Analysis of Geographical and Temporal Variation in Vehicle Classification Count Statistics

DAVID L. GREENE, PATRICIA S. HU, and GLENN F. ROBERTS

ABSTRACT

The problem of estimating highway travel by vehicle type using available traffic vehicle classification count data is discussed. The data are analyzed by using techniques of discrete multivariate analysis. It is determined that vehicle type relative frequency distributions vary importantly across regions, highway systems, seasons, day of week, and time of day, but that interactions among these factors, which would complicate estimation of travel, are not of great importance. The only important two-way interactions involve highway system type; therefore it is possible to derive unbiased estimates of vehicle travel by vehicle type and highway system even from a nonrandom sample of classification count observations, provided that total travel by system is known. Some exploratory disaggregate vehicle travel estimates are presented.

The quantities of travel by type of vehicle and type of highway system are fundamental transportation data. Such information is important for analysis and forecasting of travel and energy use and for determining exposure rates in studies of highway safety. Vehicle survey data are useful for estimating travel by vehicle type, but not by highway system (1,2). In order to obtain travel estimates disaggregated in both dimensions, vehicle classification count statistics are needed. Classification count data consist of hourly counts of vehicles by type that are recorded at a particular location on the highway system network. Determining disaggregate travel by vehicle and highway type is thus a problem of inferring vehicle miles from vehicle counts.

If there were a sufficiently large, well-designed random sample of traffic counts, deriving unbiased estimates of vehicle travel would be, in principle, a rather simple exercise. Unfortunately, while there is a great volume of classification count data, none has been collected according to a statistically designed sampling plan. The problem is then one of removing, to the greatest extent possible, the bias inherent in the existing sample. To do this effectively, the variation in vehicle type distributions across time and space must be understood. If temporal and spatial dimensions affect the distribution of vehicle types independently, then sample bias can be corrected by a simple reweighting of the data.

This paper is divided into three parts. In the first part a probabilistic model of vehicle type relative frequencies, which helps to clarify the relationship between vehicle miles and vehicle counts by vehicle type, is presented. Second, the three major sources of vehicle classification data are described, and the results of an analysis of the structure of classification count data using log-linear models are presented. The implications for using available data to estimate disaggregate vehicle travel are discussed. Finally, several preliminary estimates of travel by 13 vehicle types and 10 highway system classes are presented and discussed. In the concluding section the interpretation of these estimates is discussed and important areas for further research are recommended.

STATISTICAL MODEL OF VEHICLE TYPE COUNTS AND TRAVEL

Traffic counts do not represent vehicle travel but rather represent density at a point on a road. Thus a set of assumptions must be specified by which vehicle travel estimates can be derived from vehicle count data. It is shown that if a functional class can be divided into homogeneous systems, then an unbiased estimate of vehicle type relative frequencies can be obtained as a weighted average of the estimated system relative frequencies. This result will be used in the section Exploratory Disaggregate Estimates of Vehicle Travel to estimate relative frequencies and travel by vehicle type for functional highway classes. The systems used will be regional functional classes classified by season, day of week, and time of day. The analysis of the variability of vehicle type relative frequencies across these systems in the next section will show that a particularly simple weighting scheme can be used that permits weights for temporal dimensions to be constant across systems. That the systems defined may in fact not be homogeneous is a persistent problem that can only be solved by improved random sampling strategies.

Assume that a functional highway class (see Table 1) is divided into segments that are sufficiently small and homogeneous that vehicle miles on the segment are equal to its length times a traffic count taken anywhere on the segment. The segment then forms the basic unit of analysis because there is nothing to be gained by subdividing it.

A collection of segments with identical (in practice, similar) traffic densities and vehicle type distributions are called a system in this paper. Clearly, a given functional highway class (e.g., urban Interstate) may be made up of several different systems. In fact, the same strip of road can be considered to belong to different systems, depending on the time of day or season of the year. In this sense a functional class has no underlying parameters of its own to be estimated, but rather is merely a sum of individual systems. Because the goal is to make inferences about vehicle miles of travel on functional highway classes, these will be derived from weighted averages of inferences about the systems that compose it. In particular, for the purposes of this study, the interest is in inferring the distribution of vehicle miles by vehicle type for each functional class.

Assume that an observer, standing at a roadside recording vehicle counts for a fixed time period such as an hour, is observing a random process. In particular, if N total counts are recorded during the period, assume that the probability of observing

TABLE 1 Variables and Categories

Variable	Number	Category
Quarter	4	1st
		2nd
		3rd
		4th
Region	4	Northwest
		South
		North Central
		West
Road type	10	Interstate, rural
		Other principal arterials, rural
		Minor arterials, rural
		Major collectors, rural
		Minor collectors, rural
		Interstate, urban
		Other freeways, urban
		Other principal arterials, urban
		Minor arterials, urban
		Collectors, urban
Day	2	Weekday
		Weekend
Time of day	5	5:00-9:00 a.m.
		9:00 a.m.-3:00 p.m.
		3:00-7:00 p.m.
		7:00-11:00 p.m.
		11:00 p.m.-5:00 a.m.
Vehicle type	13	Standard and compact cars
		Subcompact cars
		Motorcycles
		Buses
		Pickups, panels, and other two-axle, four-tire trucks
		Two-axle, six-tire single-unit trucks
		Two-or-more-axle single-unit trucks
		Three-axle combination trucks
		Four-axle tractor-semicombinations
		Other four-axle combinations
		Three-axle tractor, two-axle semicombinations
		Other five-axle combinations
		Six-or-more-axle combinations

C_1 vehicles of type 1, C_2 of type 2, up to C_m of type m is given by the multinomial distribution,

$$P(C_1, C_2, \ldots, C_m) = N! \prod_{k=1}^{m} p_k^{C_k} / \prod_{k=1}^{m} C_k! \quad (1)$$

The p_k's are the probabilities of observing a vehicle of type k in a sample of one, or alternatively, the relative frequencies of type k vehicles in the total population of vehicles traveling the given system. Also, it is required that

$$\sum_{k=1}^{m} C_k = N.$$

In general, the total number of counts recorded in an hour will itself be a random variable. Assume that the number of counts observed will follow a Poisson distribution. The Poisson is widely used both in traffic engineering and elsewhere to represent random arrivals (3):

$$P\lambda(N) = e^{-\lambda} \cdot (\lambda N/N!) \quad (2)$$

The Poisson distribution has expected value (mean) and variance both equal to λ. Compounding the Poisson and multinomial distributions in this way results in a distribution in which each of the vehicle type counts is distributed Poisson with parameter $\lambda_k = p_k\lambda$ (4),

$$P(C_1, C_2, \ldots C_m) = \prod_{k=1}^{m} \left\{ e^{-\lambda} p_k [(\lambda p_k)C_k/C_k!] \right\} \quad (3)$$

From this model some useful results concerning estimators of system traffic densities and vehicle frequencies can readily be derived.

Maximum likelihood estimators of λ and p_k can be obtained by taking derivatives of the log likelihood function,

$$\max_{\lambda, p_k, k} \log(P) = \sum_{k=1}^{m} [-\lambda p_k + C_k \log(\lambda p_k) - \log(C_k!)] \quad (4)$$

setting them equal to zero and solving for λ and p_k:

$$\partial \log(P)/\partial p_k = -\lambda + C_k (1/p_k)$$

$$\partial \log(P)/\partial \lambda = -\sum_{k=1}^{m} p_k + (1/\lambda) \sum_{k=1}^{m} C_k \quad (5)$$

Setting these equal to zero, and because

$$\sum_{k=1}^{m} p_k = 1,$$

then

$$\hat{p}_k = C_k/\lambda$$

$$\hat{\lambda} = \sum_{k=1}^{m} C_k \quad (6)$$

The unbiasedness of these estimators can be shown by taking expected values:

$$E(\hat{p}_k) = (1/\lambda) E(C_k) = (1/\lambda)\lambda p_k = p_k$$

$$E(\hat{\lambda}) = \sum_{k=1}^{m} E(C_k) = \sum_{k=1}^{m} \lambda p_k = \lambda \sum_{k=1}^{m} p_k = \lambda \quad (7)$$

In general, however, the actual λ will not be known in order to be able to estimate $\hat{p}_k$, and instead $\hat{\lambda}$ will have to be used. By using an approximation from Mood et al. (5) it can easily be shown that their quotient is unbiased at least up to a second-order Taylor series approximation,

$$E(p_k') = E(C_k/\hat{\lambda}) \doteq (p_k\lambda/\lambda) - (1/\lambda 2)\, p_k\lambda + (p_k\lambda/\lambda 3)\,\lambda = p_k\lambda/\lambda = p_k \quad (8)$$

This result follows from showing that

$$\mathrm{Cov}\left(C_k, \sum_{j=1}^{m} C_j\right) = \mathrm{Var}(C_k) = p_k\lambda.$$

(The full proof is available on request from the authors.)

These estimators are appropriate for estimating the parameters of a system composed of essentially homogeneous road segments. If a random sample of segments is taken from the same system, then these estimators can be used to obtain maximum likelihood, unbiased estimates of the system parameters. A functional road class in a given region and time period will most likely be composed of several systems. In a sense, it has no underlying parameters of its own but, rather, is merely a summation of individual systems. In particular, it is clear that the relative vehicle mile frequencies (f_k) for vehicle types k = 1, ..., m are just the weighted averages of those of all systems in the class:

$$f_k = \left(1/\sum_{i=1}^{S} T_i\right) \cdot \sum_{i=1}^{S} p_{ki} T_i = \sum_{i=1}^{S} r_i p_{ki} \quad (9)$$

Systems are indexed by i = 1, ..., S, T_i to represent total vehicle miles of travel on system i, and

r_i is the proportion of total functional class travel occurring on system i.

Unfortunately, the actual travel on a system is not generally known. However, it is known that on a segment j travel is

$$T_j = \lambda_i \ell_{ij} \tag{10}$$

where ℓ is the segment length and λ is the system average traffic count rate. For the system,

$$T_i = \lambda_i \sum_j \ell_{ij} \tag{11}$$

If $\Sigma\ell_{ij}$ is known, the maximum likelihood, unbiased estimator of λ can be used to estimate T_i (this estimator will also be unbiased). Then f_k can be straightforwardly estimated by substituting Equation 11 into Equation 9.

Suppose that N samples (a sample being, for example, a 1-hr vehicle count on a segment) are taken from different systems, where n_i is the number of samples from system i. To obtain an unbiased estimate of the true weighted average for the functional class, it follows from Equation 9 that parameter estimates from each sample must be weighted in proportion to total vehicle miles from each system. This is readily done by using counts and system mileage as demonstrated in Equation 11. The important result here is that to obtain an unbiased estimate of the vehicle type distribution, it is only necessary to have traffic counts and system lengths.

STRUCTURE OF VEHICLE CLASSIFICATION COUNT DATA

Four data bases were supplied by FHWA. One data base contains estimates of total vehicle miles of travel (VMT) by state and highway class (corresponding to FHWA, table VM-2). The remaining three data bases contain vehicle type count records from (a) the Highway Performance Monitoring System (HPMS) case study (6), (b) various truck weight study (TWS) counts, and (c) various traffic counts conducted by states for their own purposes.

The TWS and HPMS data are both large data bases of equivalent size. The HPMS contains 27,070 usable hourly records and the TWS contains 32,650 such records. The distribution of these records by functional class, however, is extremely different. The HPMS cases are divided about equally: 13,246 rural and 13,824 urban. The TWS, on the other hand, is heavily biased toward rural roads, with 27,158 rural cases and only 5,492 urban ones. Geographically, the TWS used for this study is more comprehensive, with data from 22 states, with at least 1 in each of the 9 census regions. Because it is a case study, the HPMS includes data from only four states and one planning region: Arkansas, Iowa, Minnesota, Washington, and the Delaware Valley. In terms of traffic counts, the two data bases are roughly equal in size, with each having just more than 10 million counts.

The chief problem with vehicle count data is that it has not generally been gathered in accordance with statistical sampling procedures designed to produce comprehensive coverage for the entire United States. Instead, counts have been taken for different purposes and at different times under varying conditions. In short, what has been produced is a nonrandom sample. Most techniques of statistical inference are designed to be applied to a random sample. The challenge in working with a nonrandom sample lies in discovering ways to eliminate the bias inherent in the sample (e.g., weekdays may be oversampled relative to weekends, or daytime hours oversampled relative to nighttime hours). One aspect of the sample bias that cannot be corrected within the scope of this project is the choice of traffic count observation locations on the road network. In terms of theory offered in this paper, this is to say that the authors may not be able to work with homogeneous systems. From the viewpoint of this analysis, the choice of traffic count locations must be assumed to be representative of or a random sample within a particular functional class and region.

Each data base was reorganized into a table of total count frequencies classified by quarter, day, time, region, functional class, and vehicle type. The categories of each variable used are given in Table 1. Thus the cell labeled spring, weekday, 9:00 a.m.-3:00 p.m., region 4, rural Interstate, motorcycles, would contain the sum of all motorcycle counts from all observations having those attributes in the data base in question. Certain vehicle categories were combined so that no variable had more than 10 categories, a requirement of the statistical software that was used. The result is a six-dimensional table with a total of 16,000 cells, many of which are empty for any given data base. In terms of the theory, each cell is considered to be a homogeneous system.

The technique of discrete multivariate analysis using log-linear models is used to analyze tables of frequency data cross-classified by categorical variables. Consider a three-way table of traffic counts by vehicle type (V), functional highway class (C), and region (R). The lower case letters i, j, k are used to index the levels (or categories) of the variables V, C, R; and I, J, K are the number of levels in each category. Let f_{ijk} be the observed frequency (count) in cell i, j, k of the table (matrix). Log-linear modeling assumes that the logarithm of the expected cell count $[E(f_{ijk}) = F_{ijk}]$ is a linear function of certain parameters associated with individual effects of each variable and interactions of variables. If the variable symbols are used as superscripts and the variable indices are used as subscripts to indicate the level of each variable, the model can be written as

$$\ell_n F_{ijk} = \theta + \lambda_i^V + \lambda_j^C + \lambda_k^R + \lambda_{ij}^{VC} + \lambda_{ik}^{VR} + \lambda_{jk}^{CR} + \lambda_{ijk}^{VCR} \tag{12}$$

The λ's are usually called effects and the superscript identifies to which variable or interaction of variables the effect pertains. In Equation 12, λ^V, λ^C, λ^R are the main effects of variables V, C, R, in which λ^{VC}, λ^{VR}, λ^{CR} are their two-way interaction and λ^{VCR} is their three-way interaction. Clearly, the table of frequency counts contains only IJK cells, whereas Equation 12 specifies (1 + I + J + K + IJ + IK + JK + IJK) parameters. To eliminate this parameter redundancy, the following constraints are imposed:

$$\sum_i \lambda_i^V = 0, \sum_j \lambda_j^C = 0, \sum_k \lambda_k^R = 0$$

$$\sum_i \lambda_{ij}^{VC} = \sum_j \lambda_{ij}^{VC} = \sum_i \lambda_{ik}^{VR} = \ldots = \sum_k \lambda_{jk}^{CR} = 0$$

$$\sum_i \lambda_{ijk}^{VCR} = \sum_j \lambda_{ijk}^{VCR} = \sum_k \lambda_{ijk}^{VCR} = 0 \tag{13}$$

With the constraints of Equation 13, the model (Equation 12) has exactly as many parameters as there are cells in the table. If all parameters were estimated, the model would fit the table exactly. The model (Equation 12) is termed the saturated model because it includes all possible effects. In general, all effects are not statistically

significant, and thus the identification of a log-linear model consists of determining which effects are needed, and which λ terms are superfluous.

Generally, only hierarchical models are considered. In a hierarchical model, a higher-order interaction effect is included only if all lower-order effects involving the variables in the higher-order effects are also included. Thus if λ^{VR} is included, λ^{V} and λ^{R} must also be included. When only hierarchical models are considered, each model can be described as a minimal set of higher-order effects. For example, specifying the hierarchical model (VC, CR) is equivalent to the model (θ, V, C, R, VC, CR). In this fitted model, the marginal sums associated with V, C, R, VC, CR, and the table total will exactly equal those of the original table. Thus in the hierarchical model including the parameter VR is equivalent to exactly fitting the IxK marginal table formed by summing over j (the levels of the variable C).

Log-linear models are useful for understanding the relationships between variables in a table and for estimating a table of expected frequency counts using a fitted model. What needs to be known are the important relationships among functional highway class (C), season (Q), day of week (D), region (R), and time of day (T), and also the distribution of traffic counts by vehicle type (V). The procedure consists of estimating a new table that fits a subset of the six-way table margins and measuring the degree to which it fits the original table. The degree of fit is measured by means of the likelihood ratio χ^2 statistic,

$$\chi^2 = 2 \sum_{IJK} f_{ijk} \ln(f_{ijk}/F_{ijk}) \quad (14)$$

which is asymptotically distributed as χ^2 with degrees of freedom equal to the number of cells minus the number of parameters to be estimated.

In order to test the significance of a particular parameter (e.g., λ^{AB}) in the model (Equation 12), the difference in χ^2 is computed between the hierarchical model that includes this term,

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} \quad (15)$$

and the model that includes all the same terms except λ^{AB},

$$\ln F_{ijk} = \theta + \lambda_j^A + \lambda_j^B + \lambda_k^C \quad (16)$$

The difference in the two models' χ^2 is also distributed χ^2 with degrees of freedom equal to the difference in degrees of freedom of the two models [here (I-1)(J-1)]. By testing Equation 15 versus Equation 16, it is actually a test of whether A and B influence cell counts independently or whether they interact in determining cell counts.

Log-linear analysis allows simultaneous interaction of all variables. In some cases it is reasonable to consider one variable a dependent variable that is affected by the other variables but does not influence them. In the present case vehicle type should be considered the dependent variable (e.g., vehicle type does not influence the number of counts on weekends versus weekdays, rather the reverse). When one variable is considered the dependent variable and all others are independent variables, the joint marginal of the independent variables must always be fitted. In the six-way traffic count table the CDTQR margin must always be fitted. Given this, the interest is in testing hypotheses about only those terms involving V.

In general, the technique of log-linear model analysis is applied to a random sample of data. When the data have not been collected by means of a simple random sample, it is necessary to fit additional marginals to control for the fact that the sample size (marginal sums) in some combinations of categories has been determined exogenously. In the case of the traffic count data, the only factor that is in fact random is the number of counts by vehicle type for a given observation. Everything else has been determined by the peculiarities of the traffic count sample frame. This requires that the CDTQR margin be fitted exactly. Fortunately, this is the same requirement imposed when V is considered the dependent variable.

The analysis of the traffic count data proceeds by adding terms to the null model (CDTQR) to form successively more complex models involving V. A stepwise procedure of the BMDP4F statistical software package was used. Each of the three traffic count data bases (HPMS, TWS, and state data) were analyzed separately. Because of the extremely large sample sizes of these data bases (on the order of 5 to 10 million counts), every conceivable effect is significant at commonly used significance levels (e.g., 0.05, 0.01). The reason for this is that, in a very large sample, even the most trivial differences can be detected with great accuracy. To determine which parameters are important and which are trivial, some other measure is needed. Goodman (7) has suggested a quasi-R^2 (coefficient of multiple determination) based on the percentage reduction in χ^2 brought about by introducing an additional parameter. In the present case the interest is in percentage reductions in χ^2 over the null model (CDTQR, V) brought about by adding interaction terms involving V.

Because only hierarchical models are considered, a shorthand notation is used in which only the highest-order terms are mentioned. For example, the following two are equivalent:

CDTQR, VCT, VCR

and

C, D, T, Q, R, CD, CT, CQ, CR, DT, DQ, DR, TQ, TR, QR, CDT, CDQ, CDR, CTQ, CTR, CQR, DTQ, DTR, DQR, TQR, CDTQ, CTQR, CDQR, DTQR, CDTR, CDTQR, VCT, VC, VT, VCR, VR, V.

A limitation of the BMDP software (8) is that no more than 10 categories can be defined for a single variable. It was therefore necessary to combine three vehicle type categories. Single-unit truck counts, except pickups and so forth, were combined into one class, as were all four-axle combinations and five-axle combinations.

The stepwise procedure begins with the basic (null) model CDTQR, V and adds terms. Results for the HPMS data are given in Table 2. The individual effect of each variable on the vehicle type distribution is captured by the two-way interactions with V. Region and road class appear to be the most important influences. Day and time effects are only about one-third as potent and the quarter effect is almost negligible. When all the two-way effects are included in the model, the percentage of χ^2 accounted for increases to 83. Interestingly, this is almost exactly equal to the sum of χ^2 reductions the individual effects (84), an indication that interactions of higher order may not be important.

Examination of χ^2 reduction due to three factor interactions indicates that only the class-region interaction reduces χ^2 by more than 1 percent.

TABLE 2 Stepwise Analysis of HPMS Traffic Count Data

Model	Degrees of Freedom	Likelihood-ratio χ^2	Quasi-R^2
CDTQR, V	9,171	1,417,160	0.0
Individual Two-Way Interactions			
CDTQR, VC	9,054	1,018,124	0.28
CDTQR, VD	9,162	1,288,042	0.09
CDTQR, VT	9,135	1,281,126	0.10
CDTQR, VQ	9,144	1,370,291	0.03
CDTQR, VR	9,153	931,127	0.34
CDTQR, VC, VD, VT, VQ, VR	8,964	245,229	0.83
Individual Three-Way Interactions			
CDTQR, VT, VQ, VR, VCD	8,865	239,917	0.83
CDTQR, VD, VQ, VR, VCT	8,598	227,791	0.84
CDTQR, VD, VT, VR, VCQ	8,676	231,348	0.84
CDTQR, VD, VT, VQ, VCR	8,838	182,694	0.87
CDTQR, VC, VQ, VR, VDT	8,928	229,811	0.84
CDTQR, VC, VT, VR, VDQ	8,937	242,053	0.83
CDTQR, VC, VT, VQ, VDR	8,946	229,074	0.84
CDTQR, VC, VD, VR, VTQ	8,856	239,964	0.83
CDTQR, VC, VD, VQ, VTR	8,892	237,322	0.83
CDTQR, VC, VD, VT, VQR	8,910	237,410	0.83

TABLE 3 Stepwise Analysis of TWS Traffic Count Data

Model	Degrees of Freedom	Likelihood-ratio χ^2	Quasi-R^2
CDTQR, V	3,330	870,986	0.0
Individual Two-Way Interactions			
CDTQR, VC	3,224	601,158	0.31
CDTQR, VD	3,321	868,682	0.003
CDTQR, VT	3,294	690,654	0.21
CDTQR, VQ	3,303	825,040	0.05
CDTQR, VR	3,303	710,485	0.18
CDTQR, VC, VD, VT, VQ, VR	3,125	227,259	0.74
Individual Three-Way Interactions			
CDTQR, VT, VQ, VR, VCD	3,070	224,809	0.74
CDTQR, VD, VQ, VR, VCT	2,802	208,984	0.76
CDTQR, VD, VT, VR, VCQ	2,968	203,739	0.77
CDTQR, VD, VT, VQ, VCR	2,915	151,177	0.83
CDTQR, VC, VQ, VR, VDT	3,089	226,411	0.74
CDTQR, VC, VT, VR, VDQ	3,106	225,176	0.74
CDTQR, VC, VT, VQ, VDR	3,107	223,694	0.74
CDTQR, VC, VD, VR, VTQ	3,017	224,779	0.74
CDTQR, VC, VD, VQ, VTR	3,017	219,974	0.75
CDTQR, VC, VD, VT, VQR	3,035	140,305	0.84
CDTQR, VD, VT, VQR, VCR	2,825	74,329	0.92

TABLE 4 Stepwise Analysis of State Traffic Count Data

Model	Degrees of Freedom	Likelihood-ratio χ^2	Quasi-R^2
CDTQR, V	3,942	1,251,697	0.0
Individual Two-Way Interactions			
CDTQR, VC	3,799	751,830	0.40
CDTQR, VD	3,933	1,226,458	0.02
CDTQR, VT	3,906	1,138,397	0.09
CDTQR, VQ	3,915	1,186,096	0.05
CDTQR, VR	3,856	844,404	0.33
CDTQR, VC, VD, VT, VQ, VR	3,641	258,141	0.79
Individual Three-Way Interactions			
CDTQR, VT, VQ, VR, VCD	3,595	252,306	0.80
CDTQR, VD, VQ, VR, VCT	3,278	163,581	0.87
CDTQR, VD, VT, VR, VCQ	3,395	220,831	0.82
CDTQR, VD, VT, VQ, VCR	3,406	222,791	0.82
CDTQR, VC, VQ, VR, VDT	3,605	255,681	0.80
CDTQR, VC, VT, VR, VDQ	3,615	256,210	0.80
CDTQR, VC, VT, VQ, VDR	3,624	253,473	0.80
CDTQR, VC, VD, VR, VTQ	3,533	247,105	0.80
CDTQR, VC, VD, VQ, VTR	3,531	233,086	0.81
CDTQR, VCR, VTR, VD, VQ	3,310	205,838	0.84

These results suggest that spatial variation in traffic distributions dominates temporal variation. Simple two-way region and road class interactions with vehicle type reduce χ^2 by 34 and 28 percent, respectively. The effect of season only reduces χ^2 by 3 percent. This suggests that for the HPMS data base, at least, little would be lost by ignoring the seasonal variation in vehicle type relative frequencies (not total counts, because these have been accounted for by the CDTQR terms). The results also suggest that each factor, class, day, time, region, and road class can be considered approximately independent of the others in its effect on vehicle type relative frequency.

Because HPMS includes only five states covering four regions, the importance of region might be expected to be greater than in the other data sets where each regional effect is the average of several possibly different states Second, the HPMS has by far the most complete coverage across all other variables. This is simply a result of the fact that the HPMS is a systematic data-gathering program. The sampling system ensured good coverage by day of week, season, time of day, and road class. The other data sets are not systematic and generally have large gaps (e.g., weekends at night are sparsely sampled). In brief, it should be expected that the variable region in the HPMS data base is in fact representing particular states. On the other hand, variables such as time or day in the other data sets could possibly be assumed for particular states that reported data for odd times while others did not.

Log-linear model analyses of the TWS counts are summarized in Table 3. The general pattern is similar to that of the HPMS. Functional highway system class is the most important single factor. Region and time of day are considerably less important, and day of week and quarter are almost negligible. The VC term alone accounts for a 31 percent reduction in χ^2. All two-way interactions account for 74 percent as compared with 83 percent in the HPMS. Interactions are somewhat more important. The road class-region interaction and quarter-region interaction appear to be most important. Including both reduces the lack of fit by 92 percent.

The state data base analysis results show strong similarity to that of the HPMS (Table 4). Road class and region appear to be the most influential factors. Time of day is considerably less important, and day of week and quarter are again almost negligible. Two-way road class and region interactions with vehicle type reduce χ^2 by 40 and 33 percent, respectively. Including all two-way interactions accounts for 79 percent of reduction in χ^2. The road class-time interaction with vehicle type reduces χ^2 by 87 percent. This suggests that road class and time of day depend on each other in their effect on vehicle type relative frequencies.

EXPLORATORY DISAGGREGATE ESTIMATES OF VEHICLE TRAVEL

The results of the log-linear analysis imply a simple basic structure to traffic count data. The distribution of vehicle traffic among vehicle types does vary across time and space. But the only important interaction effects are two-way effects that

include highway class. In terms of the theory, this means that the effect of time of day does not vary across systems within a highway class. Therefore, to develop estimates of travel by type of vehicle disaggregated by highway class, the remaining dimensions can be weighted independently. This result was used to produce some experimental estimates of disaggregate regional travel. These estimates are experimental because biases not controlled in this analysis (e.g., network location) can, and probably do, still influence the results. It should also be noted that the authors did not have complete coverage of states in this sample.

The estimation process consists of two components: (a) weighting and collapsing categories and combining data bases, and (b) using the final processed data to estimate travel. To appreciate the role of preprocessing and weighting in the estimation process, it is useful to begin with a description of the second and final step. Assuming that there is either a random sample or that the vehicle classification count data are weighted to correct for sample bias, the estimation of disaggregate vehicle travel is relatively straightforward. Let c represent traffic counts that are indexed by i = 1, 2, ..., I for vehicle types; j = 1, 2, ..., J for regions; and k = 1, 2, ..., K for highway functional class. All other dimensions (e.g., time of day, day of week, season) have been eliminated in the weighting and dimension collapsing process. The absence of a subscript will be used to signify that counts have been summed over that dimension. For example,

$$\sum_i c_{ijk} = c_{jk},$$

which is the total count for all vehicles in region j on functional class k. From the c_{ijk} and c_{jk}, the relative frequencies are computed for each vehicle type, region, and functional class, which is represented by f_{ijk},

$$f_{ijk} = c_{ijk}/c_{jk}.$$

Recall that if there is a random sample or if the bias in the sample has been eliminated through weighting, then the vehicle miles by each vehicle type i should be proportional to f_{ijk} for all i = 1, 2, ..., I. This of course applies only to the appropriate region and road system. Given this fact, and the fact that

$$\sum_i f_{ijk} = \sum_i (c_{ijk}/c_{jk}) = c_{jk}/c_{jk} = 1,$$

the f_{ijk} can be used to distribute total VMT, on a given functional class in a particular region, among the various types of vehicles. Let T_{jk} denote travel in region j on functional class k; then

$$T_{ijk} = f_{ijk}T_{jk}$$

is the estimate of disaggregate vehicle travel. If summed across vehicle types, the analyst will get back the total vehicle travel in region j, functional class k, with which he began:

$$\sum_i T_{ijk} = T_{jk} \sum_i f_{ijk} = T_{jk} \cdot 1.$$

The key assumption made is that once the three-dimensional array of traffic counts c_{ijk} is arrived at, any bias in the data has already been removed. In general, this will not be true unless there is a reasonably well-designed sample to begin with. Bias in the sample may arise from three principal sources, only one of which can be corrected:

1. Location bias, which results from collecting counts on an atypical location on the road network;
2. Time-space bias, which results from a nonrandom allocation of observations over time, across functional classes, across regions, and even across states within a region; and
3. Missing data for any category, especially states or highway classes.

Only the second kind of bias can be mitigated. This can be done by the weighting of categories in advance.

The weighting process is best illustrated by example. Suppose that the dimensions and categories given in Table 1 are used. Statistical analyses of the three major vehicle classification count data bases indicated that vehicle type frequency distributions vary across all these dimensions and categories. For example, for any given day of week, season, region, and functional class the distribution of traffic by vehicle type will be different at different times of the day. Therefore if there are twice as many daytime as nighttime observations, the final estimate of total travel by vehicle type will be biased toward the daytime pattern. It was also noted that the vehicle type frequency distribution varies jointly by functional class and time of day and by functional class and region. Because the final estimates will be by functional class and region, this does not complicate matters. To get unbiased estimates, the weights of observations by time, day, and season need only to be corrected independently.

Suppose that half of the observations (records, not counts) were taken on weekends and half on weekdays. This represents sample bias because a uniform distribution over time would give 2/7 on weekends and 5/7 on weekdays. To correct this bias a weight of 2 for weekends and 5 for weekdays can be specified. Because it is known in advance what the distribution of samples over time in an unbiased sample should look like, it is simple to weight categories of temporal dimensions.

Unfortunately, by weighting the sample observations, there is a trade-off of a reduction in bias for a loss in efficiency. To see this, imagine that there were 10^5 weekday observations in the data but only 10 weekend observations. By using a 5:2 weighting, the bias is reduced in theory but the variance (decrease in reliability) of the estimate is greatly increased. The reason is that while there is a great deal of information about weekday travel, next to nothing is known about weekend travel, and yet the data are used as if the analyst had 0.4×10^5 weekend observations. In practice, caution should be exercised when weighting observations when the input data are extremely maldistributed. In such cases it may be better not to try to correct for sample bias at all.

In the same way that categories of a dimension can be weighted and summed, data from different data sets can also be assigned weights and combined. The weights may reflect the analyst's confidence in a particular data set or simply the actual number of observations in each. This allows several data sets to be processed (categories weighted and dimensions collapsed) individually, combined at any desired point, and then further processed as a combined set.

Three sets of disaggregate vehicle travel estimates by region and functional class were produced based on 1980 VMT by state and functional class. [Note that these data are from the FHWA, U.S. De-

partment of Transportation (1982). Used were tables of "Vehicle miles of travel classified by state and functional class highway category" for 1980 and 1981, table VM-1, "Annual Vehicle Miles of Travel and Related Data--1981," and three traffic count data tapes supplied by Paul Svercl of Highway Planning, Highway Statistics Branch.] The first two sets are based on the traffic count data from the TWS and HPMS data bases. The state data base was not used because inconsistencies in its method of vehicle classification could not be resolved. For each state, traffic counts are given a weight proportional to total state VMT. In general, this changed the results little in comparison with counts not weighted by state. Finally, a combined set of estimates was produced based on the state VMT-weighted data from both data sets. Observations for a given region from the HPMS and TWS data were given equal weight, even though the TWS always represented more states.

The estimates based on weighted traffic counts are given in Tables 5-7. Vehicle categories have been combined to reduce the size of the tables and also because four vehicle types--large cars; small cars; two-axle, four-tire trucks; and 3S-2 semitrailers (18 wheelers)--account for virtually all the vehicle travel. Some vehicle categories never achieve as much as 1 percent of total travel in any region.

Some general patterns of vehicle travel hold up across regions and data bases. For example, combinations or semitrailers are always most prevalent on rural Interstates and are less common the lower the order of the road system. Also, the distribution of total regional travel among vehicle types varies importantly, but not drastically, across regions and data bases. For example, the West and South always show the most single-unit trucks, mostly two-axle, four-tire (pickup) trucks. Finally, it appears from these data that combination trucks may account for a greater percentage of total vehicle miles than previously thought, possibly by as much as a factor of 2 (it should be noted that the tables do not include local roads, which account for 14 percent of the 1980 VMT). This holds for both the HPMS and TWS data bases. In 1980, the FHWA estimated that 3.7 percent of total U.S. highway miles were by combination trucks. The exploratory estimates from this research are considerably higher.

The travel estimates represented in these tables represent direct empirical estimates based on the available data. Because of problems with these data,

TABLE 5 Estimates of VMT by Highway Category and Vehicle Type (1981), HPMS Data Only

	Vehicle Type							
					Trucks			
	Cars and Motorcycles		Buses		Single Unit		Combination	
Road Type	VMT (10^9)	Percent	VMT (10^9)	Percent	VMT (10^9)	Percent	VMT (10^9)	Percent
Rural Interstate	70.1	59.3	0.4	0.3	24.5	20.7	23.2	19.6
Rural arterial	133.1	59.7	0.9	0.4	70.0	31.4	18.9	8.5
Rural other	79.4	56.1	1.1	0.7	51.8	36.6	9.4	6.6
Urban Interstate	109.4	68.2	0.4	0.2	33.9	21.1	16.7	10.4
Urban other	385.6	74.4	2.0	0.4	117.7	22.7	12.9	2.5
Total	777.6	67.0	4.8	0.4	297.9	25.7	81.1	6.9

TABLE 6 Estimates of VMT by Highway Category and Vehicle Type (1981), TWS Data Only

	Vehicle Type							
					Trucks			
	Cars and Motorcycles		Buses		Singe Unit		Combination	
Road Type	VMT (10^9)	Percent	VMT (10^9)	Percent	VMT (10^9)	Percent	VMT (10^9)	Percent
Rural Interstate	77.7	57.6	0.4	0.3	26.7	19.8	30.1	22.3
Rural arterial	177.9	67.8	0.5	0.2	60.6	23.1	23.3	8.9
Rural other	110.7	70.1	0.6	0.4	34.8	22.0	11.9	7.5
Urban Interstate	117.8	73.4	0.3	0.2	26.4	16.5	15.9	9.9
Urban other	355.7	76.1	0.9	0.2	96.6	20.7	14.0	3.0
Total	839.8	71.0	2.7	0.2	245.1	20.7	95.2	8.1

TABLE 7 Estimates of VMT by Region and Vehicle Type (1981), HPMS and TWS Data Combined

	Vehicle Type							
					Trucks			
	Cars and Motorcycles		Buses		Single Unit		Combination	
Region	VMT (10^9)	Percent	VMT (10^9)	Percent	VMT (10^9)	Percent	VMT (10^9)	Percent
Northeast	132.9	79.2	0.9	0.5	23.7	14.1	10.3	6.1
South	286.6	63.7	1.4	0.3	118.0	26.2	43.8	9.7
North	242.3	72.5	1.0	0.3	66.7	20.0	24.1	7.2
West	147.0	66.6	0.6	0.3	63.2	28.6	10.0	4.5
Total	808.8	67.0	3.9	0.3	271.6	23.2	88.2	7.5

it is not possible to quantify the accuracy of these estimates with any precision. Three of the four regions are missing data for one road type. The Northeast and South are missing minor rural collector data, and the West is missing data for other urban expressways. In addition, not all states are represented, and there are good reasons to believe that routes high in truck traffic were oversampled.

CONCLUSIONS

Vehicle classification count data are the sole source of information on vehicle travel by type of vehicle, highway system class, and geographical area. Although a great deal of classification count data has been collected, it has not been collected according to statistically unbiased sampling procedures, and this presents serious problems for estimation of vehicle travel. Discrete multivariate analysis of the classification count data has revealed a simple structure to the variation in vehicle type distributions across time and space. Vehicle type relative frequencies vary by region, highway system, day of week, time of day, and season. There are also important interactions between highway class and region, and highway class and time. Vehicle type relative frequencies vary most across the geographical dimensions (regions and highway systems), although temporal variations are also important. The combination of all main effects and two two-way interaction effects accounts for about 90 percent of the variation (as measured by reduction in χ^2) in vehicle type relative frequencies in three different vehicle classification count data bases.

This result implies that sample bias in classification count data along these five dimensions can be corrected relatively easily if vehicle travel by highway class and region, as well as by vehicle type, is being estimated. This can be done by appropriately weighting observations according to the time-space distribution of the road network. Region and functional highway system were the only geographic dimensions used in this analysis. Because it is not necessary to aggregate over these dimensions, there is no need to develop weights for them. Weights could easily have been computed, however, based on highway system mileage by region.

An important geographic factor not controlled in this analysis is the particular location of the traffic count on the given highway class. In principle, to obtain unbiased estimates of vehicle type relative frequencies, locations for observing classification counts should be randomly distributed on the highway system. This is the most important unknown factor in estimating vehicle travel from available classification count data. Another important issue deserving further attention is the fact that although weighting factors can remove sample bias, they also tend to increase the variance of estimators, especially when the sample is extremely maldistributed.

Experimental estimates of disaggregate vehicle travel by 4 census regions and 10 FHWA highway system classes were derived by using data from the HPMS and TWS data bases. The estimates suggest a much higher level of combination truck travel than official FHWA estimates. Because of the way the data were collected, there is reason to believe that the classification count estimates may be biased by a selection of locations on the highway network with above-average levels of truck traffic. This question deserves further attention.

The ability to estimate highway travel by vehicle type is limited by (a) a lack of comparable data for all states, (b) the gross spatial and temporal biases of existing traffic classification count samples, and (c) the unknown bias due to choice of observation location on the network. Some of the problems caused by a and b can be ameliorated and to some extent quantified by further analysis. The problem of locational bias and the final resolution of other data problems can ultimately be solved only by the use of statistically valid sampling techniques.

ACKNOWLEDGMENTS

The authors wish to thank David McElhaney, Frank Jarema, and Paul Svercl for their support and guidance from the inception of this research to its conclusion. The authors also thank Stephanie Floyd for her customary diligent yet expeditious preparation of this manuscript.

This research was sponsored by the FHWA under an Interagency Agreement under Martin Marietta Energy Systems, Inc., with the U.S. Department of Energy.

REFERENCES

1. U.S. Department of Transportation. 1977 Census of Transportation Truck Inventory and Use Survey. Report TC77-T-52. Bureau of the Census, U.S. Department of Commerce, May 1980.
2. Consumption Patterns of Household Vehicles. Report DOE/EIA-0328, Feb. 1983; Report DOE/EIA-0319, April 1982. Energy Information Administration, U.S. Department of Energy, 1983.
3. D.L. Gerlough and M.J. Huber. Statistics with Applications to Highway Traffic Analyses. Eno Foundation for Transportation, Inc., Westport, Conn., 1978.
4. W. Feller. An Introduction to Probability Theory and Its Applications, Volume 1. Wiley, New York, 1968.
5. A.M. Mood, F.A. Graybill, and D.C. Boes. Introduction to the Theory of Statistics. McGraw-Hill, New York, 1974.
6. Highway Performance Monitoring System, Vehicle Classification Case Study. Office of Highway Planning, FHWA, U.S. Department of Transportation, Aug. 1982.
7. L.A. Goodman. Analyzing Qualitative/Categorical Data. Abt Books, Cambridge, Mass., 1978.
8. M.B. Brown. Frequency Tables--P4F. In BMDP Statistical Software (W.J. Dixon, ed.), University of California Press, Berkeley, 1981.