Title
Using Section 15 Data: Adapting and Evaluating the Magnetic Tape Version for Statistical Analysis

Authors
Fielding, Gordon J.
Brenner, Mary E.
de la Rocha, Olivia

Publication Date
1983-12-01

UCI-ITS-WP-83-6

Using Section 15 Data: Adapting and Evaluating the Magnetic Tape Version for Statistical Analysis

UCI-ITS-WP-83-6

Gordon J. Fielding
Mary E. Brenner
Olivia de la Rocha

School of Social Sciences and
Institute of Transportation Studies
University of California, Irvine

December 1983

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

Abstract

Section 15 data has already proven itself a useful tool in transit decision making. Yet, its wider use has been inhibited by the difficulty of accessing it electronically. This paper describes a set of strategies for extracting, reorganizing and evaluating data originating in the electronic data files disseminated by Transportation Systems Center on magnetic tape.

The current organization of information within the files is unsuitable for most statistical software packages. Therefore, it is necessary to extract information from the Section 15 files and rearrange it in a form suitable for analysis. Different classes of missing data are also defined and remedies for the problem are addressed. Additionally, the cross-validation of values and the computation of basic transit variables are considered.

Many statistical models make assumptions about the distributional characteristics of variables. Differences of scale among transit systems on such measures as size of fleet often result in variables whose distributions violate these assumptions. Transformations which remedy the problem are recommended.

INTRODUCTION

Since its first release for fiscal year 1979, the Section 15 reporting system has proven itself a powerful tool in transit decision making. It has provided standardized definitions of transit activities and recording procedures (1); replaced burdensome and non-uniform data collection efforts by local operators (2); allowed local, regional and nationwide comparison of transit performance (3); and facilitated management, performance evaluation and the allocation of financial assistance at all jurisdictional levels (4, 5, 6, 7). In short, analysis of Section 15 data offers greater leverage for understanding transit performance than has, hithertofore, been possible.

The most complete version of Section 15 available for analysis is distributed by the Transportation Systems Center (TSC) in the form of 62 electronic data files stored on magnetic tape. Although this version promises to be the most useful in the long run, current use of the tape is inhibited by the difficulty associated with reading it and adapting the information to a form suitable for statistical analysis. Considerable time and effort must be allocated to the development of a system for accomplishing the adaptation.

As TSC adopts a new operating system and develops new software for Section 15 data, a wider variety of data tape formats may become available. However, the first four years of Section 15 data (FY 1979-1982) share the same organization described in this paper.

An alternative to the magnetic tape is the National Urban Mass Transportation Statistics Annual Report (8), which provides tabular summaries of Section 15 data. But there are two drawbacks to substituting the Annual Report, as we shall call it, for the tape version. First, the tape is a comprehensive set of data including far more information than the Annual Report. All levels of reporting are included in the tape, while only the Required Level of information is reported in the Annual Report. Entire classes of data such as operating schedules and peak loads are available only on the tape. This additional information permits the cross validation of values, a critical step in assessing the accuracy of these data. Second, for users who wish to analyze transit systems on a nationwide level or use many variables, the cost of making the Annual Report machine readable could rival or exceed that involved in adapting the tape. For example, the data to be used require keypunching. Then, a number of preliminary computational steps, such as converting percentages back to "raw" values, must be carried out before actual analysis commences. Therefore, it would be useful if a set of strategies could be outlined which would facilitate the use of Section 15 data as it originates on magnetic tape.

This paper describes such a set of strategies. A conceptual scheme underlying the conversion of the magnetic tape data to a conventional statistical format is first described. This is followed by a discussion of data preparation steps which precede statistical analysis and includes the treatment of missing data. The paper concludes with a brief evaluation of the distributional characteristics of basic transit variables for fiscal year 1980.

DATA REORGANIZATION

In this section the issue of why the tape data must be reorganized to make them acceptable to a statistical software package like BMDP or SPSS is discussed. The objective is to explain why a software package can "read" the data, but cannot, without reorganization, perform a statistical analysis on them. Why reorganization is needed and what steps are required to reorganize are the focus points.

The data reorganization process revolves around four questions concerning data files. First, what are the basic organizational features common to all numerical data files? Second, what are the distinguishing features of a data file organized for statistical analysis? Third, how are the data files on TSC's magnetic tape different from the statistical convention? Fourth, what steps are required to reorganize them?

Basic Organizational Features

All data files are organized in rows and columns. A sample file is shown in Table 1. However, the meaning of the data is not inherent in this simple physical organization, but must be conveyed to the computer by the programmer. The system or scheme used by the programmer to give meaning to the array of numbers is called the logical organization.

TABLE 1

A SAMPLE DATA SET

1001063019801011467500000
1001063019801022138100000
1001063019801031761400000
1002123119801014287100000
1002123119801025891600000
1002123119801033892500000
1002123119802015411600000
1002123119802027382700000
1002123119802039188400000
1004063019801014816500000
1004063019801021810200000
1004063019801031718400000

The specification of the logical organization is laid out in a document called a codebook. In a codebook the meaning of data is defined by the way the numbers are organized into sets of columns. A large number like $4,000,000 takes up 7 columns, for example. The assigned sets of columns are called fields.

Table 2 shows a codebook from TSC's documentation. According to the codebook, columns one through four of the number array have been reserved for Transit System numerical ID. Columns five through twelve are reserved for the fiscal year end date for the system which is identified in columns one through four. Column thirteen is assigned to the mode code. And so it goes. With the help of this scheme the computer can be informed about the meaning of the data by the way fields in the block of numbers are assigned. This process is called formatting.

In formatting space is set aside in the array and named so that any number found in that space by the computer can be presumed to have the assigned meaning. Any number found in columns one through four of the sample block of data (Table 1) will mean "Transit system ID number" to the computer, so long as it is formatted that way.

It is important to realize that there is some flexibility in the way data may be formatted. That is to say, there may be more than one, meaningful logical organization for the same physical file.

Finally, an actual line of data like 10041114676 which can be formated is called a record. A record may take up one or more rows and there may be more than one type of record in a data file.

TABLE 2

A SAMPLE CODEBOOK

COLUMN	NAME	TYPE	DESCRIPTION
1 - 4	TRSID	INTEGER	TRANSIT SYSTEM ID
5 - 12	FY	DATE	FISCAL YEAR
13 - 13	MODE	INTEGER	MODE CODE
14 - 15	EMCOD	INTEGER	EMPLOYEE CLASS CODE
16 - 21	OLABR	REAL	OPERATING LABOR
22 - 27	CLABR	REAL	CAPITAL LABOR

Statistical Files.

Statistical procedures operate by making systematic comparisons among objects. The objects are compared on those attributes which have been measured in some way. For example, in Section 15 analyses transit systems are compared on such attributes as size of fleet and speed.

In statistical data files the most important organizational units are cases (objects) and variables (attributes). A case may be thought of as the full collection of information items defined in the codebook for a single transit agency. If some defined item is missing, the statistical case is incomplete, and a place-holding code must be inserted to fill it out.

A variable, like a case, is a statistical concept. When all cases have been measured on a given attribute, the resulting collection of values is organized in a list called a variable. Statistical procedures compare these lists and depend on the fact that the cases always appear in the same order. Once again if no place-holder resides in the position of a missing item, the order is disturbed and statistical results are rendered meaningless.

One danger to be avoided in comparing all numerical data files to a smaller subset of them, i.e., statistical files, is that the distinctions between their separate terminology will blur. It is important to keep in mind the differences between the (horizontal) concepts "row," "record," and "case" on the one hand and the (vertical) concepts "column," "field," and "variable" on the other. In general usage, the members of these trios are often used interchnageably. Since understanding the data reorganization process may hinge on the distinctions among them, a glossary is provided following the text.

The Organization of the TSC Tape Files

The organization of the TSC tape files is closely linked to the reporting forms. Form No. 404, Transit System Employee Count Schedule, provides an example. Figure 1 shows Form 404 and the information for one transit system as it appears in a data file on the tape. Spaces have been inserted between the fields for ease of reading. In the actual file there are no spaces.

A comparison of the form and the data shows the first three fields TRANSIT SYSTEM ID, FISCAL YEAR ENDED and MODE coming from the top of the form, and repeating on every record in the data. The next two fields, EMPLOYEE CODE and OPERATING LABOR are taken from the "Employee Classification" and "Operating Labor" sections of the form. The Figure shows a one-to-one correspondence between the numbers assigned to employee categories on the form (11, 12, 13, etc.) and the values under EC in the data. However, the one-to-one correspondence is not quite complete. If Form 404 were used to construct a codebook which acted as the logical organization for the data appearing in the figure, then there would be a discrepancy between what the logical organization predicts and what actually appears in the physical file. There is no record appearing for category 22, Maintenance Support Personnel, in the data.

To reiterate, most statistical software packages require some entry to stand in for the missing category 22. Until a stand-in value is

Form No. 404

TRANSIT SYSTEM EMPLOYEE COUNT SCHEDULE

Transit System ID [ ] Level [R]

Fiscal Year Ended [ ] Month [ ] Day [ ] Year Mode ______ Code [ ]

a b c

LINE NO.	EMPLOYEE CLASSIFICATION	EMPLOYEE EQUIVALENTS OPERATING LABOR	CAPITAL LABOR
	EMP	EMP	SCH
01	11. Transportation Executive, Professional and Supervisory Personnel		
02	12. Transportation Support Personnel		
03	13. Revenue Vehicle Operators		
04	21 Maintenance Executive, Professional and Supervisory Personnel		
05	22. Maintenance Support Personnel		
06	23. Revenue Vehicle Maintenance Mechanics		
07	24. Other Maintenance Mechanics		
08	25. Vehicle Servicing Personnel		
09	31. General Administration Executive, Professional and Supervisory Personnel		
10	32. General Administration Support Personnel		
11	00. TOTAL TRANSIT SYSTEM EMPLOYEES		

ID	FY	M	EC	OLABR
1056	19800630	1	11	4.5000
1056	19800630	1	12	2.5000
1056	19800630	1	13	47.800
1056	19800630	1	21	2.3000
1056	19800630	1	23	5.6000
1056	19800630	1	24	.50000
1056	19800630	1	25	2.6000
1056	19800630	1	31	1.0000
1056	19800630	1	32	2.3000
1056	19800630	1	00	67.100

ID=ID NUMBER
FY=FISCAL YR END DATE
M=MODE
EC=EMPLOYEE CODE
OLABR=OPERATING LABOR
(CAPITAL LABOR VALUES WERE OMITTED)

FIGURE 1. THE CORRESPONDENCE BETWEEN REPORTING SYSTEM FORMS AND THE LOGICAL AND PHYSICAL ORGANIZATION OF TSC DATA FILES

substituted, the information cannot be said to form a complete case. Therefore, all such instances of "missing" records must be remedied before statistical analysis can proceed. Only two widely available software packages, SAS and SPSS-X, are known to have methods for dealing with this problem.

Another important consequence of the correspondence between the data and the forms is the way we know which values are being compared, i.e., which values are making up the variables. Consider once again Figure 1. In a statistical routine, the OLABR value of 4.5000 cannot be compared to the OLABR value of 2.5000 beneath it. The 4.5000 must be compared to another value, not shown here, which also has an EC of 11. OLABR, therefore, is not one variable, but eleven variables (the number of employee classifications) collected together in one field.

Informing the computer of this relationship between the values in the OLABR field requires devising a new logical organization to replace that found in the TSC codebook. For statistical purposes, OLABR is too general a category to qualify as a variable. It would not be useful to compare the number of one system's Reveneue Vehicle Operators to another system's Vehicle Servicing Personnel. Instead, Revenue Vehicle Operators must be compared to Revenue Vehicle Operators. A variable, then, would be all instances of OLABR for category 11 or all instances of OLABR for category 00, Total Transit System Employees.

The TSC data file organization is common, economical, and often used as input to management information systems using customized software. In computer science parlance, it is referred to as hierarchical ordering.

By way of summary, two major differences needing reconciliation between satistical files and TSC files are: (1) the omission of stand-ins for "missing" records; and (2) hierarchical organization. The concept of "missing" data is an important issue in its own right and is discussed more fully in a later section.

Implementing Reorganization

The main goals of reorganization are to supply stand-in values for missing records and to reformat instances where several variables have been grouped together in one field. A hypothetical example of the results of reorganizing is shown in Figure 2.

There are several noteworthy features in Figure 2. First, in File I under the field SYSTEM ID, there is no information present for system number 1003, and systems 1002 and 1004 appear to have only half the information they need.

Also in File I, the field WAGES can be seen to contain six different variables. The values in the fields MODE and EMPLOYEE CATEGORY must be used to find these variables. For example, the first WAGES value, 500, has a MODE of 1 and an EMPLOYEE CATEGORY of 0. These values indicate that the first 500 is for motor bus drivers' wages. Hence, the only other value it can be compared to is WAGES of 650, six lines down in case 1002 which also has MODE of 1 and EMPLOYEE CATEGORY of 0. There are six WAGE variables possible because in addition to the MODE and EMPLOYEE CATEGORY combination of 1 and 0 there are also the combinations of 1 and 1 or 1 and 2, etc. Because there are two values of MODE and three values

DATA FILE I. HIERARCHICAL ORGANIZATION

SYSTEM ID	MODE	EMPLOYEE CATEGORY	WAGES	
1001	1	0	500	
1001	1	1	600	
1001	1	2	600	
1001	2	0	400	
1001	2	1	700	MODE
1001	2	2	700	1 = MOTOR BUS
1002	1	0	650	2 = TROLLEY BUS
1002	1	1	600	
1002	1	2	700	EMPLOYEE CATEGORY
1004	2	0	700	0 = DRIVER
1004	2	1	000	1 = MAINTENANCE
1004	2	2	000	2 = ADMINISTRATION

DATA FILE II. STATISTICAL ORGANIZATION

SYSTEM ID	MTRBUS DRIVER WAGES	MTRBUS MAINT WAGES	MTRBUS ADMIN WAGES	TRBUS DRIVER WAGES	TRBUS MAINT WAGES	TRBUS ADMIN WAGES
1001	500	600	600	400	700	700
1002	650	600	700	999	999	999
1003	999	999	999	999	999	999
1004	999	999	999	700	000	000

999 = MISSING VALUE CODE

FIGURE 2. HYPOTHETICAL DATA FILE BEFORE AND AFTER REORGANIZATION

of EMPLOYEE CATEGORY, it takes two times three, or six, combinations to exhaust all pairs possible.

The six variables each have their own separate fields in File II. The information in MODE and EMPLOYEE CATEGORY from File I has been incorporated into the new logical organization of File II. Therefore, they disappear from File II. File II also has full sets of information (complete cases) for all transit system IDs, although missing value codes of 999 had to be inserted to make this possible. For example, even though system 1002 has no trolley busses, stand-in values of 999 were inserted in the three trolley bus variables for this case.

In summary, the basic reorganization steps can be reduced to four:

1. Using the logical organization in the TSC codebook, in which a case is not a transit system but a single record, read and write the data, eliminating unwanted information.

2. Locate the positions in the retained data needing stand-in values.

3. Insert the stand-in values.

4. Format the data with a new logical organization which considers all the records belonging to a single transit system as a case. Once the stand-in values have been inserted, this number of records will be the same for all transit systems.

Working with the Tape

The objective of this section has been to explain the reasons for data reorganization and the steps which are necessary to accomplish this task. A technical manual has been prepared which explains some of these steps in more detail (7). The complexity of the task lies not in the nature of the problems so much as in the large amounts of data which must be manipulated and the number of steps required to carry out the manipulations. Some statistics concerning the data files make this clear.

In fiscal year 1980 there were 62 data files. Twenty were text files containing labels and forty-two were numerical data files. The files ranged in size from approximately 300 records to 22,000 records, and all 62 files combined required 775,000 words or 3,800,000 characters. For comparison, the Annual Report is comprised of approximately 2,100,000 characters.

The large number of steps required to reorganize a file is quite surprising. The most complicated expense file contained 22,000 records and required the use of over 75 temporary data files during the process of inserting over 2,000 needed stand-in values.

DATA PREPARATION

Once the data have been reorganized, additional data preparation is required before analysis can commence. There are three phases to preparing the data: calculating basic variables, identifying and flagging missing information, and validating existing data.

The Section 15 database contains a wealth of information which is too detailed for many purposes. The data to be used for statistical analysis must be customized to the purpose at hand. Our purpose was a comparative analysis of motorbus performance in terms of general concepts such as labor efficiency and utilization of service. Thus we wanted to aggregate many small pieces of information into more comprehensive variables which contain only information about the motorbus mode and which are applicable to an entire year's operation.

The information about transit employees is a clear example of too much information which must be summarized into broader categories. Ten employee categories are reported--3 in vehicle operations (e.g., supervisors, revenue vehicle operators, support), 5 in maintenance and two in general administration. These ten categories are further subdivided into capital labor and operating labor. For our purposes we wanted the number of vehicle operators, the number of maintenance employees and the number of administrative employees. The first step in creating these variables was to add together operating and capital employees since we were not interested in this distinction. At this point, the number of revenue vehicle operators was ready for use. The number of maintenance employees was calculated by adding together the five categories of maintenance employees. The number of administrators was calculated by adding together the supervisory personnel in vehicle operations and maintenance to the two categories of administrative personnel.

Other variables which must undergo this aggregation process include the total number of accidents, total amounts of subsidies and the miles of roadway used on bus routes.

Estimating Annual Data

The data on service supplied by a transit agency and the service consumed by passengers must undergo a different kind of calculation before they can be used in a general analysis. While the Section 15 reporting system requires that all financial data be reported for a complete fiscal year, information on service variables such as unlinked passenger trips and revenue vehicle hours is collected by a sampling procedure and reported for an 'average weekday,' 'average Saturday' and 'average Sunday.' This information must be combined with a formula which annualizes it so that it is comparable to the financial data. A formula was used which allowed for 253 weekdays, 53 Saturdays, 52 Sundays and 7 holidays (also calculated as Sundays) with each of these numbers multiplied by the given values for average weekdays, Saturdays and Sundays.

A series of calculations were also needed to disaggregate data so that it applied only to the motorbus mode. Revenue and subsidy information are reported in the Section 15 system for entire transit systems, not by mode. In addition multi-modal systems have the option of reporting expenses as joint expenses between modes, and a few systems report most of their expenses in this way. A series of weighting formulas were designed which allow assignment of revenues or joint expenses to specific modes. For example, a proportion of passenger revenue is assigned to the motor bus mode by multiplying the system's total passenger revenues by the ratio of motorbus passengers to total passengers. Although the resulting values are only estimates, they are better than the distortions caused by using overly-large figures or dropping the multi-modal systems (32% of the systems reporting in 1980) from the analysis.

Missing Data

The second phase of preparing data for analysis is detecting which cases have missing data and which therefore must be eliminated from further analysis. A database prepared for statistical analysis will usually have a special symbol such as -9 which indicates that information is missing. However, the Section 15 data tape has no such special symbol, and the analyst must therefore insert the symbol during the process of calculating the variables. The analyst is able to detect missing data problems by considering the logical properties of specific variables, by comparing a variable to other information in the data base and by comparing the Section 15 data to other sources of information (including the analyst's own knowledge of transit systems).

For some variables, detecting missing data is straight forward and quite logical. For instance, a transit system which has zero operating expenses can readily be assumed to have a missing data problem. But most variables require more judgment on the part of the analyst. It is possible for a transit system to have zero accidents for a given fiscal year, but the larger a system, the less likely that it will have no accidents. The analyst must examine other transit systems of similar size to the one reporting zero accidents to see if zero is a possible number. A cross-year comparison of reported accidents gives the analyst further evidence on which to base a decision. For our purposes, we decided that any system with more than ten revenue vehicles could not have zero accidents, and a missing data symbol was inserted for these systems. Other systems were then judged individually--taking into account their peak vehicle size (a better measurement of size than revenue vehicle fleet), their safety record in other years as reported in Annual Reports or APTA reports (9) and the performance of like-sized systems.

Some judgments about missing data involve making decisions about whether a concept is adequately measured by a combination of several different variables. For instance, vehicle maintenance can be supplied by employees on the transit agency payroll or by contract with other organizations. Thus if a system reports zero maintenance employees, the analyst would expect to have zero maintenance wages reported but a substantial expenditure for services indicated under either the maintenance function or general administration. In the absence of wages and service expenses, a missing data symbol would be used to indicate that maintenance expenses are missing.

For some other variables, the decision is more complex because a zero value can be a real value or it can be an indication of a problem. The example of total vehicle miles will make this clear. Total vehicle miles, as noted above, is constructed from three variables--average weekday miles, Saturday miles and Sunday miles. If weekday miles are zero, it can be assumed that information is missing. However, many systems do not offer weekend service, so a zero for Saturday or Sunday miles might be real or might be an indication of a problem. Since this information is based upon a time consuming sampling procedure, there is a definite possibility that a transit system failed to collect this information, and thus has a missing data problem. The Section 15 data tape includes information about the service schedule of each system. Therefore, it is possible to determine if a system offers Sunday service or not, and whether it has a missing data problem or not. This kind of cross-checking of variables is possible only with the data tape since the Annual Report does not carry information on service schedules.

The problem of missing data has received detailed attention because it is an inevitable problem with a data base as complex as the one mandated by Section 15. Over 300 different systems must learn to interpret and fill out numerous forms--ranging from 17 pages for a small, single mode system to 90 pages for a large, multi-modal system. Since 1980 was only the second year in which this information was reported, some systems were still in the process of instituting accounting systems compatible with Section 15 requirements. While missing data will become less of a problem as transit systems become accustomed to the reporting requirements, there are always new systems which will be completing the forms for the first time. In 1980, 321 systems reported; in 1983 414 systems are expected to report.

Missing data are not evenly distributed across variables or transit systems. In 1980 the most complete data were available for economic variables such as operating expenses and passenger revenue (see

Table 3). The most incomplete information was available for passenger measures such as unlinked passenger trips and passenger miles.

The missing data situation is particularly acute for small systems--those with fewer than 25 revenue vehicles. Thirty % of these systems are missing information on passenger trips and 6% on expenses. Although it is still possible to analyze the smaller systems, since over 1/3 of all systems fall within this size category, generalizations to <u>all</u> small systems must be made cautiously.

The failure to identify missing values with a special symbol can greatly distort the results of a statistical analysis. If too many zeroes are allowed to remain in the data, the mean for a variable will be unrealistically low while the standard deviation will be too high or distorted. Unwarranted conclusions will also be drawn if care is not taken. For instance, it would look like small systems carry many fewer passengers per peak vehicle because small systems are missing 30% of the data on this measure while the large systems are missing only 13% of the data.

Data Validation

The final phase of data preparation consists of cross-checking the data for validity. Errors can enter the database in many ways--misinterpretation by a transit system of what number should be reported, miscalculation of totals, and key punching errors as data are prepared for the computer. Four major methods were used to validate the data: recomputation of totals, comparisons of redundant information,

Table 3

The Distribution of Missing Data in Selected Transit Variables

Variable	% missing values out of 304
Passenger Revenue	0.7%
Total Operating Expense	2.0%
Total Employees	2.6%
Total Vehicle Miles	8.2%
Unlinked Passenger Trips	18.1%
Passenger Miles	24.0%

comparisons of related information and comparison to feasible value ranges. An example of each of these methods with specific variables will be given.

The total number of employees reported for each system was compared to the sum of the separate categories. In about ten cases, the totals differed by more than could be accounted for by rounding errors. In most cases the differences were apparently caused by keypunching errors (e.g., reversal of digits) or simple miscalculations. For these cases, reported totals were replaced by the recalculated totals and cross-checks made with the Annual Reports.

Much of the financial data is reported in several different places. For instance, the Revenue Summary Schedule (Form 201) summarizes the information on the Revenue Subsidiary Schedule, (Form 203). Total operating expenses are also reported in two different places on the magnetic tape. A simple comparison of these numbers reveals differences and the correct number can often be identified by the other validation methods.

Different variables in the database are sometimes different measures of the same thing. For instance, employee counts and employee wages are two different measures of labor utilization. If a transit system has a large number of vehicle operators, it must have a proportionately large amount of vehicle operator wages. However caution must be used in some of these comparisons. Maintenance employee counts and maintenance wages are sub-divided into distinct, non-comparable sub-groups.

The final method of identifying mistakes is to look for values that lie outside an expected range for that specific variable. This method works best for measures that are combinations of two variables such as miles per hour or cost per passenger. Miles per hour (speed) has an expected range of about 5 miles per hour (dense urban areas) to 30 miles per hour (commuter service). Any system that falls outside this range or is in the wrong part of the range for the kind of service it offers, probably has a mistake in either its measure of miles or hours.

A variable such as cost per passenger is a little more difficult to work with since inflation and difference in fiscal years causes the feasible range to change over time and the boundaries of a feasible range are indefinite. In this instance we looked at all cases which lay more than three standard deviations from the mean as well as the largest and smallest cases. While some of these outliers had apparent, real causes, such as extremely long trip lengths, others were so different from the norm that they were obviously wrong. In these cases we looked for the correct values in other parts of the database, or in other sources. If a correction was impossible, incorrect values were designated as missing.

In the future, many of these validation procedures will be incorporated into the preparation of the Section 15 data tape. Beginning with the 1981 data, totals and internal measures of validity were checked for each transit system by TSC. However, the last validation procedure outlined above will remain a useful procedure for the next few years because it looks at a transit system in relation to other systems. TSC also compares a system to itself across years as another validity check.

Although this was done for specific problems in our validation procedure, it was not done systematically.

UNIVARIATE PROPERTIES OF VARIABLES

To be useful for statistical analysis, a data set must meet the basic assumptions of the specific technique to be used. One common assumption is that a variable is normally distributed. Another common assumption is that the variances of two variables are equal. Tables 4 and 5 show a set of variables from the Section 15 data base and the statistics which show whether or not they are normally distributed.

The most striking characteristic of the Section 15 data is the great variation of values for many variables. The major reason for this is the great range in size of transit systems. As the number of peak vehicles in Table 4 shows, transit systems can be very small or very large. Most other variables such as expenses or passenger trips will have correspondingly wide variation.

A normal distribution can be described in terms of a few characteristics. The mean is an average value for a variable. Most values will be quite close to the mean. In fact, 95% of the cases will be within two standard deviations larger or smaller than the mean. As the value of a variable gets farther from the mean, fewer cases will have that value. Additionally, there are just as many values larger than the mean than smaller in a normal distribution.

Table 4

Extreme Ranges of Typical Transit Variables

Variable	Mean	Stan. Dev.	Range min	-	max
Unlinked Passenger Trips per Dollar Cost	1.35	.57	.203	-	3.55
# Passengers (in 1000s) per Peak Vehicle	9.57	5.18	1.029	-	36.0
# Peak Vehicles	124.7	316.4	1	-	3,378
Operating Expense (in Thousands)	12,462	41,560	10	-	441,060
Unlinked Passenger Trips in Thousands	22,118	88,655	10	-	1,139,560

Table 5

A Comparison of Normal and Non-normal Variables

	Variable Name	Variance	Skewness	Kurtosis
Most Normal	Unlinked Passenger Trips (1,000's) per Dollar Cost	.32	.81	1.27
	Unlinked Passenger Trips (1,000's) per Peak Vehicle	27.0	1.57	4.47
	# Peak Vehicles	100,109	5.98	46.40
	Operating Expense (1,000's)	1,727,233,600	7.80	76.75
Least Normal	Unlinked Passenger Trips (1,000's)	7,859,709,000	9.42	105.68

The skewness of variable shows, relatively, how many of the cases are either larger or smaller than the mean. A normal distribution has a skewness of zero because there is no difference between the number of larger and smaller cases. The kurtosis shows, relatively, how many cases are closely bunched together. A normal kurtosis is also zero.

As Table 5 shows, the Section 15 variables vary greatly in terms of how normal they are. In fact, most common variables which describe aspects of a transit system--such as number of peak vehicles, operating expense and unlinked passenger trips--deviate greatly from normality. The distributions of these variables are very skewed because many more systems fall below the mean than above it. The distributions have a high kurtosis because the small systems are quite similar to each other while the large systems are more disparit.

Figure 3 shows this graphically. The solid line shows a normal distribution. The segmented line shows a typical transit variable distribution. The high skewness of the distribution can be seen in the way most cases fall to the left of the mean. The high kurtosis can be seen in the way the transit variable's peak is higher than that of the normal distribution.

Although many statistical techniques can tolerate some departure from normality, our work has shown that the direct use of these variables produces meaningless results. For instance, a near perfect regression correlation can be obtained between passenger revenue and unlinked passenger trips, but predictions are wrong by as much as 10,000% for small transit systems.

Figure 3.

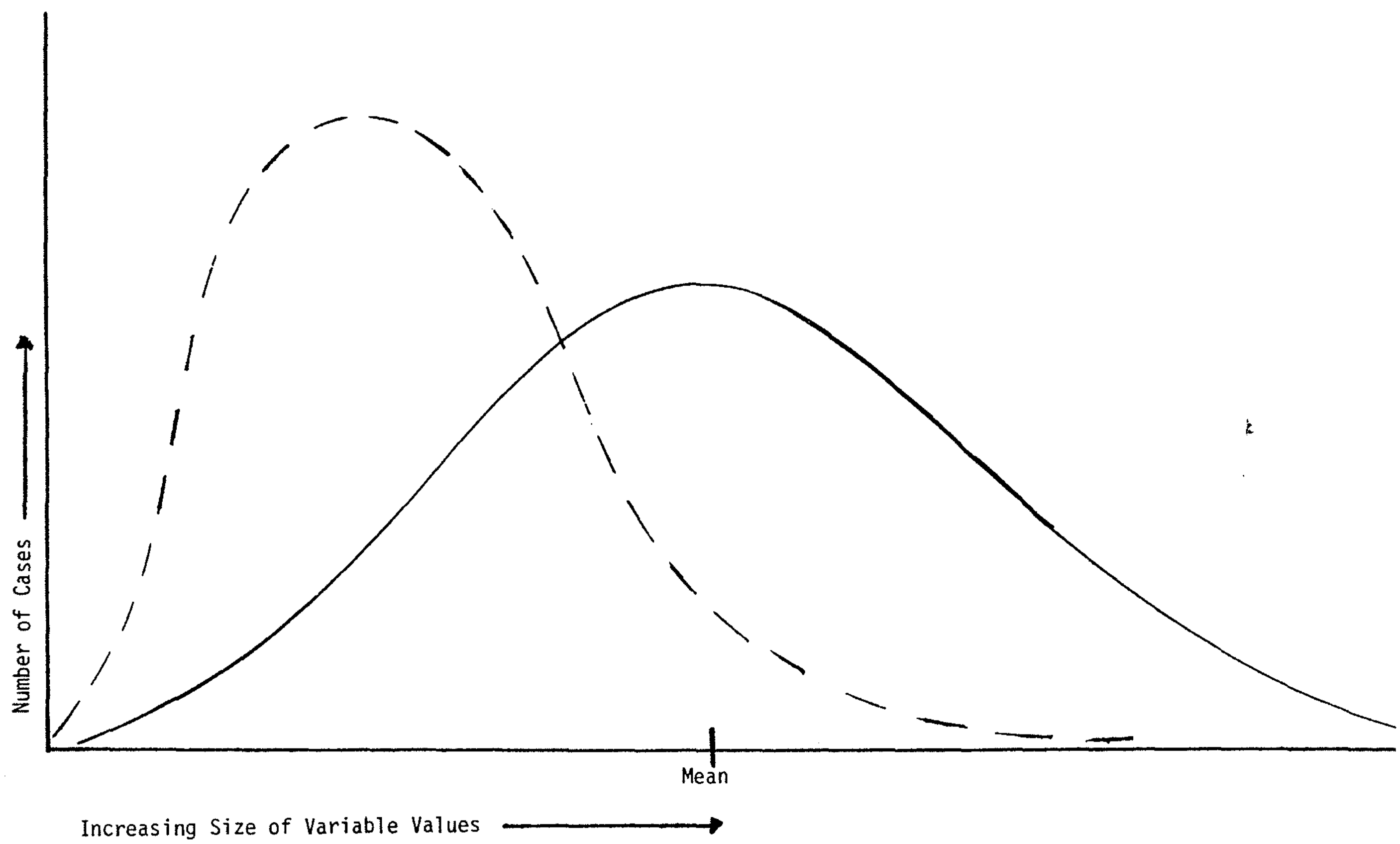

Figure 3. A Comparison of a Normal Distribution and the Distribution of a Non-normal Transit Variable

Table 5 also shows that the variances of different variables are quite different. While transit systems show great variation in their size, this variation is exponentially increased in the variance measure. Thus great care must be used when combining variables in statistical analyses. For some purposes, standardization will take care of variance problems. But for other purposes, the entire distribution must be transformed.

Since the departures from normality are a consequence of the great range in size of transit systems, any correction for size will make a more normal distribution. The first variable on Tables 4 and 5, passenger trips per dollar cost, shows this effect in action. Large numbers of passengers and high expenses tend to go together so the ratio of the two corrects for the largeness or smallness of a transit system. This ratio variable is more normal than either of the variables that were used to compute it. However some ratios such as passengers per peak vehicle are less normal because the original variables were not equally distorted by the effects of size, as shown in the greater differences in their variance, skewness and kurtosis.

Another technique which corrects for non-normality is a logarithmic transformation of a variable. This transformation causes the outlying, very large systems to be more proportionately scaled to the rest of the transit systems. Other methods for coping with the non-normality can be devised, including elimination of large outliers and analysis with smaller peer groups of transit properties that are relatively homogeneous with respect to size. However, these methods reduce the sample size and potentially eliminate important variance in the data. The method chosen should depend on the goals of the statistical analysis.

Summary

The Section 15 reporting system has created a rich, new source of data for analyzing the performance of the transit industry for both researchers and transit managers. For those who want a limited amount of information on a few systems, the published Annual Reports provide easy access to basic information. However, for those who wish to use large samples, information in great detail or information reported at the A, B or C levels, the magnetic data tape provided by TSC is the better source for data.

This paper has outlined methods for using the magnetic tape, including the reorganization of data for use with statistical software, calculating basic variables, identifying missing information and validating the data values reported. In addition, some cautions are given for using the data because the pattern of missing data makes the existing data nor perfectly representative of the transit industry and many of the data variables are not normall distributed.

In coming years, access to valid, reliable data on the transit industry will become increasingly available. Missing data problems will decrease as the transit industry becomes familiar withthe Section 15 reporting requirements. Beginning with the FY 1981 data, TSC has begun extensive validation checks. In addition, they have begun to distrubite the data in new ways. The same information that is reported in the

Annual Report for 1981 is now available on diskettes for mini-computers. A magnetic data tape in a 'sequential' format is also available for the 1981 data. Although this data tape reduces the 62 file structure into two files, it has the same formatting problems delineated in this report.

Beginning with the FY 1983 data, TSC will be using a new operating system and will begin to explore new ways of distributing the data for specific purposes such as statistical analysis. However, analysts who wish to work with the first four years of data will need to reorganize and clean the data before beginning further analyses.

ACKNOWLEDGMENT

Research for this paper has been supported by the Urban Mass Transportation Administration, U.S. Department of Transportation, Research Contract No. CA-11-0026. The U.S. Government asumes no liability for the contents or use thereof.

GLOSSARY

Case: A statistical concept; the full collection of information items defined in a codebook for a single transit company; to be distinguished from a "row" or a "record."

Codebook: The scheme by which data are organized in sets of columns; the logical organization of a data file. A sample codebook appears in Table 2.

Field: A set of columns reserved for one and only one kind of information, to be distinguished from a "column" or a "variable."

Format: The imposition of a logical organization on a physical file; the act of communicating to the computer how data are defined.

Hierarchical ordering: A data file organization scheme wherein a field may contain more than one variable and missing records are permissible.

Logical organization: The scheme contained in a codebook by which data are broken up into fields. There may be more than one, meaningful logical organization for a given data file.

Record: A formattable string of numbers. Not to be confused with "case" or "row."

Variable: A statistical concept; when all cases have been measured on a given attribute, the resulting collection of values organized into an ordered list.

REFERENCES

1. J.M. Holec, D.S. Schwager, A. Fandialan. Use of Federal Section 15 Data in Transit Performance Evaluation: Michigan Program. Transportation Research Record No. 746, 1980.
2. E.L. Owens. Utilization of Section 15 Data by State Governments. Paper prepared for the 59th Annual Meeting of the Transportation Research Board, Washington, D.C., January 1980.
3. Iowa Department of Transportation. Uniform Data Management System: System Development and Testing. Department of Transportation RPT-3715(001)-93-52, DOT-I-81-2, October 1980.
4. F.W. Sherkow. Billing and Accounting by Use of a Computerized Data Reporting System: The Iowa Experience. Transportation Research Record No. 823, 1981.
5. Post, Buckley, Schuh & Jernigan, Inc. and McDonald Transit Associates, Inc. Florida Transit System Performance Measures and Standards. Tallahassee, Florida: Florida Department of Transportation, Division of Public Transportation Operations, 1979.
6. J.M. Holec, Jr., D.S. Schwager, and M.J. Gallagher (Peat, Marwick, Mitchell and Company). Improving Usefulness of Section 15 Data for Public Transit. Transportation Research Record No. 835, 1981.
7. G.J. Fielding, M.E. Brenner, and O. de la Rocha. Using Section 15 Data for Transit Performance Analysis. Interim Report prepared for the U.S. Department of Transportation, Urban Mass Transportation Administration, January, 1983.

8. U.S. Department of Transportation, Urban Mass Transportation Administration. National Urban Mass Transportation Statistics: Second Annual Report, Section 15 Reporting System. Washington, D.C., U.S. Department of Transportaton, July 1982.

9. American Public Transit Association. Operating Statistics Report 1981: Transit System Operating Statistics for Calendar/Fiscal Year 1980. American Public Transit Association, Washington, D.C., October 1981.