8.40\mathrm{E}-08 & -1.01\mathrm{E}-07 & 7.83\mathrm{E}-08\end{bmatrix}$$
$$\times\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}18014\\ 15207\\ 39511\end{pmatrix}\right],$$

$$D_2^2(x_0)=\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}17712\\ 17512\\ 43231\end{pmatrix}\right]'\begin{bmatrix}1.88\mathrm{E}-07 & 4.25\mathrm{E}-08 & -8.40\mathrm{E}-08\\ 4.25\mathrm{E}-08 & 2.15\mathrm{E}-07 & -1.01\mathrm{E}-07\\ -8.40\mathrm{E}-08 & -1.01\mathrm{E}-07 & 7.83E-08\end{bmatrix}$$
$$\times\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}17712\\ 17512\\ 43231\end{pmatrix}\right],$$

$$D_3^2(x_0)=\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}14166\\ 21337\\ 44575\end{pmatrix}\right]'\begin{bmatrix}1.88\mathrm{E}-07 & 4.25\mathrm{E}-08 & -8.40\mathrm{E}-08\\ 4.25\mathrm{E}-08 & 2.15\mathrm{E}-07 & -1.01\mathrm{E}-07\\ -8.40\mathrm{E}-08 & -1.01E-07 & 7.83\mathrm{E}-08\end{bmatrix}$$
$$\times\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}14166\\ 21337\\ 44575\end{pmatrix}\right],$$

$$D_4^2(x_0)=\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}13524\\ 23478\\ 47986\end{pmatrix}\right]'\begin{bmatrix}1.88\mathrm{E}-07 & 4.25\mathrm{E}-08 & -8.40\mathrm{E}-08\\ 4.25\mathrm{E}-08 & 2.15\mathrm{E}-07 & -1.01\mathrm{E}-07\\ -8.40\mathrm{E}-08 & -1.01E-07 & 7.83\mathrm{E}-08\end{bmatrix}$$
$$\times\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}13524\\ 23478\\ 47986\end{pmatrix}\right],$$

$$D_5^2(x_0)=\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}8629\\ 18353\\ 33716\end{pmatrix}\right]'$$
$$\times\begin{bmatrix}1.88\mathrm{E}-07 & 4.25\mathrm{E}-08 & -8.40\mathrm{E}-08\\ 4.25\mathrm{E}-08 & 2.15\mathrm{E}-07 & -1.01\mathrm{E}-07\\ -8.40\mathrm{E}-08 & -1.01E-07 & 7.83\mathrm{E}-08\end{bmatrix}\left[\begin{pmatrix}x_0^{\text{a.m.}}\\ x_0^{\text{p.m.}}\\ x_0^{\text{off}}\end{pmatrix}-\begin{pmatrix}8629\\ 18353\\ 33716\end{pmatrix}\right]. \qquad (12)$$

For Los Angeles the two modified equations were

$$
\begin{aligned}
D_1^2 = {} & \left[\begin{pmatrix} x_0^{\mathrm{a.m.}} \\ x_0^{\mathrm{mid}} \\ x_0^{\mathrm{p.m.}} \\ x_0^{\mathrm{nite}} \end{pmatrix} - \begin{pmatrix} 19605 \\ 26612 \\ 25078 \\ 7499 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 1.12\mathrm{E}-07 & -1.19\mathrm{E}-07 & 5.33\mathrm{E}-08 & 1.32\mathrm{E}-08 \\ -1.19\mathrm{E}-07 & 1.58\mathrm{E}-07 & -8.31\mathrm{E}-08 & -2.65\mathrm{E}-08 \\ 5.33\mathrm{E}-08 & -8.31\mathrm{E}-08 & 6.59\mathrm{E}-08 & 6.40\mathrm{E}-09 \\ 1.32E-08 & -2.65\mathrm{E}-08 & 6.40\mathrm{E}-09 & 6.00\mathrm{E}-08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\mathrm{a.m.}} \\ x_0^{\mathrm{mid}} \\ x_0^{\mathrm{p.m.}} \\ x_0^{\mathrm{nite}} \end{pmatrix} - \begin{pmatrix} 19605 \\ 26612 \\ 25078 \\ 7499 \end{pmatrix}\right],
\end{aligned}
$$

$$
\begin{aligned}
D_2^2 = {} & \left[\begin{pmatrix} x_0^{\mathrm{a.m.}} \\ x_0^{\mathrm{mid}} \\ x_0^{\mathrm{p.m.}} \\ x_0^{\mathrm{nite}} \end{pmatrix} - \begin{pmatrix} 16007 \\ 27315 \\ 31040 \\ 8773 \end{pmatrix}\right]' \\
& \times \begin{bmatrix} 1.12\mathrm{E}-07 & -1.19\mathrm{E}-07 & 5.33\mathrm{E}-08 & 1.32\mathrm{E}-08 \\ -1.19\mathrm{E}-07 & 1.58\mathrm{E}-07 & -8.31\mathrm{E}-08 & -2.65E-08 \\ 5.33\mathrm{E}-08 & -8.31\mathrm{E}-08 & 6.59\mathrm{E}-08 & 6.40E-09 \\ 1.32\mathrm{E}-08 & -2.65\mathrm{E}-08 & 6.40\mathrm{E}-09 & 6.00\mathrm{E}-08 \end{bmatrix} \left[\begin{pmatrix} x_0^{\mathrm{a.m.}} \\ x_0^{\mathrm{mid}} \\ x_0^{\mathrm{p.m.}} \\ x_0^{\mathrm{nite}} \end{pmatrix} - \begin{pmatrix} 16007 \\ 27315 \\ 31040 \\ 8773 \end{pmatrix}\right].
\end{aligned}
\tag{13}
$$

Using the second testing procedure, the misclassification rates for San Diego and Los Angeles were 0.448 and 0.268. These values are almost identical to those determined using the first testing method. Tables 13 and 14 summarize the results for test procedure 2. For observed (correct) clusters 3 and 4, the same numbers of links were classified into clusters 3 and 4. However, both data sets were very small.

Upon completion of the matched link analysis, Eqs. (10) and (11) were used to classify the unmatched links. We also determined the probability that any given unmatched link was properly classified. The probability P that the unmatched link with model volumes x_0 was properly classified into cluster i was calculated as

$$
P\langle i|x_0\rangle = \frac{\exp\left(0.5D_i^2(x_0)\right)}{\sum_j \exp\left(0.5D_j^2(x_0)\right)}, \tag{14}
$$

where $D_k^2(x_0)$ is the standardized squared distance from the new link to cluster k. For San Diego, 30% of the unmatched links were classified into cluster 1, 21% to cluster 2 and 25%, 7%, and 17% to clusters 3, 4, and 5, respectively. The actual number of links in each cluster is shown in Table 15. For Los Angeles, 1261 (68%) links were classified into cluster 1 and 599 (32%) into cluster 2.

Table 13
Classification table for San Diego (test procedure 2)

Correct cluster	Model predicted cluster					Total
	1	2	3	4	5	
1	5 83.3%	1 16.7%	0 0.0%	0 0.0%	0 0.0%	6 100%
2	3 25.0%	6 50.0%	2 16.7%	0 0.0%	1 8.3%	12 100%
3	1 14.3%	1 14.3%	2 28.6%	2 28.6%	1 14.3%	7 100%
4	0 0.0%	0 0.0%	1 50.0%	1 50.0%	0 0.0%	2 100%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 100.0%	2 100%

Table 14
Classification table for Los Angeles (test procedure 2)

Correct cluster	Model predicted cluster		Total
	1	2	
1	86 70.5%	36 29.5%	122 100%
2	17 22.4%	59 77.6%	76 100%

Table 15
Unmatched links classified into each San Diego cluster

	Cluster classification					Total
	1	2	3	4	5	
Number	303	214	260	76	174	1027
Percent	29.5%	20.8%	25.3%	7.4%	16.9%	100.0%
Avg. prob.	0.544	0.528	0.428	0.550	0.490	

Summaries of the values are shown in Table 16. Also shown in Tables 15 and 16 are the average probabilities for each cluster. Recall that these values represent the estimated probability that the link was properly classified. For San Diego, these values ranged from 0.42 for cluster 3 to 0.55 for cluster 5. The average probabilities for Los Angeles were 0.78 for cluster 1 and 0.71 for cluster 2. The larger Los Angeles probabilities and smaller San Diego probabilities reiterate the original findings that Los Angeles had a smaller misclassification rate (26%) and San Diego had a larger rate (45%). Intuitively, this is reasonable; with only two clusters in Los Angeles, there is already a 50% chance of classifying the link properly.

Table 16
Unmatched links classified into Los Angeles clusters

	Cluster classification		Total
	1	2	
Number	1261	599	1860
Percent	67.8%	32.2%	100.0%
Avg. prob.	0.776	0.714	

5. Conclusions

This study extends the MMR methodology to disaggregate travel demand model period-based volumes into hourly volumes for all network links, matched or unmatched. The benefit of such a procedure is increased resolution for estimation of gridded running stabilized mobile emissions. The procedure extends prior theoretical work in estimating allocation factors, by incorporating spatial variation and devising and testing a new technique for incorporating unmatched network locations.

Fig. 8 summarizes the procedure developed to disaggregate both matched and unmatched model link volumes into hourly profiles. First, the necessary data is collected, including freeway DM link volumes and observed volumes. For each model location, a determination must then be made as to which links match an observed count location and which remain unmatched. Second, all matched links are spatially grouped, or clustered, based on the observed daily temporal traffic patterns on the link. Third, one set of allocation factors, say β_c^i for hour i is estimated for all links within cluster c. In other words, instead of estimating one set of allocation factors for all links, separate factors are estimated for clusters of locations with similar daily traffic distributions.

The methodology was applied to the SCOS97 data set and resulted in two classification models, one for San Diego and one for Los Angeles. Testing of the new methodology indicated that the San Diego model had a predicted correct classification rate of 55%. Misclassifications can, however, be partially acceptable if the link was incorrectly clustered into an adjacent cluster exhibiting a similar pattern to the appropriate cluster. For Los Angeles, the predicted misclassification rate was estimated as 26%. In both cases, we found that links exhibiting lower volumes were slightly more likely to be misclassified. Although this analysis used LA and San Diego, the full methodology can be easily implemented, and appears well suited for application to other regions as a means for improving the resolution of the gridded running stabilized mobile emissions.

Further extensions to the method should include development of a statistically-based framework for disaggregating non-highway travel demand volumes to hourly profiles. The current method can generally be applied only to those facilities with automatic counters, typically highway level facilities. Although counts may be taken on arterial level network links, these will usually be of limited duration (i.e., a day or week versus an extended period). Additional research would also be useful in identifying the minimum number of day counts that are necessary to ensure that the mean volumes are both diurnally and spatially representative. This would help metropolitan planning agencies and local governments to budget count days for off-highway travel demand model links.

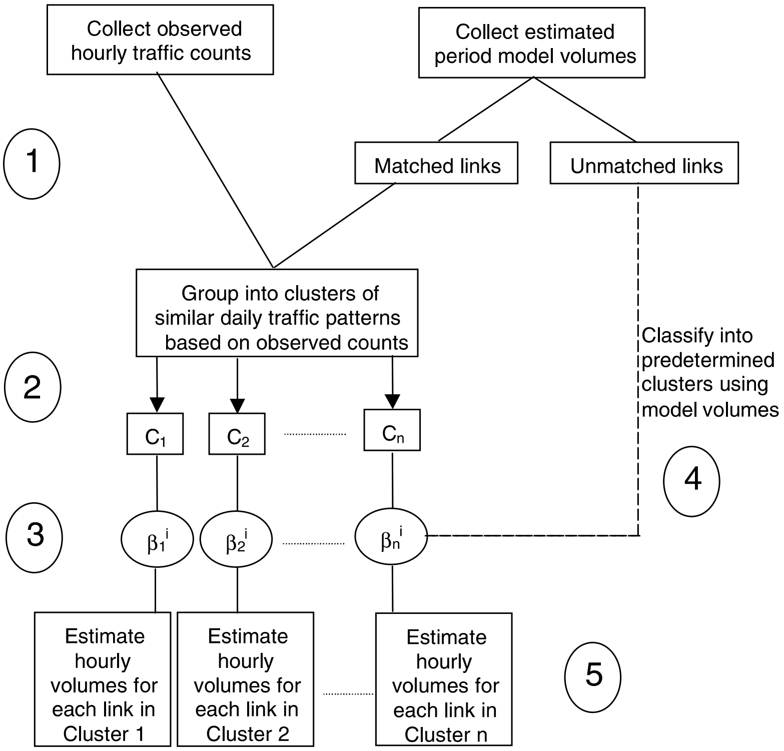
Collect observed hourly traffic counts
Collect estimated period model volumes
1
Matched links
Unmatched links
Group into clusters of similar daily traffic patterns based on observed counts
2
Classify into predetermined clusters using model volumes
C1
C2
Cn
4
3
β1i
β2i
βni
Estimate hourly volumes for each link in Cluster 1
Estimate hourly volumes for each link in Cluster 2
Estimate hourly volumes for each link in Cluster n
5

Fig. 8. Extended methodology for improving hourly profiles.

Acknowledgements

The authors wish to thank Dr. Sung-Eun Kim for statistical assistance and Matthew Korve for assistance with the initial data collection and analyses. Staffs from Caltrans, SANDAG, and SCAG were helpful in providing the real-time data and model volumes necessary to complete the empirical research. The California Air Resources Board provided funding for this research. The views reflect those of the authors alone.

References

California Department of Transportation, 1998. Direct Travel Impact Model 3 (DTIM3): User's Guide. Sacramento, CA.

California Environmental Protection Agency, 1997. California Environmental Protection Agency's Air Resources Board Launches Comprehensive Southland Smog Study. CEPA News Release, http://www.arb.ca.gov/newsrel/nr061197.htm. Accessed 4/29/99.

Johnson, R.A., Wichern, D.W., 1992. Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Lin, K., Niemeier, D., 1998. Using multivariate multiple regression models to improve the link between air quality and travel demand models. Transportation Research 3 (6), 375–387.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Niemeier, D., Hicks, J., Korve, M., Kim, S., 1999a. Estimation of Allocation Factors for Disaggregation of Travel Demand Model Volumes to Hourly Volumes for Highways in the South Coast Air Basin. Institute of Transportation Studies, University of California, Davis.

Niemeier, D.A., Lin, K., Utts, J., 1999b. Using observed traffic volumes to improve fine-grained regional emissions estimation. Transportation Research 4, 313–332.