MODULAR PROGRAMMING PROCESSES NEW TRAFFIC DATA INTO SAS* DATABASE

Barbara Hertzog and Gary Cowan
Washington State Department of Transportation

ABSTRACT

New high tech methods of capturing traffic information generate very large quantities of raw data which must be checked for accuracy and put into standard format before being made available for analysis and research.

This paper describes the modular methodology developed by the Washington State Department of Transportation (WSDOT) to shape a powerful SAS traffic database updated monthly with data gathered electronically at various locations statewide. Program modules allow for judgement and intervention at appropriate stages of data reduction, diagnostic procedures, and for SAS to detect and remove equipment malfunction errors.

The process illustrated uses data from permanent recorder stations that operate continuously, 365 days a year, counting all passing traffic, by direction and lane, and classify that traffic into one of four groups by length of vehicle. Totals in each group are taken and recorded by preset intervals. Data from the recorders are transferred by telemetry and processed by mainframe SAS into two databases for routine reports and departmental research using SASGRAPH and other SAS processes. Data can be made available to other users such as economic groups, accident analyses, universities and federal and state agencies.

INTRODUCTION

Working with large volumes of electronically captured data gives rise to special processing problems as revealed during the construction and development of an exciting new tool for traffic research in the State of Washington. Almost at once answers to long-standing questions needed for the technical advancement of an industry began to emerge. The database has been built from a wealth of hereto unavailable data and has been an eagerly awaited event.

BACKGROUND

Traffic information is avidly sought by many users, not only for the design of highways, but because it reflects economic growth, changes in population, trends in urban vs rural growth, impacts on large shipping industries, accident insurance rates and many other events that generate major economic investments by Federal or State agencies, taxpayers, large industries, and construction investors.

"Information" is an end product of a process. The quality of the product depends on the gathering of accurate, representative, sample data to analyze, synthesize and eventually apply to the solution of practical problems.

Traffic, itself, is a very dynamic, interactive collection of field data. With field data we can not control the multitude of variables as is recommended for research in a laboratory setting. However, against a background of what may look like chaos we can still follow many of the rules of good research design to gather data for highway planning.

Even the casual observer can detect that there are different characteristics of traffic (such as urban, commuter, industrial, rural, residential) and that these can exist at different locations even on the same highway.

At any single location there are constantly shifting but predictable patterns in quantity and quality of the vehicle mix that vary by hour of the day, day of the week, month of the year and year to year. These regular patterns may be altered by transient events, such as weather, holidays, festivals, harvests and unexpected happenings such as accidents or acts of God.
In the past Washington State Department of Transportation (WSDOT) collected data for highway planning and design purposes, but these were fragmented and could seldom be used except for the explicit project for which they were collected. Longitudinal or comparative data was not available except where volumes (only) could be counted by automatic recorders at a single location. The types of vehicles which made up those volumes was a question that was answered expensively by sending personnel out in the field to manually classify or "count cars".

**CURRENT EFFORT**

With the application of high technology to the traffic industry we now have electronic equipment which can classify vehicles. The equipment we will be discussing determines the length of the vehicle passing over the equipment. This is done by field computers which apply a formula based on the time it takes the vehicle to pass a single point on the roadway and its speed between two points. This value will register to the nearest inch but is only accurate to one foot. For Department purposes this value is recorded in one of four classes. The recorders count by preset interval (usually by hour), round-the-clock, 365 days a year.

We are now able to measure with precision the dynamics of daily, weekly and seasonal patterns of different kinds of vehicles at a these locations. We can begin to provide critical information about the traffic stream which has been needed on a national basis by a broad network of users who have had only fragments of information in the past to use for their serious decision-making.

Through SAS processing and the use of SAS graphics capabilities we can illustrate the long-awaited seasonal (annual) patterns for large trucks that carry interstate and local heavy trucking loads. We can determine factors to translate short samples of traffic at other locations to estimates of annual averages. These advancements improve traffic forecasting for design of the roadways that must carry these loads and for which taxpayers must pay. We can also supply good, quality information and graphics illustrations to researchers at the University of Washington Transportation Research Center (TRAC) to enable them to advance the state of the art for the entire industry.

**THE STATE OF THE DATA**

We now have 19 months of continuous traffic volumes at 18 locations on Washington State highways, the first of some 75 locations to be upgraded with this kind of equipment. The 18 locations have collection systems that use electromagnetic induction loops in the pavement and a small computer with modem connected to a telephone line. The stations are powered by batteries that generate a very small current of 0.012 amps (12 milliamps), or 0.06 watts, to activate sensors in the pavement and operate the computer. The small 6 volt, 10 amp-hour batteries are recharged from nearby electrical lights or by solar panels. Equipment problems can arise from failure of the batteries or sensors, and from interference from radio transmitters (CBs) or lightning.

Each location can be dialed up by the Traffic Data Office in Olympia and data transferred by telemetry and can be sampled periodically to detect any problems with the equipment in the field. From the retrieved data a single, concatenated data set which includes all locations is transferred monthly by mainframe computer to the Operations Planning Office and is used to build the SAS databases. Each month's set has 4 different data configurations when received.

Each time a station is contacted it transmits a data set containing a 2 line header followed by a series of datalines. Depending upon its configuration the dataline is made up of fields containing from one to three hours of data, multiple lanes and/or alternate directions of traffic travel. Contained in each field are volumes for that time interval for 4 vehicle classes: Single Unit Vehicles, 3 and 4-Axle Combinations, 5-Axle Combinations, and Over 5 Axles.
The datalines must be first given identifiers obtained from the headers. It is then sorted into the 4 different configurations groups, each of which is reduced individually to a standard format and eventually merged into a single database. Along the way judgment must be used as to the condition of the data and assessment made for equipment malfunction which has not been eliminated, for duplication of data from multiple retrievals, or other errors that can occur in the system.

Judgment must include knowledge of the behavior of the traffic at the various locations, weather or other local conditions which can cause legitimate anomalies. Since the State of Washington has a wide variety of geography and climate, this background knowledge is only gained by personnel who have served in the field doing statewide traffic studies and must be integrated into any statistical research. Thus the person designing the programs and those who handle the data are field experienced statewide as well as SAS knowledgeable.

During the first year or so, personnel handling the field installations, those doing telemetry retrieval and the Operational Planning Analysts had to gather experience and each develop their own methods. During that time the quality of the data was often problematic and much extra energy had to go into the effort. With time and experience this is improving and the process is smoothing out.

INCOMING RAW DATA WITH MIXED FORMATS AND HEADER LINES

STRATEGY:

1. INFILE each line as a single Character variable 'ALINE'. Use the $CHAR to retain leading blanks.
2. Use the Truncator Operator to identify Headers vs Datalines. (i.e. If ALINE='*' then do; ... etc.)
3. Use the SUBSTRING Function to break out variables, calling out the Numeric variables in the LENGTHS to change from Characters.

DATA TEMP;
INFILE SEGIN;
INPUT A1 ALINE @ CHAR75.;
DATABASE DEVELOPMENT

When high tech equipment for traffic data collection was authorized and purchased it had to be tested in the field so equipment could be calibrated and a standard selected. In September of 1987 the first actual data began coming through and it became apparent that the mass of data would provide a rich source for new knowledge.

It became apparent also that processing the data could not be entirely machine-dependent. New malfunctions take place and have to be handled individually according to the problem. One characteristic of the equipment is to give zero reading if certain conditions take place. However, a zero reading may by quite legitimate if no traffic passes over the recorder during a time interval, if the road is closed due to severe weather, etc. Fairly intimate knowledge is required to discern problems, and "best guess" is sometimes needed.

The evolution of the 2 SAS databases went through the following phases:

PILOT STUDY

WRITE PROGRAMS & BUILD DATABASES (2)

DEVELOP MONTHLY UPDATE PROCESS

1. Pilot Study. This required designing a number of programs that were modified as new knowledge about the mechanical conditions that affected the data became known. Because quantity of data is massive, diagnostic procedures were designed to provide for checking and debugging.

2. Development of 2 Main Databases using all data obtained to that date (8 months). It was decided to retain 4 states of file development to allow rework, if necessary, and to retain the original data with all errors uncorrected since it was not known what other hidden errors might be in the electronically captured data. Keeping this mass of data required setting up a large mainframe data set, and all processing to date has taken place on the mainframe, although it is hoped some can eventually be done by our IBM PS/2 Model 80 with 115 meq hard disk to store the finished data files.

3. An update method was developed so that each month's traffic could be added. At the time of update, provision is made to reprocess any past data which may have been determined to be flawed.

4. Problems of duplication or random sampling errors were introduced, had to be tracked down and backed out of the databases. We now generally receive only the main retrieves, however necessary mechanisms alert us to this kind of problem before errors are introduced.

ADVANTAGES OF MODULAR PROCESSING

1. Analytic diagnostic programs can be run on the data at various stages of development of the database.

2. Data problems detected in a diagnostic can be corrected in a module by writing intervention (editing) commands. The are documented on the program printouts.

3. Rerunning to correct a single module is easier and less expensive than going through an entire process before finding something has gone wrong and then needing to repeat the entire process.

4. Debugging is easier in smaller program increments.

5. If some data already in the database has to be reworked, it can be merged into the monthly update cycle at the appropriate place and done at the same time since programs are user determined and not machine dependent.

6. We copy each month's programs and adapt them to the new month. This gives us clear documentation for previous work, in case later problems show up.

7. With mainframe processing an entire month's update can be handled in a single day if the mainframe is relatively free. While one module is waiting and/or running the analyst/programmer is making the necessary modifications to the next update module.

901
The several program modules of the data reduction system do the following:

1. Get Identifiers from fixed fields in the header using SAS Substring function, retaining them on each dataline. Compute Hour, Day, Month, Year for each dataline using the beginning Time/Date and time interval from each header, advancing the day, month, etc. Delete header lines and subset into the 4 configurations. Diagnostics print out duplicate datalines.

2. Run 4 individual programs to break datalines into hours, directions, etc. per configuration using SAS Substring function. Change the Substring resulting character variables to numeric by Length callout in the Data Step. Compute Increasing & Decreasing Direction of travel using the State Route number. Record any duplicate hours where datalines may overlap.

3. From diagnostic find zero hours that are malfunctions and write code to remove. Proc Means sums all lanes and intervals to single hour for each direction using N to determine incomplete data (not all lanes, intervals) to remove. Concat 4 configurations to single standardized data set, adding descriptive identifiers and Macro. Result: hourly by direction of travel.

4. Proc Means sums all hours in each day, giving the N of hours. If N does not equal 48 (24hrs in 2 directions) the day is discarded.

5 & 6. Concat #3 and #4 onto final databases, sort and print lists.

DATABASES BUILT

Two final databases were built to allow us to provide (nearly) raw data and in a summarized form. Each record contains 13 variables: 8 identifiers, volumes by 4 vehicle types and a total volume.

HOURLY FILE: CLS2DIR. This file holds hourly data, by directions of travel, at the 18 locations. From this file we can determine daily 24 hour patterns of the 4 vehicle types which vary by direction due to the flow of commuters (inbound in the AM and outbound in the PM), the time of day of business trucking, or the 24 hour round-the-clock flow of long-distance interstate trucking.

24-HOUR SUMMARY FILE: CLS24HRS. This file holds daily 24-hour totals for each of the 4 vehicle types with both directions combined together. All days have 24 hours of data in both directions. This summary information is sought by many users and shows seasonal trends, holiday variations, and annual volumes which can be used to estimate the weights of vehicles using the roads as well as economic growth. (See Figure 1 for annual patterns for large trucks.)
The first year of data showed us some variations that alerted us to possible repeating patterns which we have not detected in the past. Annually repeating are situations such as harvest hauling of crops or events such as hunting season, holidays, etc. Non-typical data can be gathered because of such things as construction on the highway or severe weather. Other things can skew an entire season or year if the event has ongoing consequences such as forest fires or the eruption of Mt. St. Helens.

From data we can build models to determine the optimum time to take traffic samples at other locations to get the most representative mix of traffic for the year or during the day. The locations can be used to find patterns that can be applied to other parts of the state highway system to upgrade our entire traffic analysis and growth projections.

STANDARD DISCLAIMER

The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Washington State Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

REFERENCES


SAS and SAS/GRAPH® are registered trademarks of SAS Institute, Inc., Cary, NC, USA.