UC Irvine
Research Reports

## Title

Review Current Methods for Allocating Average Weekday Mobile Source Emissions to Weekend Days

## Permalink

https://escholarship.org/uc/item/4vd249bf

## Authors

Rindt, Craig R.
Marca, James E
Recker, Will

## Publication Date

2012-06-30

eScholarship.org Powered by the California Digital Library
University of California

# Review Current Methods for Allocating Average Weekday Mobile Source Emissions to Weekend Days

Final Report

Craig R. Rindt
James E. Marca
Will W. Recker

University of California, Irvine
Institute of Transportation Studies

June 30th, 2012

# Summary

This report provides a review of the California Air Resources Board's current method for adjusting estimated emissions from on-road motor vehicles. These estimates are a critical feature of CARB's air quality modeling efforts, which are an important tool for maintaining and improving California's air quality.

The central approach of the method adjusts "average weekday" emissions estimates using factors derived from measurements of diurnal and day-of-week variations in multi-class vehicle activity provided by Caltrans from automated vehicle classifer (AVC) data. The diurnal factors are used to adjust the inputs to the Direct Travel Improvment Model (DTIM), which provides gridded, hourly estimates that are assumed to represent general spatio-temporal emissions patterns on an average weekday. The day-of-week factors are used to adjust the average weekday county-wide emissions estimates produced by EMFAC, CARB's tool for estimating emissions from on-road vehicles. The day specific spatio-temporal DTIM patterns are then used to produce hourly, gridded emissions from the county-wide EMFAC totals.

The body of this report begins with a review the general methodology in the context of the broader model. It continues with a review of the SAS programs that compute the diurnal and day-of-week factors from Caltrans AVC data. We conclude with some recommendations for how the method might be improved over the near-term and long term.

The general findings from our review are that the current method for allocating average weekday mobile source emissions to weekend days makes use of the readily available data sources in a methodologically sound manner. Given the available data, we could not find significant fault with the approach. Our review of the code also concluded that it appears to be consistent with the method, that is, it does what it is supposed to. Because our review was limited to the code for computing the factors, we cannot comment on how these factors are integrated into the EMFAC/DTIM factorization.

We offer two near-term and two long-term recommendations for further investigations into improving the method. In the near term, we first suggest looking into methods for bounding the uncertainty in the emissions estimates as a function of the uncertainty of the available data. Though non-trivial, such an analysis could strengthen the models and enhance their power for supporting policy decisions.

Our second near-term suggestion is to make better use of currently available data sources for developing factors. One readily available source of such data is the California Vehicle Activity Database (CalVAD), currently under development by CARB. This system already fuses available data sources to produce "best-estimate" of vehicle activity by class. Linking CalVAD and other data sources should improve the computation of daily and diurnal factors, while also providing a way of integrating new data sources as they become available.

This leads to our first long-term recommendation, which is to obtain better vehicle activity data for computing the weekend and diurnal adjustment factors. We believe there are many sources of vehicle activity data that could significantly improve the estimation of spatio-temporal activity

patterns. In particular, the explosion of mobile communications systems and their services offer significant potential for high-fidelity data on personal and commercial vehicle travel across all vehicle classes.

Finally, we suggest that the ultimate solution is the development of a fully integrated model system that incorporates live data sources, such as PeMS and the aforementioned commercial data, with MPO-developed activity-based travel demand models. This complete model system would combine the annual reporting features of EMFAC with the spatio-temporal resolution and policy sensitivity of the DTIM model. By combining this system with live data sources, it could be maintained with significantly less effort than the current MPO-based surveys used to develop EMFAC.

# Table of Contents

# Nomenclatures

| | |
|---|---|
| $AVC$ | Automated Vehicle Classifier Data |
| $ITN$ | Integration Transportation Network Data |
| $V$ | volume |
| $VMT$ | vehicle miles traveled |
| $a$ | AVC wim station |
| $ij$ | grid cell index |
| $cnty$ | is a county index |
| **core** | A core day $\in Tu, We, Th$ |
| $\Gamma$ | ratio of day VMT to core day VMT |
| $d$ | day of the week |
| $DF$ | distribution factor for a given county and day |
| $D$ | caltrans district |
| $DTIM$ | DTIM-computed emissions |
| $EMFAC$ | EMFAC-computed emissions |
| $EMFAC'$ | Distributed but unscaled EMFAC-computed emissions |
| $E$ | emission value |
| $E$ | Gridded emissions |
| $h$ | hour of day |
| $\beta$ | hourly ratio of VMT |
| $L$ | length of link |
| $\ell$ | link index |
| $NF$ | normalization factor for a given county and day |
| $p$ | pollutant (category) |
| $c$ | vehicle profile |
| $RH$ | relative humidity |
| $SCC$ | (pollutant) source classification code |
| $S$ | vehicle speed |
| $T$ | temperature |
| $avc$ | vehicle class (type) |

# 1 Introduction

Air quality modeling is a complex domain both in its political and institutional motivations as well as its implementation for specific analysis. It has its theoretical origins near the beginning of the last century with the development of the first models of air movement (Taylor, 1915 and Sutton, 1932), extending through the development of models of pollutant dispersion (Hewson, 1945 and Gifford, 1960) followed by continuing advancement toward the latter half of the century (Durham, 2006). With an increasing number of major pollution events during the 1950s and 60s and their consequent impacts on public health, federal and state governments took action to regulate the emissions causing air pollution (California Air Resources Board, 2012). With the development of modern digital computers during the 1970s and 1980s, the earlier dispersion models were coded into simulations that could be used to evaluate policy. As computing technology continues to advance, so too does the sophistication and power of air quality models. Indeed, the use of these models for air quality analysis has been formalized by the Environmental Protection Agency (EPA) under the Clean Air Act so that:

> *. . .[for] several of the criteria pollutants, regulatory requirements call for the application of air quality models to evaluate future year conditions as part of State Implementation Plans to achieve and maintain the National Ambient Air Quality Standards (NAAQS)* (Environmental Protection Agency, 2004).

Among its many responsibilities, the California Air Resources Board (CARB) is tasked with performing such air quality and emissions studies. These studies are used to evaluate California's air quality relative to state and federal air quality standards and to support the development of strategies and policies for its continuing management.

In an effort to improve its modeling of mobile source emissions that are a critical feature of the air quality modeling system, CARB has developed methods for estimating and modeling temporal variations in heavy-duty vehicle activity and in weekend travel activity for all vehicle classes. This report reviews these methods and the computer code implementing them and offers recommendations on how these methods might be improved. Our review does not extend to the complete air quality modeling system used by CARB and focuses only on how particular mobile source emissions are allocated across time and space for particular days (weekends) and/or for particular classes of vehicles (heavy-heavy duty trucks).

The next section provides a brief overview of air quality modeling to provide some context. Section 3 describes how on-road mobile sources fit into the big picture. Section 4 offers a detailed review of the current method used to estimate on-road mobile source emissions and we offer some comments on this method in Section 5. This is followed in Section 6 by an annotation and review of the SAS computer code used to generate adjustment factors for time of day and day of week from Caltrans data sources. We then offer a set of near- and long-term recommendations for improving the generation of mobile source emissions inventories in Section 7.

# 2 Air Quality Modeling

Modern air quality modeling involves three major steps: emissions modeling, meteorological modeling, and photochemical modeling. The photochemical model, which models the hourly changes in the atmosphere over a region of interest as a three dimensional grid of boxes. To do this, it models the movement of air, mixture of pollutants between boxes, and the chemical processes within each box as a function of the chemical composition and incoming solar radiation. The emissions and meteorological models provide the inputs and boundary conditions to the photochemical model (Texas Commision on Environmental Quality, 2012).

The current review is focused on the modeling work of the Central California Ozone Study (CCOS) (California Air Resources Board, 2007). However, the CCOS study uses a model system that will be applied statewide to demonstrate how and when California will attain air quality standards established under both the federal and California Clean Air Acts. Our review has therefore been conducted with this more general purpose in mind.

As noted, the scope of this review is limited to consideration of the particular methods used in computing weekend and heavy-duty vehicle mobile source emissions. While alternatives exist to the choices of particular components for the complete air quality modeling system, California's models are consistent with the state of the art and with EPA regulations (Environmental Protection Agency, 2005). Thus, we assume that these models will perform adequately and now turn our attention to the specifics of emissions modeling.

## 2.1 Emissions model overview

As both a major input and the primary target for regulation, emissions are a critical feature of air quality modeling. Air quality studies rely on emissions inventories detailing temporally and spatially resolved emissions estimates categorized into:

- stationary sources, both elevated and low-level, ranging from power generation to sources of agricultural waste,
- on-road mobile sources,
- area sources (which include off-road mobile sources), and
- biogenic sources

Of the above inputs, mobile source emissions are a significant contributor to ground-level emissions and present a particular challenge to estimate. Furthermore, emissions from trucks and other heavy-duty vehicles are the primary contributor of particulates and other pollutants that play an important role in the atmospheric chemistry that causes unhealthful air quality and other negative environmental impacts. The importance of mobile source emissions to air quality means that as

much care as is possible must be taken to produce quality estimates of these emissions (National Research Council, 2004).

There are two main approaches to mobile source emissions modeling: *average speed models* and *modal models* (Bigazzi and Bertini, 2009). The former use estimated average speeds of the vehicle fleet along with emissions factors computed for various classes of vehicle to determine emissions for the fleet. The most commonly applied models fall into this category, including CARB's EMFAC model as well as the EPA's MOBILE model used by most of the United States and various international models such as COPERT (Gkatzoflias et al., 2012),

The latter attempt to characterize vehicle activity in different *modes* using a combination of roadway, driver, and traffic factors to provide more detail and sensitivity. Though more complex, these models are gaining popularity. Prominent models include CMEM (Barth et al., 1996), IVE (Davis et al., 2005), and the EPA's MOVES model (Environmental Protection Agency, 2010).

The standard emissions model in California is the EMFAC model, which estimates emissions for a representative weekday for each California county. The EMFAC model is maintained by CARB and is updated regularly based upon reporting of speed-binned vehicle miles traveled (VMT) received from regional transportation planning agencies across the state (California Air Resources Board, 2007, Appendix D). EMFAC is specific to California, capturing its unique vehicle fleet arising from its unique emissions standards and mix of out-of-country vehicles (Lents et al., 2012).

As with its general air quality modeling process, CARB's EMFAC model is a standard, accepted method whose use is well established for policy analysis. Though an eventual shift to a modal or other model may deserve consideration in the future, EMFAC is uniquely suited to its current applications.

## 2.2 Spatio-temporal allocation

The emissions models discussed above produce emission estimates for particular levels of vehicle activity. The average speed models, such as EMFAC, are generally applied to average activity over larger regions, typically the county level. An assumption underlying this aggregation is that "variations in the emission rate can be expected to even out so that the emission factor adequately reflects the average emission rate across the activity range" (Miller et al., 2006).

A problem introduced by this aggregation, however, is that the resulting spatial (by county) and temporal (daily) resolution is too coarse to be used directly in the photochemical air quality models, which require gridded, hourly emissions. The standard approach to handling this problem is to perform some sort of spatio-temporal allocation using direct measures or indirect surrogates representing how activity is distributed. Though there is some variation in the exact models used for this allocation, the methods are similar across all applications we reviewed, which included various applications of the SMOKE modeling system along with the Emissions Preprocessing System (EPS) (Boulton et al., 2002, Gaffney, 2002, Lindhjem et al., 2010, Institute for the Environment, 2011 and Texas Commision on Environmental Quality, 2012). In these methods, levels of vehicle activity are projected onto a geo-coded representation of the transportation network. Some rule is selected

for determining how to allocate these links to specific grid cells. Sometimes this is done by placing a link in the cell of one of its endpoints and sometimes a link's activity is divided proportionally to all grid cells it traverses. Geographic Information System (GIS) techniques are used to perform this allocation. For temporal allocation, common practice is to collect representative patterns of diurnal travel activity as a proxy for fleet activity during the day and use these to spread daily emissions estimates across the hours of the day. Typically, different patterns will be determined for weekdays and weekend days.

Because this report is primarily concerned with CARB's method for performing this spatio-temporal allocation of county-wide, daily EMFAC emissions estimates, the next section will discuss California's current approach in more detail.

# 3 California's approach

California's approach to spatio-temporal allocation is similar to approaches used elsewhere in that spatial surrogates are combined with diurnal profiles to allocate EMFAC-estimated emissions. The 2007 CCOS report (California Air Resources Board, 2007) details California's end-to-end process for air quality modeling in support of developing an 8-hour ozone State Implementation Plan (SIP). This process draws heavily on the work done as part of the 2000 Central California Ozone Study (CCOS), particularly for the development of the gridded emissions inventory, of which the mobile source emissions model is a key component.

To perform allocation, the Direct Travel Impact Model (DTIM) model was developed specifically to provide 4km gridded, hourly emissions using:

- episodic meteorological data;
- emission rates from EMFAC for specific temperatures, relative humidity, and speed bins; and
- levels of vehicle activity from the The Integrated Transportation Network (ITN).

The ITN combines the work of regional transportation planning agencies to produce a single network of streets and highways along with related traffic analysis zones (TAZs) for the entire state of California. The network and TAZ data include VMT for 13 vehicle classes for a representative peak-summer weekday.

The output of DTIM is gridded, hourly mobile-source emissions. However, DTIM's emissions modeling differs from the standard EMFAC approach (California Air Resources Board, 2007). Furthermore, because of its relatively high level of disaggregation, the ITN upon which DTIM relies is costly to develop and maintain, and therefore is out of date compared to the EMFAC vehicle activity numbers. Finally, since general activity levels are already collected as part of the regular updates to the EMFAC model, it is preferable that the gridded emissions are consistent with the EMFAC county-wide estimates.

As a result, as part of developing a standard protocol for photochemical modeling, CARB developed a method using DTIM outputs to produce spatial-temporal factors that are applied to the countywide, weekday EMFAC emissions estimates to produce 4km gridded, hourly emissions estimates that can be used by the photochemical models (California Air Resources Board, 2007). This approach is based on the reasoning that general spatial and temporal patterns of vehicle activity will be relatively stable over time, even if the magnitude of the activity increases. We discuss the implications of this reasoning in Section 7.

The first version of the ITN estimated weekend day VMT, volumes and trips, including the use of weekend-specific diurnal variation to produce hourly activity (California Air Resources Board, 2007, Appendix D). Due to problems with this method, subsequent versions of the ITN make no attempt to develop weekend activity estimates and CARB has proposed an updated method for making these allocations. The next section reviews this method in detail.

# 4 Updated method for generating gridded, hourly mobile-source emissions

The work under present review implements an new method for developing weekend-specific activity and emissions estimates as a replacement for the older reliance on estimates embedded in ITN. It also focuses on producing updated hourly activity distributions for heavy-duty truck activity based upon Caltrans Weigh-In-Motion (WIM) data.

A useful flowchart detailing the relationship between EMFAC and DTIM is available in the report for SCAG's MATES-III study (SCAQMD, 2008, see fig 3-1 on p. 3-10) Using this as a basis, we have created a similar flowchart representing the CCOS protocol for producing gridded emissions as shown in Figure 1. Note that instead of the SCAG travel demand model, SIP uses ITN for link activity data.

In general terms, the modeling process involves the use of the DTIM model as a baseline representation of how emissions are distributed in time and space. The process requires the following inputs:

- A meteorological episode to represent conditions for which to model air quality performance
- Statewide vehicle activity from the Integrated Transportation Network, (currently on version 3).
- Caltrans automated vehicle classifier (AVC) data collected from WIM sites in each county.

Given these inputs, mobile source emissions are produced using the steps described below. The first two steps, labeled A and B, come from the main CCOS report (California Air Resources Board, 2007). They are omitted from the *Proposed Method* document (Gridded Inventory Coordination Group, 2006) provided by CARB to describe the method for review in this report, but we show them here to provide context.

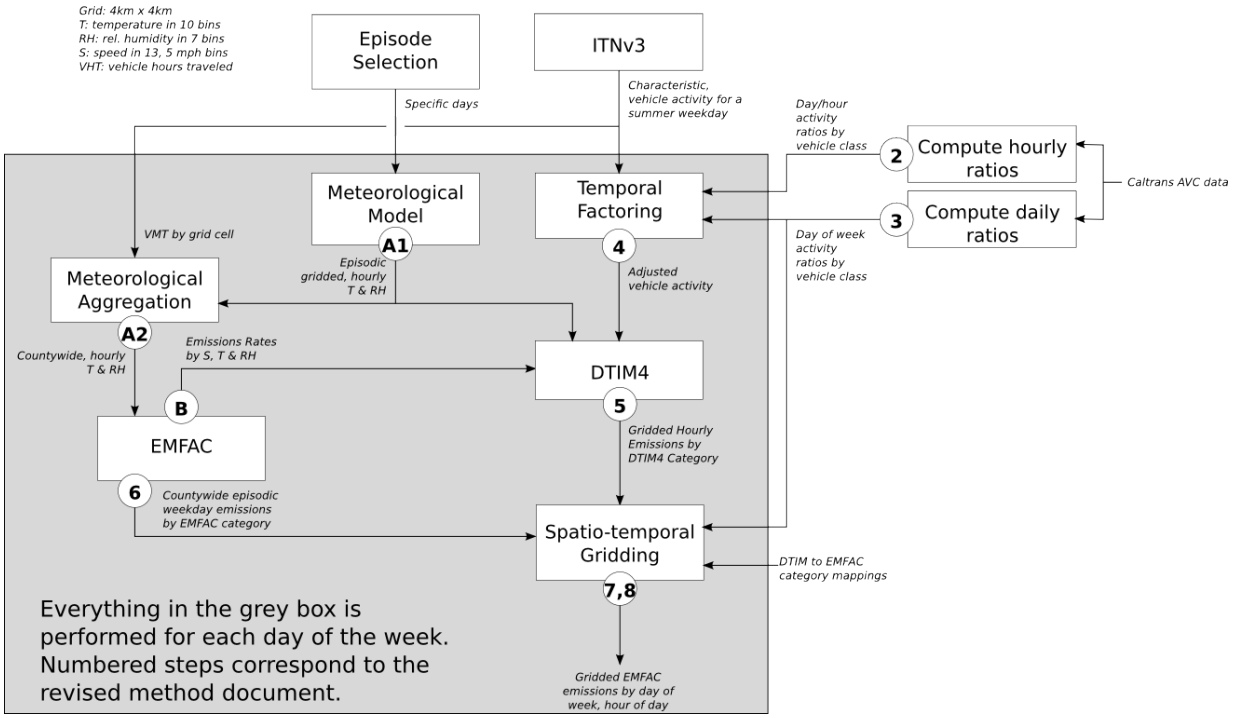


**Figure 1** The emissions modeling process

A. Compute temperature ($T$) and relative humidity ($RH$) values for the episode. This is performed in two steps to accommodate the different models:

   - Run the meteorological model to get gridded temperature and relative humidity for the episode.

   - Perform *meterological aggregation* to compute county-wide averages for hourly $T$and $RH$for the episode.

B. Run EMFAC to get for each county the various emissions rates by $T$, $RH$, and speed bins. These rates are used by DTIM to compute its gridded emissions estimates.

The remaining steps are numbered from 1 through 8 to correspond with the Gridded Inventory Coordination Group (2006) document.

1. Sum the hourly volumes by vehicle type and county on the ITN network. These are used in step 4 below to redistribute ITN volumes to match day-specific hourly distributions. Specifically, compute:

$$\text{Total class avc core day volume on link } \ell = \sum_{\text{h}\in[0,23]} \text{V}_{\text{ITN}}(\ell,\text{h},\text{avc}) \tag{1}$$

2. Using Caltrans AVC data, compute *hourly ratios* $\beta(h,d,c,cnty)$ for each county *cnty*. These ratios represent for each vehicle profile *c*and day of week $d$, the fraction of the day's total activity

**Table 1** Vehicle profiles for grouping Caltrans AVC classes

| Profile | Description | AVC Classes |
|---|---|---|
| LD | light duty vehicles | LDA, LDT1, LDT2, MDV, UBUS, SBUS, OBUS, MCY, MH |
| LM | light and medium heavy duty vehicles | LHD1, LHD2, MHD |
| HHD | heavy-heavy duty vehicles | HHD |

that occurs during hour $h$. There are different set of ratios per Caltrans district $D$. However, each county is contained by a single Caltrans district, so there is a unique mapping $cnty \rightarrow D$, and a county's hourly ratios simply equal those of its parent district. These ratios are computed from district-wide AVC data collected from all AVC locations $a$in the district $D$. There are three vehicle profiles that group together the separate vehicle classes measured by Caltrans AVC sites to define a mapping $avc \rightarrow c$ as shown in Table 1. The computation is as follows:

$$\beta(h,d,c,cnty) = \frac{\text{Measured V during hour h}}{\text{Measured V during day d}} = \frac{\sum_{a\in D}\sum_{avc\in c} V_a(h,d,avc)}{\sum_{h\in[0,23]}\sum_{a\in D}\sum_{avc\in c} V_a(h,d,avc)} \tag{2}$$

3. Using Caltrans AVC data, Compute *daily ratios* $\Gamma(d,c,cnty)$ for each county $cnty$. These ratios represent for each vehicle profile $c$, the ratio of day $d$'s total activity to the total activity on a core (Tu-Th) day. Again, this is computed using AVC data so that the same $cnty \rightarrow D$ and $avc \rightarrow c$ mappings apply. Note that since core days are treated as a single day, $\Gamma = 1.0$ for all core days.

$$\Gamma(d,c,cnty) = \frac{\text{Measured V during day d}}{\text{Measured V during a core day}} = \frac{\sum_{h\in[0,23]}\sum_{a\in D}\sum_{avc\in c} V_a(h,d,avc)}{\sum_{h\in[0,23]}\sum_{a\in D}\sum_{avc\in c} V_a(h,\mathbf{core},avc)} \tag{3}$$

4. Perform *temporal factoring* on the ITN link and TAZ activity by applying temporal adjustments to ITN link and TAZ activity. Depending on the vehicle class $avc$and day of week $d$, ITN Link or TAZ activity can be adjusted for the daily totals (**daily**) and/or the hour of the day (**hourly**) as shown in Table 2. Heavy duty truck activity is always redistributed to AVC-based hourly distributions AND on non- core days, all vehicle activity is redistributed to AVC-based hourly distributions and daily totals are adjusted to reflect lower total activity on weekends. Note that while it is not shown here, an equivalent process is used to factor TAZ-based trip ends used by EMFAC.

$$V'_{ITN}(\ell,d,h,avc) = \beta(d,h,c,cnty)\cdot\Gamma(d,c,cnty)\cdot\sum_{h\in[0,23]} V_{ITN}(\ell,h,avc) \tag{4}$$

**Table 2** Temporal adjustments made to ITN link and TAZ activity.

| | Passenger Cars | Light- and Medium-Trucks | Heavy-duty Trucks |
|---|---|---|---|
| Core Days ($d \in \{Tu,W,Th\}$) | No adjustments | No adjustments | **hourly** only (i.e., $\Gamma(d) = 1.0$) |
| Weekend Days ($d \in \{F,Sa,Su,M\}$) | **total**, **hourly** | **total**, **hourly** | **total**, **hourly** |

This raises an area of concern for the overall method. Per the California Air Resources Board (2007, Appendix D, pg 3-4) , the ITN network provides for each link:

- the hourly volumes ($V$) by vehicle class ($avc$) and
- the hourly speeds ($S$) on the link.

Presumably, VMT for each hour ($h$) is computed as:

$$VMT(\ell,h) = V(h) \cdot L(\ell) \tag{5}$$

It stands to reason that as the volumes rise towards the capacity of each link, the hourly average speeds on each link should decrease, and that if the volumes fall away from capacity (as might be expected on a weekend day) the hourly average speeds should increase.

5. Run DTIM4 using EMFAC emissions rates from step B with adjusted vehicle activity from ITN ($V$' from step 4) to get gridded emissions by pollutant ($p$) and emission source ($SCC$) as $DTIM(p, SCC, ij, h)$.

6. Run EMFAC to get emissions estimates $EMFAC(p, SCC, cnty)$ for each pollutant ($p$) from each (mobile) source category ($SCC$) for each county ($cnty$).

At this point, we now have the following:

- The (weekday) EMFAC emissions by county for the modeled day and
- DTIM emissions by pollutant and source category (mapped to EMFAC pollutant/categories) for each 4km x 4km grid cell $ij$, day of the week $d$, and hour of the day $h$ combination.

The remaining step is to combine the step 6 EMFAC emissions ($EMFAC$) with the step 5 DTIM ($DTIM$) emissions. The former is assumed to give the best estimate county-wide daily emissions totals for the episode. The latter assumed to give the best within-day spatial-temporal distribution. The described method breaks this into two steps: *spatio-temporal distribution* and *daily factorization*.

7. (***spatio-temporal distribution***) Distribute the EMFAC daily, countywide totals across constituent grid cells and each hour of the day according to the relative proportions computed by DTIM. In

short, compute a distribution factor $DF(p, SCC, ij, h)$ for each grid cell and hour of the day and then distribute the $EMFAC(p, SCC, cnty)$ by that factor for each cell in the county to produce an intermediate gridded emissions estimate $EMFAC'$.

$$DF(p, SCC, ij, h) = \frac{DTIM(p, SCC, ij, h)}{\sum_{\forall ij \in cnty} \sum_{\forall h} DTIM(p, SCC, ij, h)} \tag{6}$$

$$EMFAC'(p, SCC, ij, h) = DF(p, SCC, ij, h) \times EMFAC(p, SCC, cnty), \forall ij \in cnty, \forall cnty$$

8. (***daily factorization***) Factor the intermediate emissions estimates $E'$ from step 7 to reflect daily activity variations by county. Here, we use the *daily ratios* $\Gamma(d,c,cnty)$ from step 3. Thus, for each county *cnty*, we factor each intermediate emission estimate to get the final adjusted emissions estimates $E$ for the day being modeled.

$$E(p, SCC, ij, d, h) = \Gamma(d, avc, cnty) \times EMFAC'(p, SCC, ij, h), \forall ij \in cnty, \forall cnty \tag{7}$$

At this point in the modeling process, the whole process is run again for the next day. Note that steps 1 through 3 are run outside this loop.

# 5 General Comments on the Adjustment Process

This section summarizes the two issues we noted with the adjustment process we describe above. We summarize them here for ease of reference and in Section 7, we offer suggestions for mitigating these issues.

## 5.1 Speeds change as volumes change

Per the California Air Resources Board (2007, Appendix D, pg 3-4) the ITN network provides for each link:

- the hourly volumes ($V$) by vehicle class ($avc$) and
- the hourly speeds ($S$) on the link.

Presumably, VMT for each hour ($h$) on each link ($link) is computed as:

$$VMT(\ell, h) = V(h) \cdot L(\ell) \tag{8}$$

It stands to reason that as the volumes rise towards the capacity of each link, the hourly average speeds on each link should decrease. This must have been considered, but we cannot find any discussion of this point.

Perhaps the change in speeds on each link doesn't matter because DTIM emissions aren't used directly, but only as a means for spatio-temporal disaggregation of county-wide EMFAC emissions. That said, the question is still open and needs to be adequately addressed: Does a failure to adjust speeds distributions downward in the DTIM inputs impact the distribution of DTIM's gridded emissions in any significant way that will negatively impact the quality of the EMFAC disaggregation?

## 5.2 Stability of patterns

The entire factoring approach is based on the reasoning that general spatial and temporal patterns are stable enough such that factoring them to reflect general rising and falling of activity will sufficiently capture variations from the original average mid-summer weekday activity represented in the ITN network. This concern applies to the assumptions that:

- Friday through Monday patterns are sufficiently similar to Tuesday through Thursday patterns.
- Seasonal patterns are sufficiently similar, even if the magnitude of the vehicle activity increases.
- Long-term patterns are sufficiently similar, such that general patterns of traffic in the base ITN year can represent patterns of traffic in future years.
- The underlying network will not change significantly, or that any major network changes will be reflected in the DTIM model through updates to the ITN.

In our opinion, each successive application of a factoring step should be backed up with an even stronger rationale than the previous one. The use of multiple factoring steps is a sign that there may be a more direct route from the input data to the desired output. If, as is the case here, no suitable models or data are available in the near term as a replacement for factoring, efforts should be made to concentrate resources on developing those capabilities over the medium to long term.

# 6 Code review of SAS programs

The overall analysis of the weekend factors method goes hand in hand with a review of the code used to generate the weekend factors. The following review steps through the code line by line and examines what the code does, points out possible sources of bugs, and identifies important assumptions built into the code. For the impatient, the overall conclusion is that the code should work as advertised—it will correctly compute weekend and time of day factors to apply to midweek data that conform to the definition of the factors laid out in the overall method. Those interested in the details of the code, and those tasked with maintaining or updating the code may find the following write up useful.

## 6.1 CalcNoSta.sas and CalcNoStaR.sas

These two programs have the same header bit, and have a lot of duplicated code, but the CalcNoStaR.sas has a lot of commented out parts. Most of the analysis was done to the CalcNoSta.sas file.

## 6.2 CalcNoSta.sas

The basic structure of the file is a data statement that loads data from a file; a proc summary statement that generates some mean values; and then another data statement that further manipulates the in-memory data. Thorough analysis of the statments and the comment blocks is given below.

### 6.2.1 Comment block

This file starts with a big comment block describing what it does. On the other hand, since the comment is identical to the one in CalcNoStaR.sas, it has probably been cut and pasted from some other source, and should not be taken literally.

### 6.2.2 More comments

Next comes a proc format statement that is commented out. The format statement defines a mapping between values that TuWeTh can take and both the logical value and the string representation of the value.

After that is a commented out macro statement that would iterate over districts, setting D to 1 through 12. This is replaced with a let statement that sets D to 12. My guess is that the macro approach didn't work quite right so the program author decided to go for manually adjusting the D value with each run of the program.

### 6.2.3 Data statement line 34

SAS programs start off with a DATA statement. Here the data statement defines a SAS data set called stables.AVC&D., where the &D. is replaced with the current value of D. For example, when processing District 12, the data set would be named stables.AVC12. The period after the &D tells SAS the boundary of the interpolated variable.

The data statement reads in a file named `dist&D.avc.txt`, where the `&D` is once again an interpolated variable that changes into the current District being processed.

The `infile` statement includes the `missover` option. The `missover` option assigns missing values to variables for records that contain no data for those variables.

A format for the time variable is defined next.

Then each record is defined and read using the `input` statement. This line gives variable names and also instructs SAS what each value is. The `@` symbol followed by a number specifies an exact column in the input data for the next variable. A character (non-numeric) value is identified by dollar sign ($) after the variable name in the INPUT statement. So for example:

```
@4 CT_D $ 2.
```

will read a two-character variable as a string, and name it `CT_D`. Obviously this will read and interpret the Caltrans District.

The full input statement is:

```
input @4 CT_D $ 2. C_Sta 3. @10 Dir $ 1. @12 YrMoDa YYMMDD6. Hr 2. @25 (C01-C15)(5.);
```

It defines the variables listed in Table 3.

**Table 3** Variables defined in the first data statement

| Variable | Start col | End col | Type | Description |
|---|---|---|---|---|
| CT_D | 4 | 5 | string | Caltrans district |
| C_Sta | 6 | 8 | number | The detector station (?) |
| Dir | 10 | 10 | string | The direction |
| YrMoDa | 12 | 17 | date | The date of the observation |
| Hr | 18 | 19 | number | The hour of the day |
| C01 | 25 | 29 | date | Class 1 vehicle counts |
| C02 | 30 | 34 | number | Class 2 vehicle counts |
| ... | ... | ... | number | ... |
| C15 | 95 | 99 | number | Class 15 vehicle counts |

The next statement, `Day=weekday(YrMoDa);` creates a day of the week variable, which is critically important to this program.

Finally, the line:

```
if '01Jun04'd<=YrMoDa<='31Aug04'd & ^('02Jul04'd<=YrMoDa<='05Jul04'd);
```

will bracket the data to contain only the days between June 1 and August 31, inclusive, while skipping the July 4$^{th}$ weekend.

The `run;` command executes the data statement, reading the input file and creating the variables.

### 6.2.4 Proc summary, line 43

SAS uses `proc summary` to summarize data by computing descriptive statistics for variables. These statistics can be computed across all variables, or within groups of observations. An example is to run through a list of data and compute a mean value.

The first line is the `proc summary` statement itself, which typically defines the input data and options for the procedure. The line is:

```
proc summary data=stables.AVC&D. order=formated Mean; *Std;
```

Here the input data is defined as the `stables.AVC&D` that was defined in the previous data statement. Again the `&D` part is just the variables `D` interpolated, so for example `stables.AVC12`. The `order=formatted` option tells SAS that the desired sort order of the unique combinations of variables is defined by the ascending formatted values. The `Mean` statistic keyword asks for the mean value to be computed. The `std` keyword has been commented out, and so the standard deviation is not computed.

The next line specifies the format for the `day` variable in the dataset, setting it to be the `TuWedTh` value. Presumably this is the same as the `TuWedTh` format that is defined in the commented-out `proc format` statement at the top of this program listing. This will order the output such that Tuesday, Wednesday, and Thursday are combined into one class and listed first, followed by Sunday, Monday, Friday, and Saturday.

The `class day hr;` line tells the summary statement to compute means for each unique combination of day and hour. In addition, the `class` statement causes `proc summary` to generate another count statistic called `N Obs`, which gives the total number of observations of the FREQ variable that `proc summary` processes for each class level. In the output data set, the value of `N Obs` is stored in the `_FREQ_` variable.

After the `class` statement comes the `var` statement `var C01-C15`. The `var` statement identifies the analysis variables and their order in the output, in this case, the fifteen vehicle class count variables. Now it is clear that the `proc summary` statement is computing the mean value of the vehicle class counts for each day and hour combination.

The last statement of the `proc summary` invocation is the `output` statement:

```
output out=stables.D&D.C01_15 mean=mean01-mean15; *std=std01-std15;
```

This statement defines the output dataset as `stables.D&D.C01_15`, or for the case of District 12, `stables.D12C01_15`. The computed mean values are stored in this output set as the variables `mean01` through `mean15`. The standard deviation output is commented out, as before.

The `run;` statement ends the `proc summary` invocation and runs the procedure, producing the output dataset.

In total, this `proc summary` statement is computing the **mean hourly count** for each of the vehicle classes, over all of the input records that match the different class variables combinations.

One of the features of `proc summary` and `proc means` is that they produce an automatic variable called `_type_` that describes the type of the interaction that is being summarized. The overall summary (mean over all observations) is a `_type_=0` interaction; the first level sums have `_type_=1`, and so on. From the `proc summary` call, the `class day hr;` line tells the SAS processor to group the output by all combinations of day and by hour (all, hour, day, day*hour). The class line dictates the ordering of `_type_`, so `_type_=0` for the overall means, `_type_=1` for the means over unique hours, `_type_=2` for means over unique days, and `_type_=3` for means over unique combinations of day and hour.

Note that the mean value isn't summed anywhere. It is always the mean of the input records, or in this case, the mean hourly count for each vehicle class. What changes is the range of records over which the average is calculated.

In the code that follows, two types of means are evaluated further: `_type_=2` and `_type_=3`, or the mean hourly counts for each day, and the mean hourly counts for each hour in each day. The idea is that by knowing how the hourly mean of hourly counts differs from the daily mean of hourly counts, you know how to adjust estimates of mid week hourly volume average counts to the hourly average for any hour of any day.

### 6.2.5 Data statement line 50

The last operation of the SAS program is to execute the second `data` statement. This statement is a bit more complicated than the first, as it combines and manipulates the loaded counts and the computed means to create the District factors file into the final output.

The first line is the data statement that defines the dataset and the keep variables.

```
data stables.stables&D(keep=Day DOW Hr StableLD StableLM StableHH SumLD SumLM
SumHH);
```

The dataset will be named `stables.stables&D`, where the `&D` is once again the interpolation of the district variable. (Note that the period is not needed after `&D` because the open parenthesis is enough to tell SAS that the variable name is completed.) The `keep` option limits the variables in the output to those listed, no matter what other variables might be defined inside of the `data` statement.

The next three lines are array statements.

```
Array LDSum(5) LD3 LD1 LD2 LD6 LD7;
Array LMSum(5) LM3 LM1 LM2 LM6 LM7;
Array HHSum(5) HH3 HH1 HH2 HH6 HH7;
```

These statements create arrays to hold data inside of the loops. Each array position writes to the specified variable. So `LDSum(1)=10` will assign 10 to the variable `LD3`.

The `retain` statement,

```
retain BaseLD BaseLM BaseHH LD3 LD1 LD2 LD6 LD7 LM3 LM1 LM2 LM6 LM7 HH3 HH1 HH2 HH6
HH7;
```

tells the SAS processor to keep the values of the indicated variables from one iteration of the data step to the next. These kinds of variables are typically used for things like running sums or maintaining constant values for all rows of data. The reason these variables are kept here is to make the `_type_=2` mean hourly count values available when processing the `_type_=3` mean hourly count values, so that the adjustment factors can be computed properly.

Next is a `filename` statement associates the *fileref* variable `factors` with the output file indicated.

```
filename factors "I:\SIPGrid\STABLES\FactorsD&D..txt";
```

The next line in the script is a `set` statement, a flexible and somewhat difficult to understand statement.

```
set stables.D&D.C01_15;
```

In this case, the `set` statement is being used to read observations from the existing SAS data set stables.D&D.C01_15 into the current data statement. So for example if the script is processing Caltrans District 12, and `D=12`, then the set statement reads in `stables.D12C01_15`, which was defined in the previous `proc summary` statement as the means of each hourly count variable, plus automatically generated variables like the `_type_` variable.

Within a `data` statement, only one record is read in at a time, corresponding to the iteration step. So the first iteration step the first record is read in, and so on. Because you can only look at one record at a time, the `proc summary` step is one way to create an aggregate value you can use with each row. Furthermore, the `retain` statement combined with the assignment into the arrays will make one kind of mean value available to the other kind of mean value.

The `stables.D&D.C01_15` data set only has one row per each combination of day of week and hour of day, containing the mean hourly count values of each vehicle class. Thus the `set` statement will read each of these mean value records in turn with each iteration of this `data` statement.

The next line creates a day of week variable DOW by manipulating the day variable of the current mean value record.

```
DOW=put(put(day,TuWedTh.),$TuWedTh.);
```

The trick with the double invocation of put merely calls both of the format statements defined at the top of the file. The first format will recode the day (which ranges from 1 to 7) into a value from 0 to 7, with Tuesday (3), Wednesday (4), and Thursday (5) transformed into a zero value. The outer put takes the result of the inner put (0,1,2,6,7) and transforms it into a string value ('TuWedTh', 'Sun', 'Mon', 'Fri', 'Sat'). I don't understand why the programmer didn't create separate format statement that could do it in one step but this should work fine.

### 6.2.5.1 _type_=2 processing

The following block of code operates on the mean hourly count values for each day, over all hours in the day (_type=2).

```
if _type_=2 then do;
  file factors;
  if i=0 then
    put "Dist" @6 'DOW' @13 'Day'
    @20 ' LD Factor' ' LM Factor' ' HH Factor' '   Base LD' '   Base LM' '   Base
HH';
  if day=3 then do;
    BaseLD=Sum(of mean01-mean03);
    BaseLM=Sum(of mean07-mean08);
    BaseHH=Sum(of mean09-mean14);
  end;
  i+1;
  LDSum(i)=Sum(of mean01-mean03);
  LMSum(i)=Sum(of mean07-mean08);
  HHSum(i)=Sum(of mean09-mean14);
  LDFactor=LDSum(i)/BaseLD;
  LMFactor=LMSum(i)/BaseLM;
  HHFactor=HHSum(i)/BaseHH;
  put "&D" @6 DOW $TuWedTh. @15 day 2. @20 (LDFactor LMFactor HHFactor BaseLD
BaseLM BaseHH)(10.3);
end;
```

First, the file handle `factors` is invoked. The effect is to assign the output of this iteration of the data statement to the file, rather than to the dataset `stables.stables&D`.[1]

On the first row of the iteration, the `put` statement will set out the header row for the output file.

The next line is a possible bug, because while it relies on the fact that the `proc summary` step will output its data in the order of the formatted output (as requested in that step), it also implicitly expects that the first day will be day 3, although nothing guarantees that this will be the case, except convention. The day assigned to the first group (`'TuWeTh'`) could conceivably be any of 3, 4, or 5.

But if everything is ordered without surprises, then the first day will be formatted as zero (`day=3`, `day=4`, or `day=5`), *and* given a value of 3, and so the very first iteration with `_type_=2` will have `day=3`. In that case, the very first iteration will also generate the base factors that are needed for the output.

Another assumption here that may cause a bug in the future is that the `proc summary` step will output `_type_=2` rows before `_type_=3` rows. I didn't find anything in the SAS documentation that guaranteed this would be the case, although it *does* follow the convention of not doing anything surprising.

The program defines the following three classes of factors.

1. *LD* consists of class 1, 2 and 3 vehicles
2. *LM* consists of class 7 and 8 vehicles
3. *HH* consists of class 9, 10, 11, 12, 13, and 14 vehicles

Vehicle classes 4, 5, and 6, although actual vehicles, are not included in any of the scaling factors output from this model.

The `i` value provides dead reckoning access into the `LDSum`, `LMSum`, and `HHSum` arrays. More expressive programming languages might use an associative array for the same effect. When `i=1` (the first pass through the `data` statement after the `i+1` line), `LDSum(i)` is assigning a value to the first element, `LD3`.

Note that these are not daily sums, even though at first glance they might appear to be. Rather, they are the sums of the mean hourly values for each of the vehicle classes specified for each variable.

Once the sums of mean hourly values are defined for each group of vehicle classes, the adjustment factor for that day of the week is determined by dividing the current day of the week by the base values (the sum from `day=3`, which is really the sum of the average hourly counts over all Tuesday, Wednesday, and Thursday hourly observations).

These factors can be used as follows. Given that one has a predicted value for the mean hourly volume of a certain type of vehicle during the midweek period, multiplying by the factor will adjust the hourly midweek mean into the particular day's mean.

---

[1] I haven't verified this by reading the SAS manuals, but it seems reasonable and matches what is in the output files.

Then the factors are dumped to the output file, along with the current day of the week and the relevant base values with the statement

```
  put "&D" @6 DOW $TuWedTh. @15 day 2. @20 (LDFactor LMFactor HHFactor BaseLD
BaseLM BaseHH)(10.3);
```

As before, the `@6` type constructs are merely identifying the output column for each record, while the `(10.3)` statement at the end of the line describes the output formatting for the real-valued data.

Table 4 describes each variable at this point, so as to clarify what is being generated by the output of this data statement. Some of these variables are carried to `_type_=3` records examined next.

#### 6.2.5.2 `_type_=3` processing

As noted above, the output of the `proc summary` block is assumed to be sorted at the highest level according to the `_type_` variable. Thus the previous block of code will complete first, followed by the following block that only runs for `_type_=3` summaries. Most importantly, the `LDSum`, `LMSum`, and `HHSum` arrays (which reference day of week sums of hourly means for each vehicle class) are available for use here.

The `_type_=2` summaries (mean values) give the daily hourly mean for each quantity. That is, the `proc summary` code looks at all of the input records with the specific day (or days in the case of Tuesday, Wednesday and Thursday), and computes the mean *hourly* count of each vehicle for the day. Now we turn to the `_type_=3` records, in which the hourly means are computed for each unique combination of day of the week and hour of the day. So if there are 10 Fridays in the input data, that means that there should be 10 rows in the original input file observed for Friday at 5pm, and so the `*type*=3_ record is the mean of those 10 observations.

The `_type_=3` code block is as follows.

```
if _type_=3 then do;
    if hr=0 then do; j+1; sumLD=0; sumLM=0; sumHH=0; end;
        StableLD=Round((10000/24)*Sum(of mean01-mean03)/LDSum(j));
        StableLM=Round((10000/24)*Sum(of mean07-mean08)/LMSum(j));
        StableHH=Round((10000/24)*Sum(of mean09-mean14)/HHSum(j));
        sumLD+stableLD;
        sumLM+stableLM;
        sumHH+stableHH;
        output;
end;
run;
```

**Table 4** Variables defined in the `_type_=2` code block.

| Variable | Definition |
|---|---|
| `day` | Effectively, an associative array mapping day of the week (Sunday through Saturday) onto an ordered list such that 3=Tuesday, Wednesday, Thursday, and comes first in the sort; 1=Sunday; 2=Monday; 6=Friday; and 7=Saturday. |
| `mean01` | An array of values. For each unique combination of day of the week and hour of the day, compute the mean over all data for vehicle class 1 |
| `meanNN` | The mean hourly value for the $\texttt{NN}^{th}$ vehicle class |
| `BaseLD` | The sum of the mean counts for classes 1, 2, and 3 on Tuesday, Wednesday, or Thursday. Note that even though midweek means are over multiple days, they are still hourly means. |
| `BaseLM` | The sum of the hourly mean counts for classes 7 and 8 on Tuesday, Wednesday, or Thursday. |
| `BaseHH` | The sum of the hourly mean counts for classes 9 through 14 on Tuesday, Wednesday, or Thursday. |
| `LD3` | The sum of the mean counts for classes 1, 2, and 3 on Tuesday, Wednesday, or Thursday. Note that even though midweek means are over multiple days, they are still hourly means. The same as `BaseLD` |
| `LD1..LD7` | The sum of the mean counts for classes 1, 2, and 3 on Sunday (LD1), Monday (LD2), Friday (LD6), or Saturday (LD7). |
| `LDSum(i)` | An array that is really pointing at one of `[LD3,LD1,LD2,LD6,LD7]`, depending on the value of `i`. The result is preserved at the end of each iteration. The value is only changed in the first few iterations of the `data` statement when the `type=2` means are evaluated. |
| `LM3` | The sum of the mean counts for classes 7 and 8 on Tuesday, Wednesday, or Thursday. The same as `BaseLM`. |
| `LM1..LM7` | The sum of the mean counts for classes 7 and 8 on Sunday (LM1), Monday (LM2), Friday (LM6), or Saturday (LM7). |
| `LMSum(i)` | An array that is really pointing at one of `[LM3,LM1,LM2,LM6,LM7]`, depending on the value of `i`. The result is preserved at the end of each iteration. The value is only changed in the first few iterations of the `data` statement when the `_type_=2` means are evaluated. |
| `HH3` | The sum of the mean counts for classes 9—14 on Tuesday, Wednesday, or Thursday. The same as `BaseHH`. |
| `HH1..HH7` | The sum of the mean counts for classes 9—14 on Sunday (HH1), Monday (HH2), Friday (HH6), or Saturday (HH7). |
| `HHSum(i)` | An array that is really pointing at one of `[HH3,HH1,HH2,HH6,HH7]`, depending on the value of `i`. The result is preserved at the end of each iteration. The value is only changed in the first few iterations of the `data` statement when the `_type_=2` means are evaluated. |
| `LDFactor` | The ratio of each `LD` variable to `BaseLD`. `LDFactor(1) == 1` because the first day analyzed (`i=1`) is `day=3`, or Tuseday, Wednesday, Thursday. Other days' values (`i=2 .. 5`) represent how to adjust the base day value (a prediction for Tuesday, Wednesday, or Thursday) to the non-midweek day in question. A value greater than one means there are more `LD` hourly vehicles, and a value less than one means you should expect fewer on the given day. |
| `LMFactor` | The ratio of each `LM` variable to `BaseLM`. See the above discussion of `LDFactor`. |
| `HHFactor` | The ratio of each `HH` variable to `BaseHH`. See the above discussion of `LDFactor`. |

The first line executes conditional code each time the `hr` variable equals 0. This "dead reckoning" approach to iteration can be buggy, but if the data remains well ordered after the `proc`

summary statement, then it should be an inelegant but effective way to do something with each new day in the data set.[2] First the index j is incremented by one, moving to the next day of the week in the three arrays. Then the three vehicle class sums are reset to zero.

The next three lines compute the "stables" values for each vehicle class (StableLD, LM, and HH). The definition is somewhat odd, but my assumption is that 10000/24 (which equals 416.667) is some factor determined by ARB. From the comment block at the top: "The STABLES values are 10000 times the hourly means divided by 24 times the same day's daily means."

The ratios computed for each line reflect the adjustments needed to shift the mean hourly volume per day to the mean volume for that particular hour. As with the earlier code block, the range mean01-mean03 is allocated to LD vehicles; mean07-mean08 is allocated to the LM class; and mean09-mean14 is assigned to HH vehicles. As before, classes 4, 5, and 6 are ignored.

The the cumulative sums of each vehicle class are tallied up in the next three lines. sumLD starts out with the zero$^{th}$ hour's StableLD, and then tacks on the next hour with each subsequent iteration, resetting when the hr variable hits zero at the start of the next day. The same thing is done for sumLM and sumHH.

One item to note here is that the programmer changed the capitalization of the stable variables from StableLD, etc, to stableLD, etc. This is bad practice, but is not a bug. To quote from the SAS manual on this topic:

> A variable name can contain mixed case. Mixed case is remembered and used for presentation purposes only. (SAS stores the case used in the first reference to a variable.) When SAS processes variable names, however, it internally uppercases them. You cannot, therefore, use the same letters with different combinations of lowercase and uppercase to represent different variables. For example, cat, Cat, and CAT all represent the same variable.

Therefore, stableLM is internally strictly identical to StableLM. While mixing up the case in the two variables is not good practice, there is no bug introduced.

The output; line causes the variables listed in the keep option of the data statement to be emitted to the current row of the data set being constructed with this data statement. Thus the data statement will only contain the results of processing the _type_=3 records, with the _type_=2 variables included only in that they are included in the computed values. The output set should only include as many rows as there are _type_=3 rows.

The data statement ends (and is executed) with a run; statement.

```
options pageno=1;
/*proc print data=stables.stables&D; by DOW notsorted; sum stablelD stablelm sta-
blehh; sumby DOW; pageby DOW;
run; */
```

[2] The problem with this approach comes when the code is moved, or when other code is run after the proc summary statement, possibly changing the ordering of the data.

**Table 5** Variables defined in the `_type_=3` code block.

| Variable | Definition |
|---|---|
| `StableLD` | The ratio of the current hour's hourly mean LD count to the the daily hourly mean LD count. |
| `StableLM` | The ratio of the current hour's hourly mean LM count to the the daily hourly mean LM count. |
| `StableHH` | The ratio of the current hour's hourly mean HH count to the the daily hourly mean HH count. |
| `sumLD` | The cumulative sum of hourly mean counts of LD vehicles up to and including the current hour. |
| `sumLM` | The cumulative sum of hourly mean counts of LM vehicles up to and including the current hour. |
| `sumHH` | The cumulative sum of hourly mean counts of HH vehicles up to and including the current hour. |

The program ends with an `options` statement, and a commented out print statement. Presumably this is because `CalNoSta.sas` is run manually, and then other programs are called to save the results by printing to a file.

## 6.3 Overall review

After the review of the code, it appears that it will perform exactly as described. While a manual code review will rarely catch all syntax bug, it can verify that the overall logic is correct and that no edge cases are overlooked. Because these programs were used in practice already, it is safe to assume that there are no syntax errors that will result in runtime bugs.

# 7 Directions for Methodological Improvements

The general findings from our review are that the current method for allocating average weekday mobile source emissions to weekend days makes use of the readily available data sources in a methodologically sound manner. The method makes use of various factors computed from vehicle counts to adjust average weekday mobile-source emissions estimates to reflect altered weekend travel activity patterns. Given the available data, we could not find significant fault with the approach. In the review in Section 5, we identified two concerns with the method:

- Stability of activity patterns underlying the model
- That speeds are likely to change as volumes change

For the most part, these issues are unavoidable given the availability of data. Nonetheless, their impacts may be significant and attempts made to mitigate them. In the following sections, we offer near-term and long-term suggestions toward this end.

## 7.1 Near term improvements

We believe the following suggestions are feasible lines of inquiry that could be carried out today, and without signficant cooperation of organizations outside of CARB.

### 7.1.1 Model the propagation of uncertainty

An immediate concern mentioned above is that the method relies on the refactoring of emissions estimates derived from combining two separate models (EMFAC and DTIM), both of which are also the product of factored estimates collected from diverse sources with variable quality controls. The major concern at this level is that these potentially error-prone data may lead to errors that propagate throughout the model system in a manner similar to that identified for travel demand models as described by Zhao and Kockelman (2001) , who concluded that:

> *uncertainty is likely to compound itself – rather than attenuate – over a series of models. Mispredictions at early stages (e.g., trip generation) in multi-stage models appear to amplify across later stages.*

The outputs of any model have significantly more value if they are qualified by estimates of the uncertainty associated with the prediction. We recommend a deeper exploration of how errors will propagate within the mobile-source emissions model. Ideally, this will lead to techniques for bounding the uncertainty of its estimates.

In support of this recommendation, we note that in their extensive review of emissions modeling, Miller et al. (2006) recommended systematically continuing "to improve emission inventories by applying sensitivity and uncertainty analyses" as well as developing "guidance, measures, and techniques to improve uncertainty quantification, and include measures of uncertainty (including variability) as a standard part of reported emission inventory data". Though CARB has conducted such analyses in the past (Wagner et al., 1992, for example), these have not been tailored to the specific vehicle activity factoring techniques applied in the method under review.

For large-scale modeling, sensitivity analysis is both practical and critical. For instance, the Econometrics and Applied Statistics Unit at the Joint Research Centre (JRC) of the European Commission recommends sensitivity analysis when the assessment being conducted makes use of some form of mathematical model (EAS-JRC, 2012) and thus:

> *You want to be sure that your model-based inference can stand in court, that nobody can prove you wrong by flagging an incomplete exploration of the input assumptions (E.g. 'You treated X as a constant when we know it is uncertain by at least Y%') * You want to be sure that small but plausible changes in the input assumption do not lead to results antithetic to those you have presented (E.g. 'It would be sufficient a modest change in X to make your statement about Z fragile').*

Clearly such concerns are important in this modeling context and warrant consideration. A starting point for implementing such uncertainty bounds is a series of papers regarding errors in modeling road transport emissions (Kühlwein and Friedrich, 2000,2005). In their earlier work, they recommend using sensitivity analysis on the model outputs as a function of uncertain inputs. A similar method could be applied to CARB's emissions models to identify its sensitivity to variations to the plausible range of inputs. Though this review is concerned primarily with the impacts of factoring average weekday emissions estimates, the factoring occurs at the level of vehicle activities that serve as inputs to the EMFAC model. Thus, any sensitivity analysis should be performed on EMFAC.

This analysis could be used in two ways. First, the sensitivity of the EMFAC rates could be computed as a function of speed and VMT measurements in a manner similar to that applied by Tang et al. (2003) for the MOBILE6 model[3]. At the very least, this sensitivity could be qualitatively assessed in relative terms to identify some metric of model reliability. Second, if the related uncertainty is outside acceptable bounds, the analysis could be inverted to produce minimum standards that could be used as the basis for guidance to MPOs on the reporting VMT/speed data used to develop EMFAC.

Since DTIM also relies on EMFAC, the sensitivity analysis could be extended to the spatio-temporal distribution of emissions as well as their impacts on magnitudes. It is not obvious whether such sensitivities would lead to significant changes in the distribution, but if the sensitivities are nonlinear, as Haagen-Smit (1958) and Miller et al. (2006) suggest, it is quite possible and therefore deserves consideration.

Summarizing our suggestions related to model sensitivity, we recommend performing sensitivity analysis on the following model outputs:

1. EMFAC emission rates as a function of speed bins

2. EMFAC county-wide emissions estimates as a function of VMT/speed inputs

3. DTIM spatio-temporal variations

The sensitivity of the models should be assessed relative to the anticipated errors associated with the model inputs. Of principal concern is determining the impact of variations in the time-of-day and day of week factors. Though the AVC count data from Caltrans WIM stations is fairly reliable, the WIM locations are sparse relative to the complete network. A better understanding of how this sparse sensor network might lead to errors is needed. One approach would use fleet vehicle models to identify how well fleet traffic is sampled by the WIM network. It is unclear, however, the extent to which such models are available. It is likely that spending the same effort on obtaining direct fleet measurements through alternative data sources (see section 7.2.1) would lead to better long-term results, especially if such measurements could be integrated into an automated system such as CalVAD.

[3] It is possible that such a sensitivity analysis has been performed. If so, that analysis might be used directly or extended to the particular purpose described here.

Secondly, the MPO VMT/speed surveys serve as the basis for EMFAC forecasts. Understanding the quality of these estimates is therefore crucial to the overall quality of CARB's emissions estimates. Because these are combined with the factors above, however, the prior sensitivity analysis could be used to assess what levels of uncertainty are tolerable from these inputs.

Third, because the MPO counts and computed factors are combined to produce DTIM inputs, the impact of each input should be assessed separately as well as together. Generally, all three of these analyses involve evaluating the three model outputs across a representative range of inputs. The art of the sensitivity analysis lies in identifying these ranges systematically.

The results of these analysis would guide the next actions. These might include:

1. **Nothing:** The sensitivity of the model to inputs might be deemed appropriate for the policy applications.

2. **Improving data collection guidance:** If the model sensitivities raise concerns, they could be used to define data collection goals. Goals could be developed separately for each class of input data. The guidance might also indicate certain inputs are simply insufficient for their use in the model. In this later case, new data sources might be required (e.g., in the case of measuring HDT activity).

3. **Improving core models:** If the quality of data cannot be improved in the near term, another step would be to consider improving the rigor of the underlying models to improve their accuracy.

### 7.1.2 Incorporate additional data sources

An obvious approach to improving the confidence of the methodology is to seek additional data sources that can offer improved measurements of travel activity patterns, particularly in areas where the currently available sources of data are limited in their scope. In particular, the reliance upon AVC/WIM data to compute daily and diurnal factors, combined with scarcity of WIM stations collecting that data raises concerns about the quality of estimates produced using those factors. The sensitivity analysis recommended above would clarify the degree to which this is a concern, but we continue the current discussion with the assumption that it is a concern.

There are many sources of public and private data that aren't currently being used that offer great potential for measuring general patterns of activity. For instance Caltrans' Performance Measurement System (PeMS) offers some measurements of daily and diurnal patterns across different vehicle classes. Though these typically are not direct measurements of different vehicle classes, the underlying algorithms might be improved for specific application in this context.

Notably, the CalVAD system (Marca et al., 2012), a project at UC Irvine sponsored by CARB, adopts this approach and includes a data fusion algorithm for merging all available data sources into a "best-available" estimate of activity by vehicle class. Further, though CalVAD only estimates highway activity, it is currently being expanded to include arterials.

It is easy to conceive of a path by which this system might be expanded to develop a complete and continuously updated source of data for providing vehicle activity estimates to CARB's mobile source emissions models. We suggest the following steps:

1. Embed the calculations to compute hourly and daily ratios (steps 2 and 3 of the gridding process) directly into CalVAD. Add a CalVAD interface for obtaining the adjustment factors for average weekday adjusting vehicle activity and EMFAC county-wide emissions to any specific day or range of days requested by the analyst.

2. Extended to use CalVAD's existing data-fusion algorithms for computing VMT by vehicle-class for instrumented highways. This would offer the ability to generate county *and* grid-specific factors using the best data from its fusion algorithm.

3. Continued refinement of the adjustment factor formulas if and when new data sources become available.

## 7.2 Long term improvements

The following suggestions require cooperation from external entities, advancements in the state-of-the-art, or some combination of both of these. Though they are speculative in nature, we believe they represent promising directions of development that have potential for significantly supporting CARB's mission.

### 7.2.1 Getting better data

As today's telecommunications systems continue to evolve, and the public, commercial, and private use of connected mobile computing system expands, it is increasingly clear that data is everywhere. The current reliance of formally collected travel activity data in the development of the EMFAC and DTIM model omits a broad class of "softer" data sources that may offer substantial insight into travel activity patterns across every class of vehicle.

In particular, we believe the private sector has enormous potential for providing data regarding general activity patterns. Wireless providers such as AT&T, Verizon, T-Mobile and the like have huge subscription bases of individuals using smart phones and other mobile devices that are inherently aware of people's activity over the course of a day. In addition, the major players in end-to-end consumer services such as Apple and Google, along with their various partners, have complete ecosystems that collect high-fidelity information regarding their user's behavior. Indeed, these providers are developing entire business models based upon knowing what people are doing when. As impressive as these systems are today, their continued growth is almost certain along with increased sophistication.

Furthermore, most major fleet operators employ sophisticated logistics tools to optimize their operations. These tools rely on sophisticated and pervasive telemetry systems. In short, the major fleets know where their assets have been, where they are, and where they are going. These data can clearly identify the major patterns of heavy-duty vehicle operations.

Clearly, such high-quality data offers vast potential improvement over the limited point counts offered by the AVC counts used in the current method. Of course, with data quality comes proprietary value. Thus the major challenge is identifying mutually beneficial public/private partnerships that would provide this data for public modeling while preserving its proprietary value to its owners.

A significant factor on the use of external, and specifically private or commercial data sources, is that such data has proprietary value to its owners. Thus, on the open market the sort of data we're thinking of can be pretty expensive to obtain. However, some companies might be willing to enter into partnerships for various reasons. Companies that value sustainability may be willing to offer certain types of data free of cost because it's consistent with their corporate mission. Alternatively, companies might be willing to discount their costs in exchange for other data that CARB or Caltrans may have that is of value.

Companies might worry about their proprietary information being used in ways contrary to their interests. For instance, fleet operators might be reticent to give CARB detailed information on the activity of their specific fleet. Such concerns might be alleviated through the use of third-party data clearinghouse(s) that collect information from individual fleets and aggregate them for CARB's purposes.

In general, the California's State government is a significant presence in the marketplace and such weight should convey potentially favorable bargaining positions. Further, it is likely that existing partnerships with private companies might offer leads for obtaining new data. For instance, Caltrans currently has a partnership with Google to develop the CT-Earth mapping and asset tracking system. In is possible that this partnership might be extended to include data from Google's own cloud regarding the activity of various vehicle fleets. Another existing partnership that might bear fruit is Nokia's work with Caltrans on the mobile millennium project. Undoubtedly, similar partnerships exist with other entities that may provide useful leads.

### 7.2.2 A fully integrated model

The ideal long-term solution would be the development of an integrated statewide travel demand model capable of generating on-the-fly multi-class vehicle activity estimates of hour-by-hour travel activity for any given day of the year. Ideally, this approach would merge EMFAC and DTIM into a single model system serving policy analysis needs ranging from traditional transportation planning applications to CARB's emissions modeling needs.

This concept would lead to a model that could do away with the somewhat convoluted process used today whereby California MPOs aggregate locally collected link-level data to produce county-wide VMT estimates. These data are used to make countywide emissions estimates that are, in turn, disaggregated down to the 4km grid by using planning model outputs made at the link level. Ideally,

the aggregation could be deferred until after the activity estimates are made in a manner similar to how DTIM was originally designed.

As above, the major requirements are obtaining better data regarding the travel patterns of the population. These, in turn, must feed into models that are sensitive to the potential impacts of policies that may be evaluated with the models. The state of the art, and the focus of research, in the developing travel forecasting models continues to be systems based upon activity-based travel demand models.

Though this is an ambitious goal, there are some models that hold promise, particularly the development of the California Statewide Travel Demand Model (CSTDM) (ULTRANS and HBA Specto Inc., ). This model system includes the ability to perform activity-based, hour-by-hour forecasting of multi-fleet vehicle activity. Though it is a recent model, it contains many of the necessary features for improving statement travel activity estimates.

Beyond the CSTDM work, the continued development of activity-based models by specific MPOs deserves attention. Davidson et al. (2011) provides a recent review that includes specific mention of the San Diego Association of Government's efforts in incorporating activity-based analysis into their core model systems. As California MPOs continue to build out such systems for their regional planning initiatives, the ability to obtain higher resolution, policy sensitive estimates of vehicle activity will improve. The major challenges would include building on the original DTIM work to stitch together these separate models into a single system capable of producing estimates.

Continuing this optimistic thread, we envision tying the statewide model to specific travel activity data collected by public and private agencies, to bound the general travel patterns predicted by the model to best fit these available observations. Significant challenges exist to producing such a model as it would require coordination across multiple agencies and advancements in the state-of-the-art. Still, the progress made in recent years toward improving travel demand models offers some indication that this line of inquiry could bear significant fruit.

Because this long-term solution is dependent on advancements in the state-of-the-art, we can only offer limited recommendations for supporting this long-term goal. Most important is that CARB's emissions modeling efforts remain flexible enough to leverage new models and new data as they become available. Focusing on open data standards and modeling platforms is one approach to this. Another is to architect component-based model systems that can be connected using manageable transformations of data. EMFAC easily meets these needs because it is a more aggregate model. The DTIM system is necessarily more complex and, because of its more disaggregate nature, is more sensitive to the upstream travel demand models that are used in its construction. As CARB looks to improve DTIM, careful consideration should be given to potential advancements in travel demand modeling. Any improvements in the representation of spatio-temporal variability in travel activity—including start-stop behavior that is an important component of emissions modeling—could offer significant benefit to emissions modeling quality.

Any work in this direction should consider the caution offered by Wang et al. (2009) that:

> *emissions calculated based on the traditional travel demand modeling process tend to be underestimated compared to EMFAC and are not sufficiently accurate for air quality research*

# 8 Conclusion

Given the available documentation, our review of the relevant literature, and our direct analysis of the methods used in CARB's emissions models, we cannot identify significant fault with the methods currently used for allocating weekend and heavy-duty vehicle emissions. The approach is consistent with the state-of-the-art in air quality modeling and is reasonable given the available data.

Nonetheless, models of complex systems can always be improved. We have offered two near-term and two long-term suggestions that will benefit CARB's modeling approach. In the near term, we suggest:

- performing rigorous sensitivity and uncertainty analyses of the EMFAC model and the spatio-temporal disaggregation methods using DTIM and temporal factors derived from Caltrans measurements and
- incorporating additional data sources that are currently available for identifying diurnal patterns of activity, particularly those associated with heavy-heavy duty vehicles.

In the longer term, we suggest:

- pursuing additional data sources that are not currently available but that have the potential to offer significant improvements—particularly private fleet data and
- developing a fully integrated statewide model using activity-based travel demand models to provide much better temporal sensitivity to the emissions allocation process.

# 9 References

Barth, M., An, F., Norbeck, J. and Ross, M. (1996). Modal emissions modeling: A physical approach. *Transportation Research Record No. 1520*, pp. 81–88.

Bigazzi, A. Y. and Bertini, R. L. (2009). Adding green performance metrics to a transportation data archive. In *Proceedings of the 88th Annual Meeting of the Transportation Research Board, January 11–15, 2009* . http://www.its.pdx.edu/upload_docs/1249573439.pdf.

Boulton, J. W., Siriunas, K. A., Lepage, M. and Schill, S. (2002). Developing spatial surrogates for modelling applications. In *Proceedings of the 11th International Emission Inventory Conference, Atlanta, GA, April 15–18, 2002* . http://www.epa.gov/ttnchie1/conference/ei11/modeling/boulton.pdf.

California Air Resources Board (2007). Photochemical modeling protocol for developing strategies to attain the federal 8-hour ozone air quality standard in central California: Draft report. Technical Report California Air Resources Board, Planning and Technical Support Division.

California Air Resources Board (2012). Key events in the history of air quality in california. . http://www.arb.ca.gov/html/brochure/history.htm, viewed on June 25th, 2012.

Davidson, B., Vovsha, P. and Freedman, J. (2011). New advancements in activity-based models. In *Australasian Transport Research Forum 2011 Proceedings, 28 - 30 September, 2011, Adelaide, Australia*. Australasian Transport Research Forum. http://www.atrf.info/papers/2011/2011_Davidson_Vovsha_Freedman.pdf.

Davis, N., Lents, J., Osses, M., Nikkila, N. and Barth, M. (2005). Development and application of an international vehicle emissions model. In *Proceedings of the 81st Annual Meeting of the Transportation Research Board January, 2005* . Washington, D.C..

Durham, M. (2006). Dispersion modeling 101: ISCST3 vs. AERMOD. . http://www.tecenv.com/awma_iowa/2006.

EAS-JRC (2012). Who needs sensitivity analysis?. . Published on the world wide web at http://sensitivity-analysis.jrc.ec.europa.eu/docs/who needs SA.htm, viewed June 25th, 2012.

Environmental Protection Agency (2004). *Risk Assessment and Modeling - Air Toxics Risk Assessment Reference Library*, volume 1, chapter Chapter 9. Assessing Air Quality: Modeling Environmental Protection Agency. http://www.epa.gov/ttn/fera/risk_atra_vol1.html.

Environmental Protection Agency (2005). Revision to the guideline on air quality models: Adoption of a preferred general purpose (flat and complex terrain) dispersion model and other revisions; final rule. *Federal Register, 70*(216). 40 CFR Part 51.

Environmental Protection Agency (2010). *Motor Vehicle Emission Simulator: User Guide for MOVES2010a*. .Environmental Protection Agency EPA-420-B-10-036.

Gaffney, P. (2002). Developing a statewide emission inventory using geographic information systems (GIS). In *Proceedings of the 11th International Emission Inventory Conference, Atlanta, GA, April 15–18, 2002* . http://www.epa.gov/ttnchie1/conference/ei11/modeling/gaffney.pdf.

Gifford, F. A. (1960). Atmospheric diffusion calculations using the generalized Gaussian plume model. *Nuclear Safety, 2*(2), 56.

Gkatzoflias, D., Kouridis, C., Ntziachristos, L. and Samaras, Z. (2012). *COPERT: Computer programme to calculate emissions from road transport, User Manual (version 9)*. .ETC/AEM

Gridded Inventory Coordination Group (2006). Proposed method to improve temporal distribution of gridded on-road motor vehicle emissions. Technical Report California Air Resources Board.

Haagen-Smit, A. (1958). Air conservation. *Science, 128*, 869–878.

Hewson, E. W. (1945). The meterological control of atmospheric pollution by heavy industry. *Quarterly Journal of the Royal Meterological Society, 71*, 226.

Institute for the Environment (2011). *SMOKE v3.0 User's Manual*. .The Institute for the Environment - The University of North Carolina at Chapel Hill http://www.smoke-model.org/version3.0/html/.

Kühlwein, J. and Friedrich, R. (2000). Uncertainties of modelling emissions from road transport. *Atmospheric Environment, 34*(27), 4603 - 4610. http://www.sciencedirect.com/science/article/pii/S1352231000003022.

Kühlwein, J. and Friedrich, R. (2005). Traffic measurements and high-performance modelling of motorway emission rates. *Atmospheric Environment, 39*(31), 5722 - 5736. http://www.sciencedirect.com/science/article/pii/S135223100500138X.

Lents, J., Walsh, M., He, K., Osses, N. D. M. and Tolvett, S. et al. (2012). Handbook of air quality management. . This is an incomplete manuscript published at http://www.aqbook.org/ but offers some useful summaries of air quality modeling methods..

Lindhjem, C. E., DenBleyker, A., Jimenez, M., Haasbeek, J. and Pollack, A. K. et al. (2010). Use of MOVES2010 in link level on-road vehicle emissions modeling using concept-mv. In *Proceedings of the 19th International Emission Inventory Conference, San Antonio, Texas - September 27- 30, 2010* . http://www.epa.gov/ttnchie1/conference/ei19/session2/celindhjem.pdf.

Marca, J., Rindt, C. and Recker, W. (2012). The California Vehicle Activity Database (CalVAD) viewer. Technical Report Institute of Transportation Studies at UC Irvine. http://calvad.ctmlabs.net/index.html.

Miller, C. A., Hidy, G., Hales, J., Kolb, C. E. and Werner, A. S. et al. (2006). Air emission inventories in north america: A critical assessment. *Journal of the Air & Waste Management Association*. http://pubs.awma.org/gsearch/journal/2006/8/miller.pdf.

National Research Council (2004). *Air Quality Management in the United States*. Washington, D.C.: National Academy Press.

SCAQMD (2008). Mates III final report. Technical Report South Coast Air Quality Management District. http://www.aqmd.gov/prdas/matesIII/MATESIIIFinalReportSept2008.html.

Sutton, O. G. (1932). A theory of eddy diffusion in the atmosphere. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 135*(826), 143–165.

Tang, T., Roberts, M. and Ho, C. (2003). Sensitivity analysis of MOBILE6 motor vehicle emission factor model. Technical Report FHWA Resource Center AIR QUALITY TEAM. http://www.fhwa.dot.gov/resourcecenter/teams/airquality/tinja.cfm.

Taylor, G. I. (1915). Eddy motion in the atmosphere. *Philosophical Transactions of the Royal Society, 215*, 1–26.

Texas Commision on Environmental Quality (2012). Introduction to air quality modeling. Texas Commision on Environmental Quality. Accessed June 25th, 2012.

ULTRANS and HBA Specto Inc. *CSTDM09 - California Statewide Travel Demand Model*. .ULTRANS

Wagner, K. K., Wheeler, N. J. M. and McNerny, D. L. (1992). The effect of emission inventory uncertainty on urban airshed model sensitivity to emission reductions.. In *Proceedings of the Air & Waste Management Association Inter- national Specialty Conference: Tropospheric Ozone: Nonattainment and De- sign Value Issues, Boston, MA, October 1992.* . http://gate1.baaqmd.gov/pdf/0455_Error_Propagation_California_Air_Resources_Board_Airshed_Model_Volume1_1989.pdf.

Wang, G., Bai, S. and Ogden, J. M. (2009). Identifying contributions of on-road motor vehicles to urban air pollution using travel demand model data. *Transportation Research Part D: Transport and Environment, 14*(3), 168 - 179. http://www.sciencedirect.com/science/article/pii/S136192090800151X.

Zhao, Y. and Kockelman, K. M. (2001). The propagation of uncertainty through travel demand models: An exploratory analysis. *The Annals of Regional Science*. http://www.springerlink.com/content/q8ywu35h0fe7hnp9/.