United States
Environmental Protection
Agency

Air and Energy Engineering
Research Laboratory
Research Triangle Park, NC 27711

Research and Development

EPA/600/SR-94/003 March 1994

EPA

# Project Summary

# Evaluation and Reporting of County Gasoline Use Methodologies

Sharon L. Kersteter

**The Emissions and Modeling Branch (EMB) of EPA'a Air and Energy Engineering Research Laboratory (AEERL) has been investigating improvements in allocating state-level gasoline sales to counties in order to improve annual county-level emissions estimates from this source category. This report reviews two EMB studies on improving estimates of county gasoline sales. The approaches given in these studies are compared with the current approach prescribed by EPA.**

**The studies reviewed in this report attempted to develop improved procedures for estimating county-level gasoline sales using data for several states and counties. The first study developed regression equations using county-level data to estimate county gasoline sales, while the second study analyzed proportional allocation methods using state and county-level data to estimate county gasoline sales. Equations were developed using various demographic and vehicle-characteristic variables, and were based on 1986 data.**

**Allocating state-level gasoline sales to the county level using the regression equations was generally closer to actual sales than the values estimated using the existing EPA approach. However, since some coefficients used in the regression equations were not statistically significant and since only 1 year of data were analyzed, these equations may not apply to years other than 1986. Using the proportional allocation approach, several variables were found to perform as well as the current EPA methodology. When comparing the results using the EPA methodology to actual gasoline sales, the EPA methodology consistently underestimated actual gasoline sales.**

***This Project Summary was developed by EPA's Air and Energy Engineering Research Laboratory, Research Triangle Park, NC, to announce key findings of the research project that is fully documented in a separate report of the same title (see Project Report ordering information at back).***

## Introduction

Over the past 2 years, EMB has been investigating improvements in allocating state-level gasoline sales to counties in order to improve annual county-level emissions estimates from this source category. This project reviewed results of two EMB studies on improving estimates of county gasoline sales. In addition, the approaches given in these studies were compared to the current approach prescribed by EPA.

## Existing EPA Methodology

Current EPA guidance for estimating emissions from gasoline distribution activities is based on county-level fuel consumption estimates. The suggested method for estimating fuel consumption at the county level is to collect county-level gasoline tax revenues or supplier data. For example, since tax is collected on each gallon of gasoline sold, actual total gasoline sales within a county can be back-calculated with tax formulas. In general, it is assumed that county-level gasoline sales equal county-level gasoline consumption. If these data are unavailable, data from various national publications can

be used to estimate state gasoline consumption. Countywide estimates can be determined by apportioning these statewide totals by the percent of state service station sales occurring within each county.

Countywide service station gasoline sales data are available from the Bureau of the Census which reports sales data by Standard Industrial Classification Code (SIC) for counties containing more than 300 establishments in the SIC. Other apportioning variables, such as registered vehicles or vehicle miles traveled (VMT), can be used if the inventorying agency feels that their use results in more accurate distributions of state totals to the county level.

The use of fuel tax or supplier data depends on both the availability of the data at the county-level and the manner in which the data are compiled. For example, reported county fuel tax revenues may not represent actual fuel sales, but rather the portion of total state sales revenues assigned or apportioned to that county. In addition, fuel sales taxes may vary from county to county within a state, resulting in biased estimates of fuel sales and consumption.

If sales data are unavailable, the inventorying agency may consider surveying county suppliers; however, this process is time-consuming and costly. Many suppliers may not respond to a survey, causing the agency to develop procedures to "scale-up" the survey results to account for the nonrespondents.

The alternative approach, using state-level data to estimate county fuel sales, also has advantages and disadvantages. These state-level data are easily obtained from national publications and are updated regularly. However, this type of apportionment assumes that the variables affecting fuel sales in each county are the same from county to county and have the same effect in all counties. The two studies reviewed in this project reflect EPA's research into improving the estimating methodologies and assumptions.

## General Description of the Studies

In the studies, arbitrarily identified Studies 1 and 2, regression analyses and allocation methodologies are used to identify the demographic and geographic variables (singly and in combination) which most closely estimate actual county-level gasoline sales for 1986 for several states. The equations are developed at the state level and fit is evaluated by the resulting $R^2$ values.

The data used in both studies were initially collected for the Study 1 analyses and were provided for the Study 2 analyses. These data included demographic variables (e.g., population, number of licensed drivers) and geographic variables (e.g., land area, miles of highways). All 50 states were contacted in Study 1 to identify states collecting county-level highway vehicle gasoline sales data. Some data were available for only ten states; however, only 6 states had sufficiently complete data for 1986. Results of the regressions were compared to these county-level data.

### *County-Level Motor Vehicle Fuel Consumption (Study 1)*

The purpose of this study was (1) to develop an equation to estimate county-level fuel consumption using demographic and geographic variables as correlates, and (2) to compare the results of this equation with the current EPA methodology.

For this study, fuel sales were considered an appropriate surrogate for fuel consumption. The base year for the study was 1986. Only six states were included in the study, due primarily to the availability of state-collected county gasoline sales data: Arizona, Florida, Hawaii, Nevada, New York, and Washington. Candidate county-level variables for estimating county-level gasoline consumption included: taxable gasoline gallonage or gasoline sales in gallons (GASOLINE); total population (POP); total population, aged 18 to 64, inclusive (AGE); number of persons per square mile of land (DENSITY); number of persons aged 18 to 64 (inclusive) per square mile of land (RATIO); total number of licensed drivers (DRIVER); land area in square miles (AREA); total number of miles of paved roads (MILEAGE); miles of roads classified as interstate highways (INTERSTATE); miles of roads classified as principal arterials (PRINART); miles of roads classified as minor arterials (MINART); miles of roads classified as collectors (COLLECTOR); total number of registered vehicles (REGIST); eight weight classes of registered gasoline vehicles (RGVW1 through RGVW8); and average engine size in liters for gasoline vehicles for each of the eight weight classes (SGVW1 through SGVW8). The statistics on numbers of registered vehicles and average engine size in various weight classes were obtained from R.L. Polk and Co.

Linear regression analyses were performed to examine various combinations of the variables and their ability to predict county-level gasoline sales, with the goal of developing a single equation for a state which could be applied to all the counties in the state. Equations were evaluated by counting the number of counties for which predicted gasoline sales deviated from actual gasoline sales by more than 20%. Because this is an atypical approach to developing regression equations, the resulting equations were not tested for multicollinearity, heteroscedasticity, and autocorrelation.

### Study 1 Results

The analyses were performed for all counties in Arizona, Hawaii, Nevada, and Washington, 50 of the 67 counties in Florida, and 53 of the 67 counties in New York. In addition, a regression analysis was performed on the combined state data in an effort to identify national trends. Approximately one third of the 182 counties included in this analysis exceeded the 20% deviation between actual and predicted fuel sales.

The variable SGVW2 (representing average engine size in liters for gasoline vehicles weighing between 6,001 and 10,000 lb) appears in the equations of most of the analyses, followed by POP. Population factors were present in all equations, represented either by POP or AGE, but no equations used both POP and AGE. (Since POP (total population) and AGE (population between ages 18 and 64) are collinear, it is not advisable to develop an equation that includes both variables.) The R.L. Polk data were included in all equations, either as total gasoline-powered vehicle registrations by a vehicle weight class or as average engine size by a vehicle weight class. Highway mileage categories were not strongly represented in the equations.

A case study using the Florida data included sales data (represented as the variable SALES) as an independent variable. The resulting equation included the following variables: POP, PRINART, COLLECTOR, MILEAGE, SGVW2, SGVW8, and SALES. No county had a variance in excess of 20 %. This equation is judged to be superior to the earlier equation which did not use sales data for Florida, which had errors as large as 31%.

Finally, the EPA methodology using SIC 554 data and the best fit regressions were compared to actual consumption for three states: Florida, New York, and Washington. In Florida, the regressions compare favorably with the EPA methodology, with the regressions yielding a significant reduction in outliers (i.e., counties with deviations greater than 20% from actual con-

sumption). In New York, 19 counties had estimates which deviated by more than 20% using the EPA method; the regression equation had only five such outliers. Neither the EPA nor the regression model predicted Washington county-level fuel sales well. However, for Washington, as for Florida and New York, the regression analysis was more accurate at predicting county-level fuel consumption.

## Study 1 Conclusions

Overall, the state-level gasoline sales regressions analyses demonstrated that, for a given state, equations may be developed that predict gasoline sales better than the current EPA allocation methodology, with correlates varying by state. However, this statement can be made with confidence only for the year 1986. A comparison of the state studies and the combined national study show that the factors in the national equation included correlates that were seldom used in the state-level studies. Using this comparison, Study 1 suggests that the correlates in this study are insufficient to develop a single national equation for estimating fuel sales. An additional analysis of the combined data showed the marginal effect of adding variables to the equation. In this analysis, a regression equation was developed using only two variables (SGVW1 and AGE), with an $R^2$ of 0.940. Adding four additional variables (SGVW6, RGVW1, DENSITY, and SGVW2) increased the $R^2$ to 0.964. Study 1 suggests that this slight increase in the fit of the equation ($R^2$) resulting from the addition of the four variables emphasizes the dominance of the first two variables in the equation.

## *Predicting County-Level Gasoline Sales (Study 2)*

The objective of this study was to identify a generally applicable allocation equation or set of equations that could be reliably applied to estimate county-level gasoline sales, given state gasoline sales and relevant county-specific information such as population, number of registered drivers, total highway mileage, and SIC 554 sales data. As in the Study 1, gasoline sales are considered to be a surrogate for gasoline consumption. The equation developed should be applicable across states; i.e., the equation should not be state-specific. A limiting factor in this study was the availability of data for identifying and validating prediction methods. This study focused on relatively simple allocation methods, since such simple methods are more likely to satisfy the criterion of general applicability.

## Study Design and Data

Twelve potential variables were identified from Study 1 for use in allocating state gasoline sales to counties: SIC 554 revenue data (dollars) (SIC554 Sales); county population estimates for 1986 (Population); county land area in square miles (Area); miles of roads classified as principal arteries (Artery); miles of roads classified as collectors (Collector); miles of roads classified as collectors, principal arteries, or minor arteries (Mileage); number of licensed drivers (Drivers); total number of gasoline vehicles in all size classes (Gas Fleet); combined engine size (liters) of all registered gasoline vehicles (Total Engine Size); combined total engine size of all registered gasoline vehicles divided by total number of gasoline vehicles in all size classes (Average Engine Size); number of vehicles registered as passenger cars, trucks, or buses (Total Registrations); and number of registered passenger vehicles (Total Passenger Registrations).

Four states were included in this study: Florida, Hawaii, Nevada, and Washington. These states were chosen based on the availability and completeness of the variables identified above. While data on all 12 variables were available for Florida, only five variables (SIC554 Sales, Population, Area, Mileage, and Total Population) were available for the remaining states (Hawaii, Nevada, and Washington).

The simple allocation methods evaluated in Study 2 are proportional allocation methods similar to the current EPA methodology, which is a proportional allocation method based on SIC 554 Sales. The proportional allocation method takes the form:

$$Y_{county} = \frac{X_{county}}{X_{state}} \times Y_{state}$$

where $Y_{county}$ is the predicted gasoline for the county, $X_{county}$ is the value of the variable X for the county, $X_{state}$ is the value of variable X for the state, and $Y_{state}$ is the state gasoline total.

Study 2 evaluated the potential allocation methods in terms of their relative errors of prediction (REs), defined as:

$$RE = 100\left(\frac{\text{predicted gasoline} - \text{actual gasoline}}{\text{actual gasoline}}\right)$$

For a given allocation method, the distribution of REs across all counties indicates the method's performance. To compare allocation methods, Study 2 used differences between the absolute values of the relative errors. A statistical test of the average of the differences, $\bar{D}$, was obtained by calculating $z_D$ as:

$$z_D = \frac{\bar{D}}{S_D / \sqrt{n}}$$

If $z_D$ was near zero, it was concluded that there was no significant difference between the two prediction methods. Specific critical values were obtained from tables of the standard normal distribution.

The study noted that the data are incomplete with respect to variables (i.e, not all variables are available for all states) and observations (information may be available for most, but not all, counties). Study 2 states that, while missing data for select counties probably have a slight effect on the identification of feasible state-level allocation rules, these counties are commonly smaller, less populated, and have low gasoline sales. Low gasoline sales are inherently more difficult to estimate when the RE is the criterion used to judge performance. Missing data for these counties may have a more significant effect on the estimated performance of the allocation equations.

## Simple Allocation Method Results

Simple allocation methods were investigated for Florida alone and for the combined data for Hawaii, Nevada, and Washington. The analysis of the Florida data included 48 of the 67 counties, since complete county sales data were available for only 48 counties. The 12 potential variables were plotted against actual county gasoline sales, represented as the variable GASOLINE, from the state files and displayed as scatterplots. Seven potentially useful variables were identified from visual inspection of the scatterplots SIC554 Sales, Population, Drivers, Gas Fleet, Total Engine Size, Total Registrations, and Total Passenger Registrations The method derived from the SIC554 Sales is the basis of the current EPA methodology and was employed as the benchmark in the analysis; i.e., the remaining six allocation methods were compared to the SIC554 Sales method by comparing their relative errors of prediction. REs were calculated using predicted sales based on the variable and the actual sales (GASOLINE). In general, Study 2 concludes that methods based on Population, Drivers, or Total Registrations are the most reasonable alternatives to the SIC554 Sales method.

Hawaii, Nevada, and Washington had relatively complete information for five predictors: SIC554 Sales, Population, Area, Mileage, and Total Registrations. The five potential variables were plotted against county gasoline sales from state files and displayed as scatterplots. Only three po-

tentially useful variables were identified from visual inspection of the scatterplots: SIC554 Sales, Population, and Total Registrations. All counties in Hawaii, 14 of 17 counties in Nevada, and 34 of 39 counties in Washington had complete data and were included in the analyses. According to Study 2, Population or Total Registrations methods are potential alternatives to the EPA (SIC554 Sales) method.

Study 2 concludes, from analyses of simple allocation methods, that several predictors in addition to SIC554 Sales can be used. Statistical analyses suggest that Population, Drivers, Gas Fleet, Total Engine Size, and Total Registrations are not much different than SIC554 Sales for allocating state-level gasoline sales to counties. Predictors such as Population are readily available and can be used in place of SIC554 Sales (i.e., the EPA methodology) with little or no loss of accuracy. All of the predictors analyzed, however, fail to yield allocation equations with uniformly small relative errors. Larger magnitude errors are always associated with small counties.

The Study 2 analysis of the Florida data shows that REs from the SIC554 Sales and Population allocation methods were generally less than 50%. However, since small counties were excluded from the Florida data, these results may be misleading. For the combined data including Hawaii, Nevada, and Washington, small counties were better represented and REs as large as 100% were not uncommon. Study 2 indicated that this result is probably more representative of the performance of the simple allocation methods in general.

## Other Prediction Methods

Study 2 also investigated whether there are prediction equations depending on two or more predictors that significantly outperform the best simple allocation rules and are generally applicable. Three forms of two-variable allocation equations were investigated: (1) weighted averages of two simple allocation equations; (2) general linear combinations of two simple allocation equations; and (3) linear combinations of two simple allocation equations including an intercept. The allocation equations investigated have parameters that must be estimated from the data. For each model, parameter estimates were obtained by minimizing the sum of the squared relative errors of prediction. This method of estimation ensures that no other parameter values can result in better overall performance in terms of relative prediction errors. The equations were limited to two-variable equations since the more parameters that are estimated from a given state's data, the greater the likelihood that the resulting equation that works well for that state will not work well for other states.

The Florida data were analyzed first. The intent of this analysis was to identify useful variables and equations for the Florida data and to establish a 'best' equation (or set of equations) as a benchmark to compare with simpler allocation equations with the combined data. The analyses show that the equations depending on SIC554 Sales and Population have estimated parameters that are very similar. This suggests an equation that is obtained by averaging the simple allocation equations based on SIC554 Sales and Population. Since Population is highly correlated with Drivers and Total Registrations, either predictor could be substituted for Population in the equation with similar results. Statistical analyses of the relative errors of the equations indicated that the three-parameter equation yielded slightly smaller absolute relative errors than the two-parameter equation, and that the two-parameter equation slightly outperformed the one-parameter equation.

The best allocation equations identified using the Florida data were then applied to the combined data for Hawaii, Nevada, and Washington. However, since fewer variables were available for the combined data, only certain Florida equations were used. The results of the comparisons indicate that only the simple allocation equations based on SIC554 Sales, Population, and Total Registrations, and two averaged allocation equations (SIC554 Sales and Population, and SIC554 Sales and Total Registration) perform well for both sets of data. The more complicated allocation equations determined by fitting equations to the Florida data result in better estimates for the Florida data, but result in worse estimates for the combined data. Comparisons of the results of the one-, two-, and three-parameter equations show that, when applied to the combined data, the three-parameter equation results in the least accurate estimates.

## Per Capita Modeling of All Data

An additional analysis of the data was performed in which variables were normalized for state-to-state comparison by creating per capita versions of the variables and Gasoline (i.e., all variables were divided by Population). The Florida data set and combined data set were merged, and the five variables common to these data sets were investigated. In this analysis, the equation that best predicts per capita gasoline for all the data was sought. County Gasoline was then obtained by multiplying the per capita prediction by county population. The purpose of this exercise was not to derive an allocation equation, but to confirm that the equations identified by this analysis were similar to those obtained in the previous analyses.

Study 2 applied standard methods of linear-model estimation and variable reduction to the per capita data. Relative prediction errors were used to judge the equations' applicability. The results of this exercise indicate that equations that fit data from all states well do not need to include more than two variables (SIC 554 Sales and Population or Total Registrations), and the results are in general agreement with the conclusions of the other equation analyses.

## *Study 2 Conclusions*

Study 2 states that the analyses described in the previous sections suggest two major conclusions. First, if SIC554 data are not available for a particular county, any one of the simple allocation equations based on Population, Total Registrations, or Drivers can be used. The resulting estimates are comparable to the SIC554 Sales allocation method. In addition, there is no evidence in the data analyzed that the allocations can be improved significantly by using more complex estimation schemes. This contrasts with Study 1 which adopted very complex, input variable-intensive equations.

Second, if SIC554 Sales data are available, one of the averaged allocation equations (SIC554 Sales and Population or SIC554 Sales and Total Registrations) should be used. There is evidence that these equations yield better estimates across states than any simple allocation equation. There is no evidence that any other allocation equation will work as well for all states.

# Conclusions

The two studies reviewed for and included in this report attempted to develop improved procedures for allocating state-level gasoline sales to the county level using data for several states. Study 1 developed regression equations using county-level data to estimate county gasoline sales, while Study 2 analyzed proportional allocation methods using state and county-level data to estimate gasoline sales. Equations were developed using various demographic and vehicle-characteristic variables. These equations were based on the 1986 data.

## Data and Study Design

The variables used in these studies were initially identified and collected during Study 1 and provided for Study 2. No additional data were collected during Study 2. Although Study 1 did not explain how the variables were chosen, subsequent contact with the Study 1 researchers indicated that the variables were chosen based on the data Study 1 identified as being inexpensive, easily obtained, and regularly updated, and only includes demographic (e.g., population) and vehicle-characteristic (e.g., number of registered gasoline-powered vehicles by weight class) data.

Both studies were based on available county-level gasoline sales data for several states. Equations were developed using various demographic and vehicle-characteristic variables that most closely predicted the available county gasoline sales data. Study 1 does not discuss the reliability of these state-supplied county-level data. The manner in which these data are collected and reported may differ between states; in fact, some states do not report actual county gasoline sales, but rather the tax revenues received or assigned to each county. Gasoline taxes that differ between and within counties may not be accurately accounted for. In addition, county tax revenues may not reflect actual gasoline sales in that county, but rather the amount of revenues received from the state by that county based on highway mileage or some other characteristic. Revenue data thus recorded and used may bias the equations and not reflect actual conditions and activity.

Finally, the biases that may be introduced into the equations by excluding from the analyses those counties with missing data have not been adequately addressed in either study. It is likely that these counties are small, rural counties and would generally not be of concern in State Implementation Plan (SIP) inventories. However, some of these counties may be part of a nonattainment area, and neither study provides guidance or suggestions for handling this situation.

## Overall Conclusions

Table 1 presents comparisons of the Study 1 regression to the EPA method and the Study 2 allocations to the EPA method. Due to the nature of Studies 1 and 2, results can be reasonably compared for only one state—Florida. Since the Study 1 term "deviation" and the Study 2 term "RE" are equal to

$$(100) \times \left( \frac{\text{predicted} - \text{actual}}{\text{actual}} \right) \text{percent}$$

the results can be compared directly. Study 2's use of the EPA methodology for the Florida data results in 25% of the counties deviating from actual gasoline sales by more than 20%. Study 1 shows that 19% of the counties deviated from the actual value by more than 20%. This difference may be due to the number of Florida counties included in the EPA methodology (Study 1–53 and Study 2–48).

The results given in Table 1 appear to suggest that the Study 1 regression equation provides the best estimates of actual gasoline sales. However, this conclusion is misleading since, in developing the regression equations, Study 1 kept some statistically non-significant coefficients. It is not known exactly what effect this has on the results. In addition, since only one year of data was used in the analyses, these resulting equations may not work well for years other than 1986.

Table 2 presents more detailed information on the deviations from actual seen in Study 1. This analysis suggests that the EPA method consistently underestimates actual county-level gasoline sales. Sixty percent of the counties analyzed by Study 1 for the EPA methodology resulted in underestimates of actual sales. This may be an artifact of the retail outlet sales data which may not be complete or may include sales of items other than gasoline.

Based on the problems outlined above, it is difficult to draw conclusions on reasonable alternate methods for estimating county-level gasoline sales. For lack of a proven alternative, the simple approach of the existing EPA allocation methodology may be best, although it may underestimate gasoline sales. However, if the inventorying agency plans to use the estimates for modeling, more detailed data will be needed; i.e., the existing EPA methodology may not be acceptable and will not provide the necessary level of detail.

***Table 1.*** *Comparisons of Studies' Results to the EPA Methodology*

| | ***Study 1*** | | ***Study 2*** | |
|---|---|---|---|---|
| | *Percent of counties analyzed with deviations from actual > 20%* | | *Percent of counties analyzed with RE > 20%* | |
| *State* | *EPA Methodology* | *Study 1 regression* | *EPA Methodology* | *Study 2 allocation (average of proxies)* |
| *Florida* | *19 (w/o sales)*<br>*19 (w/sales)* | *4 (w/o sales)*<br>*0 (w/sales)* | *25* | *29* |
| *New York* | *45* | *9* | *na* | *na* |
| *Washington* | *63* | *18* | *na* | *na* |
| *Combined Hawaii, Nevada, and Washington* | *na* | *na* | *50* | *58* |

**Table 2.** *Comparisons of Deviations in Study 1*

| | EPA Methodology | | | Study 1 Regressions | | |
|---|---|---|---|---|---|---|
| | | Percent of counties analyzed with deviations | | | Percent of counties analyzed with deviations | |
| State | No. of counties analyzed | Above actual | Below actual | No. of counties analyzed | Above actual | Below actual |
| Florida | 53 | 43 | 57 | (w/o sales) 50 | 56 | 44 |
| | | | | (w/sales) 48 | 46 | 54 |
| New York | 42 | 33 | 67 | 53 | 51 | 49 |
| Washington | 16 | 44 | 56 | 39 | 44 | 56 |
| Overall average | 111 | 40 | 60 | 190 | 49 | 51 |



☆U.S. GOVERNMENT PRINTING OFFICE: 1994 - 550-067/80228

*S. Kersteter is with Southern Research Institute, P.O. Box 13825, Research Triangle Park, NC 27709-3825.*
***Charles C. Masser** is the EPA Project Officer (see below).*
*The complete report, entitled "Evaluation and Reporting of County Gasoline Use Methodologies," (Order No. PB94-145455/AS; Cost: $27.00, subject to change) will be available only from:*
*National Technical Information Service*
*5285 Port Royal Road*
*Springfield, VA 22161*
*Telephone: 703-487-4650*
*The EPA Project Officer can be contacted at:*
*Air and Energy Engineering Research Laboratory*
*U.S. Environmental Protection Agency*
*Research Triangle Park, NC 27711*

United States
Environmental Protection
Agency
Center for Environmental ResearchInformation
Cincinnati, OH 45268

Official Business
Penalty for Private Use $300

EPA/600/SR-94/003

BULK RATE
POSTAGE & FEES PAID
EPA
PERMIT NO. G-35