U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA
People Saving People
www.nhtsa.dot.gov

DOT HS 809 005 **December 1999**

NHTSA Technical Report

Control Charts as a Tool in Data Quality Improvement

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 809 005		

4. Title and Subtitle	5. Report Date
Control Charts as a Tool in Data Quality Improvement	December 1999
	6. Performing Organization Code

7. Authors	8. Performing Organization Report No.
Carl E. Pierchala, Ph.D. and Jyoti Surti, B.S.	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
State Data Reporting Systems Division National Center for Statistics and Analysis 400 7th Street, SW Washington, DC 20590	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Research and Development National Highway Traffic Safety Administration 400 7th Street, SW Washington, DC 20590	NHTSA Technical Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Control-charting has been successfully applied to two National Highway Traffic Safety Administration data systems to help improve and assure the quality of their data. This paper describes the methods used, illustrates the approach through various examples, and discusses various technical issues in applying control charts to these traffic safety data. The paper also explains the rationale of the methods in terms of statistical process control logic. Finally, an example of nonrandomly missing data is given.

17. Key Words	18. Distribution Statement
Data quality assurance, motor vehicle traffic safety data, statistical process control, hierarchical data, nonrandomly missing data, nonindependent observations	This document is available to the public through the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161

19. Security Classif. (of the report)	20. Security Classif. (of this page)	21. No. Of Pages	22. Price
Unclassified	Unclassified	25	

Form DOT F 1700.7 (8-72) (facsimile) **Reproduction of completed page authorized**

ACKNOWLEDGMENTS

The authors thank Charles J. Thronson for extensive editorial assistance in preparing this paper, and Michael P. Cohen for comments on drafts of this paper.

1. INTRODUCTION

Data of adequate quality are key to the successful use of diverse tools such as computer technology, data bases and statistical methods. Redman (1992, 1996), a pioneer in the application of modern quality methods to data quality, gives many reasons to be concerned about data quality. This paper describes an approach using control charts that can help improve data quality, which adds value to a data base. This approach can also help to assure that a data base does not have major inadequacies in data quality. The method advocated in this paper does not require the collection of additional data. The approach can be used on its own to complement more traditional quality assurance practices, or it can be used in conjunction with or in transition to the more intensive "data tracking" approach recommended by Redman (1996). By using the approach outlined in this paper, problems in data quality can be detected and dealt with far in advance of the release of a data base to the general data user community.

Statistical quality control has received much attention from industry in the last decade, in part due to the work of Deming (1986a, 1993). That work led us to hypothesize that control charts can be used to improve the quality of large data systems, such as the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS). The opportunity to test this arose just prior to the release of the 1993 FARS Annual Report File in July 1994, when it was discovered that many values for Restraint Use (i.e., seat belt usage) were incorrectly coded for one state. Because this is an important data element for NHTSA and various FARS data users, it was imperative to determine the source of this incorrect coding since some preliminary results had already been released. In follow-up, it was demonstrated that had the data been routinely control-charted, it would have been known months in advance that some sort of a problem existed. Consequently, management determined that a program of control-charting of the data should be established. This paper describes program features, gives examples of successes to date, and proposes ideas for enhancement.

Whereas Deming (1986a, 1993) provides a fine overview on the use of control charts from a management perspective, Western Electric's (1956) *Statistical Quality Control Handbook* provides a more detailed 'how to.' Following Western Electric's handbook as a guide, we have employed p-charts of attributes over time. A p-chart is one type of control chart, used when the basic data are summarized as percentages. We developed software using Base SAS® (SAS Institute Inc.1988) and SAS/GRAPH® (SAS Institute Inc.1990) to produce p-charts based on the Western Electric model. We follow the four rules for identifying 'out-of-control' points as recommended in the handbook (Western Electric 1956, pp. 25-28).

We found few reports documenting the use of control charts for data quality control. Deming and Geoffrey (1941) employed control charts of error rates in transcription, coding and card punching in the 1940 population and housing census. Hansen, Fasteau, Ingram, and Minton (1962) mention the use of control charts to help ensure the proper spacing of microfilm negatives used for computer data entry via optical sensing in the 1960 Census. Naus (1975) elaborates upon this particular application of control charts. Liepins (1989) mentions the possibility of using p-charts for repeated surveys. More recently, Redman (1992, 1996) and colleagues (Huh, Keller, Redman, and Watkins 1990; Huh, Pautke, and Redman 1992; Pautke and Redman 1990) applied control charts in conjunction with a technique called 'tracking' to improve data quality in information management systems. Tracking involves randomly sampling a subset

of observations when they first come into a data system, and following those observations through the various subprocesses in the system to pinpoint the occurrence and rates of errors. The resulting data are both control-charted and used in other ways as outlined in standard quality and process improvement theory (e.g., Redman 1992, 1996).

Neter's (1952) paper on statistical methods in auditing includes a substantial section on the control of clerical accuracy. Neter (1952, p. 14) states that statistical control gives the auditor the best assurance of reasonably satisfactory clerical accuracy. Analogously, the adequacy of data quality is best assured by statistical control of data quality. Indeed, Deming (1986a, pp. 268-269 and p. 332) states that it is necessary to keep the measurement process in a state of statistical control. Control charts are a fundamental tool for attaining statistical control. In addition, control charts can be a tool for quality improvement. One illustration of this is Neter's (1952, pp.8-9) example of a mail order business that used control charts to reduce errors in filling orders by 58% through the use of control charts.

2. DATA AND METHODS

2.1 Data

We have applied control chart methodology to data from two of NHTSA's data systems: the Fatality Analysis Reporting System (FARS) (U.S. Department of Transportation, 1997) and the General Estimates System (GES) (U.S. Department of Transportation, 1998b). These data systems can be viewed as having a three-level hierarchical data structure. The basic unit of observation is the motor vehicle crash. Data elements (variables) describing the crash comprise the base level of the data structure. The second level consists of data elements that describe each vehicle in the crash. The third level consists of data elements that describe persons. Within a vehicle, there are usually one or more occupants – the driver and passengers. A crash may also involve pedestrians or occupants of non-motorized vehicles such as bicycles. Data elements are referred to as crash-level, vehicle-level, or person-level, depending on their level in the hierarchy. Note that driver data elements are often viewed as vehicle-level variables since there is almost always a one-to-one correspondence between vehicle and driver. For example, in FARS the data elements related to Driver License Status are stored in the Vehicle File (U.S. Department of Transportation, 1998a, pp. V.22-V.24).

The FARS (formerly, the Fatal Accident Reporting System) is an annual census of all fatal crashes in the United States. NHTSA has cooperative agreements with all 50 states plus the District of Columbia to provide data on fatal crashes. (Data are also collected for Puerto Rico, but these are not considered in the current work.) Each state has one or more data collectors called FARS analysts. They obtain existing state and local records concerning fatal crashes, interpret and mentally reconstruct the relevant events, code the data in the standardized FARS format, and key the data into microcomputer files that are periodically transferred to headquarters. There are a variety of quality assurance steps in the process of obtaining the data and building computerized files. Four data releases are made for a calendar year, with the first release covering the first six months of the year and the final release being the most complete.

The GES is a complex probability sample of all police-reported crashes in the United States, and is the

basis for national estimates on a variety of data from such crashes. Sixty Primary Sampling Units (PSUs) were randomly sampled from 12 strata based on geographic region and degree of urbanization. The selected PSUs are located in 26 states. Within a PSU, police jurisdictions are randomly selected. Crashes are listed at the police jurisdiction. There is an additional stratification into a variety of crash types of interest to NHTSA, based on vehicle type, tow status and existence of injuries. Listed crashes are then sampled on a probability basis within these strata. For these crashes, photocopied police reports are sent to a central location where data entry/coding specialists code the data following the standard GES format. A file used for analytical efforts is built and updated quarterly during each calendar year, with the year's finalized file being completed around the middle of the following year.

Note that in building the FARS and GES files, data are entered in an ongoing fashion into what is called a master file. Then, at certain points in the data cycle for a calendar year, the accumulated data are 'frozen' in so-called analysis files in SAS data sets. At the later stages of the data cycle, the analysis files are released for public use. The earlier analysis files are used for in-house purposes only, since they are substantially incomplete data sets.

2.2 Methods

We use control charts in two ways. First, as part of our quality assurance in preparing for a data release, we examine control charts to check that there are no major anomalies in selected important data elements. Second, we use control charts in the quality management sense, to identify opportunities for data quality improvement. In this paper, we focus on the later use of control charts.

2.2.1 Control Chart Construction

We follow the Western Electric model in producing p-charts (e.g., Figure 1), each with time (in months) on the x-axis, and percentage (per month) of observations that have a specific attribute on the y-axis. (An attribute is a categorical characteristic which can be determined as either existing or not existing for each observational unit. For additional explanation, see Western Electric (1956, p.17). For example, in FARS if a person is coded 1 (Totally Ejected) or 2 (Partially Ejected) for the data element Ejection, then that person is affirmative for the attribute 'ejected'. As a practical matter in the current work, an attribute is defined by a data element plus one or more specific coded values for that data element.) The actual date of the crash is used in grouping crashes by time in months. Data from three calendar years are used -- the data from those months in the current data release that are expected to be reasonably complete, plus all the data from the latest releases for the two previous years. For each attribute charted in FARS, we produce a chart for each state and the District of Columbia; this yields 51 charts per attribute charted. In the GES we produce a chart for each PSU. This yields 60 charts per attribute. However, we use an automatic screening algorithm (described below) to suppress the printing of many control charts. The attributes charted are those judged important to the respective data system. Recently, we have been charting 23 attributes in FARS and 17 attributes in GES. Thus, there is the potential to produce 1173 charts for FARS, and 1020 charts for GES.

Figure 1 shows a p-chart in good statistical control in one PSU from the GES. For each month over three years, we have plotted the percent of drivers reported by police as using alcohol (code 1) or, in

jurisdictions where the distinction cannot be made from the crash report, alcohol/drugs (code 7). This control chart was produced using the 1998 nine month file data plus the 1997 and 1996 final file data. Note that an alias is employed for the PSU (or state) name in this and all other charts shown in this paper. Each plotted point for a month is the monthly percentage of passenger vehicle drivers whose value of PER_ALCH (Police Reported Alcohol Involvement) was 1 or 7 (recoded as 11 for programming purposes), with drivers having unknown values excluded from the computations. The monthly sample size, M, is not fixed, since the number of drivers involved in crashes varies from month to month. To simplify comparisons between p-charts, we always let the y-axis run from 0% to 100%.

In addition to plotting the monthly percentages, the chart has a centerline and 'stairstep' control limits. We call the first two years in the chart the *base years*, and the last year the *extension year*. The centerline is a percentage, 100 times the proportion, P, computed by taking the cumulative number of passenger vehicle drivers having the value 11 for PER_ALCH in the two base years, and dividing by the total number of passenger vehicle drivers in the two base years having nonmissing PER_ALCH. Control limits for each month are displayed around the centerline at plus and minus three standard errors (or 'three sigma'). The standard error is taken as that of a binomial distribution, so $SE=\sqrt{P(1-P)/M}$. The lower control limit for a month is then LCL = 100(P-3SE), and the upper control limit for a month is UCL = 100(P+3SE). These are so-called three-sigma control limits. Since the sample size M varies from month to month, the control limits vary, giving rise to the term 'stairstep' control limits.

2.2.2 Control Chart Theory

When the fluctuations in the charted data appear random (see Figure 1), the system producing the data is termed 'stable.' The data behave as if they are coming from a binomial random process with parameter P. The points are predictable in terms of the random process, and all fall within three-sigma control limits. All points are in 'statistical control.' The PSU whose data are charted in Figure 1 is moderate in terms of an average monthly sample size of 134, but its P of 0.072 is small, hence the narrow control limits. In addition, the fluctuations seen from month to month in the control limits are small, indicating low variability in the monthly sample sizes.

The use of two base years to estimate the centerline and control limits is based on Deming's (1993, p.180) notion that a system in statistical control is predictable, at least in the short run. If the system producing the data is in statistical control, then the data in the extension year should continue to fall between the control limits determined from the base years.

In Figure 2, for percent of unrestrained passenger vehicle occupants from FARS, the data are not in statistical control. In January of the third year charted, the point goes "out-of-control" beyond the lower control limit. Each successive point is also out-of-control. It appears that the process producing the data stabilizes around March at a much lower value than in the preceding two years. In short, the process becomes inconsistent with a binomial model with a fixed P for all three years.

Western Electric (1956) recommends four rules to identify patterns (and implicitly, points) in control charts as out-of-control. The first is the three-sigma rule illustrated in Figure 2; that is, the chart has at least one point falling outside of the three-sigma control limits. The other rules are: rule 2, two out of three

consecutive points more than two sigma away from the centerline (with the two points on the same side of the centerline); rule 3, four out of five consecutive points more that one sigma away from the centerline (with all four on the same side of the centerline); and rule 4, eight consecutive points on the same side of the centerline.

When a point's value violates one of the four rules, it is marked as 'unnatural' or 'out-of-control.' The Western Electric handbook does this by putting an 'x' above or below each unnatural point. For programming simplicity, we plot an unnatural point with an asterisk '*' rather than with a large dot as is used to plot a 'natural' or 'in control' point. Note that rules 2-4 apply to control charts where time or some other ordered variable is on the x-axis. At each point and for a given rule, we take the required number of preceding points, if they exist, and apply the rule. If the set of points satisfies the condition of the rule, the point is unnatural.

Figure 3 illustrates rules 2 and 4. The two points for April 1992 and May 1992 are both more than two sigma above the centerline, so the point for May is marked as out-of-control because of the two-out-of-three rule. The point for June 1993 is marked because it violates the eight-in-a-row rule. Finally, the point for May 1994 is marked because it violates the two-out-of-three rule, in this case below the centerline. It appears that a slight reduction in nonuse of restraints (i.e., an increase in restraint use) occurred around late 1992 or early 1993.

As explained by Deming (1993), the reason for using a control chart in industry is to determine whether a process or system is in statistical control. If not, then the process is not predictable, and cannot be relied on. A responsibility of workers and management is to bring the process into statistical control. If a process is not in statistical control, it is affected by 'special cause' variability. The workers operating the process are usually in the best position to identify and eliminate the special causes of variability. If a process is in statistical control, its variability is called 'common cause.' In this case, management needs to assess the adequacy of the performance of the process. Efforts to improve performance are typically beyond the control of the workers. So if improvement is needed, it is usually the responsibility of management to find ways to change the process to improve its performance.

Although FARS and GES traffic safety data are neither industrial nor commercial, a statistical process control approach to thinking about them seems appropriate. Such data are a product of multiple systems. In addition to the environment, roadway, and other systems and factors that lead to crashes, there are also systems for highway safety management, notification of crashes, police reporting, recapture (by FARS or GES staff), recoding, collection of supplementary data, data entry, quality control, and electronic data processing. The purpose of control charts is to detect situations that need special investigation. The results of the investigation may point to a problem in data quality, or may point to an actual change in highway safety. If data quality, then some action is needed to remedy the data quality problem. If highway safety, then the results of the special investigation may suggest other action.

By way of example, suppose highway safety management practices for influencing restraint usage are changed and lead to a different usage level. This is a special cause of variability, and the points in a control chart for restraint usage are expected to go out of statistical control. Thus, although errors or problems in one of the subsystems may be a reason for a control chart having points out-of-control, it is also the case

that changes over time in highway safety management or other highway/traffic factors can affect a control chart.

2.2.3 Control Chart Enhancements

An early refinement in our use of control charts for data quality improvement and assurance was partial automation in the review and selection of charts to be referred for special investigation to the data collection staff. Based on experience gained with our data, we do not want to refer to the data collection staff every chart that has an out-of-control point, because it is time consuming and expensive to carry out special investigations. We want to refer to them only those charts having more dramatic patterns of out-of-control points. So we conduct a manual review of the charts and apply judgement before selecting the charts to be referred to them. Many control charts are in statistical control or close to statistical control. Since the manual review of control charts is time consuming, at the encouragement of management we developed an algorithm to eliminate from consideration charts that are not substantially out of statistical control. Many charts have no points or just one point out of statistical control, so we decided not to print those charts. We modified our software to print out two categories of charts for manual review. The first category of charts, termed 5S charts, consists of those having two or more points more than 5 standard errors (5 sigma) from the centerline. The second category of charts, termed 2O (numeric *2*, capital letter O), consists of non-5S charts that have two or more points out-of-control based on any one or more of the four Western Electric rules. The number of 5S charts is normally small, so we are able to review them quickly. The number of 2O charts is larger, hence it takes more time to review those. Nonetheless, with this procedure only about a half of the charts need manual review. When applicable, we identify a chart as 5S or 2O by marking it so in the upper right-hand corner of the graph. Figure 2 shows a 5S chart.

In the manual review of the charts, our objective is to select out-of-control charts that are highly likely to reflect underlying data quality problems. Such selected charts are passed on to the data collection staff so that they may conduct special investigations into the cause of the lack of statistical control. In the past when selecting charts, we prioritized them in terms of apparent need for additional investigations as high, medium, low, and none. In a typical run for a data release, we found roughly 30 high priority charts that we passed on to the data collection staff. The special investigations are time consuming, and usually not all of the out-of-control charts were due to a data quality problem. Thus, in view of the work required to do a special investigation, and a presumed relatively low likelihood of data quality problems being the cause of these additional out-of-control charts, we did not feel warranted to pass on medium or low priority charts to the data collection staff. Currently, we only select high priority charts to be forwarded to the data collection staff, and do not identify low or medium priority charts.

2.2.4 Miscellaneous

In choosing the attributes that are charted, we conferred with subject matter specialists to select attributes of high importance to the data system. For FARS, we had been control-charting 18 attributes, but recently we increased the number to 23; for GES, we control-chart 17 attributes. We produce control charts at every major data release for these systems, four times per year both in FARS and in GES. If several data releases show that an attribute is rarely out-of-control in the states or PSUs, then that attribute is usually removed from the set being charted and we substitute another attribute in its place.

Regarding unknown (i.e., missing) data, the unknown values are included in the denominator only when charting percent unknown for a data element. For the percentages for other categories, the unknown values are not used when determining the counts for the denominator. The rationale for this approach is given below. See the comments in Section 3.2, paragraph 8, and the discussion in Section 4, paragraphs 7 and 8.

3. RESULTS

3.1 Substantive Results

At NHTSA, we find the control-charting process useful. The charts help identify a variety of data quality issues. Data system managers find the results useful, and they support the ongoing production and use of these control charts as a complement to their other quality assurance efforts.

Three examples illustrate the usefulness of control charts in identifying problems. First, in Figure 2, we show the chart for the predicament described in the Introduction. It has a striking pattern of out-of-control points, caused by a FARS analyst who misunderstood revised coding conventions. Had we been using control charts earlier, we would have discovered the difficulty sooner and fixed it more easily. Second, in Figure 4, License Compliance code 3 (valid license for this class vehicle) has a dramatic pattern of out-of-control points. This was due to a programming error which put certain data into the wrong FARS data element. Third, see Figure 5 for Ejection codes 1 and 2 (total and partial ejection, combined together as code 12). This chart has points out-of-control beginning in April 1994. These were associated with employment of a new FARS analyst, who apparently needed more training. In summary, the charts help the data system managers to pinpoint specific issues, such as data processing inadequacies and needs for training.

Another finding is that many of the control charts are in statistical control. The data from the two base years give control limits between which the data usually fall in the third (extension) year. Almost no seasonal variability appears in the charted percentages. In other words, the charted attributes are stable, in the quality management sense of the word (Deming 1986a, 1993), in very many of the states and PSUs over the three year time periods utilized. Thus, the charted attributes by and large appear predictable.

A pervasive finding in FARS was that the slow accrual of data caused the percentage of unknown values for certain data elements to go out-of-control toward the end of the extension period. Efforts are being made to improve the timeliness. In some states the local systems producing the data are inherently slow and cannot easily be speeded up (e.g., obtaining blood alcohol values from medical examiners). One management effort was adding an earlier "three month" analysis file to the data cycle, to encourage the FARS analysts to promptly begin entering data into the system at the beginning of a new year. Previously some states allowed a long time lag before they began entering data into the system.

Our focus to date in the GES has been on the data entry process that produces the data in the standard GES format. Some data entry problems have been found via the special investigations instituted based on the control chart results, leading to the reentry of portions of the data. But only a minority of the charts with

out-of-control points were due to causes related to data entry. Nonetheless, according to the GES staff, enough data entry problems have been detected using the control charts that it continues to be worthwhile to use them for this purpose.

Some charts with out-of-control points are due to recent positive safety changes in the highway management system. Figure 6 for one state's FARS data has out-of-control points beginning in September 1993 that are associated with the initiation of enforcement of a new restraint use law in that state on September 1, 1993. In this state, public information and education programs accompanied the new law.

Lastly, there are charts with out-of-control points that remain unexplained despite our special investigations. Perhaps a wider scope of investigation is needed. Nonetheless, we find the control charts a useful addition to the quality control efforts carried out for the FARS and GES data systems.

3.2 Technical Results

In this section we focus specifically on methodological issues.

Prior to producing p-charts, we expected to see seasonal variation in many charts. This would have called for making seasonal adjustments and producing control charts of the resulting residuals. Surprisingly, there was no apparent seasonable variability in the attributes charted. Many charts were in or close to statistical control (see Figure 1.) For example, in the February 1997 runs for the 1996 FARS early assessment file, there were 918 potential charts for 18 attributes for 50 states and the District of Columbia. Of these, 436 were excluded by the automatic selection routine. Of the 482 printed charts, 464 were judged as not warranting special investigation. Of the 18 charts forwarded to the FARS staff for additional investigation, none appeared influenced by seasonal variability.

For our highway safety data, we found it necessary to apply a substantial amount of judgement in interpreting charts with out-of-control points. In the normal use of control charts, such points indicate the need to search for and eliminate any special cause of variability. But many of our charts had one or a few isolated points out-of-control, which, compared to more striking out-of-control patterns in other charts, made it appear that a special investigation would not likely be cost effective. For our data, it is clear that the decision to carry out a special investigation cannot be made simply based on the existence of out-of-control points as determined by standard control chart rules. Since special investigations are time consuming and costly to carry out, it is more practical to apply judgement when selecting the control charts referred to data collection staff for special investigations. For example, when there is a continued pattern of out-of-control points, as in Figure 2, then a special investigation seems more warranted.

Regarding the four rules for determining out-of-control points, the three-sigma and eight-in-a-row rules appear more useful than the two-out-of-three and four-out-of-five rules. However, the latter two rules do sometimes add to a pattern of out-of-control points suggesting a more compelling need for a special investigation.

There is technical justification for ignoring isolated out-of-control points. The underlying binomial model for the control limits is based on the assumption that observations are statistically independent. While this

assumption may be reasonable for crash-level variables, it is not reasonable for vehicle- and especially for person-level variables. For example, factors affecting the crash may cause two or more vehicles in the crash to have unusually high or unusually low values for certain data elements. The same reasoning applies at the person-level. For example, suppose there is a tendency in some vehicles for everyone to wear a seat belt, while in other vehicles no one is using a seat belt. If one of those vehicles is involved in a fatal crash, then all of the occupants will have the same value for restraint use. If all occupants are restrained, no one may be ejected in the crash. But if no occupants are restrained, everyone may be ejected. The data for persons in the same vehicle are clearly correlated, violating the assumption of statistical independence of the observations.

We examined some charts with isolated out-of-control points to check whether they could be explained by a lack of statistical independence. As an example, Figure 7 is the chart for percent ejected for a small state (in terms of monthly average sample size), reflected by wide control limits. Note the out-of-control point in December 1993. Detailed investigation showed that only 24 vehicles with 56 persons were involved in fatal crashes that month. However, one vehicle, apparently a van, had 11 occupants, 7 of whom were ejected. A second vehicle had 6 occupants, all of whom were ejected. Thus, two vehicles had 13 (45%) of the ejections that month. Had these two fatal crashes not occurred, the monthly percentage would have been inside of the three-sigma control limits. Clearly, the observations (persons) in these vehicle are correlated, resulting in the unusually high value in December 1993.

In short, the binomial model is not strictly warranted for vehicle- and person-level variables, lending justification to our practice of ignoring isolated out-of-control points in our charts. Related to this, it appears that states or PSUs with small monthly sample size are more prone to have isolated points beyond three-sigma control limits. However, we have not studied this systematically.

At times it is useful to compare two or more charts. Figure 8 shows three charts for the GES data element Manner of Leaving Scene. The upper chart displays percent unknown. The two lower charts display the percent towed due to damage, but use different methods of handling the unknown values. Although the percent unknown increased substantially beginning in April 1997, the percent towed due to damage remains in statistical control when the unknowns are counted in the denominator in computing the percent towed. However, when the unknowns are excluded in computing percent towed, a lack of statistical control is associated with the increase in unknown values. Apparently, the additional unknown values occurring in the later months are not occurring at random.

Attention to the width of the control limits can help identify unknown value problems, because the width is inversely proportional to the square root of the monthly sample size. In some charts the control limits become much wider in the last few months of data collection. This appears to be almost always due to slow data accrual causing many incomplete observations. For example, in Figure 7 note the extremely high three sigma upper control limit for September 1994, and in Figure 5 in the last few months note the high upper control limits. In Figure 3, where monthly sample sizes tend to be moderate, averaging 46 per month, note the variability in control limits. This reflects months with sample sizes that are only one quarter of those for months having the largest sample sizes.

4. DISCUSSION

In summary, control charts can usefully be applied to data quality. In this paper, we described their use as a tool to identify data elements that appear to need investigation in terms of the possible existence of data quality problems. The control charts help us identify data elements that 'just don't look right.' Additional investigations are then made; action is taken as appropriate. Sometimes this is at the micro level, as in the case of a chart with out-of-control points found to be due to a programming error. In response, the program was corrected. In other cases action is taken at the management level. For example, in one instance it was decided to add an earlier version of the analysis files. This was prompted by investigations into why unknown values tended to go out-of-control in certain states. It was concluded that certain states were very slow to begin their coding. The idea was that building the first analysis file at an earlier date would motivate slow-starting states to initiate the coding earlier in the cycle.

In addition, we can use control charts to identify meaningful changes in data (Spiring, 1994). The example on pp. 77-84 of Western Electric's (1956) *Statistical Quality Control Handbook* shows that control charts sometimes allow insight that is lost using conventional statistical methods. As Deming (1986a, p.132) stated, "Analysis of variance, t-test, confidence intervals, and other statistical techniques taught in the books, however interesting, are inappropriate because they provide no basis for prediction and because *they bury the information contained in the order of production*." (Italics added.) Stated otherwise by Deming (1986b, p. i.), "Thus, the mean, standard deviation–in fact any moment–is in most applications inefficient, as it causes the loss of all information that is contained in the order of observation."

Various issues have prompted ideas for possible enhancements and additions to the type of control chart we have been using to date. For example, we want to explore using new types of control charts in an attempt to better identify problems in FARS and GES. We anticipate charting results across states or PSUs, rather than across time, as a way of identifying anomalous stratification units. Also, a lack of independence might be circumvented by redefining the statistic charted so that it is determined on a crash-level basis rather than a vehicle-level or person-level basis. For example, we could look at the monthly proportion of *crashes* having no belted occupants rather than looking at the monthly proportion of unbelted occupants. However, such restructuring might lose needed detail.

Other issues include the selection of control charts forwarded to data collection staff for further investigations. We need to obtain better data on the usefulness of the charts to improve their selection. Also, our control-charting software needs maintenance and user documentation. Next, in some cases the Gaussian approximation is not appropriate for obtaining the control limits for the binomial-based p-charts; the software should be generalized to obtain better control limits when M and P are small. Finally, for the GES, we have been ignoring the sampling design in producing control charts. We need to investigate the appropriateness of the simple binomial control limits, and to examine the possibility of stratum-dependent quality issues.

There are some interesting questions related to the computation of the probability that a chart will be out-of-control when using the four rules to test for unnatural patterns, given the system is stable. Western Electric (1956, pp. 180-183) provides some related computations, but these do not actually give the probability that at least one point will be out-of-control in a chart. This probability is clearly not simple to calculate.

We have skirted the issue in part by relying on the reputation of Western Electric's handbook. This is combined with the fact that exact calculations of probabilities would require independence of the observations so that the usual formula can be used to determine the control limits. But we know that the individual observations are not statistically independent for many of our charted attributes. Nonetheless, it may be useful to consider this computation. Perhaps the probability is too high and adjustments to the rules would be useful to avoid too many 'false positives.'

Work comparing the NHTSA approach with Redman's data tracking approach, described earlier, may be fruitful. The Redman approach, although comprehensive, appears to require a great deal of work to collect the needed data. This is particularly true when some of the subsystems are primarily manual as opposed to automated processes. None of the references given above to Redman's work discussed the sample sizes and costs needed to successfully implement his approach. The NHTSA approach requires no additional data to produce the control charts. However, data to identify the root cause of a problem suggested by a control chart may be lacking.

Another issue is the possibility of a test of the randomness of an incremental increase in unknowns. As mentioned above, we currently *exclude* unknown values before computing monthly percentages (except when charting a data element's percent unknown). Previously, we *included* unknown values when computing the monthly percentages. However, a problem occurred at some data releases when due to slow data accrual some data elements had substantial increases in later months in the percent of unknown values. These increases in percent unknown caused the percentages for other data element categories to decrease. This often caused points in charts for those other categories to go out-of-control in those last few months. We inferred from this that if both the incremental increase is occurring at random and the unknown values are excluded when computing monthly percentages, then the other categories will remain in statistical control. This should hold provided the system does not coincidentally become unstable due to other reasons. Thus, by excluding unknown values, fewer charts may need special investigations.

Under the assumption that the system did not become unstable for other reasons, a randomness test of the added unknown values in the latter months appears possible. When unknown values are excluded in computing the monthly percentages and points go out-of-control for some category in the later months, that suggests that the extra unknown values are not occurring randomly (see Figure 8). It should be fruitful to study this approach as a methodology for assessing whether unknowns are occurring at random.

Besides technical problems in applying control charts as discussed above, there have been a few management issues. For example, additional and more systematic staff training is needed regarding the purpose and interpretation of control charts. Also, there has been a tendency to use the control charts as an *acceptance sampling* tool to decide if the data appear ready for release. It will be better to increase the use of the control charts as a tool to identify quality problems, which then can be rectified to improve the system. However, further discussion of management issues takes us beyond the scope of this paper. An extensive literature has appeared over the last decade and a half on quality management. The best (Deming 1986a, 1993; Redman, 1996) argue that for maximum effectiveness top management needs to take an active interest in quality.

In conclusion, control charts were found useful as an aid in the improvement and quality assurance of FARS

and GES data. Various data problems have been noted using the control charts, giving specific quality issues to be addressed. The charts also provide additional quality assurance for the data items that are graphed. Management satisfaction with the benefits of the control charts has resulted in their routine use since 1995 in the quality control of the FARS and GES data systems. It is anticipated that resolution of technical and management issues discussed above will lead to increased use of control charts to identify and rectify data quality problems.

REFERENCES

Deming, W.E. (1986a), *Out of The Crisis*, Cambridge, MA: MIT Press.

_____(1986b), "Forward," in *Statistical Method from the Viewpoint of Quality Control* by W.A. Shewhart, New York: Dover, pp. i-ii, (new in Dover Edition).

_____(1993), *The New Economics for Industry, Government, Education*, Cambridge, MA: MIT Press.

Deming, W.E., and Geoffrey, L. (1941), "On Sample Inspection in the Processing of Census Returns," *Journal of the American Statistical Association*, 36, 351-360.

Hansen, M.H., Fasteau, H.H., Ingram, J.J., and Minton, G. (1962), "Quality Control in the 1960 Censuses," in *New Frontiers in Administrative & Engineering Quality Control: Proceedings of the 1962 Middle Atlantic Conference*, Milwaukee, WI: American Society for Quality Control, pp. 311-339.

Huh, Y.U., Keller, F.R., Redman, T.C., and Watkins, A.R. (1990), "Data Quality," *Information and Software Technology*, 32, 559-565.

Huh, Y.U., Pautke, R.W., and Redman, T.C. (1992), "Data Quality Control," in *International Software Quality Exchange (ISQE 92) Proceedings*, Wilton, CT: Juran Institute, Inc., Session 7A, pp. 1-27.

Liepins, G.E. (1989), "Sound Data Are a Sound Investment," *Quality Progress*, 22, 61-64.

Naus, J.I. (1975), *Data Quality Control and Editing*, New York: Marcel Dekker.

Neter, J. (1952), "Some Applications of Statistics for Auditing," *Journal of the American Statistical Association*, 47, 6-24.

Pautke, R.W., and Redman, T.C. (1990), "Techniques to Control and Improve Quality of Data in Large Databases," in *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, pp. 319-333.

Redman, T.C. (1992), *Data Quality: Management and Technology*, New York: Bantam.

_____ (1996), *Data Quality for the Information Age*, Boston: Artech House.

SAS Institute Inc. (1988), *SAS® Language Guide for Personal Computers* (Release 6.03 Edition), Cary, NC: author.

_____ (1990), *SAS/GRAPH® Software: Reference, Version 6* (1st ed.,vols.1 & 2), Cary, NC: author.

Spiring, F.A. (1994), "A Bill's Effect on Alcohol-Related Traffic Fatalities," *Quality Progress*, 27(2), 35-38.

U.S. Department of Transportation (1997), *Fatality Analysis Reporting System, 1998 Coding and Validation Manual*, National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.

_____ (1998a), *FARS Analytic Reference Guide, 1975-1998*, DOT HS 808 792, National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.

_____ (1998b), *General Estimates System Coding Manual, 1997*, National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington, DC: author.

Western Electric Company, Bonnie B. Small, Chairman of the Writing Committee (1956), *Statistical Quality Control Handbook*, Indianapolis, IN: AT&T Technologies (Select Code 700-444, P.O. Box 19901, Indianapolis 46219) .

FIGURES

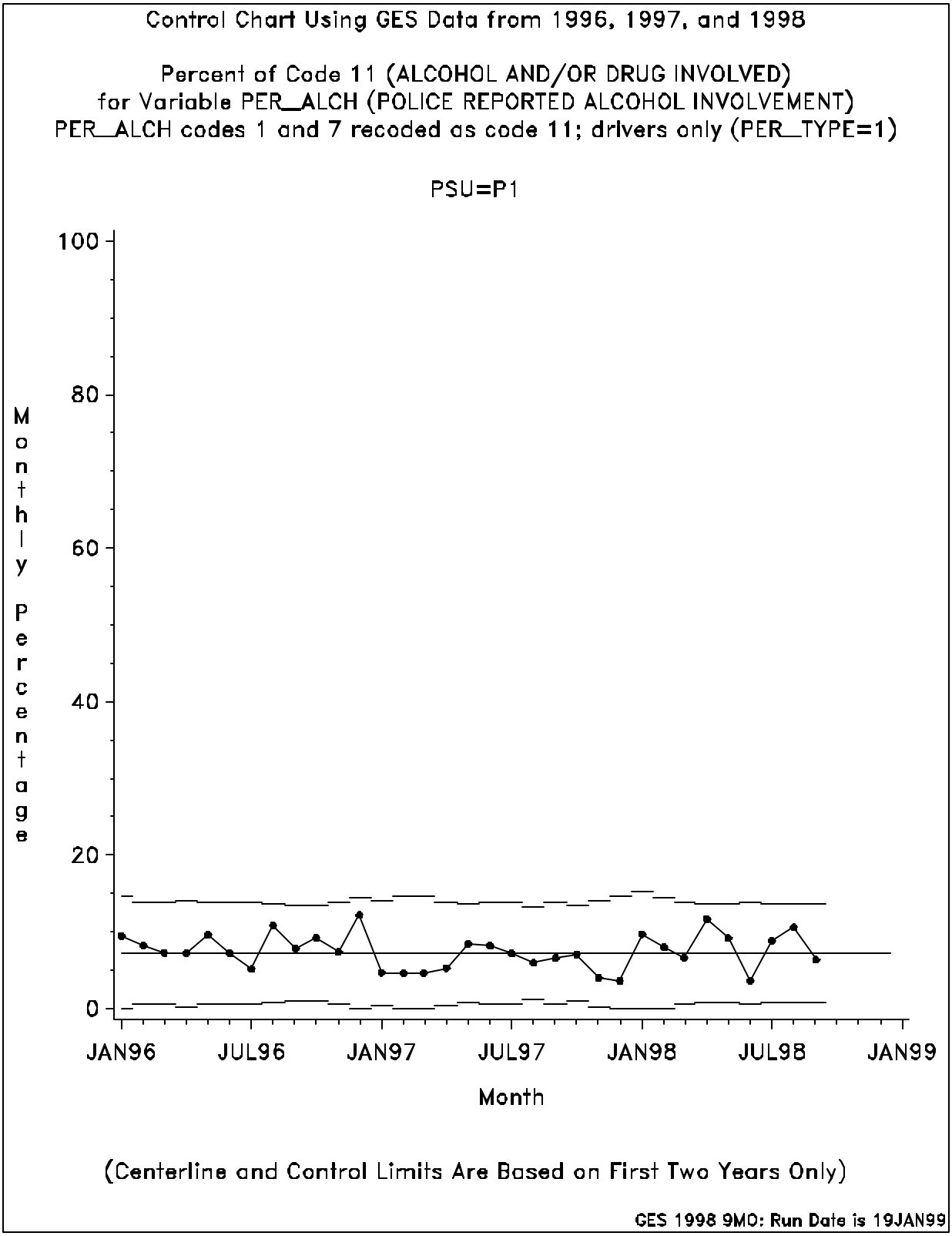


Figure 1. An Example of a Control Chart in Statistical Control.

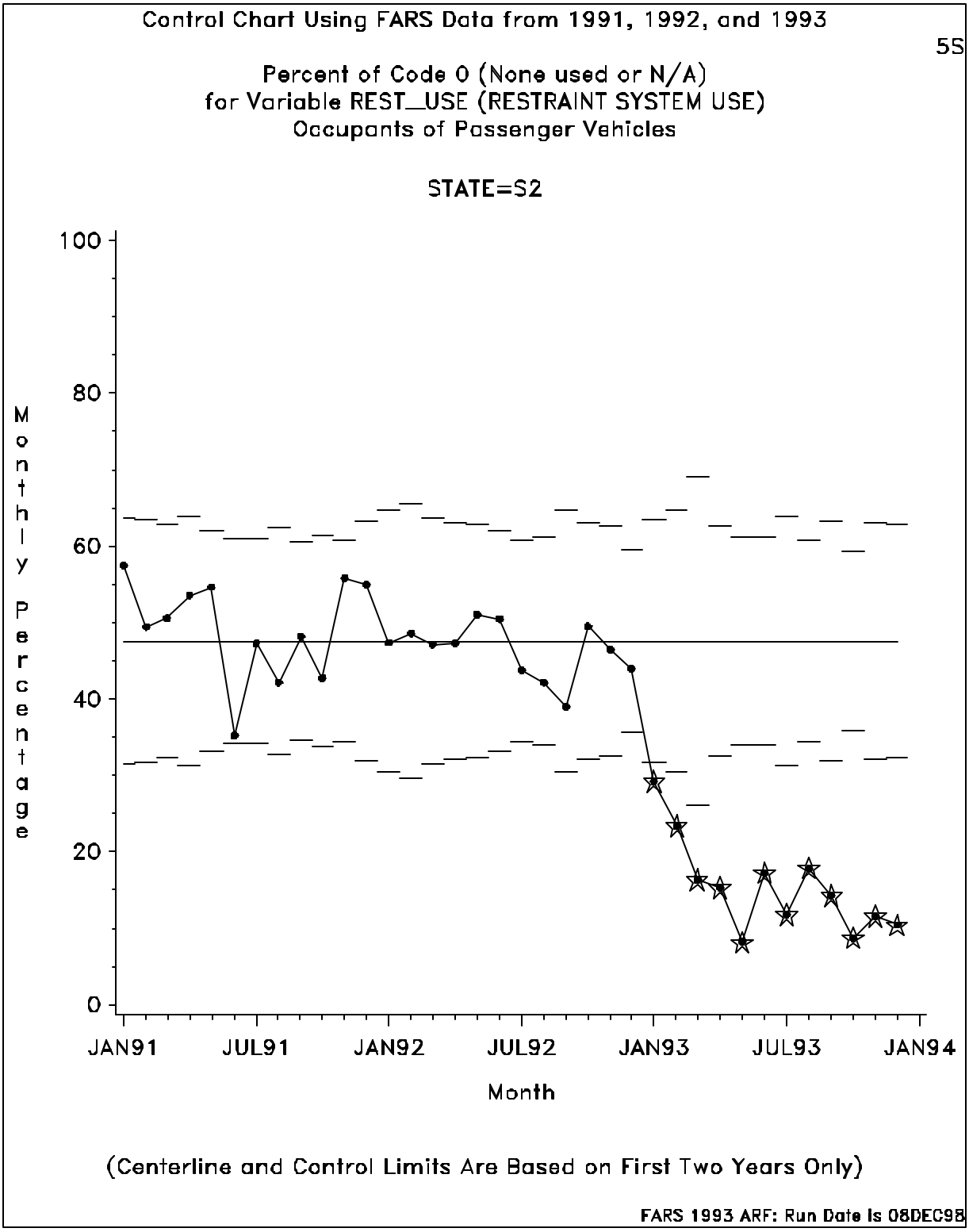


Figure 2. A Control Chart Out of Statistical Control.

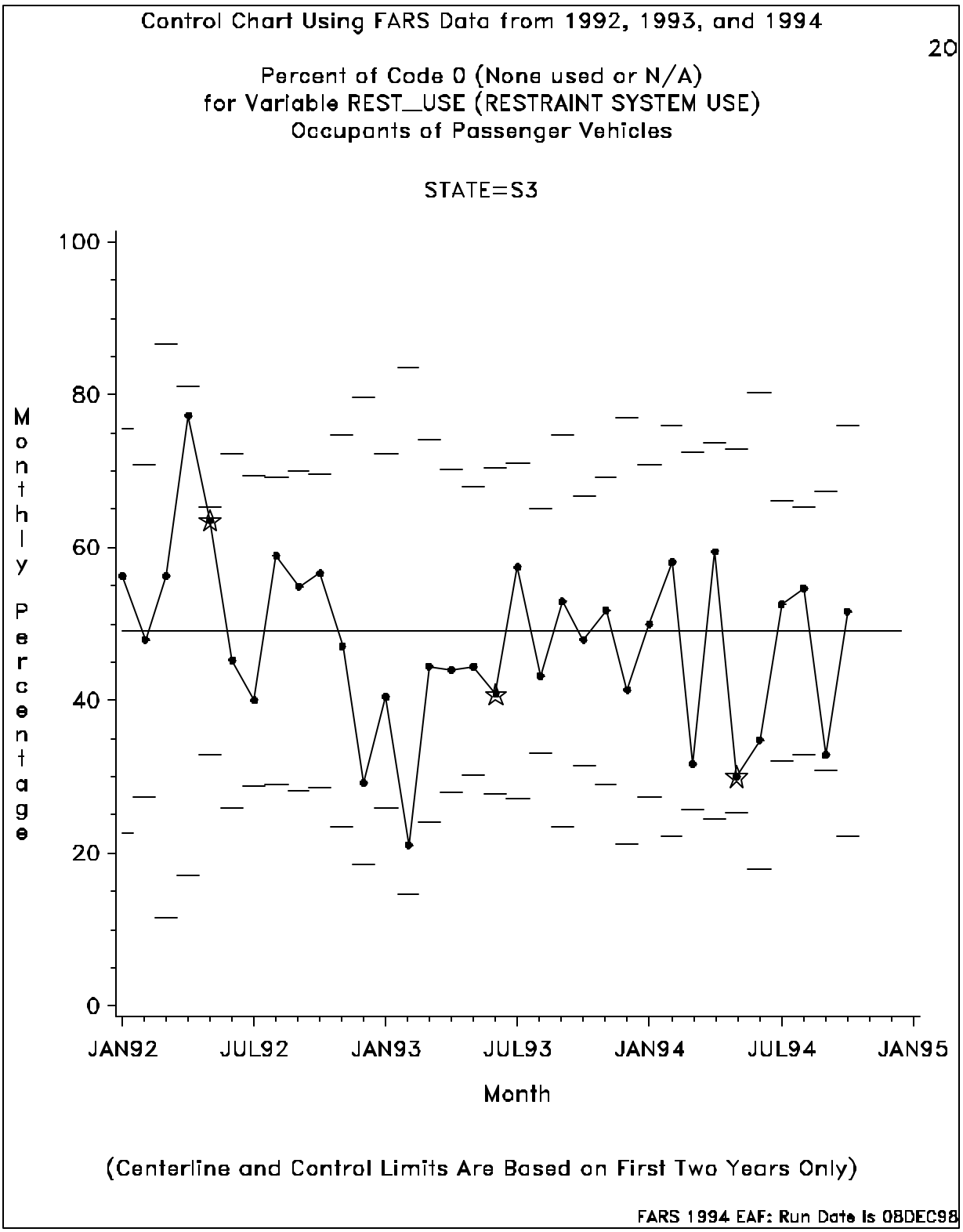


Figure 3. A Control Chart More Subtly Out-of-Control.

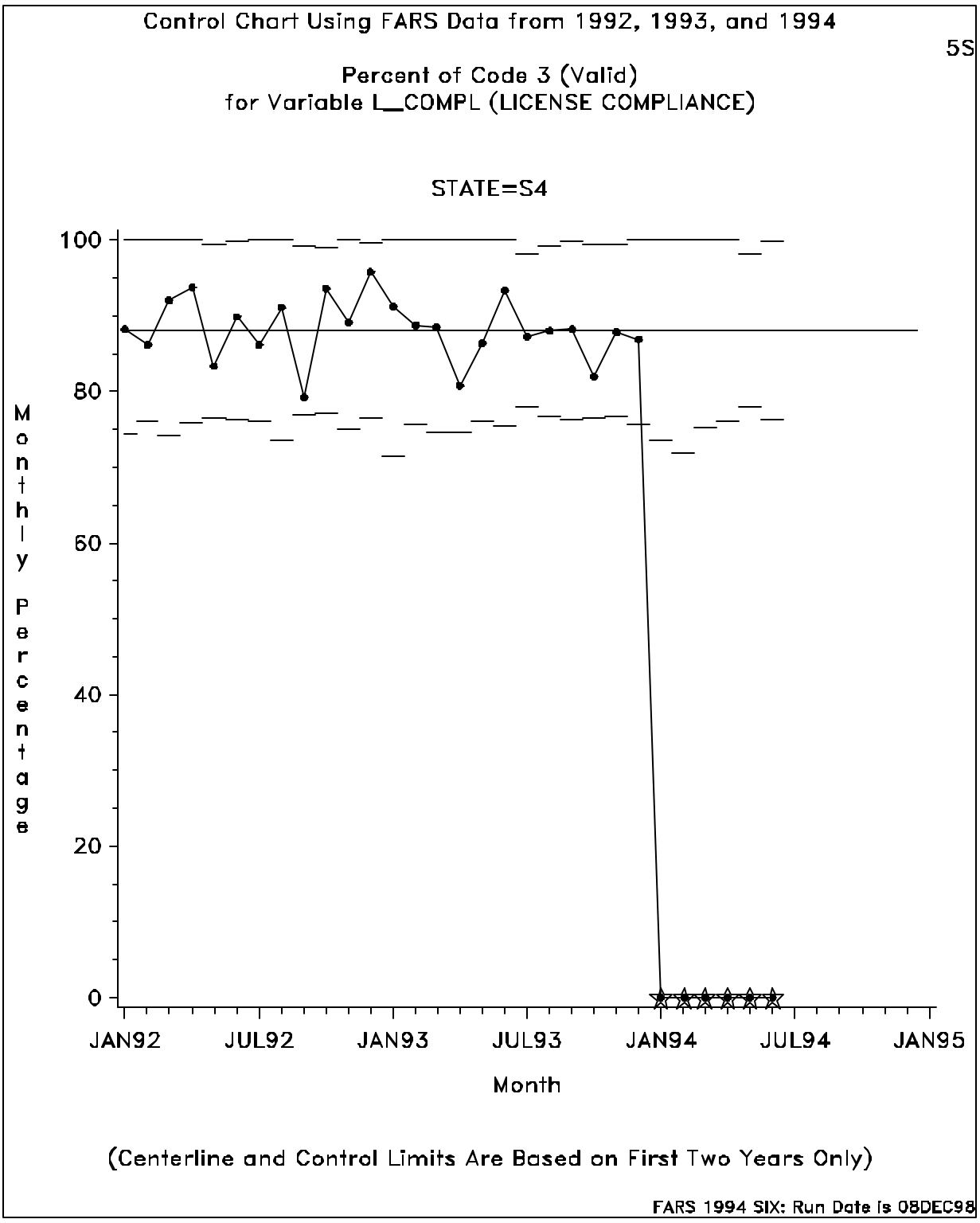


Figure 4. A Chart Out-of-Control Due To a Programming Error.

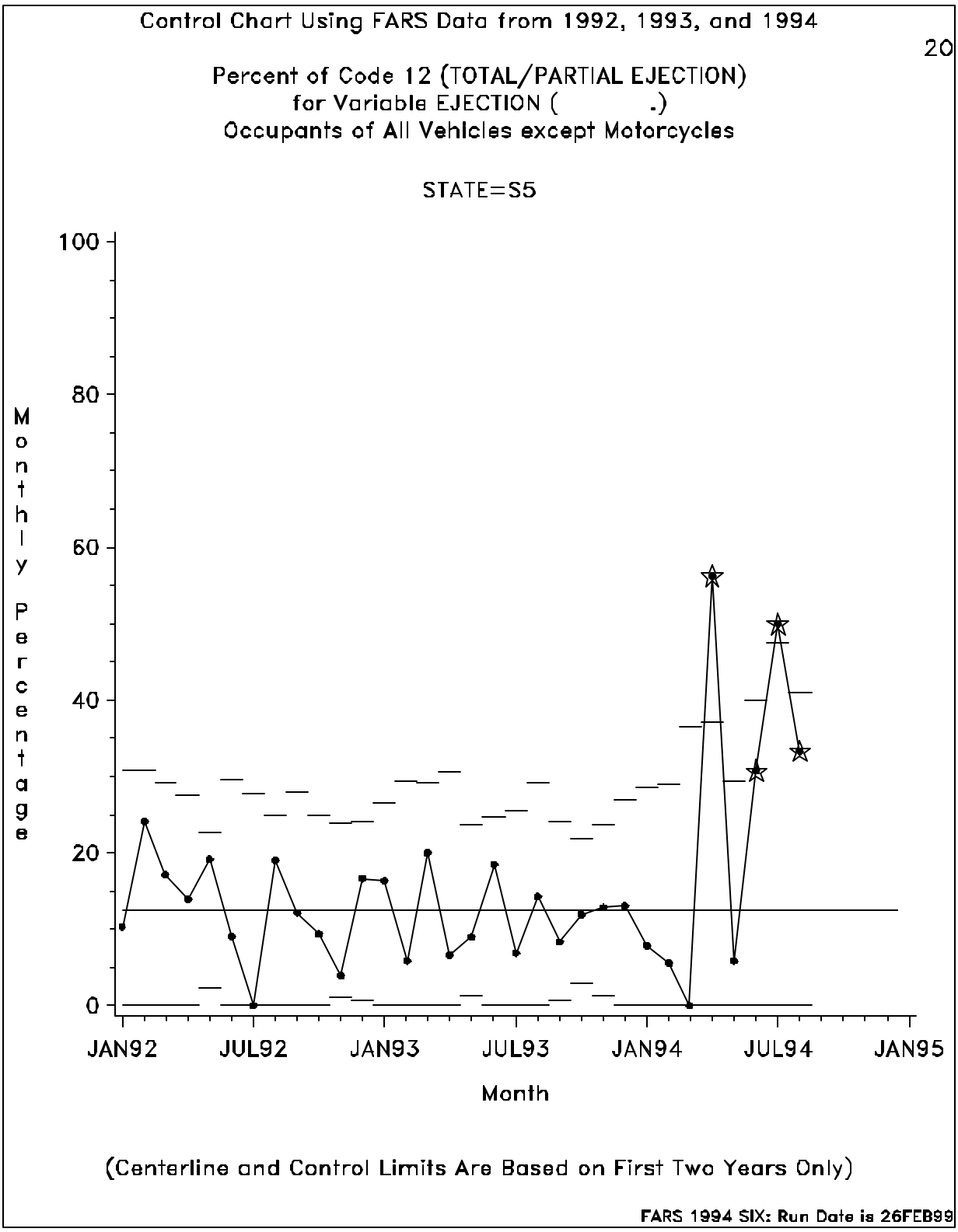


Figure 5. A Chart with Out-of-Control Points Associated with A New Data Collector.

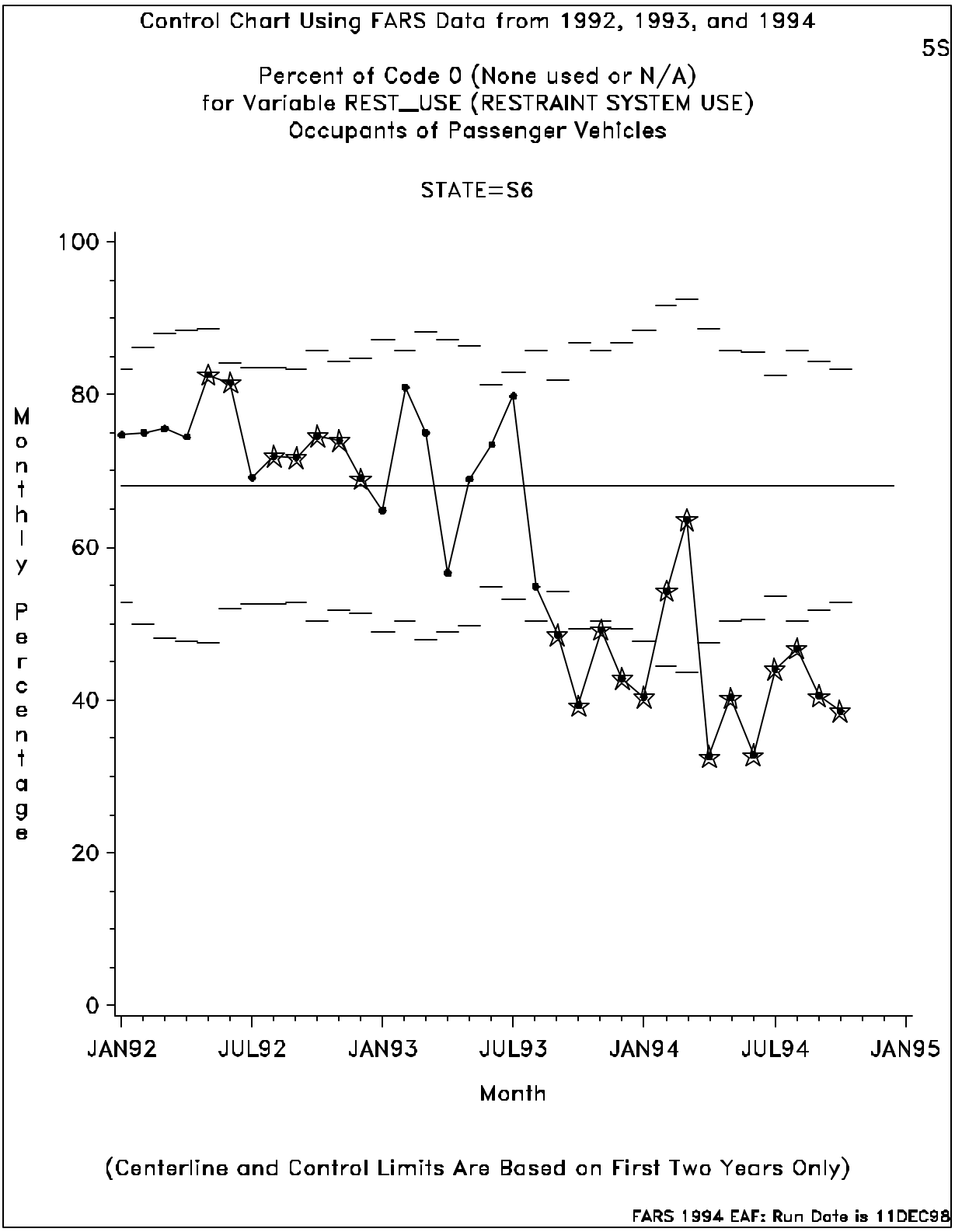


Figure 6. A Chart with Out-of-Control Points Associated with A New Restraint Use Law.

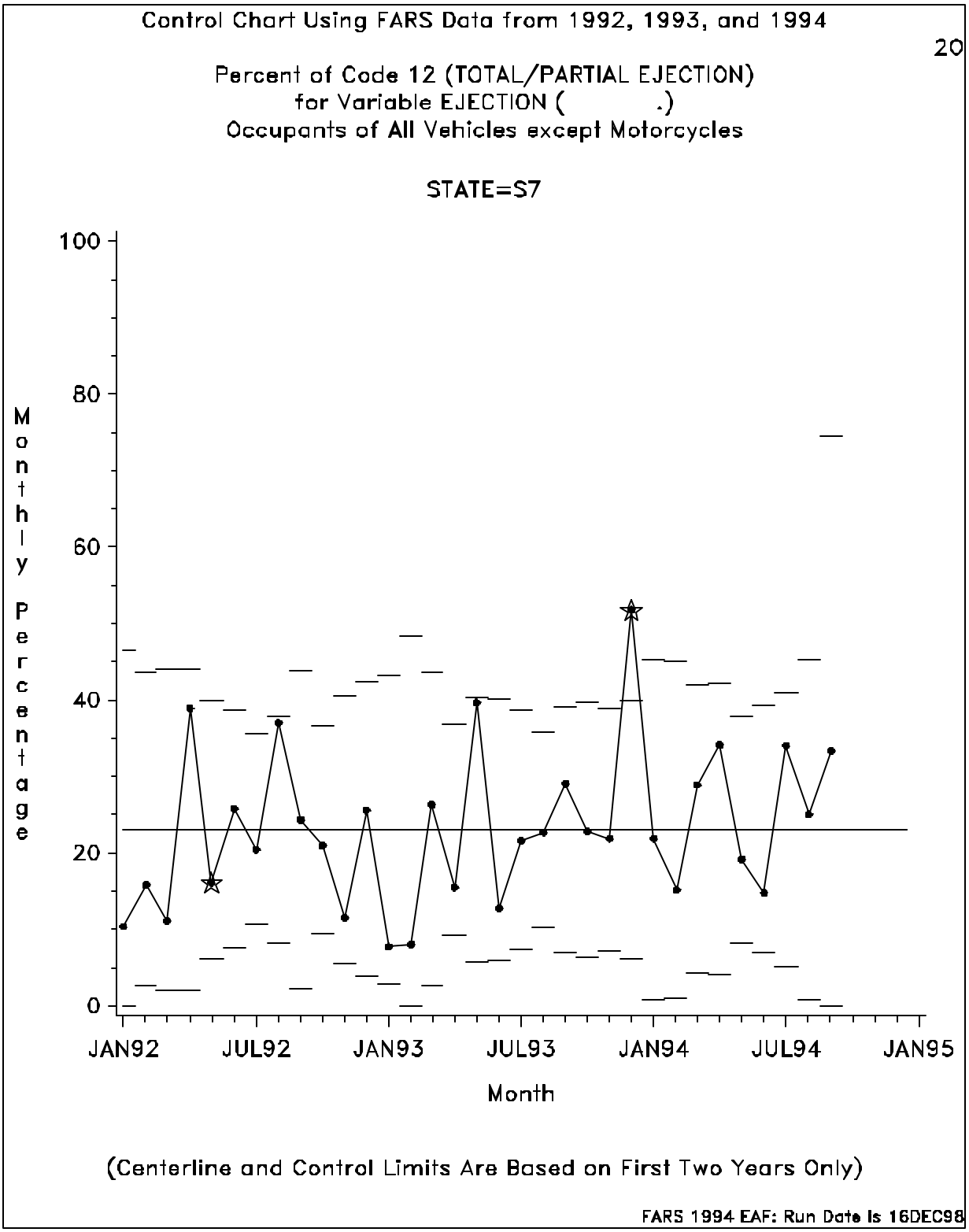
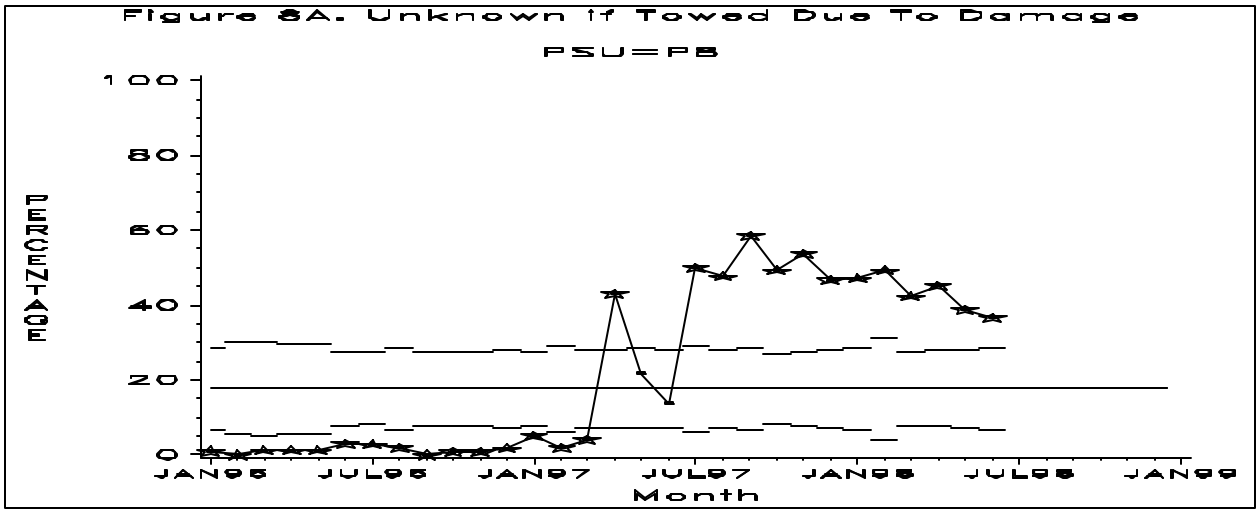
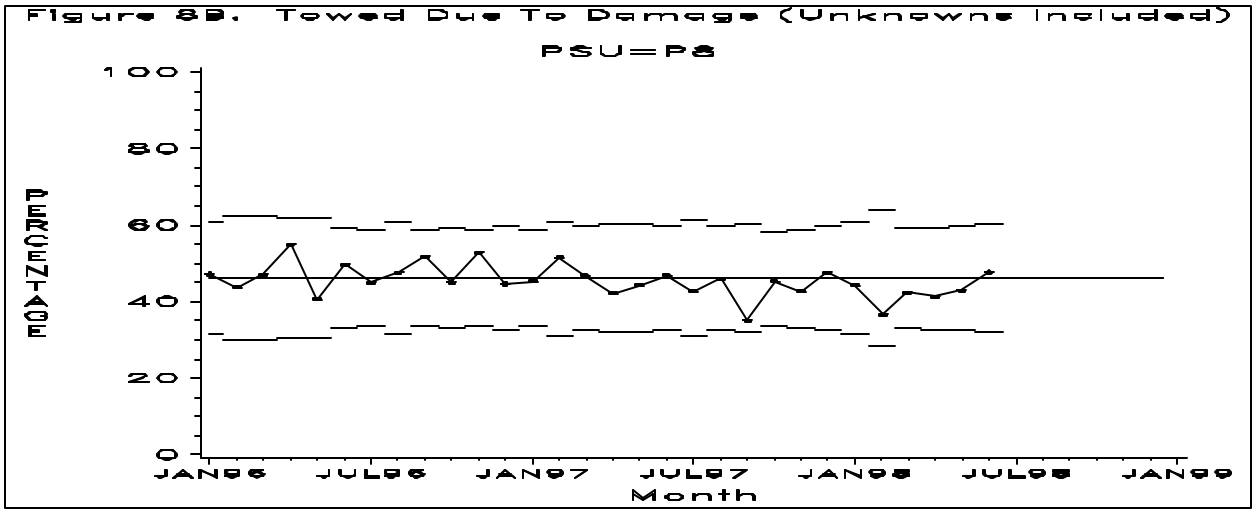


Figure 7. A Chart with an Out-of-Control Point Due to Failure of the Independence Assumption.





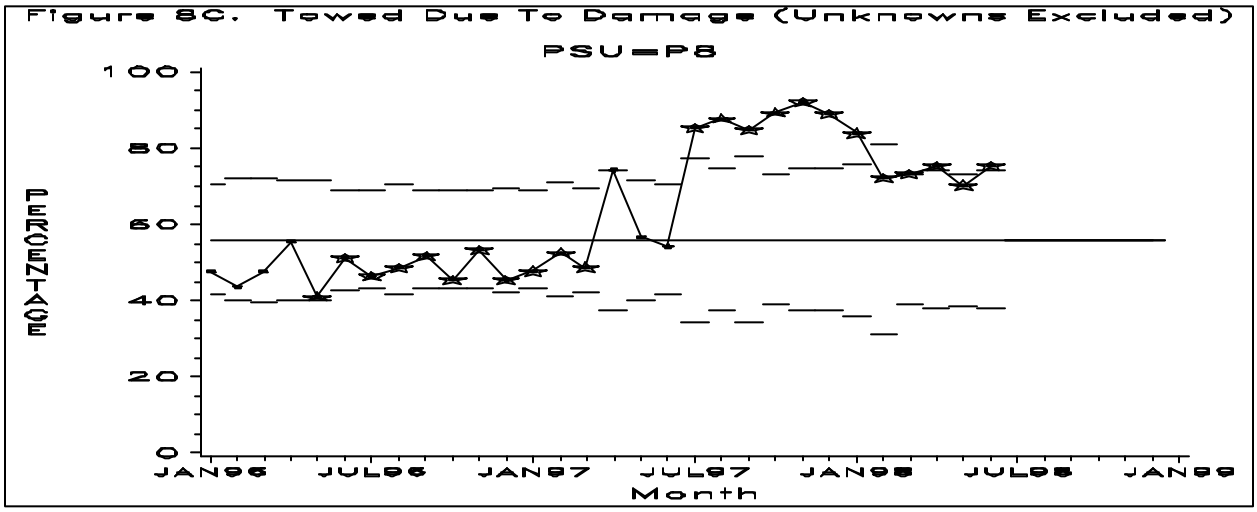


Figure 8. Unknown Values Not Occurring at Random Appear to Explain these Control Charts.