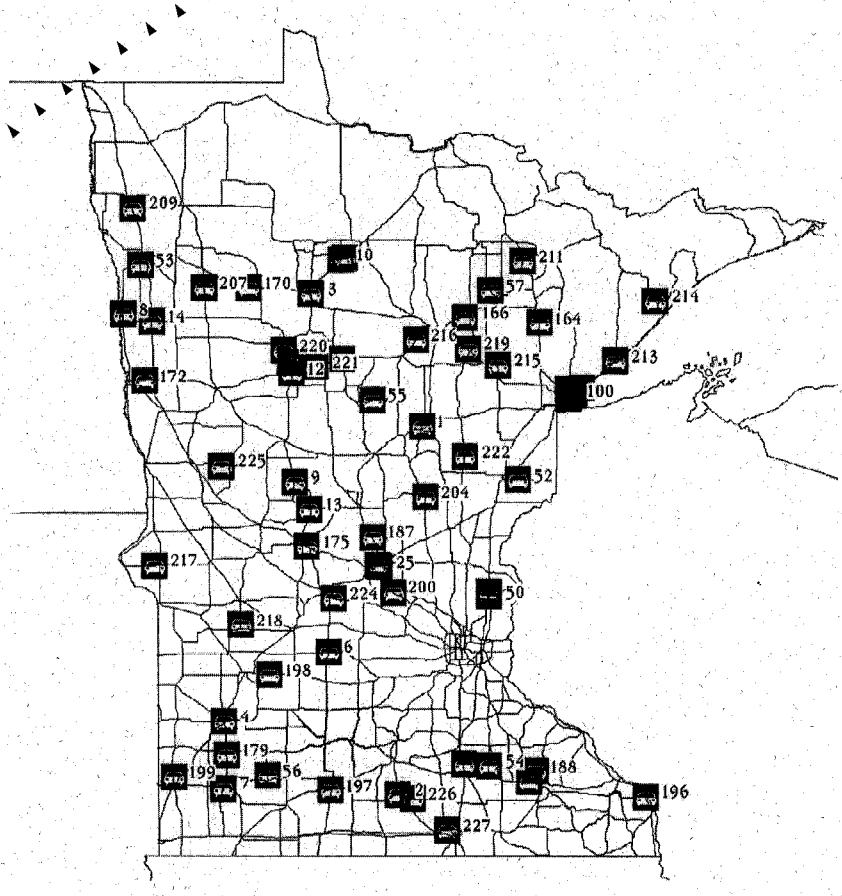
Report Number 97-05

research





# Estimation Theory Approach to Monitoring and Updating Average Daily Traffic

UNIVERSITY OF MINNESOTA
CENTER FOR TRANSPORTATION STUDIES

# Technical Report Documentation Page

| 1. Report No. | 2. | 3. Recipient's Accession No. |
|---|---|---|
| MN/RC - 97/05 | | |

| 4. Title and Subtitle | 5. Report Date |
|---|---|
| ESTIMATION THEORY APPROACH TO MONITORING AND UPDATING AVERAGE DAILY TRAFFIC | January 1997 |
| | 6. |

| 7. Author(s) | 8. Performing Organization Report No. |
|---|---|
| Gary A. Davis | |

| 9. Performing Organization Name and Address | 10. Project/Task/Work Unit No. |
|---|---|
| University of Minnesota<br>Department of Civil Engineering<br>122 Civil Engineering Building<br>500 Pillsbury Dr. S.E.<br>Minneapolis, MN 55455-0220 | 11. Contract (C) or Grant (G) No.<br>(C) 70488 TOC #107 |

| 12. Sponsoring Organization Name and Address | 13. Type of Report and Period Covered |
|---|---|
| Minnesota Department of Transportation<br>395 John Ireland Boulevard Mail Stop 330<br>St. Paul, Minnesota 55155 | Final Report 1994-1997 |
| | 14. Sponsoring Agency Code |

15. Supplementary Notes

16. Abstract (Limit: 200 words)

This report describes the application of Bayesian statistical methods to several related problems arising in the estimation of mean daily traffic for roadway locations lacking permanent automatic traffic recorders. A lognormal regression model is fit to daily count data obtained from automatic traffic recorders, and this model is then used to develop (1) a heuristic algorithm for developing traffic sampling plans which minimize the likelihood of assigning a site to an incorrect factor group, (2) an empirical Bayes method for assigning a short-count site to a factor group using the information in a sample of traffic counts, and (3) an empirical Bayes estimator of mean daily traffic which allows for uncertainty concerning the appropriate factors to be used in adjusting a sample count. An evaluation of these methods confirmed results reported in other work, in which a sample consisting of two, 1-week counts was found to be adequate for overcoming prior uncertainty concerning the correct adjustments for a site. The empirical Bayes method produced sample-based estimates of mean daily traffic that on the average differed by 5%-6% from estimates based on daily counts for an entire year. The report concludes with suggestions for agencies wishing to implement these methods.

| 17. Document Analysis/Descriptors | 18. Availability Statement |
|---|---|
| Traffic Counts<br>Average Daily Traffic<br>Empirical Bayes | No restrictions. Document available from:<br>National Technical Information Services,<br>Springfield, Virginia 22161 |

| 19. Security Class (this report) | 20. Security Class (this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| Unclassified | Unclassified | 90 | |

# ESTIMATION THEORY APPROACH TO MONITORING AND UPDATING AVERAGE DAILY TRAFFIC

**Final Report**

Prepared by

Gary A. Davis, Ph.D.

Department of Civil Engineering
University of Minnesota

**January, 1997**

Published by

Minnesota Department of Transportation
Office of Research Administration
200 Ford Building Mail Stop 330
117 University Avenue
St. Paul, Minnesota 55155

## ACKNOWLEDGEMENTS

The author would like to thank the Office of Transportation Data Analysis of the Minnesota Department of Transportation for support of this research, and particularly Mark Flinner, who provided the traffic data used here. The author would also like to thank Yuzhe Guan, who prepared the highway network data base, developed the traffic volume data files, and conducted the initial classification probability analyses.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# EXECUTIVE SUMMARY

This report considers some of the statistical issues arising when one attempts to estimate an annual mean daily traffic (MDT) from a sample of daily traffic volumes. Two sources of estimation error are identified, one caused primarily by the inherent random variability in daily traffic volumes, the other due to bias in the estimation procedure. A review of early research indicates that traditional estimates of the likely error arising from using short counts to estimate MDT were computed under experimental conditions which minimized or eliminated the effect of bias. On the other hand, a computational example shows that bias caused by the use of incorrect monthly and day-of-week correction factors could be the dominant source of error. A review of more recent research leads to the conclusion that if bias due to incorrect seasonal adjustment could be eliminated, a 48-hour short count could give a reasonable estimate of MDT, but that the site's monthly and day-of-week adjustment factors must be identified with an accuracy equivalent to that provided by two 1-week counts in order to eliminate this bias.

Using automatic traffic recorder (ATR) data supplied by the Minnesota Department of Transportation (Mn/DOT), it is then shown that a class of lognormal regression models provides defensible statitistical models for the daily traffic counts obtained from 50 ATRs in outstate Minnesota. These regression models also provide a method for estimating the monthly and day-of-week adjustment terms which characterize a factor group, and the estimates of these adjustment terms were computed for three factor groups used by Mn/DOT in 1992. Bayesian decision methods are then used to derive formulas for computing the probability a count sample obtained from some highway site is consistent with the monthly and day-of-week pattern characterizing a factor group. This analysis is then used to design counting samples which minimize the likelihood of misclassifying a site. Finally, Bayesian methods are used to derive an estimator of MDT which used the factor group assignment probabilities developed earlier, and had the property that on average its estimates should be closer to the true MDT than those produced by other procedures. Both the classification and estimation methods were evaluated using actual ATR data.

The following conclusions can be drawn from this research. First, the class of lognormal regression models provides an accurate yet tractable description of the statistical properties of bi-directional daily traffic counts, at least for outstate Minnesota. This is an especially important

development since possession of such a model allows an analyst to apply a wide range of sophisticated statistical methods to solve problems in traffic estimation.

Second, assuming that the three factor groups used in this study are representative of all outstate highway sites, then about two weeks worth of well-selected daily traffic counts from a site should be sufficient to match that site to a factor group, and to compute a reasonably accurate estimate of the site's mean daily traffic. "Well-selected" in this case means that the count sample focuses on those days for which the different factor groups give the most discrepant predictions of traffic volumes.

Third, if one can characterize each factor group of interest with a set of monthly and day-of-week adjustment terms, well-selected count samples can be easily determined using a heuristic algorithm. Although the "optimal" sample design found by the heuristic tends to be difficult to implement because it scattered the sampling days throughout the year, a simpler design consisting of two 1-week counts, selected to roughly correspond to the "optimal" design, performed as well as the optimal design.

# CHAPTER 1
# INTRODUCTION

Reliable estimates of traffic demand are essential for a wide range of design and operations activities. The AASHTO *Guidelines for Traffic Data Programs* (AASHTO, 1992, p. 63) identifies 22 different transportation engineering activities which require estimates of the expected traffic volume on a roadway segment during a typical day, so ideally a jurisdiction would have automatic traffic recorders (ATRs) on all important roadway segments. This expected typical volume could then be estimated by the annual average daily traffic (AADT), an average of a year's worth of daily traffic counts. In reality though, it is financially unfeasible to maintain ATRs on all roadway segments in a jurisdiction, and this means that the typical daily expected traffic must be estimated using temporary traffic counters. Since the demand for traffic counts often outstrips an agency's counting resources, there is a continuing interest in finding ways to use these resources most effectively.

## PROJECT HISTORY

As originally proposed, this project intended to exploit the cross-correlation between traffic volumes at different roadway sites in order first, to estimate expected daily traffic at a site using only existing traffic counts from other sites and second, to develop an optimal update of this estimate if and when new traffic counts became available. The key to this endeavor was quantification of the correlation between not necessarily adjacent highway sites, and the original proposal planned to investigate two different methods for estimating these correlations. For the Twin Cities region, a link-to-link correlation matrix was to be computed by first estimating an origin-destination matrix, then assigning the trips of the origin-destination matrix to routes, and finally computing the resulting correlation matrix from the trip assignments using the method given in Sheffi (1985). The origin-destination matrix was to be estimated from ATR traffic counts using the method described in Davis (1993). For the outstate area, a geostatistical technique based on modeling the correlation between two sites as a function of the distance separating the sites was to be used. Once this correlation versus distance relation was specified, an estimate of average daily traffic volumes at the uncounted sites could be computed as a weighted average of daily traffic at ATR sites, using a procedure known as

Kriging. Unfortunately, neither of these approaches turned out to be practically useful. For the outstate region, the TRANSCAD GIS software (Caliper, 1990) was used to develop a network model of the outstate highway system, with ATR sites located as nodes on this network. TRANSCAD procedures were then used to compute a matrix of distances between any pair of ATRs. ATR data for 1991 and 1992 were obtained from the Minnesota Department of Transportation (Mn/DOT), and a geostatistical software package was used to estimate the correlation between the AADT at two ATRs as a function of the distance separating them. It turned out that sites greater than about 50 miles apart were essentially uncorrelated, while the large difference in traffic volumes between for example, a freeway and its adjacent frontage road, implied that some of the Kriging method's underlying assumptions were violated (Guan, 1995). For the Twin Cities region, a network model of the area's road system was obtained from the Metropolitan Council, containing some 1200 nodes and several thousand links. Repeated attempts were made to abstract from this network a subnetwork small enough so that the available ATR data could conceivably be used to estimate an origin-destination matrix, but this was not successful. In addition, although the method for estimating origin-destination matrices worked well on very small example networks, it was not possible to obtain reasonable results in a reasonable amount of time for larger, more realistic networks. This was true even after the original approach of estimating each origin-destination element separately was replaced by a method which used a gravity model to give the origin-destination matrix (Hua, 1994; Hultgren, 1996).

As a result of these efforts, the work plan for this project was modified by adding an additional set of tasks, whose aim was to investigate methods for designing short-counts in order to match a site to one of an existing set of factor groups, and to develop a method for estimating mean daily traffic when it was unclear as to what correction factors were appropriate. Results related to the project's earlier efforts are documented in the Master's theses by Hua (1994), Guan (1995) and Hultgren (1996), so this report will focus on the results for the latter group of tasks. The remainder of this chapter reviews relevant literature on traffic counting methodology, while Chapter 2 describes development and testing of a statistical model for daily traffic counts. Chapter 3 first describes a Bayesian method of matching a site to a factor group given a short traffic count, and then develops a heuristic for finding short count sampling plans which minimize the probability of mis-

classifying the site. Chapter 4 describes a Bayes estimator of mean daily traffic, and compares its performance to more traditional estimators. Finally, Chapter 5 presents conclusions and suggestions for implementation.

## STATISTICAL ISSUES IN TRAFFIC ESTIMATION

Before proceeding, it will be useful to distinguish a model's parameters from methods for estimating these parameters ( known as estimators), and from the outcomes of estimating methods applied to given data samples (known as estimates). The term "mean daily traffic" (MDT) will be used to refer to the expected daily traffic volume on a typical day, where "typical day" means a day for which all seasonal, daily and other correction factors equal 1.0. Thus MDT is a (generally unknown) parameter characterizing the distribution of daily traffic volumes at a given location, while the terms "average daily traffic" (ADT) and "annual average daily traffic" (AADT), as normally defined, can be viewed as estimators of MDT, the latter producing estimates based on an entire year of daily counts. This terminology corresponds with the standard usage in statistics, where "mean" denotes an expected value, while "average" is a quantity computed from a sample in order to estimate the mean.

The basic issues arising in estimating MDT from short counts can be illustrated by considering estimation using a count from a single day. Let z denote this count, which will vary from day-to-day both systematically, due to predictable seasonal and day-of-week trends, and randomly, due to the unpredictable actions of individual trip makers. The effect of predictable trends on daily mean traffic is usually described using a multiplicative model, that is, if the count z is obtained during month number i, i=1,..,12, and on day of week j, j=1,..,7, then the expected value of z can be represented via

$$E[z] = M_i W_j z_0 \qquad (1.1)$$

where

E[.] denotes the operation of computing the expected value (that is, the theoretical mean) of the random variable appearing inside the square brackets,

$z_0$ = mean daily traffic,

$M_i$ = adjustment for month i,

$W_j$ = adjustment for day-of-week j.

In statistics, an estimator of some parameter is said to be unbiased if the expected value of that estimator equals the parameter's true value. Equation (1.1) above tells us that a single day's count will be an unbiased estimator of the MDT $z_0$ only if the product $M_iW_j$ of the monthly and day-of-week terms equals 1.0. On the other hand, if the monthly and day-of-week adjustments are known in advance, it is straightforward to verify that dividing the traffic count by these terms produces an unbiased estimator of $z_0$. That is,

$$E\left[\frac{z}{M_iW_j}\right] = z_0 \qquad (1.2)$$

Similar properties are shown by samples with more than one count, and these sample properties have led to a traffic counting methodology which is fairly uniform across the United States and Canada, where a jurisdiction maintains a relatively small number of ATRs, which are grouped into clusters or "factor groups" according to commonalities in their monthly and/or day-of-week variation patterns. Each highway site for which an estimate of MDT is needed is then assigned to one of these factor groups, and the estimates of the monthly and daily adjustment terms characterizing that group are used to correct count sample estimates from that site. Most temporary counts are short "coverage" counts, of 1-3 days duration, usually made on weekdays.

The importance of using unbiased estimators of MDT arises from the desire to produce estimates that tend to be close to the unknown MDT, $z_0$. If we let z denote some estimator of $z_0$, a commonly used measure of the difference between an estimate and the true MDT is the percent error (PE), defined as

$$PE = \frac{\hat{z}-z_0}{z_0}\times 100 \qquad (1.3)$$

Since z is a random outcome depending on a sample of traffic counts, some samples will yield small

values for PE , while other samples will yield large values. A measure of the theoretical tendency of the estimator $\hat{z}$ to produce estimates close to $z_0$ is the root-mean squared percent error (RMSPE) defined as

$$RMSPE = \sqrt{E[PE^2]} \tag{1.4}$$

An RMSPE=10% can be interpreted as telling us that on the average, the estimator $\hat{z}$ produces estimates that differ from $z_0$ by 10% of $z_0$'s value. Now letting $\bar{z}=E[\hat{z}]$ denote the expected value of the estimator $\hat{z}$, it can be shown the RMSPE of $\hat{z}$ is given by

$$RMSPE = \frac{\sqrt{E[(\hat{z}-\bar{z})^2] + (\bar{z}-z_0)^2}}{z_0} \times 100 \tag{1.5}$$

The first term under the radical sign in (1.5) is the variance of the estimator $\hat{z}$ (sometimes referred to as sampling error), while the second term under the radical sign is the square of the bias of $\hat{z}$, so that sampling error and bias each contribute to the estimator's mean distance from $z_0$. The variance of $\hat{z}$ depends on the the random day-to-day variability in the traffic counts, and usually decreases as the sample size increases. Bias on the other hand is usually unrelated to sample size, but as illustrated above, can result from an incorrect accounting of seasonal or day-of-week trends. Equation (1.5) has two important practical consequences. First, if the estimator $\hat{z}$ is biased, then increasing the sample size and thus reducing the sampling error may lead to the RMSPE being dominated by the bias term. At this point further increases in sample size will produce only marginal improvements in estimator accuracy. Second, if $\hat{z}$ is unbiased, so that the bias term in (1.5) vanishes, then RMSPE is simply the coefficient of variation for the estimator $\hat{z}$, or equivalently, the standard deviation of PE. If PE is normally distributed (as has been argued in the literature), then tables of the normal distribution can be used to compute precision intervals for $z_0$. If $\hat{z}$ is biased however, this convenient technique is not available.

It will be useful to divide the following literature review into two sections, roughly corresponding to the two components of RMSPE appearing in equation (1.5). Thus the first, and

most extensive, set of studies can be viewed as attempts to estimate RMSPE when bias due to uncertainty concerning seasonal or day-of-week trends can be neglected. The second set of studies treats problems arising in assigning short count sites to factor groups, and include attempts to estimate the increase in RMSPE caused by poor assignment. The review paper by Albright (1991a) provided an introduction to much of the early literature, especially the work of Boris Petroff.

## ESTIMATION OF SAMPLING ERROR

Reliable portable automatic traffic counters became available for use by traffic agencies in the late 1930's, and one can follow the development of current traffic estimation methods in a series of papers written by analysts from the Bureau of Public Roads (BPR). In 1946, Boris Petroff described how estimates of RMSPE could be obtained for different sample sizes if one had available continuous counts of the type provided by ATRs (Petroff, 1946). Petroff first obtained year-long daily traffic counts from 20 low-volume (*i.e.*, expected traffic volumes less than 200 vehicles/day) rural sites located in eight different states (including Minnesota). From these count records weekday samples of 24, 48, and 72 hours duration were generated, with only the weekdays Monday through Thursday being included. The daily average computed from these samples was then compared to an average weekday traffic (AWDT) computed for that station, for each month of the year. This means that the $\hat{z}$ appearing in equation (1.2) was the sample average, while $z_0$ was replaced by an average computed from all weekdays in a given month. For 48-hour samples Petroff reported estimates of RMSPE ranging from 9.38% to 25.58%, depending on the month and whether or not the average was computed for Northern or Southern states. To interpret these results, note that since each short count average was compared to a monthly average weekday traffic, bias due to seasonal variation should not be present in the computations of RMSPE. Also, day-of-week factors for Monday through Thursday tend to fairly similar, so bias due to day-of-week effects is most likely minimized, implying that Petroff's RMSPE estimates should be dominated by the sampling error. Thus these results can be reasonably interpreted as giving likely ranges of RMSPE if one has found a way to eliminate bias due to seasonal and day-of-week variation, on low-volume rural roads.

A later paper by Petroff and Blensly (1954) was primarily devoted to describing a traffic counting program implemented in Oregon, but Figure 1 of this paper (Petroff and Blensly, 1954, p.

364) summarizes a combination of the results from Petroff (1946) together with results obtained by analysts working for the state of Ohio. This figure shows expected RMSPE for estimating AWDT as a function of AWDT and duration of count. In particular, for a 48-hour count on a road where the AWDT is greater than 500 vehicles/day, this figure predicts an RMSPE of about 6% or less. Assuming that the computational procedure followed by the Ohio analysts was similar to that used in Petroff (1946), these estimates should again be interpreted as showing what one might expect if bias due to seasonal and day-of-week effects can be eliminated. Petroff and Blensly's Figure 1 turns out to have a long life, appearing most recently on page 3-3-4 of FHWA's *Traffic Monitoring Guide* (FHWA, 1985).

In Petroff (1956) one finds an early description of the current methodology for estimating MDT. In particular Petroff argues that ATR stations should be clustered into seasonal factor groups according to similarity of the ratios between monthly AWDT and AADT for the station, and he describes a clustering procedure. He also computes estimates of RMSPE using ATR data where now $\hat{z}$ in equation (1.2) is an estimate of MDT and $z_0$ is replaced by the AADT. In computing these estimates, the ATRs were first clustered into factor groups, and average adjustment factors were computed for each factor group. Short count samples where then generated from the ATR data, and the short count averages from each ATR were expanded using the seasonal factors for that ATR's factor group. Thus these RMSPE estimates contain a component arising from using an average adjustment factor to correct the short count, but can be interpreted as what one would get if one had a very good prior idea as to which factor group a site belongs. For 48-hour counts, the estimates of RMSPE ranged between 7.2% and 12.1%, depending on the state, and for 7-day counts the range of RMSPE was 6.1% to 12.3%. The lowest estimated RMSPE of 1.9% was obtained when 7-day counts where repeated every month throughout the year.

Bodle (1967) conducted a replication and extension of Petroff's work. Bodle identified three sources error in estimating MDT, namely (i) sampling error, (ii) error due to the actual seasonal factor for the coverage count site not being identical to the average factor of the assigned factor group, even if the assignment is good , and (iii) error due to incorrect factor group assignment. Bodle assumed that error from the third source is negligible, and like Petroff was primarily interested in estimating the magnitude of the sampling error component. Bodle also used year-long records from

ATRs in several states, but unlike Petroff, he had access to mainframe computing equipment. Thus Bodle's estimates of RMSPE were based on computing the PE from all possible samples of a given size and type, but otherwise the methodology employed was similar to that described in Petroff (1946). For 48-hour counts, Bodle found RMSPEs for average monthly weekday traffic ranging from 6.1% to 19.4%, depending on month and state, with an overall average of 9.5%. For 7-day counts, the overall average RMSPE was 5.5%, where these averages resulted from aggregating across AWDT levels. Bodle concluded that although Petroff had argued that a 24-hour count was sufficient to obtain a sampling error RMSPE of 10% or less when AWDT was greater than 500 vehicles/day, his own results indicated that a 48-hour count would be needed for this degree of precision.

This body of work appears to have provided the basis for the traffic volume counting practice recommended by the Federal Highway Administration (FHWA, 1985), but in the mid-1980's researchers began showing a renewed interest in counting methodology, motivated in part by the counting requirements of the Highway Performance Monitoring System and the Strategic Highway Research Program. Much of this work was concerned with methods for clustering factor groups and determining and determining the precision of estimated correction factors (Ritchie, 1986; Albright, 1987; Memmott, 1987; Faghri et al., 1992) but in 1991 Albright challenged the usefulness of the early research by Petroff and Bodle (Albright, 1991a). The basis of this challenge was the observation that many states replace apparently bad ATR counts with estimated, or imputed, values and that the inclusion of imputed data can lead to underestimation of RMSPE. To see this, simply imagine the extreme case where the count for each day at an ATR has been replaced by a common imputed value. All short count averages taken from this record will be exactly equal to the AADT, and the corresponding PE will always be zero. Since the ATR data used in Bodle's and Petroff's work could have contained imputed values, the RMSPE estimates are suspect. In other papers, Albright also reported results from estimation experiments conducted in New Mexico, where only "edit-accepted" ATR data were used (Albright, 1991b, 1991c). Rather than compute RMSPE, Alright generated nonparametric estimates of the width of a 90% precision interval, and for rural highways in New Mexico, this 90% precision interval for 48-hour count estimates of AWDT was ± 21%. That is, in 90% of the samples, an estimate of AWDT computed from a 48-hour short count was within 21% of the estimate computed using all good weekdays in the year. To compare this to Petroff's work,

Petroff (1956) reported an RMSPE of 9.5% for 48-hour estimates of ADT, using New Mexico data. Assuming that PE is normally distributed, multiplying the RMSPE by 1.645 gives a 90% precision interval of ±15.6%. Although it is tempting to attribute this difference in RMSPE to the use of imputed data, there are other differences in Albright's and Petroff's studies which could also explain the gap. In particular, Albright's results are for simple averages for which no seasonal correction has been applied, so it can be plausibly argued that Albright's ±21% spread contains a larger component of bias error than does Petroff's ±15.6%.

In summary, the above cited research appears to provide a justification for the *Traffic Monitoring Guide's* recommendation of 48-hour count intervals. If data errors and bias can be ignored, one can expect an RMSPE for a 48-hour estimate of AWDT of 10% or less, but Albright's point concerning the use of imputed data is well taken, and in practice perfect matching of a site to a factor group should not be expected. So at present, the estimates of RMSPE described above must be taken as lower bounds rather than realistic expectations, while determining the tightness of these bounds requires assessing the magnitude of data and factor group assignment errors.

## RESEARCH ON FACTOR GROUP ASSIGNMENT

As noted earlier, much of the research on the accuracy of short count estimates assumes that errors due to assigning a site to the wrong factor group can be neglected. Before reviewing research on factor group assignment, it will be informative to consider an example of the magnitude of assignment error. In Chapter 2 of this report it is argued that daily traffic volumes can be modeled as lognormal random variables, and for Mn/DOT's ATR #51, fitting the lognormal model using 1992 ATR data yielded estimates of the mean of the logarithms of the traffic counts, u= 6.2960, and the standard deviation of the log counts, $\sigma$=0.19215. Taking these estimates as true values for a hypothetical highway site, the MDT for this hypothetical site would be $z_0=\exp(u+\sigma^2/2)=552.5$. Now imagine that we wish to estimate this MDT using a single 24-hour count made on a Saturday in July, denoted as z. ATR #51 was assigned to factor group 4 in 1992, and the estimates of the monthly and day-of-week terms for July and Saturday, for factor group 4, were $M_7$=1.6516, $W_7$=1.0916. Taking these as the true values for the hypothetical site it follows that z will be a lognormal random variable with a mean equal to

$E[z]=z_0 M_7 W_7=996.1,$

and variance

$V[z]=E[z]^2 (\exp(\sigma^2)-1)=37,307.3.$

Using the factored estimator (1.2) and assuming correct assignment of the hypothetical site to factor group 4 gives the expected value of the estimator as

$$\bar{z} = E\left[\frac{z}{(1.6516)(1.0916)}\right] = 552.5$$

and the variance of estimator as

$$V[\hat{z}] = E[\hat{z}-\bar{z}]^2 = 11,477.8$$

Plugging these values into equation (1.5) the gives an RMSPE=19.4%. But suppose we incorrectly assign the hypothetical site to factor group 1. The corresponding monthly and day-of-week factors for group 1 are $M_7$=1.2158, $W_7$=0.9733. The mean and variance of the estimator of MDT will now be

$$\bar{z} = E\left[\frac{z}{(1.2158)(0.9733)}\right] = 841.8$$

$$V[\hat{z}] = E[\hat{z}-\bar{z}]^2 = 26,642.9$$

and plugging these values into equation (1.5) gives RMSPE=60.1%. Note that even if the sampling error component were reduced to zero, the bias component alone would still give RMSPE=52.4%.

For this example, incorrect assignment of the hypothetical site led to a tripling of the RMSPE, suggesting that in the absence of evidence to the contrary, factor group assignment errors should be viewed as being at least of comparable magnitude to other error sources, if not as the dominant source of error. How are non-ATR sites assigned to factor groups? In the past, selection of

correction factors was often based on physical proximity of the short count site to an ATR, tempered by engineering judgement (Ritchie, 1986; Albright, 1987). Petroff (1956) described a map-based method in which a highway segment lying between two ATRs in the same factor group was assigned to that factor group, with judgement being used to resolve ambiguous cases. Ritchie (1986) described a combination of functional and geographic-based assignment used in the state of Washington. A problem arises when one has a set of ATRs with a common monthly or daily pattern of variability that cuts across functional or geographic classifications, so that it may be unclear as to whether or not a given non-ATR site belongs to this group. This often happens for ATRs near recreational areas, and a decision as to whether or not a non-ATR site should be assigned to such a group can sometimes be made by considering nearby land use, but the appropriate use of such groups is an unresolved problem in both the *Traffic Monitoring Guide* and the *AASHTO Guidelines*. If a state's construction of factor groups has been entirely data-driven, with no reference to external classifications, this assignment problem is likely to be even more difficult.

An extensive investigation of the factor group assignment problem has been conducted by Satish Sharma, of the University of Regina, and his graduate students. In Sharma and Allipuram (1993), a hierarchical grouping technique (Sharma and Werner, 1981) was applied to data from 61 ATR sites in the province of Alberta to generate seven seasonal factor groups, each group being characterized by 12 ratios of monthly ADT to AADT. The authors then demonstrated how similar factors could be computed from a "seasonal" count of at least one week's duration repeated in two or more different months, and developed a measure of the difference between the monthly factors computed from the seasonal count and the corresponding monthly factors characterizing a factor group. The seasonal count site could then be assigned to the factor group with the closest monthly factors. The authors then investigated the relation between the length and frequency of the seasonal count and the resulting assignment accuracy. They found that for most of the factor groups, a one-week count repeated in four different months was much more accurate than a one-week count repeated twice, but that repetition beyond four different months brought little additional improvement in assignment accuracy. To quantify assignment accuracy, they developed an index of assignment effectiveness (IAE), a percentage which equaled 100% when a seasonal count always assigned the site to the correct factor group.

In Sharma and Leng (1994), ATR data from 52 sites in the state of Minnesota were clustered into four seasonal factor groups, based on monthly variation patterns. The authors then investigated the problem of specifying a seasonal count which would give a PE less than some specified precision, 95% of the time. By noting that the percent errors appeared to be normally distributed (as was also held by Petroff), they argued that the size of a 95% precision interval could be found by estimating the RMSPE generated by a particular seasonal count pattern and then multiplying this by 1.96, a quantity which they called PI95. They found that two, one-week counts generated a PI95 of between 10% and 21%, depending on the factor group. These PI95 values correspond to RMSPE estimates of 5.1% and 10.7%, but recall that equation (1.5) decomposes the square of RMSPE into two components, one due to sampling error and one due to bias. Formula [3] in Sharma and Leng, which was used to estimate their RMSPEs, and ultimately PI95, estimates only the sampling variability component, and ignores the bias component. Thus like Petroff and Bodle's work, these results should be interpreted as what one might hope for when bias due to seasonal variation can be ignored. Sharma and Leng's results are roughly comparable to those in Petroff (1956), which gave an estimated RMSPE for 7-day counts on rural roads, repeated four times during the year, of 4.2%, while the PI95 of 7.5% for four 1-week counts on rural roads, reported in Sharma and Leng's Figure 2, corresponds to an RMSPE of 3.8%. Finally, it is known that the 1991 Minnesota ATR file contains imputed values in place of apparently erroneous recorder measurements, and it is unclear how Sharma and Leng handled these imputed values. As noted earlier, treating them as actual measurements could lead to overestimating the accuracy of MDT estimates.

Finally Gulati (1995), working under Sharma's supervision, conducted a detailed investigation of the relation between the accuracy of assignment of a non-ATR site to a factor group and the resulting accuracy of estimates of MDT. Using additional ATR data from Minnesota, he found that when the assignment effectiveness was greater than 90%, indicating a good assignment of a site to a factor group, the PI95 for a 48-hour count ranged between 10% and 23%, depending on the roadway type. This PI95 range corresponds to RMSPE ranging between 5.1% and 11.7%, and compares favorably with the values reported in Petroff's and Bodle's work. As the assignment effectiveness declined however, the size of the PI95 interval increased markedly (to greater than 70%). Since Sharma and Allipuram's work indicated that two or more one-week counts would be

needed to give an assignment effectiveness greater than 90%, these results together suggest that a 48-hour count will produce a reliable estimate of MDT only when one can assign the count site to factor group with an accuracy equivalent to that achieved by at least two one-week counts.

## SUMMARY OF THE REVIEW

This review suggests that if bias due to failure to account for seasonal or day-of-week effects can be ignored (or corrected), a 48-hour count is likely to give an estimate of MDT with an RMSPE of about 10% when compared to what one would obtain from an entire year of counts. Failure to correct for seasonal bias can lead to a much wider range of uncertainty, but this effect is situation-specific, depending on the actual patterns of seasonal and/or day-of-week variation exhibited by traffic in a state. Unfortunately, it appears that counts substantially longer than 48-72 hours appear necessary for a good factor group assignment. Sharma and Allipuram recommend one-week counts repeated in 6-11 different months, and Gulati's results indicate that at least two one-week counts is the minimum necessary to get an estimate of MDT with an RMSPE of 10%. Thus if a state seeks to operate a purely data-driven counting program, the minimum count duration and frequency may have to be considerably longer than the 48-hour minimum recommended by the *Traffic Monitoring Guide*.

In what follows, this report will provide answers to three questions arising in the practical operation of a data-driven counting program. First, it appears that an appropriate assignment of a short-count site to a seasonal factor group is necessary to obtain good estimates of MDT without unreasonably large samples. How should the counting sample be designed so as to minimize the likelihood of mis-classifying the site? Second, it may happen that a sample count does not provide an unambiguous assignment of the site to a factor group, but that it is still necessary to use the count to estimate MDT. How should this be done? Third, the *AASHTO Guidelines* emphasize "Truth-in-Data", and recommend that the expected precision of estimates of MDT be reported along with the estimates themselves, but little guidance is given on how this could be done. How should the uncertainty attached to an estimate of MDT be characterized? A consistent set of answers to these questions are developed using Bayesian statistical methods. In particular, Bayesian classification methods are used to compute the probability a site belongs to given factor group using a sample of traffic counts from the site. Next, the problems of designing a sample count so as to minimize the

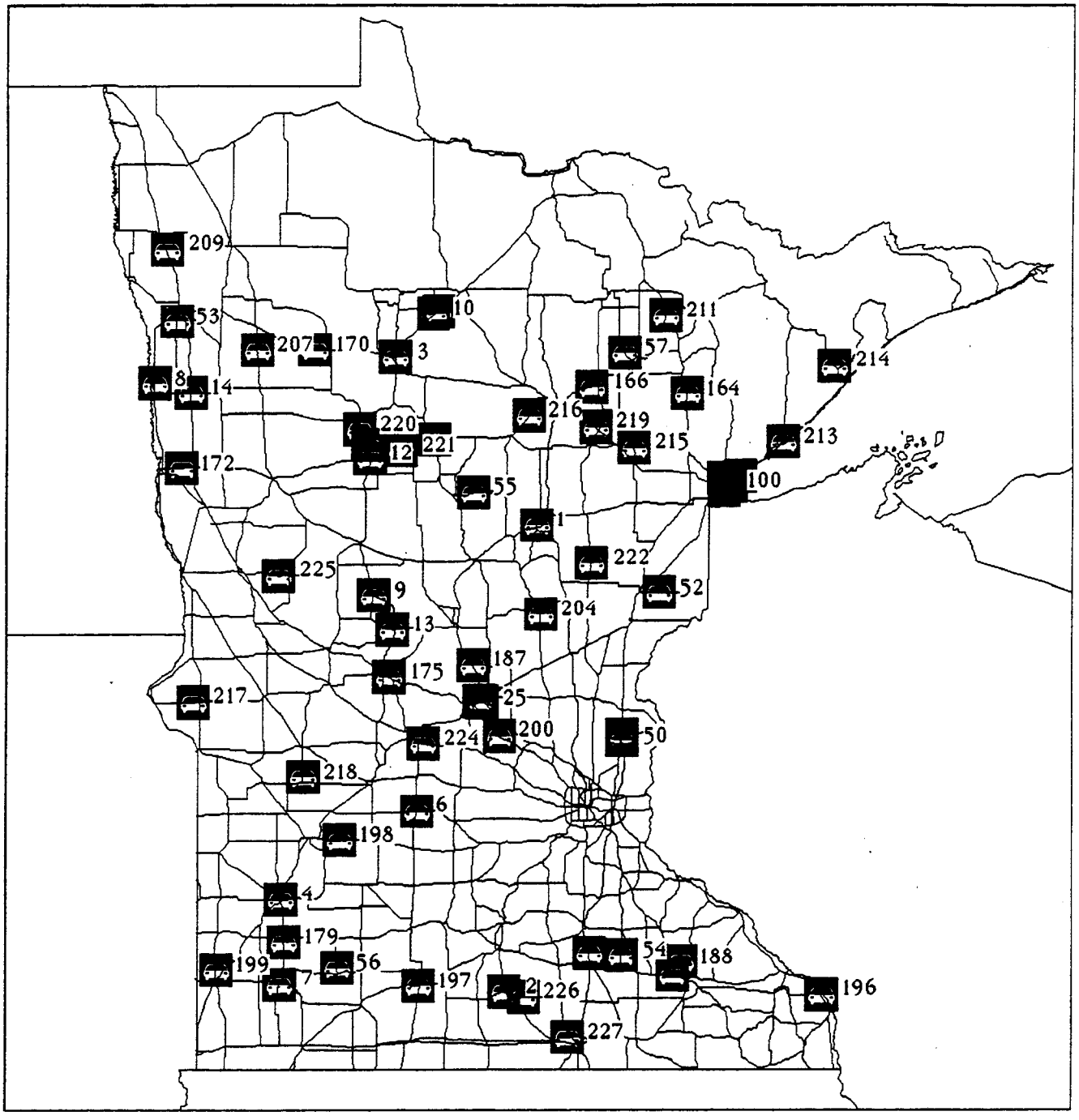
probability of misclassifying the site is taken up, and although an exact solution is difficult to obtain, a usable heuristic solution can be found by solving a particularly tractable special case. Finally, a Bayes estimator of MDT is described which does not require that a site's factor group be known beforehand, along with procedures for computing the posterior standard error of this estimator. These methods are evaluated using ATR data from Minnesota.

# CHAPTER 2
# A STATISTICAL MODEL OF DAILY TRAFFIC COUNTS

The term "statistical model" can be interpreted in two complementary ways. On the one hand, one has a well-defined statistical model of some data set when one is able to generate on a computer simulated data sets that have statistical properties identical to those of the original data. For example, if one's model of vehicle spot-speeds on a road is that these speeds are the outcomes of independent, identically distributed normal random variables, with mean equal to 56 km/h (35 mph), and standard deviation equal to 8 km/h (5 mph), then a normal random number generator on a computer could be used to simulate the effect of repeated sampling of speeds. On the other hand, one has a well-defined statistical model if one can compute the probability (or probability density) of obtaining a particular data sample, given values for the model's parameters. Using the spot speed example, first assuming that the mean spot speed equals 56 km/h (35 mph) and the standard deviation is 8 km/h (5 mph), one could compute the probability of obtaining a given set of speed measurements using the joint product of normal probability density functions evaluated at those speeds. The function which gives the probability of obtaining a set of data, treated as a function of the model parameters is also known as the likelihood function, and in classical statistical inference its specification often allows one to develop optimal statistical procedures. Again using the spot speed example, the normal likelihood function can be used to show that the minimum variance estimate of the mean speed is the sample average, and that the shortest confidence interval having a pre-specified confidence can be constructed using Student's t-distribution (Rohatgi, 1976). Also, since the data likelihood appears in Bayes' Theorem, the likelihood function also plays a fundamental role in Bayesian statistical inference (Berger, 1988).

Given that daily traffic counts provide the primary data for inferring MDT, and given the centrality of likelihood functions in statistical inference, it is somewhat surprising to find that to date no source appears to have determined an appropriate class of likelihood functions for daily traffic counts. The closest one can come in either the *Traffic Monitoring Guide*, the *AASHTO Guidelines* or in research reports is that the clustering and sample-size computations actually performed assume normally distributed data, although the connection between derived data, such as computed monthly

**Figure 2.1. Locations of ATRs in Outstate Minnesota**



factors or percent errors, and the underlying traffic counts is left unspecified. Thus before the three questions listed in the previous chapter can be answered, it is necessary to develop a plausible statistical model and corresponding likelihood function for daily traffic counts.

## MODEL DEVELOPMENT

To support the model development, two files containing ATR traffic counts for the years 1991 and 1992 were obtained from Mn/DOT, along with supplementary information such as the ATR factor groups used by Mn/DOT in 1992 and lists of the days during the year where the daily traffic volumes were effected by holidays. In addition to hourly traffic counts throughout the year, the ATR files also included codes which indicated whether the hourly counts consisted entirely of raw data or if some imputation of the hourly counts had taken place. Figure 2.1 shows the ATR stations located outside the Twin Cities region, and in 1992 Mn/DOT personnel classified 52 of these ATRs into three factor groups, labeled here as Group 1, Group 3 and Group 4. Table 2.1

**Table 2.1. Factor Group Assignment of Outstate ATRs in 1992**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ATRs in Group 1** | | | | | | | | | |
| 2 | 3 | 7 | 8 | 9 | 10 | 12 | 14 | 28 | 50 |
| 54 | 56 | 100 | 102 | 103 | 104 | 110 | 164 | 166 | 170 |
| 172 | 179 | 196 | 197 | 199 | 210 | 211 | 213 | 216 | 217 |
| 219 | 225 | 226 | 227 | | | | | | |
| **ATRs in Group 3** | | | | | | | | | |
| 52 | 175 | 187 | 200 | 204 | 207 | 208 | 215 | 218 | 224 |
| **ATRs in Group 4** | | | | | | | | | |
| 1 | 51 | 55 | 57 | 214 | 220 | 221 | 223 | | |

lists the ATRs which were included in each of these groups. The daily traffic volumes for each day and each ATR in these three groups , throughout the years 1991 and 1992, were then constructed from the hourly counts, and for those ATRs which provided directional counts the two directional volumes were added to produce total daily counts. It turned out that yearly records for some ATRs

were incomplete, and deleting these from the data set left a total of 48 ATRs in the 1991 data set and 50 ATRs in the 1992 data set. Two codes were assigned to each daily count to indicate (1) whether or not that count consisted entirely of unedited raw data, and (2) whether or not that count was subject to holiday effects. Finally, for each ATR, those counts containing imputed data were coded as missing. Because there were more ATRs available in the 1992 data set, this data set was selected for model development and estimation.

A convenient starting place for specifying a daily count model is the common assumption that the effect of seasonal or day-of-week variations on is multiplicative, rather than additive, that is

$$E[z_t] = z_0 M_i W_j \tag{2.1}$$

where,

$z_t$ = traffic volume on day t, during month i and on day of week j,

$z_0$ = mean daily traffic,

$M_i$ = multiplier accounting for deviation of month i from MDT,

$W_j$ = multiplier accounting for deviation of day-of-week j from MDT.

As discussed in Chapter 1, the connection between equation (2.1) and standard practice can be illustrated as follows. Imagine that we have available a single daily traffic count z, taken during month i and on day of week j, and wish to estimate the MDT, $z_0$. If we take z as our estimate, then (2.1) tells us that this estimate will be unbiased only if the product $M_i W_j$ equals 1.0. On the other hand, if the multipliers $M_i$ and $W_j$ are known, it can be verified that the estimator $z/(M_i W_j)$ is unbiased for $z_0$, and the multipliers $M_i$ and $W_j$ appearing in equation (2.1) are the reciprocals of the monthly and daily factors used to adjust short count averages to estimate AADT (e.g. McShane and Roess, 1990, p. 102, Eq. (6.5)).

Equation (2.1) is a loglinear model, since taking the logarithms of both sides of the equation produces a linear relationship between the logarithm of the expected daily traffic, the MDT, and the monthly and daily effects. Probability distributions which support convenient analysis of loglinear models include the Poisson, gamma, negative binomial and the lognormal, and a comparison of the

ability of these models to fit the daily count data was conducted using the generalized linear modeling (GLIM) software (Francis, 1993) on the University of Minnesota's Epix mainframe computer. This preliminary analysis yielded the following conclusions:

(1) the traffic counts tended to show greater day-to-day variability than would be expected if the counts were Poisson distributed,

(2) failure to account for holiday effects decreased the quality of a model's fit,

(3) the lognormal, negative binomial and gamma models tended to produce fits of roughly equal quality, but none of these was satisfactory,

(4) the residuals (the difference between actual counts and the counts predicted by a given model) tended to show serial correlation.

It was felt that result (3) above was at least in part due to GLIM's tendency to fit a common variance term to all the ATRs in a factor group, rather than fitting a separate variance term for each ATR. In addition, finding (4) implies that the count data violated one of GLIM's underlying assumptions, which is that the data be sequentially independent. Analysis of negative binomial and gamma models with correlated data is very difficult, but analysis of lognormal models is much easier, since the powerful methods developed for fitting time-series regression models can be applied to the logarithms of the daily traffic counts. More specifically, the lognormal model assumes that the log traffic counts follow a linear regression of model of the form

$$y_t = u + \sum_{i=1}^{12} \Delta_{t,i} m_{k,i} + \sum_{j=1}^{7} \delta_{t,j} w_{k,j} + e_t \qquad (2.2)$$

where

$y_t = \log_e(z_t)$, the natural logarithm of the traffic count on day t,

u = expected log traffic count on a typical day,

$\Delta_{t,i}$ = 1, if the count $z_t$ was made during month i, i=1,..,12,

0, otherwise;

$m_{k,i}$ = correction term for month i, characteristic of factor group k,

$\delta_{t,j}$ = 1, if the count $z_t$ was made on day-of-week j, j=1,..,7,

0, otherwise;

$w_{k,j}$ = correction term for day-of-week j, characteristic of factor group k,

$e_t$ = random error (residual).

If we let $\beta_k=(m_{k,1},..,m_{k,12}, w_{k,1},..,w_{k,7})'$ be a column vector (the ′ symbol denotes the action of taking an array's transpose) containing the monthly and daily correction terms characteristic of factor group k, and if we let $\mathbf{y}=(y_1,..,y_N)'$, $\mathbf{e}=(e_1,..,e_N)'$ be column vectors containing the log traffic counts and random error terms respectively, then the above model can be expressed in a vector form as

$$y = u\mathbf{1}_N + X\beta_k + e \qquad (2.3)$$

Here $\mathbf{1}_N$ denotes an N-dimensional column vector containing the value 1.0 in each position and $\mathbf{X}$ is a matrix, with dimensions Nx19, whose elements contain only the values 0 or 1 depending on the month and day-of-week of the corresponding daily count. The lognormal model assumes that the residual vector $\mathbf{e}$ is a normally distributed random vector, with all components having a mean value of 0, and a covariance matrix $\sigma^2\mathbf{V}$ where

$\sigma^2$ = the common unconditional variance of $y_t$,

$\mathbf{V}$ = NxN matrix of correlation coefficients such the element in row s and column t, $V_{s,t}$ is the correlation coefficient for $y_s$ and $y_t$.

The objective then was to fit the model (2.2) in such a way that monthly and day-of-week terms $m_{k,i}$ and $w_{k,j}$ were constant for all ATRs in a specified factor group, but each ATR in the factor group was allowed to have its own mean value parameter u, and residual covariance matrix $\sigma^2\mathbf{V}$. This was accomplished using a two-stage procedure, where the first stage consisted of estimating the mean-value, monthly and day-of-week terms using iteratively re-weighted least-squares, and the second stage consisted of fitting time-series models to the regression residuals obtained from the first stage. The first stage computations were performed using MINITAB (Minitab, 1995), in which a special MINITAB macro was written which iterated among the three main steps:

(1) estimate mean, monthly and day-of-week parameters via a weighted least-squares using MINITAB's GLM procedure,

(2) compute the residual variance for each ATR in a group given the current regression parameter estimates and,

(3) use these variance estimates to compute separate weighting vectors for each ATR.

This weighted least-squares approach was chosen over ordinary least-squares in order to accommodate the heteroscedasticity caused by different ATRs showing different day-to-day variances. A listing of this macro is given Appendix B. In the second stage, the regression residuals were first obtained by computing the differences between the log daily traffic counts and the value predicted using model (2.2) with the final estimates of the mean-value, monthly and day-of-week parameters. Standard methods of model identification, model estimation and diagnostic checking were then used to fit a time-series model (Box, Jenkins and Reinsel,1994 ) to each set of residuals. The form of the time-series model then determined the form of the autocorrelation matrix **V** for each ATR.Before discussing the second modeling stage, some of the first-stage results may be of interest. The lognormal model implies that if a daily count $z_t$ was made during month i and on day-of-week j, the expected value for the count is given by

$$E[z_t] = (e^{u+\frac{\sigma^2}{2}})(e^{m_{k,i}})(e^{w_{k,j}}) \qquad (2.4)$$

so that the multiplicative relationship between the expected traffic volume on a given day and the monthly and day-of-week multipliers shown in equation (2.1) is maintained, with $z_0=\exp(u+\sigma^2/2)$, $M_{k,i}=\exp(m_{k,i})$, and $W_{k,j}=\exp(w_{k,j})$. Exponentiating the estimates of $m_{k,i}$ and $w_{k,j}$ thus produces estimates of $M_{k,i}$ and $W_{k,j}$, and these estimates a displayed in Figure 2.2. If we regard the ATRs in factor group 1 as being "typical", then the ATRs in group 3 differ by tending to have higher than average traffic volumes on Sundays, while the ATRs in group 4 differ by tending to have higher summer volumes. Validation of the lognormal model in stage 2 will mean, among other things, that regression methods could be used to estimate monthly and day-of-week adjustment factors, and regression theory could be a starting point for an investigation of the statistical properties of these estimators.

## SECOND STAGE MODELING AND GOODNESS-OF-FIT

The chief objective of the second stage of analysis was to determine a plausible form (or forms) for the correlation matrix **V**, and to verify the appropriateness of the lognormal model by comparing the distribution of residuals with what would be expected had they been normally distributed. In time-series analysis the estimated autocorrelation function (ACF) computed from the residuals provides the key tool for identifying patterns of temporal dependency, and hence the appropriate form for **V** (Box, Jenkins and Reinsel, 1994, ch. 6). The ACF is estimated by first computing the correlation coefficient of observations separated by one day, then the correlation coefficient of observations separated by two days, and so forth. These estimated autocorrelations are then plotted against the lag used in computing them, and the pattern shown by this plot is used to identify the underlying time-series model. Sample autocorrelation functions were computed for each of the 50 sets of ATR residuals obtained from the 1992 data, using Dunsmuir and Robinson's (1981) procedure which allows for missing data. Figure 2.3 shows the estimated ACF for ATR #207.

Part of the art in time-series analysis lies in being able to identify an underlying statistical model from an autocorrelation function. The pattern shown in the upper half of Figure 2.3, where the autocorrelation coefficients begin damping to zero after the first lag but then rise and fall periodically at lags 7, 14, and 21 is consistent with a seasonal autoregressive (AR) model of the form

$$e_t = \phi_1 e_{t-1} + \phi_7 e_{t-7} - \phi_1 \phi_7 e_{t-8} + a_t \qquad (2.5)$$

where $\phi_1$ and $\phi_7$ are (unknown) autoregressive coefficients, and the $a_t$ are independent, identically distributed normal random variables with zero mean and common variance $\sigma_a^2$. If model (2.5) is appropriate, then by first computing estimates of the coefficients $\phi_1$ and $\phi_7$, and then computing second stage residuals via

$$\hat{a}_t = e_t - \hat{\phi}_1 e_{t-1} - \hat{\phi}_7 e_{t-7} + \hat{\phi}_1 \hat{\phi}_7 e_{t-8} \qquad (2.6)$$

should produce a sequence of second-stage residuals which: (a) show no significant autocorrelations

and (b) pass a goodness-of-fit test for normally distributed data. To test this, nonlinear least-squares estimates of $\phi_1$ and $\phi_7$ were computed from ATR #207's first stage residuals using MATLAB's FMINS (Mathworks, 1995) routine, and the second stage residuals were computed using equation (2.6). The autocorrelation function for these second stage residuals was computed, along with critical values for testing the hypothesis that no estimated autocorrelation coefficient is significantly different from zero (Madansky, 1988, p. 105). The lower half of Figure 2.3 displays the results of these computations, and it can be seen that the seasonal AR model did in fact account for the autocorrelation shown in the upper half of Figure 2.3. To test the normality of these second stage residuals, MINITAB's NORMPLOT macro was used to prepare a normal probability plot of the standardized stage-2 residuals, and to compute the correlation between the residuals and their normal scores. Figure 2.4 shows these results, with the interpretation that if the data are normally distributed, the scatter plot in Figure 2.4 should be approximately linear. The correlation coefficient R=.9915 turns out to be significantly different from 1.0 ($p < .01$), and we should reject the hypothesis that these second stage residuals are normally distributed. Inspection of Figure 2.4 shows however that the deviation from normality is caused by a few extreme values at the tails of the distribution. If we consider the three most extreme values as atypical outliers and remove them from the data, we obtain the goodness of fit results shown in Figure 2.5, which suggests we not reject the hypothesis of normally distributed residuals.

The results obtained for ATR #207 turned out to be typical of those obtained for the other 49 ATRs in the 1992 data set. The autocorrelation functions for the first stage residuals showed patterns similar to that in Figure 2.2, and this autocorrelation was essentially removed after fitting a seasonal AR model of the form (2.4). The second stage residuals all either passed a normality test without consideration of outliers, or passed the test after removal of a small (i.e. 1-6) number of extreme values. If we are willing to make the plausible assumption that these extreme values were caused by undetected atypical events such as data errors, special events or weather effects, then the lognormal regression model given in (2.2), with residuals satisfying a seasonal AR model of the form (2.4) offers a defensible statistical model of daily, non-holiday, two-directional traffic counts in outstate Minnesota. We note that a more general model incorporating holiday, weather or other effects could be developed by including additional predictor variables in the regression model (2.3).

To summarize, we have provided fairly strong evidence that a lognormal regression model describes the statistical properties of daily traffic counts, at least in outstate Minnesota. The connection with regression methods and time-series analysis means that a wide range of powerful statistical tools can be brought to bear on problems of traffic estimation. An exhaustive account of these possibilities is well beyond the scope of this report, but the model developed in this Chapter will play a crucial role in our later work.

**Figure 2.2. Estimated Monthly and Day of Week Multipliers Using 1992 ATR Data.**

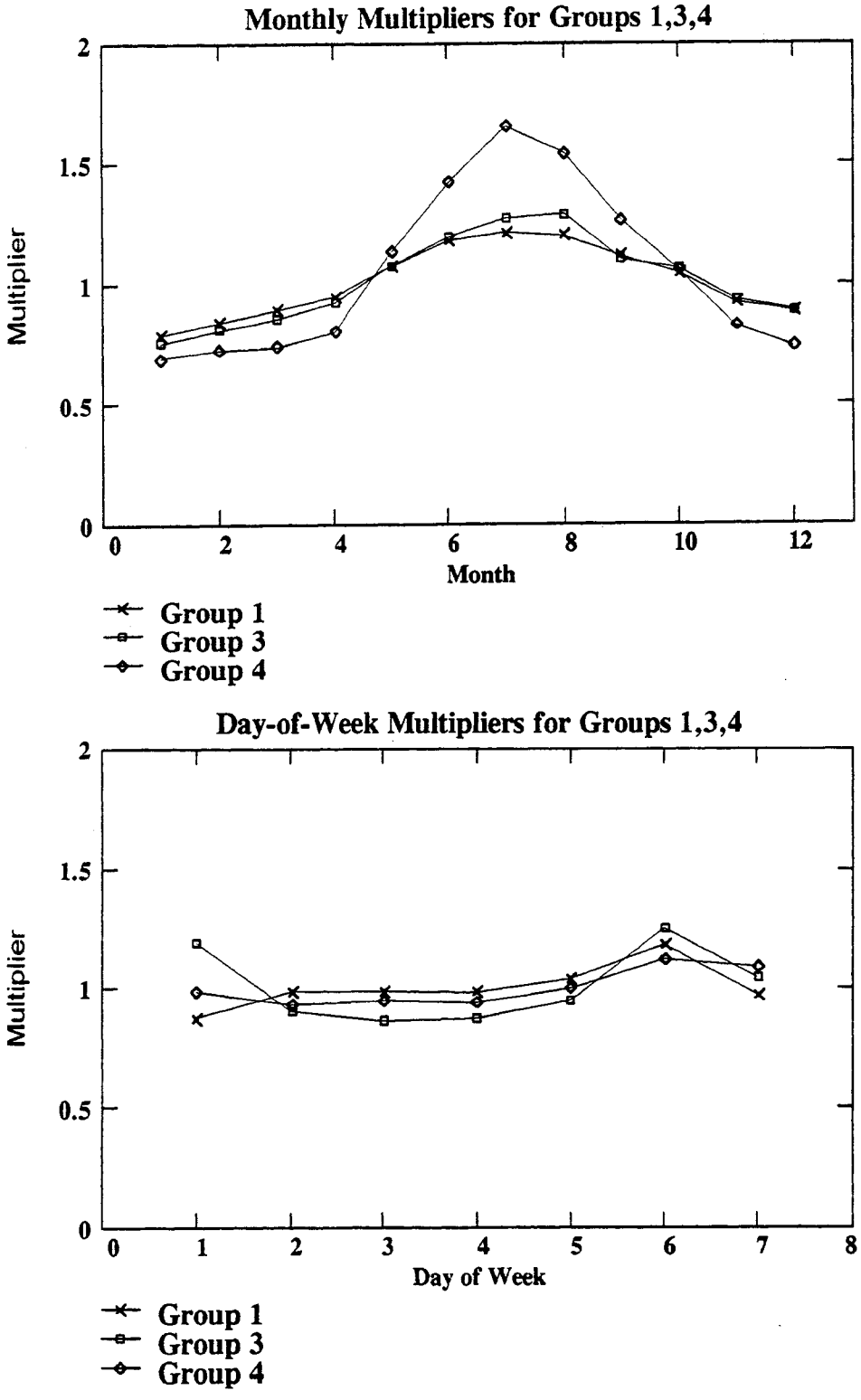
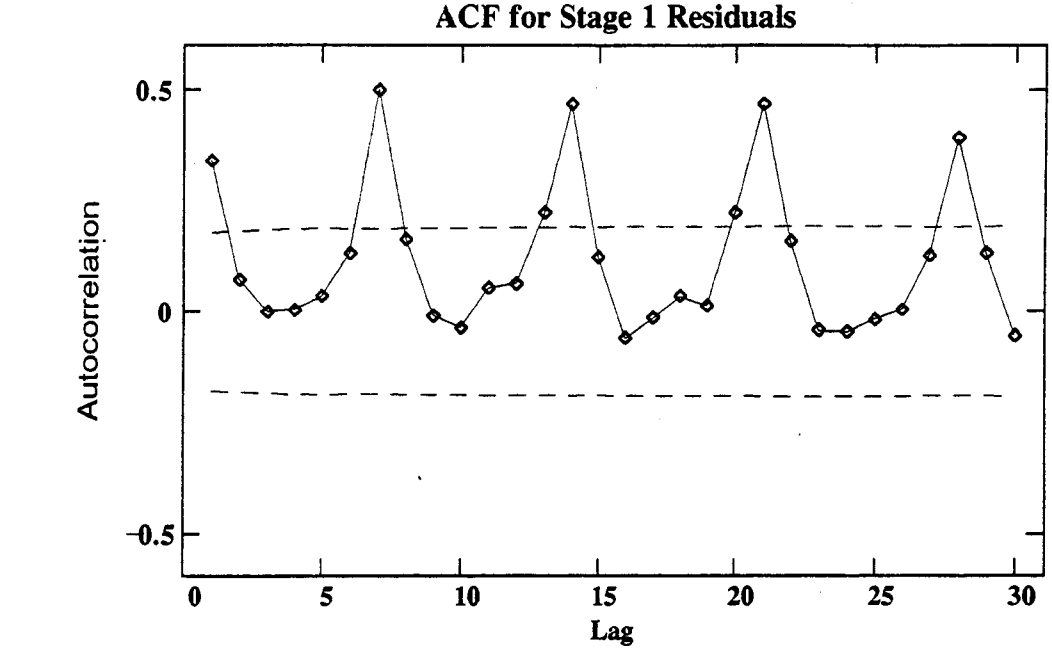


**Figure 2.3. Stage 1 and Stage 2 Autocorrelation Functions**

**ATR #207.**



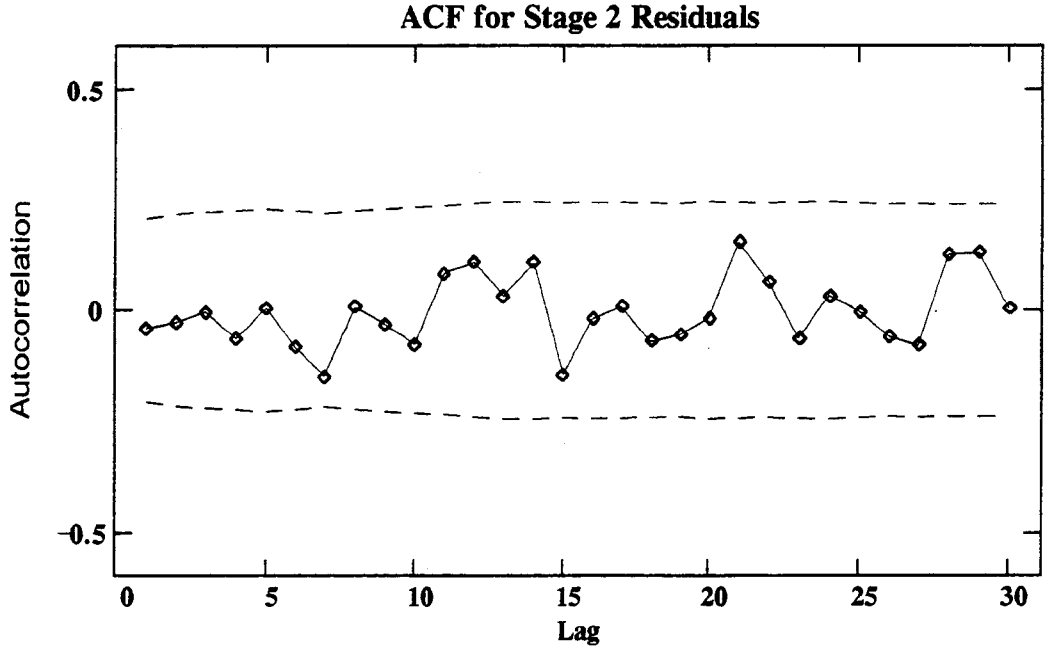



**Figure 2.4. Normality Test of Stage 2 Residuals for ATR #207.**

**No Outliers Removed**

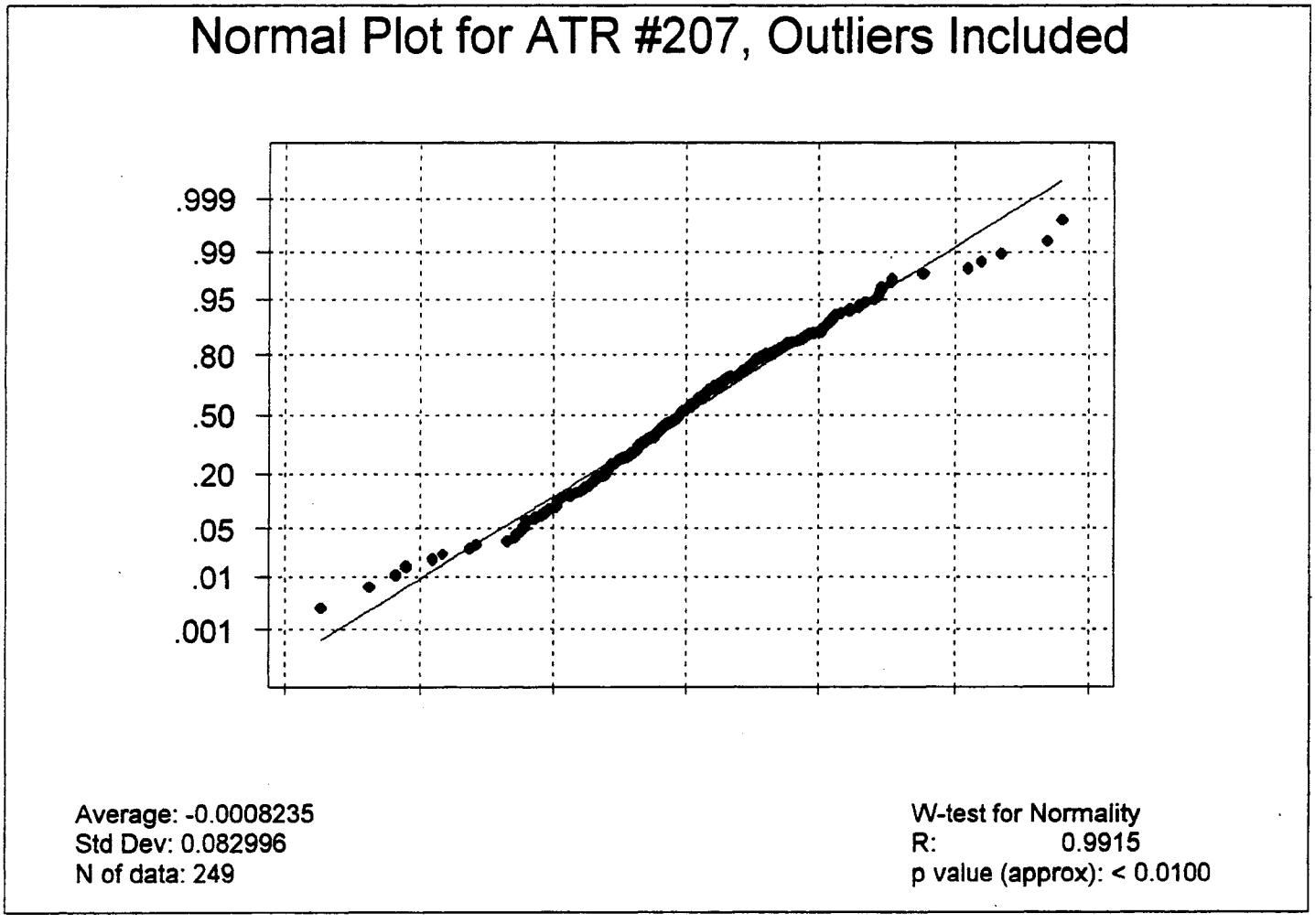
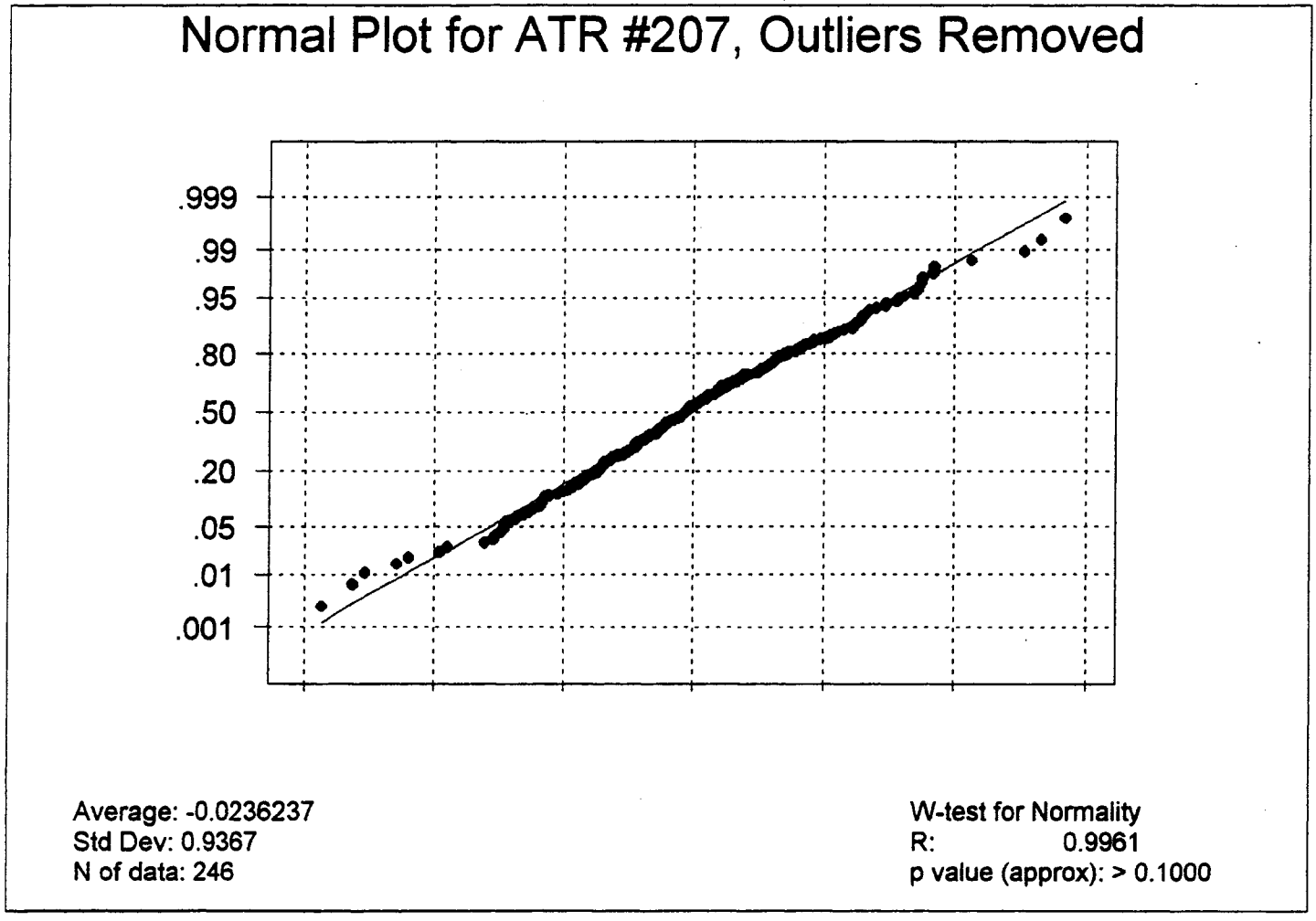


**Figure 2.5. Normality Test for Stage 2 Residuals for ATR #207**

**Three Outliers Removed**



# CHAPTER 3
# BAYESIAN ASSIGNMENT TO FACTOR GROUPS AND SAMPLE DESIGN

As noted in Chapter 1, seasonal and day-of-week variations in daily traffic cause simple averages of short-count data produce to biased estimates of MDT, so that obtaining unbiased estimates from short counts requires applying appropriate monthly and day-of-week correction factors. Methods for grouping ATRs into factor groups and estimating correction factors can be found in the literature, and it was noted at the end of Chapter 2 that at least in principle, the regression methods described there could be extended to perform simultaneous clustering and estimation of factors. A more difficult problem involves assigning short count sites to factor groups, and at present there exists no completely satisfactory, data-driven method for performing this assignment.

## FACTOR GROUP ASSIGNMENT VIA BAYESIAN DECISION THEORY

Even before any short-count data are available, it is possible to have an opinion regarding the plausibility of assigning a given site to each of a set of factor groups, and it is easy to imagine quantifying this opinion by assigning each factor group in the set a probability value, with a high probability being given to a factor group when one believes it is more likely that the chosen site belongs to this factor group. If one is completely uncertain as to a site's factor group membership, one can assign an equal probability to each group. Count data, when available, provide additional evidence concerning factor group membership, and the rational individual will modify his or her prior views on group membership in light of this data. This problem of modifying prior viewpoints in light of data in the most efficient manner is the subject of Bayesian statistics, where Bayes theorem from probability theory is used to compute the optimal updating of one's prior probability assessment after obtaining new data.

To provide a formal statement of the factor group assignment problem, imagine that an analyst has identified (through cluster analysis of ATR data, or some other method) a total of n

different factor groups, such that ATR locations assigned to a factor group show similar patterns of monthly and day-of-week variations in daily traffic counts. There should be good reason to believe that any other highway site within the state has a monthly and daily variation pattern similar to that of one of these groups since if this assumption is not plausible, the state's set of ATR locations is incomplete. Each of these groups is characterized by a set of monthly and day-of-week multipliers, and the natural logarithms of these multipliers have been collected into vectors $\beta_k$, k=1,..,n, one for each ATR factor group. The vectors $\beta_k$ as defined here are identical to the vectors of regression coefficients defined in equation (2.3). The analyst also has available a sequence of N daily traffic counts $z_1,z_2,..,z_N$, taken at some highway site that is not one of the ATR locations. Although for convenience these daily counts are indexed by t=1,2,..,N, there is no assumption made that the counts actually come from consecutive days, and could in fact be generated by any arbitrary sampling plan. Now let $\theta$ denote the vector of monthly and day-of-week terms appropriate for the highway site where the counts were taken. By assumption $\theta$ must one of the known vectors $\beta_k$, and it is unclear which, but the optimal estimate of $\theta$ can be ascertained if the analyst is able to provide three additional items of information:

(1) A set of prior (before data collection) classification probabilities $\alpha_k$, k=1,..,n, such that $\alpha_k$=probability $\theta$ equals $\beta_k$, (which we will abbreviate as $\alpha_k$=Prob[$\theta=\beta_k$]);

(2) A set of likelihood functions, one for each factor group, denoted by $f(z_1,..,z_N|\theta=\beta_k)$, which give the probability (or probability density) of obtaining the count sample had the site actually belonged to group k;

(3) A loss function, denoted by $L(\theta,\beta_k)$ which gives the penalty incurred if we select $\beta_k$ as the correction terms for the site.

The factor group assignment problem can now be solved using Bayesian decision theory (Berger, 1988). A particularly tractable, and intuitively meaningful, result can be developed if we use the "0-1" loss function:

$L(\theta,\beta_k)$ = 0, if we select $\beta_k$ and $\theta$ actually equals $\beta_k$,

1, if we select $\beta_k$, but $\theta$ actually equals some other $\beta_l$, $l\neq k$.

Thus if the site actually belong to group k, and we assign it to group k we incur a loss of 0, while if the site belongs to some other group and we assign it to k, we incur a loss of 1.0. After we collect data, an application of Bayes Theorem gives the posterior distribution of the parameter $\theta$

$$Prob[\theta=\beta_k|y] = \frac{f(y|\theta=\beta_k)\alpha_k}{\sum_{l=1}^{n} f(y|\theta=\beta_l)\alpha_l} \tag{3.1}$$

where $\mathbf{y}=[\log_e(z_1),..,\log_e(z_N)]'$, denotes the vector containing the natural logarithms of the sample counts. The expected posterior loss for selecting $\beta_k$ as the correction vector can then be shown to be

$$E[L(\theta,\beta_k)] = \sum_{l=1}^{n} L(\theta,\beta_k)Prob[\theta=\beta_k|y] = 1-Prob[\theta=\beta_k|y] \tag{3.2}$$

For each vector $\beta_k$, the right-hand side of the above equation gives the loss we can expect to incur if we use $\beta_k$ after observing the data sample $\mathbf{y}$. If the objective is to *minimize* this expected posterior loss, then the optimal decision rule is to set $\theta$ equal to the $\beta_k$ having the *largest* posterior probability. The vector $\beta_k$ having the largest posterior probability is, for the 0-1 loss function, the optimal estimate of the unknown $\theta$, and since this decision depends on the sample data $\mathbf{y}$, it will be useful later on to describe this estimate as a function of $\mathbf{y}$,

$$\hat{\beta}(y) = \textit{vector } \beta_k \textit{ such that } Prob[\theta=\beta_k|y] > Prob[\theta=\beta_l|y],\ l\neq k \tag{3.3}$$

In principle, the above estimation rule provides a complete solution to the factor group assignment problem, but only if the prior classification probabilities and likelihoods have been specified. In most practical applications, the prior classification probabilities can be taken to be $\alpha_k=1/n$, k=1,..,n, indicating complete prior uncertainty as to which group the site belongs, although any strong beliefs the analyst has concerning a site's group membership can be captured by using non-uniform $\alpha_k$'s. Specification of the likelihood function is a bit more difficult, but the model development work described in Chapter 2 provides a starting point. If we restrict our attention to two-directional traffic volumes on outstate Minnesota highways, and if we assume that the ATRs in

outstate Minnesota are representative of all relevant non-ATR sites, then the results from Chapter 2 imply that the daily traffic counts at the site of interest are generated by a stochastic process described by a lognormal regression model similar to that described in equation (2.2), with error terms following a seasonal AR model of the type described in equation (2.5). Letting **y** denote the vector containing the logarithms of the sample counts and **V** denote the correlation matrix for the count sample (which can be computed from once one knows the value of the AR parameters $\phi_1$ and $\phi_7$), then if we knew the site specific parameters u, $\sigma$, $\phi_1$, and $\phi_7$, the likelihood of the sample could be computed using the appropriate multivariate normal density

$$f(y|\beta_k,u,\sigma,\phi_1,\phi_7) = (2\pi)^{-N/2}\sigma^{-N}|V|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(y-X\beta_k-u\mathbf{1}_N)'V^{-1}(y-X\beta_k-u\mathbf{1}_N)\right) \quad (3.4)$$

As in equation (2.3) , **X** is a matrix, of dimension Nx19, each row having elements equal to 0 or 1 according to the month and day-of-week of the corresponding sample count.

At this point, the likelihood functions used in equation (3.1) depend only on knowing factor group membership, while the likelihood in equation (3.4) requires this plus knowledge of the site-specific parameters u, $\sigma$, $\phi_1$, and $\phi_7$. The Bayesian method of deriving the likelihoods appearing in equation (3.1) from the likelihood in (3.4) is to multiply (3.4) by prior probability distributions for u, $\sigma$, $\phi_1$, and $\phi_7$, to produce the joint distribution of the parameters and sample data, and then integrating this joint distribution with respect to the site-specific parameters to produce the "marginal" likelihoods appearing in equation (3.1). This prior distribution of the site-specific parameters must also be specified by the analyst, and in this project we investigated two different approaches to developing such priors. The first employs commonly used "noninformative" priors for all four site-specific parameters, while the second retains the noninformative prior for u, but uses the results from fitting the models to the ATR data, described in Chapter 2, to construct an estimate of the prior for $\sigma$, $\phi_1$ and $\phi_7$.

An extensive discussion of Bayesian analysis using normal probability models with uncorrelated observations is given in the book by Box and Tiao (1973), and discussion of Bayesian analysis of time-series regression models can be found in Zellner (1987, ch. VII) and Monahan (1983). Following Box and Tiao (1973), a noninformative prior distribution for the parameters u and

$\sigma$ is $g(u,\sigma)=1/\sigma$. In order that the time-series model be stationary it is necessary to constrain the autoregressive parameters $\phi_1$ and $\phi_7$ to have absolute value less than 1.0, but in the analysis of the ATR data reported in Chapter 2 it was found that for all 50 locations studies, the estimates of the AR parameters were non-negative. Thus a noninformative prior for the AR parameters which is consistent with these conditions is the product of two uniform probability densities, each with lower and upper bounds at 0 and 1.0, respectively. The resulting prior for u, $\sigma$, $\phi_1$, and $\phi_7$ is simply

$$g(u,\sigma,\phi_1,\phi_7) = \begin{cases} 1/\sigma, & \text{if } -\infty<u<\infty,\ \sigma>\infty,\ 0<\phi_1,\phi_7<1.0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

Multiplying the right-hand side of (3.4) by (3.5) and applying the integration formulas in Box and Tiao (1973, p.154) we obtain, after canceling out terms common to the numerator and denominator, the formula for the posterior classification probabilities

$$Prob[\theta=\beta_k|y] = \frac{\alpha_k S_k}{\sum_{l=1}^{n} \alpha_l S_l} \tag{3.6}$$

where

$$\begin{aligned} S_k &= \int_0^1\int_0^1 |V|^{-1/2}((1_N)'V^{-1}1_N)^{-1/2}(s_k)^{-(N-1)/2}d\phi_1 d\phi_7 \\ s_k^2 &= (y-X\beta_k-\bar{y}_k 1_N)'V^{-1}(y-X\beta_k-\bar{y}_k 1_N) \\ \bar{y}_k &= \frac{1_N'V^{-1}(y-X\beta_k)}{1_N'V^{-1}1_N} \end{aligned} \tag{3.7}$$

As noted earlier, the correlation matrix **V** is a function of the AR parameters $\phi_1$ and $\phi_7$. To determine the elements of **V** for given values for the AR parameters, one first computes the theoretical ACF for the seasonal AR model using the procedure described in Brockwell and Davis (1991, pp. 95-96). The element $V_{s,t}$ in the row s and column t of **V** is then theoretical autocorrelation coefficient for lag |s-t|.

As no closed forms for the integrals appearing in (3.7) are known, they must be evaluated numerically.

A noninformative prior is a reasonable choice for the mean parameter u since this will ultimately cause the posterior estimate of MDT to be determined primarily by the sample data. However the remaining parameters are nuisance parameters, and the ATR modeling described in Chapter 2 produced estimates of $\sigma$, $\phi_1$, and $\phi_7$ for each of 50 ATR sites. If the ATRs are typical of other highway sites these estimates provide considerable prior information concerning likely values for the nuisance parameters, and it is desirable to exploit this prior information in order to sharpen decisions concerning factor group membership. The marginal likelihoods appearing in equation (3.1) are mixtures of probability distributions, with the prior distribution for the site-specific parameters providing the mixing distribution. Lindsay (1983) has shown that the nonparametric maximum likelihood estimate of a mixing distribution takes the form of a probability mass function with finite support, while Padgett and Robinson (1978) use a mass function estimate of a prior distribution to computes Bayes estimates of reliability. If we let $\Phi_p=(\phi_{1p}, \phi_{7p})$ and $\sigma_p$ denote the estimates obtained for ATR number p, p=1,.., M, we can approximate the prior distribution with a probability mass function that assigns probability 1/M to each ordered triple of estimates from the ATRs. Combining this with the noninformative prior for u, we obtain the prior for u, $\sigma$, $\phi_1$, $\phi_7$ as

$$g(u,\sigma,\phi_1,\phi_7) = \frac{1}{M}\sum_{p=1}^{M} I_{(\sigma_p,\phi_{1p},\phi_{7p})}(\sigma,\phi_1,\phi_7) \qquad (3.8)$$

where $I_b(x)$ denotes a Kronecker delta function

$I_b(x) = 1$, if $x=b$,

0, otherwise.

Multiplying the likelihood in equation (3.4) by (3.8) and integrating out u gives a somewhat different formula for the classification probabilities

$$Prob[\theta=\beta_k|y] = \frac{\alpha_k S_k}{\sum_{l=1}^{n} \alpha_l S_l} \tag{3.9}$$

where now

$$\begin{aligned} S_k &= \sum_{p=1}^{M} |V_p|^{-1/2} (1_N' V_p 1_N)^{-1/2} \sigma_p^{-(N-1)} e^{\frac{s_{pk}^2}{2\sigma_p^2}} \\ s_{pk}^2 &= (y - X\beta_k - \bar{y}_{pk} 1_N)' V_p^{-1} (y - X\beta_k - \bar{y}_{pk} 1_N) \\ \bar{y}_{pk} &= \frac{1_N' V_p^{-1} (y - X\beta_k)}{1_N' V_p^{-1} 1_N} \end{aligned} \tag{3.10}$$

and $\mathbf{V_p}$ denotes the sample correlation matrix computed using these estimated AR coefficients for ATR number p, $\phi_{1p}$ and $\phi_{7p}$.

## DESIGNING SAMPLING PLANS

In the preceding section we used Bayesian decision methods to determine an optimal assignment of a highway site to a factor group, given a sample of daily traffic counts from that site. In order to specify what was meant by optimal, we introduced the "0-1" loss function which assigns a uniform penalty for misclassifying the site, and no penalty for a correct classification. This led to the intuitively appealing decision rule of simply assigning the site to the factor group with the highest posterior probability, or equivalently, selecting as the monthly a day-of-week correction terms that vector which the data most favors. Using the lognormal likelihood of daily traffic counts developed in Chapter 2, two formulas for computing these posterior classification probabilities were derived, one using noninformative prior distributions for the site variance and autocorrelation parameters, the other using the results of ATR model fitting to develop an informative prior for these parameters. Application of either formulas (3.6-3.7) or formulas (3.9-3.10) require that a sample of daily traffic counts be available, but it may very well be that traffic counts on some days are more informative than

counts made other days. In this section, we describe how to design sampling plans which, for a given sample size, are more likely to lead to correct assignment decisions.

The first step is formalize what we mean by a sampling plan. We assume that the traffic analyst has decided to count on a total of N different days, N thus being the sample size, so the remaining decision concerns what days to count. If there are a total of 365 days during the upcoming year, this is equivalent to specifying the values of a column vector **d**, of dimension 365x1, whose elements are defined via

$d_t =$ 1, if a count is taken on day t, t=1,..,365
0, of a count is not taken on day t.

The set of different vectors **d**, defined as above, with $\Sigma_t d_t = N$, then describes the set of all possible feasible sampling plans.

Selecting the optimal sample requires that we be able to assign to each sampling plan a quantitative measure reflecting how "good" that plan is. As in he previous section the "0-1" loss function will be used, which leads to sampling plans minimizing the probability of misclassifying the site. To develop this criterion, suppose the site actually belongs to factor group k (that is, $\theta=\beta_k$), and that we are committed to following the decision rule (3.3). The "0-1" loss experienced if we follow the rule and $\theta=\beta_k$ can be expressed as

$$L(\beta_k,\hat{\beta}(\mathbf{y})) = \begin{array}{l} 0, \text{ if } \hat{\beta}(\mathbf{y})=\beta_k \\ 1, \text{ if } \hat{\beta}(\mathbf{y})\neq\beta_k. \end{array} \qquad (3.11)$$

This loss depends on the actual sample count which has not yet been obtained. However, the expected loss when $\theta=\beta_k$ can be computed using the data likelihood via

$$E[L(\beta_k,\hat{\beta}(y))] = \int_y L(\beta_k,\hat{\beta}(y))f(y|\beta_k)dy = 1-Prob[\hat{\beta}(y)=\beta_k|\theta=\beta_k] \qquad (3.12)$$

and the overall loss can then be obtained by applying the prior classification probabilities $\alpha_k$, k=1,..,n

$$E[L] = 1 - \sum_{k=1}^{n} \alpha_k Prob[\hat{\beta}(y)=\beta_k | \theta=\beta_k] = h(\beta_1,..,\beta_n,d) \qquad (3.13)$$

This right hand side of (3.13) simply indicates that the expected loss depends on the parameters characterizing the factor groups, $\beta_1,..,\beta_n$, as well as the sampling plan **d**. A mathematical programming problem whose solution gives the optimal sampling plan would then be

minimize: $h(\beta_1, ..,\beta_n, \mathbf{d})$

subject to:

$d_t = 1$, if day t is counted, 0 otherwise;

$\Sigma_t d_t = N$.

This problem takes the form of a nonlinear knapsack problem, in which we have a "knapsack" which can hold at most N items, and we want to select the best combination. Since there are only a finite number of different sampling plans, in principle this problem could be solved by simply evaluating the function h(.) for each, and then selecting that plan with the smallest value. In practice though this exhaustive enumeration is not likely to work due to the very large number of different sampling plans. For example, if there are 365 days in a year, and we are restricted to sampling plans with 10 counting days, then there are over $1.021 \times 10^{19}$ different sampling plans. If for each plan the function h(.) could be computed in 1 microsecond, then over 300,000 years would be needed for a complete evaluation. In certain simple special cases fast algorithms exist for solving knapsack problems, but the nonlinear nature of the function h(.) precludes this in our case. The alternative is to abandon the search for the guaranteed optimal sampling plan, and instead concentrate on finding heuristic solutions which although they lack a guarantee of optimality nonetheless have acceptable practical performance.

To develop our heuristic, we suppose that we have very good prior knowledge of the site's

variance parameters, in particular that $\phi_1=\phi_7=0$ (that is, no serial correlation is present), and that the variance parameter is a known value $\sigma$. Also suppose that we have only two factor groups, and that we are completely uncertain as to which group the highway site should belong, that is, that $\alpha_1=\alpha_2=0.5$. Under these conditions it can be shown that

$$Prob[\hat{\beta}(y)=\beta_1|\theta=\beta_1] = Prob[s_1^2<s_2^2|\theta=\beta_1] \tag{3.14}$$

where

$$\begin{aligned} s_k^2 &= (y-X\beta_k-\bar{y}_k 1_N)'(y-X\beta_k-\bar{y}_k 1_N) \\ \bar{y}_k &= \frac{1_N'(y-X\beta_k)}{N} \end{aligned} \tag{3.15}$$

Thus the objective of the knapsack problem becomes

$$h(\beta_1,\beta_2,d) = 1-\frac{Prob[s_1^2<s_2^2|\theta=\beta_1]+Prob[s_2^2<s_1^2|\theta=\beta_2]}{2} \tag{3.16}$$

and using the lognormal likelihood model, it can be shown after some lengthy algebra, that

$$Prob[s_1^2<s_2^2|\theta=\beta_1] = Prob[s_2^2<s_1^2|\theta=\beta_2] = \Phi\left(\frac{\sqrt{(\beta_1-\beta_2)'\tilde{X}'\tilde{X}(\beta_1-\beta_2)}}{2\sigma}\right) \tag{3.18}$$

where $\Phi(.)$ denotes the cumulative standard normal distribution function, and the elements of the matrix $\tilde{\mathbf{X}}$ are given by

$$\tilde{X}_{t,i} = X_{t,i} - \frac{1}{N}\sum_{s=1}^{N} X_{s,i} \tag{3.19}$$

Since $\Phi(.)$ is a monotonically increasing function, minimizing $h(\beta_1, \beta_2, \mathbf{d})$ turns out in this case to be equivalent to maximizing the prediction difference

$$S(\beta_1,\beta_2,d) = (\beta_1-\beta_2)'\tilde{X}'\tilde{X}(\beta_1-\beta_2) \tag{3.20}$$

and after a bit more algebra, it can be shown that this is equivalent to

$$h2(\beta_1,\beta_2,d) = \sum_{i=1}^{12}\sum_{j=1}^{7} n_{ij}(\tilde{m}_i+\tilde{w}_j)^2 - N(\bar{m}+\bar{w})^2 \tag{3.21}$$

where

$$\begin{aligned} \tilde{m}_i &= m_{i,1}-m_{i,2} \\ \tilde{w}_j &= w_{j,1}-w_{j,2} \\ n_{ij} &= \sum_{t=1}^{365} d_t\Delta_{t,i}\delta_{t,j} \\ \bar{m} &= \frac{1}{N}\sum_{i=1}^{12}\tilde{m}_i(\sum_{j=1}^{7} n_{ij}) \\ \bar{w} &= \frac{1}{N}\sum_{j=1}^{7}\tilde{w}_j(\sum_{i=1}^{12} n_{ij}) \end{aligned} \tag{3.22}$$

The quantities $n_{ij}$ appearing above are simply the numbers of days for month i and day-of-week j included in the sampling plan, and one very interesting property of the objective function h2(.) is that the actual sampling plan **d** contributes to h2(.) only through the terms $n_{ij}$. This means that all sampling plans with the same $n_{ij}$ are equivalent, or if a sampling plan requires counting two Tuesdays in July, it does not matter which two Tuesdays in July are chosen. Another interesting property of h2(.) is that the resulting optimization problem is a special case of the problem of finding T-optimal experimental designs treated by Atkinson and Federov (1975a). Since the quantities $n_{ij}$ are integer values, finding

the optimal sampling plan still requires solving a nonlinear integer programming problem, which will still be difficult to solve. Following the usual practice in optimal design theory (e.g. Atkinson and Federov, 1975a; Silvey, 1980), we can generate approximate solutions by solving the simpler problem which results from allowing the $n_{ij}$ to take on non-integer values. This produces the nonlinear programming problem

maximize: $h2(\beta_1, \beta_2, \mathbf{n})$

subject to:

$$0 \leq n_{ij} \leq \bar{n}_{ij}$$

$$\Sigma_i \; \Sigma_j \; n_{ij} = N$$

where $\bar{n}_{ij}$ is the total number of days-of-week j occurring during month i during the period of interest. The fact that the $n_{ij}$ are now only subject to bounding constraints makes this problem especially easy to solve using a fixed step-size variant of the Frank-Wolfe algorithm. A listing of the MATLAB program used to implement this algorithm can be found in the Appendix B.

Finally, solution of the above problem approximates the optimal sample design when only two factor groups are present. As found by Atkinson and Federov (1975b), finding designs for more than two groups is much more difficult, so we employ the following simple heuristic:

(1) For each pair of factor groups, solve the design problem to find a design for discriminating the two groups;

(2) After solutions for each pair of factor groups have been obtained, compute the average of these solutions;

(3) Round the elements of the average solution to produce integer values.

## EMPIRICAL EVALUATION OF DESIGN AND CLASSIFICATION METHODOLOGY

The sample design problem requires as input the natural logarithms of the monthly and day-of-week multipliers characterizing each factor group, and in Chapter 2 we described how estimates of these quantities were obtained for Mn/DOT's factor groups 1,3 and 4 using 1992 daily ATR counts.

These estimates were used as input to the OPSAMP2 routine described in Appendix B to find designs for discriminating groups 1 from 3, 1 from 4, and 3 from 4, with a sample size of N=15. The resulting solutions are displayed in Tables 3.1, 3.2 and 3.3. These three tables were then averaged and rounded to produce the design displayed in Table 3.4, and because of rounding effects the sample size depicted in Table 3.4 became N*=13. One can see that the sample design tends to emphasize counts on Sundays, which Figure 2.2 indicates are good for identifying group 3, and counts in July, which Figure 2.2 indicates are good for identifying group 4.

For each of the 48 ATRs which had complete records in 1991, the sampling design depicted in Table 3.4 was converted into actual samples by simply taking the design days in order. For example, when the sampling design called for counting two Tuesdays in July, we included in our sample the first and second Tuesdays in July. If the count on a sample day was coded as missing, that day was deleted from the sample, but if possible another, similar day was chosen to replace it. This means that the actual sample could be less than the design value of N*=13. For each ATR the posterior classification probabilities were computed using both the noninformative prior formulas listed in equations (3.6-3.7) and the "Empirical Bayes" formulas listed in equations (3.9-2.10), assuming that the prior classification probabilities were $\alpha_k$=1/3. These classification probabilities, along with the resulting sample sizes, are displayed in Tables 3.5-3.7.

Assuming that Mn/DOT's three outstate factor groups are well-constructed, and that the underlying traffic generation patterns in 1991 were similar to those in 1992, the 1991 samples should give the greatest posterior probability to that group to which the ATR was assigned in 1992. Inspection of Table 3.5, which displays the results for group 3, shows that generally this was the case, and that the noninformative prior and the informative "Empirical Bayes" prior tended to produce similar classifications. Note however that in many cases using the "Empirical Bayes" prior tends to produce slightly sharper classifications than does the noninformative prior. This pattern also appears in Table 3.6, showing the results for group 4, and in Table 3.7, showing the results for group 1.

**Table 3.1. Sampling Design for Discriminating Group 1 and Group 3**

| Month | Day of Week | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 3.66 | 3.1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0.24 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.24 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 3.12 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0.46 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 3.68 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 3.2. Sampling Design for Discriminating Group 1 and Group 4**

| Month | Day of Week | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 3.76 | 0 | 0 | 0 | 3.34 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 3.78 | 0 | 0 | 0 | 0 | 0 | 3.32 |
| 8 | 0.24 | 0 | 0 | 0 | 0 | 0 | 0.18 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0.34 | 0 | 0 | 0 | 0.04 | 0 |

**Table 3.3. Sampling Design for Discriminating Group 3 and Group 4**

| Month | Day of Week | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.12 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.28 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0.04 | 4.7 | 2.34 | 0.32 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 3.72 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 3.48 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 3.4. "Optimal" Sampling Design**

| Month | Day of Week | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 0 | 2 | 1 | 0 | 0 | 1 |
| 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 3.5. Classification Probabilities for ATRs in Group 3**

**"Optimal" Sampling Design**

| ATR # | Noninformative Prior | | | Empirical Bayes Prior | | |
|---|---|---|---|---|---|---|
| | P1 | P3 | P4 | P1 | P3 | P4 |
| 52 | 0.4 | 0.554 | 0.046 | 0.337 | 0.657 | 0.01 |
| 175 | 0.05 | 0.812 | 0.138 | 0 | 0.982 | 0.015 |
| 187 | 0.079 | 0.476 | 0.446 | 0 | 0.897 | 0.102 |
| 200 | 0.035 | 0.92 | 0.045 | 0 | 0.993 | 0.01 |
| 204 | 0.036 | 0.797 | 0.167 | 0 | 0.952 | 0.047 |
| 207 | 0.076 | 0.891 | 0.032 | 0.018 | 0.974 | 0.01 |
| 208 | 0 | 1 | 0 | 0 | 0.998 | 0 |
| 215 | 0.292 | 0.408 | 0.299 | 0.187 | 0.625 | 0.188 |
| 218 | 0.273 | 0.708 | 0.019 | 0.091 | 0.908 | 0 |
| 224 | 0.079 | 0.328 | 0.593 | 0.016 | 0.556 | 0.428 |

**Table 3.6. Classification Probabilities for ATRs in Group 4**

**"Optimal" Sampling Design**

| ATR # | Noninformative Prior | | | Empirical Bayes Prior | | |
|---|---|---|---|---|---|---|
| | P1 | P3 | P4 | P1 | P3 | P4 |
| 1 | 0.762 | 0 | 0.235 | 0.798 | 0 | 0.201 |
| 51 | 0.043 | 0.016 | 0.941 | 0 | 0 | 1 |
| 55 | 0.041 | 0 | 0.958 | 0.018 | 0 | 0.982 |
| 57 | 0.102 | 0.184 | 0.714 | 0.023 | 0.086 | 0.891 |
| 214 | 0.066 | 0 | 0.933 | 0.01 | 0 | 0.993 |
| 220 | 0.018 | 0.171 | 0.81 | 0 | 0.02 | 0.98 |
| 221 | 0.029 | 0.01 | 0.965 | 0.01 | 0 | 0.989 |
| 223 | 0 | 0 | 0.997 | 0 | 0 | 0.999 |

**Table 3.7. Classification Probabilities for ATRs in Group 1**

**"Optimal" Sampling Design**

| ATR # | | Noninformative Prior | | | | Empirical Bayes Prior | | |
|---|---|---|---|---|---|---|---|---|
| | | P1 | P3 | P4 | | P1 | P3 | P4 |
| 2 | | 0.95 | 0.015 | 0.034 | | 0.977 | 0.021 | 0 |
| 3 | | 0.975 | 0 | 0.021 | | 0.985 | 0.01 | 0.01 |
| 7 | | 0.998 | 0 | 0 | | 0.998 | 0 | 0 |
| 8 | | 0.902 | 0.032 | 0.066 | | 0.927 | 0.036 | 0.037 |
| 9 | | 0.984 | 0 | 0.012 | | 0.988 | 0.01 | 0 |
| 10 | | 0.976 | 0 | 0.021 | | 0.993 | 0 | 0.01 |
| 12 | | 0.78 | 0.03 | 0.19 | | 0.926 | 0.016 | 0.058 |
| 14 | | 0.937 | 0.01 | 0.055 | | 0.977 | 0.01 | 0.013 |
| 50 | | 0.946 | 0.03 | 0.024 | | 0.947 | 0.032 | 0.02 |
| 54 | | 0.218 | 0.048 | 0.733 | | 0.258 | 0.022 | 0.72 |
| 56 | | 0.983 | 0.01 | 0.01 | | 0.988 | 0.01 | 0.01 |
| 100 | | 0.988 | 0.01 | 0 | | 0.996 | 0 | 0 |
| 102 | | 0.901 | 0.018 | 0.081 | | 0.997 | 0 | 0 |
| 103 | | 0.927 | 0.022 | 0.05 | | 0.992 | 0 | 0.01 |
| 104 | | 0.822 | 0.052 | 0.126 | | 0.951 | 0.01 | 0.04 |
| 110 | | 0.973 | 0.01 | 0.023 | | 0.995 | 0 | 0 |
| 164 | | 0.046 | 0.573 | 0.38 | | 0.038 | 0.576 | 0.387 |
| 166 | | 0.832 | 0.056 | 0.111 | | 0.916 | 0.022 | 0.062 |
| 170 | | 0.01 | 0.073 | 0.919 | | 0.01 | 0.049 | 0.944 |
| 172 | | 0.032 | 0.044 | 0.924 | | 0.018 | 0.019 | 0.963 |
| 179 | | 0.33 | 0.489 | 0.181 | | 0.34 | 0.472 | 0.188 |
| 197 | | 0.534 | 0.457 | 0.01 | | 0.531 | 0.464 | 0.01 |
| 199 | | 0.911 | 0.048 | 0.041 | | 0.907 | 0.054 | 0.039 |
| 211 | | 0.226 | 0.039 | 0.735 | | 0.159 | 0.01 | 0.836 |
| 213 | | 0 | 0.988 | 0.01 | | 0 | 0.99 | 0.01 |
| 216 | | 0.076 | 0.171 | 0.753 | | 0.024 | 0.105 | 0.871 |
| 217 | | 0.994 | 0 | 0 | | 0.997 | 0 | 0 |
| 219 | | 0.294 | 0.188 | 0.518 | | 0.285 | 0.128 | 0.587 |
| 225 | | 0.827 | 0.03 | 0.143 | | 0.902 | 0.012 | 0.086 |
| 226 | | 0.512 | 0.476 | 0.012 | | 0.385 | 0.598 | 0.017 |

In order to evaluate the utility of the above sample design method, two additional sample designs were generated. The first of these new designs called for making two week-long counts, one in March and one in July, while the second of these new designs called for making three 5-day weekday counts, one in March, one in July and one in November. These designs were selected to more closely approximate common counting practice. These designs were also converted to actual samples from the 1991 ATR data, and posterior classification probabilities were computed using the Empirical Bayes prior. These results are displayed in Tables 3.8, 3.9 and 3.10. Interestingly, the performance of the "Full-week" sample is comparable to that of our "Optimal" sample, but Table 3.8 shows that the "Weekday" sample generally does a poor job of classifying ATRs in factor group 3. This is not surprising since Figure 2.2 shows that Sunday counts are best for identifying group 3, and the "Weekday" samples contain no Sunday counts.

**Table 3.8. Classification Probabilities for ATRs in Group 3**

**Empirical Bayes Prior**

| ATR # | "Weekday" Sample | | | "Full-Week" Sample | | |
|---|---|---|---|---|---|---|
| | P1 | P3 | P4 | P1 | P3 | P4 |
| 52 | 0.941 | 0.058 | 0 | 0.754 | 0.125 | 0.121 |
| 175 | 0.256 | 0.062 | 0.682 | 0.01 | 0.957 | 0.036 |
| 187 | 0.061 | 0.04 | 0.899 | 0.01 | 0.711 | 0.279 |
| 200 | 0.519 | 0.162 | 0.319 | 0 | 0.996 | 0 |
| 52 | 0.045 | 0.031 | 0.925 | 0 | 0.973 | 0.027 |
| 207 | 0.879 | 0.11 | 0.01 | 0.062 | 0.929 | 0.01 |
| 208 | 0.099 | 0.901 | 0 | 0 | 1 | 0 |
| 215 | 0.481 | 0.151 | 0.369 | 0.037 | 0.924 | 0.039 |
| 218 | 0.892 | 0.106 | 0 | 0.184 | 0.814 | 0 |
| 224 | 0.148 | 0.023 | 0.829 | 0.01 | 0.625 | 0.366 |

**Table 3.9. Classification Probabilities for ATRs in Group 4**

**Empirical Bayes Prior**

| ATR # | "Weekday" Sample | | | "Full-Week" Sample | | |
|---|---|---|---|---|---|---|
| | P1 | P3 | P4 | P1 | P3 | P4 |
| 1 | 0.163 | 0.079 | 0.758 | 0.776 | 0.024 | 0.199 |
| 51 | 0.015 | 0 | 0.983 | 0 | 0 | 0.998 |
| 55 | 0.014 | 0 | 0.984 | 0.086 | 0 | 0.914 |
| 57 | 0.521 | 0.063 | 0.416 | 0.23 | 0.177 | 0.593 |
| 214 | 0 | 0 | 0.997 | 0.01 | 0 | 0.995 |
| 220 | 0.01 | 0 | 0.99 | 0 | 0.029 | 0.967 |
| 221 | 0.01 | 0 | 0.994 | 0 | 0 | 0.999 |
| 223 | 0 | 0 | 0.997 | 0 | 0 | 1 |

**Table 3.10. Classification Probabilities for ATRs in Group 1**

**Empirical Bayes Prior**

| ATR # | | "Weekday" Sample | | | | "Full-Week" Sample | | |
|---|---|---|---|---|---|---|---|---|
| | | P1 | P3 | P4 | | P1 | P3 | P4 |
| 2 | | 0.9 | 0.077 | 0.023 | | 0.99 | 0.01 | 0 |
| 3 | | 0.899 | 0.098 | 0 | | 0.986 | 0.01 | 0 |
| 7 | | 0.845 | 0.076 | 0.079 | | 0.994 | 0 | 0 |
| 8 | | 0.908 | 0.075 | 0.018 | | 0.823 | 0.011 | 0.166 |
| 9 | | 0.796 | 0.186 | 0.018 | | 0.998 | 0 | 0 |
| 10 | | 0.989 | 0.011 | 0 | | 1 | 0 | 0 |
| 12 | | 0.594 | 0.36 | 0.046 | | 0.97 | 0.016 | 0.014 |
| 14 | | 0.637 | 0.359 | 0 | | 0.962 | 0.034 | 0 |
| 50 | | 0.571 | 0.084 | 0.345 | | 0.986 | 0.01 | 0.01 |
| 54 | | 0.563 | 0.239 | 0.198 | | 0.228 | 0.183 | 0.589 |
| 56 | | 0.716 | 0.275 | 0.01 | | 0.998 | 0 | 0 |
| 100 | | 0.926 | 0.071 | 0 | | 0.999 | 0 | 0 |
| 102 | | 0.725 | 0.152 | 0.123 | | 1 | 0 | 0 |
| 103 | | 0.681 | 0.175 | 0.143 | | 1 | 0 | 0 |
| 104 | | 0.631 | 0.145 | 0.224 | | 1 | 0 | 0 |
| 110 | | 0.54 | 0.156 | 0.304 | | 1 | 0 | 0 |
| 164 | | 0.494 | 0.45 | 0.055 | | 0.039 | 0.749 | 0.212 |
| 166 | | 0.554 | 0.055 | 0.391 | | 0.962 | 0 | 0.036 |
| 170 | | 0.01 | 0 | 0.994 | | 0 | 0.034 | 0.965 |
| 172 | | 0.531 | 0.128 | 0.342 | | 0.024 | 0 | 0.973 |
| 179 | | 0.275 | 0.076 | 0.649 | | 0.532 | 0.123 | 0.345 |
| 197 | | 0.598 | 0.39 | 0.012 | | 0.28 | 0.712 | 0.01 |
| 199 | | 0.653 | 0.097 | 0.25 | | 0.932 | 0.038 | 0.03 |
| 211 | | 0.325 | 0.039 | 0.636 | | 0.229 | 0.047 | 0.725 |
| 213 | | 0.391 | 0.596 | 0.013 | | 0.01 | 0.992 | 0 |
| 216 | | 0.183 | 0.059 | 0.758 | | 0.011 | 0.438 | 0.552 |
| 217 | | 0.993 | 0 | 0 | | 1 | 0 | 0 |
| 219 | | 0.351 | 0.104 | 0.545 | | 0.042 | 0.873 | 0.085 |
| 225 | | 0.519 | 0.039 | 0.442 | | 0.906 | 0 | 0.09 |
| 226 | | 0.974 | 0.02 | 0.01 | | 0.658 | 0.314 | 0.028 |

# CHAPTER 4

# BAYES ESTIMATION OF MEAN DAILY TRAFFIC

In Chapter 3 we described a method for assigning coverage count locations to factor groups on the basis of information contained in traffic count samples, where the classification rule was derived using principles of Bayesian decision theory. In addition, we posed an optimization problem whose solution would give those sampling designs of a given size producing the lowest probability of mis-classifying a site. Although exact solution of this problem was not feasible, we developed a relatively simple heuristic solution which performed creditably when tested using Mn/DOT ATR data. In Chapter 1 however we pointed out that accurate factor group assignment was not an end in itself, but primarily a means for reducing the RMSPE of estimates of MDT. The most obvious way to use the methods of Chapter 3 is to first compute posterior classification probabilities using either (3.6-3.7) or (3.9-3.10), and then use the monthly and day-of-week factors characterizing the group with the highest posterior classification probability to correct the short count sample. It may be however that the appropriate factor group for a site may remain uncertain even after obtaining a count sample, so that one cannot say with high confidence which set of correction factors are appropriate. Also, because this method does not take into account uncertainty arising from mis-classification, the precision of the resulting estimates of MDT are likely to be overstated. Finally, there is no reason to believe that the resulting MDT estimates are the best that can be obtained from a given count sample. In this chapter we will develop a method for computing Bayes estimates of MDT which allow for posterior uncertainty concerning factor group membership, explicitly account for factor group uncertainty in computing posterior precision and which are optimal in the sense that on the average they will be closer to the true MDT than other estimators.

## BAYES ESTIMATOR OF MDT

Informally, developing a Bayes estimator of MDT proceeds in a manner similar to developing the Bayes classification rule. We assume the analyst begins with a some prior ideas concerning the likely values of the MDT, and then combines this prior view with the data likelihood to obtain the

posterior probability distribution which characterizes his or her uncertainty about a site's MDT after the data become available. It is well known that the point estimate of MDT which minimizes the expected squared error between the estimate and the true MDT is simply the mean of this posterior distribution. This would be our Bayes estimator of MDT.

As was stated in Chapters 1 and 2, the starting point of our analysis is that the expected daily traffic volume of day t, $E[z_t]$, occurring during month i and on day-of-week j, is given by

$$E[z_t] = z_0 M_i W_j \tag{4.1}$$

where $z_0$ is the MDT, $M_i$ the monthly adjustment term for month i, $W_j$ is the day-of-week adjustment term for day j. In Chapter 2 it was also shown that lognormal regression models provide defensible statistical models for daily traffic volumes, and that the lognormal version of (4.1) takes the form

$$E[z_t] = (e^{u+\sigma^2/2})(e^{m_i})(e^{w_j}) \tag{4.2}$$

so that $z_0=\exp(u+\sigma^2/2)$ is the expected daily traffic volume on a typical day, the MDT. Since $z_0$ is a function of the parameters u and $\sigma$, assuming a prior distribution for these parameters in turn induces a prior distribution for $z_0$, and in Chapter 3 we described how a noninformative prior for u could be combined with either a noninformative prior, or with an informative prior for $\sigma$ and the autocorrelation parameters $\phi_1$ and $\phi_7$. Thus two of the basic components for a Bayesian analysis, the data likelihood function and the prior parameter distributions are already at hand. The third necessary component is a loss function, and for this problem we will use the square of the distance between the true MDT and the estimate as our loss function, that is

$$L(z_0,\hat{z}) = (z_0-\hat{z})^2 \tag{4.3}$$

where $\hat{z}$ is an estimate of the MDT $z_0$, and we will seek the estimator which minimizes the posterior expected loss

$$E[L|y] = E[(z_0 - \hat{z})^2|y] \tag{4.4}$$

Thus our Bayes estimator of MDT, which we will denote by $z_B$, can be interpreted as producing estimates which can be expected to be closer to the true MDT than any other estimates, given the available data. Since it is well known (Berger, 1988) that the estimator minimizing (4.4) is the posterior expected value of $z_0$

$$z_B = E[z_0|y] = \int_0^\infty z_0 f(z_0|y) dz_0 \tag{4.5}$$

this estimation problem is in principle solved one we can determine the posterior distribution $f(z_0 \mid y)$.

Bayesian estimation of lognormal models for the case where no serial correlation is present (that is, where the AR parameters in our model satisfy $\phi_1=\phi_7=0$) has been treated in the statistical literature, especially in Zellner (1971) and Rukhin (1986). In particular, Zellner shows that if one uses a noninformative prior for u but assumes that $\sigma$ is known exactly, the posterior density of $z_0$ is also lognormal, and the Bayes estimator of $z_0$ is simply

$$z_B = e^{\bar{y}+\sigma^2/2+\sigma^2/2N} \tag{4.6}$$

where $\bar{y}$ is the sample average of **y**. However, when one assumes the noninformative prior for $\sigma$, $g(\sigma)=1/\sigma$, things become much more complicated. The posterior density can be expressed using modified Bessel functions, but the posterior mean of $z_0$ does not exist (Zellner, 1971, pp. 327-328). These results suggest the using the noninformative prior from Chapter 3 is not likely to produce a practical estimator of MDT. When we turn to the informative prior based on estimates of the parameters $\phi_1,\phi_7$, and $\sigma$ obtained form ATR data, things are much more promising, and it turns out that a Bayes estimator of MDT can be computed with about the same effort as the classification

probabilities derived in Chapter 3.

Rather than attempt to characterize the posterior probability density of $z_0$ and then computing the $z_B$ via equation (4.5), the discrete nature of the EB prior permits a more direct derivation. To begin with, suppose that we know the site of interest belongs to factor group k, and that $\phi_{1p}$, $\phi_{7p}$, $\sigma_p$, are the correct values for the covariance parameters. Assuming the noninformative prior for the parameter u then permits us to employ Zellner's argument to show that the posterior mean of $z_0=\exp(u+\sigma^2/2)$ is given by

$$\bar{z}_{pk} = E[z_0|y,\phi_{1p},\phi_{7p},\sigma_p,\beta_k] = e^{\bar{y}_{pk}+\frac{\sigma^2}{2}\left(1+\frac{1}{1_N' V_p^{-1} 1_N}\right)} \tag{4.7}$$

The definitions appearing in equations (3.10) also apply to equations (4.7-4.12). Using the empirical estimate of the prior for the covariance parameters, the expected MDT given the sample and that the site belongs to group k is then

$$\bar{z}_k = E[z_0|y,\beta_k] = \frac{\sum_{p=1}^{M} \bar{z}_{pk}\sigma_l^{-(n-1)}|V_p|^{-1/2}(1_N' V_p^{-1} 1_N)^{-1/2} e^{-\frac{1}{2}\frac{s_{pk}^2}{\sigma_p^2}}}{S_k} \tag{4.8}$$

Finally, the expected MDT given only the sample can be computed via

$$z_B = E[z_0|y] = \sum_{k=1}^{n} \bar{z}_k Prob[\theta=\beta_k|y] \tag{4.9}$$

The intermediate quantities appearing in (4.7-4.9) are identical to those used in computing the posterior classification probabilities, so that computing a Bayes estimate of MDT is no more difficult than computing the classification probabilities. To compute the posterior variance of $z_0$, it is first noted that

$$v_k^2 = var[z_0|y,\beta_k] = \sum_{p=1}^{M}\left(e^{2\bar{y}_{pk}+\sigma_p^2(1+\frac{2}{1_N' V_p^{-1} 1_N})}\right) Prob[\phi_{1p},\phi_{7p},\sigma_p|y,\beta_k] - \bar{z}_k^2 \tag{4.10}$$

where

$$Prob[\phi_{1p},\phi_{7p},\sigma_p|y,\beta_k] = \frac{\sigma_p^{-(N-1)}|V_p|^{-1/2}(1_N' V_p^{-1} 1_N)\exp(-\frac{s_{pk}^2}{2\sigma_p^2})}{\sum_{p=1}^{M}\sigma_p^{-(N-1)}|V_p|^{-1/2}(1_N' V_p^{-1} 1_N)\exp(-\frac{s_{pk}^2}{2\sigma_p^2})} \quad (4.11)$$

The posterior variance of $z_0$ is then given by

$$v_B^2 = var[z_0|y] = \sum_{k=1}^{n} v_k^2 Prob[\theta=\beta_k|y] + \sum_{k=1}^{n}(\bar{z}_k - z_B)^2 Prob[\theta=\beta_k|y] \quad (4.12)$$

## EVALUATION OF BAYES ESTIMATOR OF MDT

Using Mn/DOT's 1991 ATR data, the Bayes estimator of MDT was compared to two other estimators. Treating the Bayes estimator as our "first" estimator, the second estimator was a "Factored Average" of the counts in a sample, in which the three outstate factor groups used earlier were again assumed to be characterized by the monthly and day-of-week multipliers displayed in Figure 2.2. Each sample of traffic counts was first assigned to the factor group having the highest posterior probability, as computed using equations (3.9-3.10), and then the counts were divided by the appropriate monthly and day-of-week terms characteristic of that group. These factored counts were then averaged to estimate MDT. The third estimator was the simple average of the counts in a sample, without factoring or any other adjustment. The samples obtained using the three different sample designs described in Chapter 3, the "Optimal", the "Full-Week" and the "Weekday" were all used, producing a total of nine different estimates of MDT for each of the 48 ATRs having complete 1991 count data. As a reference, the AADT for each of these 48 ATRs was also computed, using the *AASHTO Guidelines* (AASHTO,1992, p. 52) successive averaging method, which accounts for missing data. The sampled estimates of MDT together with the AADT for each ATR are listed in Appendix A. Absolute percent errors (APE), the absolute value of PE as defined in equation (1.3),

were then computed for each estimator and each sampling plan, and these were then used to compute three performance measures. The first performance measure was the average APE, where the average was taken over the 48 ATR sites. The second measure was the maximum APE, so that these two measures gave the average and worst case performance, respectively. In addition, the standard deviation of the PE values was computed to give an estimate of RMSPE. Table 4.1 displays these results.

To interpret Table 4.1, consider the measures given for the "Optimal" sample design and "Bayes" estimator. On the average, the Bayes estimates generated from "Optimal" sample differed form the AADT by about 5.34%, and at most the Bayes estimates differed from the AADT by 16.98%. The estimated RMSPE for this estimator/sample combination was 5.44%.

**Table 4.1 Performance Comparison of Three Estimators of MDT**
**All Table Entries are Percentages**

| SAMPLING DESIGN | PERFORMANCE MEASURE | Bayes | Factored Average | Simple Average |
|---|---|---|---|---|
| **"Optimal":** | **Average APE:** | 5.34 | 5.80 | 9.30 |
| | **Maximum APE:** | 16.98 | 20.04 | 26.43 |
| | **RMSPE:** | 5.44 | 5.46 | 11.28 |
| **Full-Week:** | **Average APE:** | 4.64 | 4.92 | 8.95 |
| | **Maximum APE:** | 22.14 | 19.30 | 30.39 |
| | **RMSPE:** | 5.91 | 5.42 | 11.28 |
| **Weekday:** | **Average APE:** | 7.99 | 8.01 | 7.62 |
| | **Maximum APE:** | 28.83 | 29.12 | 32.48 |
| | **RMSPE:** | 9.49 | 8.80 | 9.92 |

Looking at the Average APEs, for the "Optimal" and "Full-Week" sampling designs the Bayes estimator and the Factored Average are roughly equivalent, and clearly dominate the Simple Average, while for the "Weekday" design the three estimators are roughly equivalent. This pattern of results also appears when considering the Maximum APE and RMSPE, and looking across sample designs it can be seen that the Bayes and Factored Average estimates tend to be sensitive to the quality of the

sample while the Simple Average appears more robust with regard to sample design. The similarity between the Bayes estimator and the Factored Average is partly explained by the fact that, as is shown in equation (4.8), the Bayes estimator weights its components using the posterior classification probabilities while Tables 3.8-3.10 show that for most sites, a large amount of posterior probability tended to be given to only one factor group. Particularly encouraging is the rough equivalence between the "Optimal" and "Full-Week" sampling designs, since two 1-week counts will generally be much easier to obtain than counts collected according to the "Optimal" design listed in Table 3.4.

To compare the results in Table 4.1 to those reported in other studies, it recall that Sharma and Leng reported, for samples consisting of two 1-week counts, PI95 values of 10%, 12%, 15% and 21%, for estimates of MDT from different roadway types (Sharma and Leng, 1994, pp. 22 & 24). Dividing these PI95 values by 1.96 gives estimated RMSPE values of 5.1%, 6.1%, 7.7% and 10.7%., while the range of RMSPE values in Table 4.1 is 5.42% to 11.28%. Also, Petroff (1956) reported an estimated RMSPE of 4.2% for samples consisting of 7-day counts repeated quarterly during the year. Thus the estimates of RMSPE reported in Table 4.1 are comparable to other estimates reported in the literature, but with one important difference. In Chapter 1 it was argued that the estimates of RMSPE appearing in the earlier work should be interpreted as applying only to MDT estimation situations with good prior information concerning factor group membership. In contrast, our Bayes and Factored Average estimators started with complete uncertainty concerning factor group membership, and used information in the count sample to determine the appropriate correction.

Finally, the effect of factor group assignment error can be seen clearly in Table 4.1. The Simple Average estimator ignored information concerning factor group membership, while the results in Chapter 3 showed that the "Weekday" samples had difficulty correctly assigning ATRs in factor group 3, so that estimates computed using the "Weekday" sample are more subject to assignment error. In contrast, the Bayes estimator and the Factored Average attempt to use membership information, and both the "Optimal" and "Full-Week" samples tended to provide good membership information. Thus those estimator/sample combinations which used good membership information showed RMSPE values ranging between 5.42% and 5.91%, while those combinations which either had poor membership information or did not use the information showed RMSPE values ranging between 8.80% and 11.28%.

# CHAPTER 5
# SUMMARY AND CONCLUSIONS

## SUMMARY

This report began by considering some of the statistical issues arising when one attempts to estimate an annual mean daily traffic from a sample of daily traffic volumes. Two sources of error were identified, one caused primarily by the inherent random variability in daily traffic volumes, the other due to bias in the estimation procedure. A review of early research indicated that initial estimates of the likely error arising from using short counts to estimate MDT were computed under experimental conditions which minimized or eliminated the effect of bias. On the other hand, a computational example showed that bias caused by the use of incorrect monthly and day-of-week correction factors could be the dominant source of error, and a review of more recent research led to the conclusion that if bias due to incorrect seasonal adjustment could be eliminated, a 48-hour short count could give a reasonable estimate of MDT, but that an assignment accuracy equivalent to that provided by two 1-week counts appeared necessary to essentially eliminate this bias.

Using ATR data supplied by Mn/DOT, it was then argued that a class of lognormal regression models provided defensible statitistical models for the daily traffic counts obtained from 50 ATRs in outstate Minnesota. These regression models also provided a method for estimating the monthly and day-of-week adjustment terms which characterize a factor group, and the estimates of these adjustment terms were computed for three factor groups used by Mn/DOT in 1992. Bayesian decision methods were then used to derive formulas for computing the probability a count sample obtained from some highway site is consistent with the monthly and day-of-week pattern characterizing a factor group, and this analysis was in turn used to design counting samples which minimize the likelihood of misclassifying a site. Finally, Bayesian methods were used to derive an estimator of MDT which used the factor group assignment probabilities developed earlier, and had the property that on average its estimates should be closer to the true MDT than those produced by other procedures. Both the classification and estimation methods were evaluated using actual ATR data.

## CONCLUSIONS

The following conclusions can be drawn from this research. First, the class of lognormal regression models, coupled with a time-series model to account for serial correlation in the regression residuals, provide accurate yet tractable descriptions of the statistical properties of bi-directional daily traffic counts, at least for outstate Minnesota. This is an especially important development since possession of such a model allows an analyst to apply a wide range of sophisticated statistical methods to solve problems in traffic estimation, and it is quite possible that the lack of such fundamental model development has retarded the development of optimal traffic estimation methods. This report's remaining conclusions follow directly from identification of this class of models.

Second, assuming that the three factor groups used in this study are representative of all outstate highway sites, then about two weeks worth of well-selected daily traffic counts from a site should be sufficient to match that site to a factor group, and to compute a reasonably accurate estimate of the site's mean daily traffic. "Well-selected" in this case means that the count sample focuses on those days for which the different factor groups give the most discrepant predictions of traffic volumes. In particular, since factor group 3 in this study was characterized by an unusual weekend pattern, eliminating weekend days from the count sample led to a marked decline in accuracy of assigning sites to factor groups, and a noticeable increase in MDT estimation error.

Third, given that one can characterize each factor group of interest with a set of monthly and day-of-week adjustment terms, well-selected count samples can be easily determined using the heuristic described in Chapter 3. Although the "optimal" sample design found by the heuristic tended to be difficult to implement because it scattered the sampling days throughout the year, a simpler design consisting of two 1-week counts, selected to roughly correspond to the "optimal" design, turned out to perform as well as the optimal design.

## SUGGESTIONS FOR IMPLEMENTATION

Chapters 2-4 of this report described a unified set of procedures for first estimating the monthly and day-of-week terms characterizing a factor group using regression methods, developing an empirical estimate of the prior distribution for the noise parameters characterizing a site, and then combining this prior information with a count sample to match a site to a factor group and to estimate

the site's MDT. These procedures will in almost all cases require a computer for their timely completion. The implementation developed here used a combination of MINITAB and MATLAB macros to carry out all computations, and these have been listed in Appendix B. In its final form, our analysis system used MINITAB version 10+ and MATLAB version 4.2, running under Windows 3.11 on a Pentium 90 MHZ microcomputer. A user with access to this software on equivalent equipment can simply enter the procedures listed in Appendix B as text files, and will then be able to perform any of the procedures described in this report. The initial management of the ATR data files obtained for Mn/DOT, and creation of the files of daily traffic count files used in this research, was performed using DBASE III+.

A user having access to the SAS statistical analysis system should also be able to implement all procedures described in this report, but at the cost of having to write the SAS equivalents of the macros listed in Appendix B. This should not be difficult once the nature of the computations is understood, and has the potential advantage of incorporating all data management and analysis procedures within one package. The SAS system was not used in this project due to the high annual license cost for the complete system.

# REFERENCES


AASHTO (1992) *Guidelines for Traffic Monitoring Programs*, American Association of State Highway and Transportation Officials, Washington, DC.

Albright, D. (1987) A Quick Cluster Control Method: Permanent Control Station Cluster Analysis in Average Daily Traffic Calculations, *Transportation Research Record*, 1134, pp. 57-64.

Albright, D. (1991a) History of Estimating and Evaluating Annual Traffic Volume Statistics, *Transportation Research Record*, 1305, pp. 103-107.

Albright,D. (1991b) An Imperative for,and Current Progress toward, National Traffic Monitoring Standards, *ITE Journal*, June 1991, pp. 22-26.

Albright, D.(1991c) The Development of ASTM Highway Monitoring Standards, Standardization News, 19, February, 1991, pp. 22-27.

Atkinson, A. and Federov, V. (1975a) The Design of Experiments for Discriminating between Two Rival Models, *Biometrika*, 62, pp. 57-69.

Atkinson, A. And Federov, V. (1975b) Optimal Design: Experiments for Discriminating Between Several Models, *Biometrika*, 62, pp. 289-303.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer, Verlag, New York.

Bodle, R. (1967) Evaluation of Rural Coverage Count Duration for Estimating Annual Average Daily Traffic, Highway Research Record, 199, pp. 67-72

Box, G., Jenkins, G. and Reinsel, G.(1994) *Time Series Analysis: Forecasting and Control, 3rd Edition*, Prentice-Hall, Englewood Cliffs, NJ.

Brockwell, P, and Davis, R.(1991) *Time Series: Theory and Methods*, New York, Springer-Verlag.

Caliper, (1990) *TransCAD Reference Manual, version 2.0*, Newton, MA, Caliper Corporation.

Davis, G.A. (1993) A Statistical Theory for Estimation of Origin-Destination Parameters from Time-Series of Traffic Counts, in C. Daganzo (ed.) *12th International Symposium on Transportation and Traffic Theory*, Amsterdam, Elsevier, pp. 441-464.

Dunsmuir, W. and Robinson, P. (1981) Estimation of Time Series Models in the Presence of Missing Data, *J. of American Statistical Association*, 76, pp. 560-568.

Faghri, A., Chakroborty, P. and Parameswran, J. (1992) An Analysis of a State's Traffic Counting Program, presented at 71st TRB Meeting, Washington, DC.

FHWA (1985) *Traffic Monitoring Guide*, U.S. Department of Transportation, Washington, DC.

Francis, B. (1993) *The GLIM System, Release 4 Manual*, Oxford, UK, Clarendon Press.

Guan, Y. (1995), *Bayesian Classification and Kriging in the Estimation of Average Daily Traffic*, unpublished Master's Thesis, Department of Civil Engineering, University of Minnesota.

Gulati, B. (1995) *Precision of AADT Estimates from Short Period Traffic Counts*, unpublished Master's Thesis, University of Regina, Regina, Saskatchewan, Canada.

Hua, S. (1994) *Approximate Maximum Likelihood Estimation of Origin-Destination Parameters from Time-Series of Traffic Counts*, unpublished Master's Thesis, Department of Civil Engineering, University of Minnesota.

Hultgren, P. (1996) *Calibration of Gravity Model Parameters Using Traffic Count Data*, unpublished Master's Thesis, Department of Civil Engineering, University of Minnesota.

Lindsay, B. (1983) The Geometry of Mixture Likelihoods: General Theory, *Annals of Statistics*, 11, pp. 86-94.

McShane, W. and Roess, R. (1990) *Traffic Engineering*, Prentice-Hall, Englewood Cliffs, NJ.

Madansky, A. (1988) *Prescriptions for Working Statisticians*, Springer-Verlag, New York.

Mathworks (1992) *MATLAB Version 4*, Mathworks, Inc.

Memmott, J. (1987) Adequacy of the Sample Size and Accuracy of the Highway Performance Monitoring System for Use in Texas, *Transportation Research Record*, 1132, pp. 32-39.

Minitab (1992). *Minitab Reference Manual, Release 9*, Minitab, Inc., State College, PA.

Monahan, J. (1983) Fully Bayesian Analysis of ARMA Time Series Models, *Journal of Econometrics*, 21, pp. 307-331.

Padgett, W. and Robinson, J. Empirical Bayes Estimators of Reliability for Lognormal Failure Models, *IEEE Trans. on Reliability*, R-27, 1978, pp. 332-335.

Petroff, B. (1946) Some Criteria for Scheduling Mechanical Traffic Counts, *Proceedings of Highway Research Board*, 26, pp. 389-396.

Petroff, B. and Blensley, R. (1954) Improving Traffic-Count Procedures by Application of Statistical Method, *Proceedings of Highway Research Board*, 33, pp. 362-375.

Petroff, B. (1956) Experience in Application of Statistical Method to Traffic Counting, *Public Roads*, 29, pp. 110-117.

Ritchie, S. (1985) A Statistical Approach to Statewide Traffic Counting, *Transportation Research Record*, 1090, pp. 14-21.

Ritchie, S. and Hallenbeck, M. (1985) Evaluation of a Statewide Highway Data Collection Program, *Transportation Research Record*, 1090, pp. 27-35.

Rohatgi, V. (1976) *An Introduction to Probability and Mathematical Statistics*, New York, Wiley and Sons.

Rukhin, A. (1986) Improved Estimation in Lognormal Models, *Journal of American Statistical Association*, 81, pp. 1046-1049.

Sharma, S. and Allipuram, R. (1993) Duration and Frequency of Seasonal Traffic Counts, *ASCE J. of Transportation Engineering*, 119, pp. 344-359.

Sharma, S. and Leng, Y. (1994) Seasonal Traffic Counts for a Precise Estimation of AADT, *ITE Journal*, September, pp. 21-28.

Sharma, S. and Werner, A. (1981) Improved Method for Grouping Provincewide Permanent Traffic Counters, *Transportation Research Record*, 815, pp. 13-18.

Silvey, S. (1980) *Optimal Design*, London, Chapman and Hall.

Zellner, A.(1971) Bayesian and Non-Bayesian Analysis of the Log-normal Distribution and Lognormal Regression, *J. of American Statistical Association*, 66, pp. 327-330.

Zellner, A. (1987) *An Introduction to Bayesian Inference in Econometrics, reprint edition*, Malabar, FL, Krieger Publishing Company.

# APPENDIX A

## Full MDT Estimation Results

# APPENDIX A

# FULL MDT ESTIMATION RESULTS

**Table A.1. "Optimal Design", Factor Group 1**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 2 | 1238.753 | 194.738 | 1090.167 | 1112.329 | 1305.292 |
| 3 | 4211.228 | 678.45 | 3911.769 | 3871.81 | 4271.694 |
| 7 | 649.938 | 41.937 | 658.308 | 654.284 | 701.118 |
| 8 | 437.921 | 68.245 | 436.556 | 421.748 | 527.458 |
| 9 | 1190.96 | 130.088 | 1180.909 | 1150.231 | 1173.395 |
| 10 | 1606.787 | 142.665 | 1580.9 | 1595.224 | 1731.343 |
| 12 | 1614.538 | 298.29 | 1491.692 | 1448.763 | 1549.953 |
| 14 | 1318.407 | 217.464 | 1253.077 | 1235.09 | 1357.901 |
| 50 | 2451.945 | 208.919 | 2468.154 | 2436.356 | 2582.077 |
| 54 | 186.872 | 17.627 | 196.333 | 181.365 | 217.075 |
| 56 | 386.322 | 28.057 | 374.75 | 378.654 | 394.269 |
| 100 | 34469.285 | 2449.623 | 35742.555 | 33979.316 | 33844.844 |
| 102 | 44803.055 | 4357.236 | 46156.613 | 45754.938 | 50315.621 |
| 103 | 34048.223 | 3206.679 | 34017.309 | 33865.629 | 36120.281 |
| 104 | 24812.48 | 2689.175 | 25025.4 | 24902.607 | 26806.203 |
| 110 | 17893.221 | 1730.51 | 18177.77 | 18081.4 | 19580.789 |
| 164 | 7366.551 | 704.588 | 8363.7 | 7574.802 | 7403.124 |
| 166 | 3688.991 | 442.4 | 3313.5 | 3687.521 | 4010.983 |
| 170 | 2985.289 | 208.451 | 3331.308 | 2924.02 | 2934.033 |
| 172 | 10222.561 | 678.316 | 11180.385 | 9985.415 | 10734.038 |
| 179 | 1447.816 | 168.024 | 1537.538 | 1413.942 | 1499.237 |
| 197 | 3250.902 | 261.676 | 3340.769 | 3351.815 | 3270.84 |
| 199 | 1095.692 | 75.917 | 1116.333 | 1103.498 | 1166.172 |
| 211 | 2800.623 | 251.71 | 3056.539 | 2737.392 | 2877.626 |
| 213 | 5912.8 | 328.064 | 6483.1 | 5872.791 | 6005.351 |
| 216 | 1302.849 | 120.81 | 1476.077 | 1298.246 | 1352.916 |
| 217 | 671.597 | 46.807 | 654.75 | 668.057 | 705.2 |
| 219 | 2908.112 | 307.786 | 3169.545 | 2835.091 | 2824.446 |
| 225 | 1901.514 | 206.098 | 1642 | 1871.836 | 1956.861 |
| 226 | 377.67 | 36.958 | 399.385 | 371.287 | 401.85 |

**Table A.2. "Optimal" Sample Design, Factor Group 3**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 52 | 393.656 | 26.609 | 407.636 | 380.869 | 405.58 |
| 175 | 15237.404 | 1510.168 | 17188.537 | 15396.615 | 14392.916 |
| 187 | 12385.706 | 1118.96 | 15332.583 | 13150.414 | 12590.803 |
| 200 | 24911.949 | 2430.048 | 27388.154 | 24596.926 | 24905.41 |
| 204 | 5762.903 | 580.674 | 6648.75 | 5922.787 | 6516.196 |
| 207 | 1274.802 | 122.331 | 1388.769 | 1267.642 | 1246.808 |
| 208 | 17138.121 | 775.039 | 18804.832 | 17081.312 | 18292.523 |
| 215 | 701.523 | 112.645 | 613.857 | 647.354 | 643.446 |
| 218 | 465.76 | 55.144 | 488.385 | 464.873 | 541.12 |
| 224 | 1803.384 | 187.031 | 2062.385 | 1840.09 | 1875.058 |

**Table A.3 "Optimal" Sample Design, Factor Group 4**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 1 | 1569.684 | 164.764 | 1735.154 | 1689.734 | 1734.453 |
| 51 | 429.239 | 41.773 | 558.154 | 445.96 | 466.764 |
| 55 | 564.334 | 46.362 | 668.231 | 573.982 | 621.495 |
| 57 | 1257.005 | 136.192 | 1089 | 1257.776 | 1277.032 |
| 214 | 452.965 | 33.755 | 546 | 456.421 | 502.598 |
| 220 | 2002.346 | 183.295 | 2397.615 | 2014.099 | 1896.391 |
| 221 | 1972.266 | 182.815 | 2346.154 | 1980.54 | 2003.977 |
| 223 | 2699.844 | 150.197 | 3248.182 | 2648.809 | 2651.259 |

**Table A.4. Full-Week Sample Design, Factor Group 1**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 2 | 1263.768 | 208.749 | 1246.833 | 1247.67 | 1305.292 |
| 3 | 4300.065 | 655.746 | 4381.214 | 4151.588 | 4271.694 |
| 7 | 676.263 | 64.671 | 704 | 668.638 | 701.118 |
| 8 | 410.679 | 79.934 | 521.714 | 425.643 | 527.458 |
| 9 | 1201.457 | 105.976 | 1225.692 | 1186.664 | 1173.395 |
| 10 | 1684.123 | 96.941 | 1731.083 | 1672.041 | 1731.343 |
| 12 | 1763.324 | 319.483 | 1810.786 | 1699.864 | 1549.953 |
| 14 | 1413.923 | 256.003 | 1425 | 1365.465 | 1357.901 |
| 50 | 2538.444 | 152.383 | 2665.5 | 2524.611 | 2582.077 |
| 54 | 198.822 | 24.653 | 213.214 | 190.151 | 217.075 |
| 56 | 384.683 | 28.887 | 404.643 | 380.646 | 394.269 |
| 100 | 33104.754 | 2386.608 | 33965.715 | 32483.234 | 33844.844 |
| 102 | 49566.125 | 3751.277 | 50788.715 | 48569.449 | 50315.621 |
| 103 | 35368.918 | 1978.915 | 36484.715 | 34877.066 | 36120.281 |
| 104 | 26509.891 | 984.024 | 27679.215 | 26288.482 | 26806.203 |
| 110 | 19573.686 | 1218.55 | 20520.643 | 19521.295 | 19580.789 |
| 164 | 7232.262 | 657.946 | 7854.214 | 7253.558 | 7403.124 |
| 166 | 4168.798 | 296.332 | 3736.286 | 4166.759 | 4010.983 |
| 170 | 2948.837 | 177.425 | 3457.071 | 2914.961 | 2934.033 |
| 172 | 10285.646 | 684.281 | 11823.071 | 10219.674 | 10734.038 |
| 179 | 1458.018 | 131.401 | 1570.071 | 1472.236 | 1499.237 |
| 197 | 3144.444 | 266.73 | 3278.357 | 3109.3 | 3270.84 |
| 199 | 1089.394 | 79.186 | 1169.154 | 1092.842 | 1166.172 |
| 211 | 2864.645 | 242.454 | 3224.143 | 2769.827 | 2877.626 |
| 213 | 5718.13 | 311.923 | 6105.214 | 5684.377 | 6005.351 |
| 216 | 1328.913 | 135.374 | 1509.214 | 1276.955 | 1352.916 |
| 217 | 640.072 | 50.383 | 664.857 | 634.135 | 705.2 |
| 219 | 3020.058 | 272.944 | 3237.833 | 3038.259 | 2824.446 |
| 225 | 1970.239 | 249.282 | 1682.143 | 1882.3 | 1956.861 |
| 226 | 372.762 | 31.681 | 388.214 | 369.826 | 401.85 |

**Table A.5. Full-Week Sample Design, Factor Group 3**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 52 | 377.167 | 30.077 | 394.643 | 378.12 | 405.58 |
| 175 | 15055.944 | 1532.879 | 16351.857 | 14945.757 | 14392.916 |
| 187 | 13042.625 | 1766.69 | 15620.333 | 13471.808 | 12590.803 |
| 200 | 24006.086 | 1952.106 | 25739.07 | 23799.623 | 24905.41 |
| 204 | 5700.601 | 747.898 | 6662.357 | 5806.22 | 6516.196 |
| 207 | 1269.42 | 90.169 | 1353.357 | 1261.808 | 1246.808 |
| 208 | 17603.496 | 697.194 | 18766.5 | 17509.598 | 18292.523 |
| 215 | 665.94 | 82.861 | 563.143 | 647.292 | 643.446 |
| 218 | 458.637 | 65.971 | 472.714 | 467.679 | 541.12 |
| 224 | 1931.633 | 200.854 | 2140.286 | 1952.984 | 1875.058 |

**Table A 6. Full-Week Sample Design, Factor Group 4**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 1 | 1806.539 | 250.277 | 1953.071 | 1790.321 | 1734.453 |
| 51 | 454.363 | 63.083 | 599.143 | 454.872 | 466.764 |
| 55 | 586.57 | 47.646 | 698.929 | 577.797 | 621.495 |
| 57 | 1150.062 | 126.302 | 889 | 1202.426 | 1277.032 |
| 214 | 468.538 | 33.409 | 579.429 | 468.217 | 502.598 |
| 220 | 1909.615 | 223.835 | 2458.357 | 1928.101 | 1896.391 |
| 221 | 2047.992 | 116.749 | 2528.857 | 2046.75 | 2003.977 |
| 223 | 2529.054 | 117.518 | 3086.429 | 2537.13 | 2651.259 |

**Table A.7. Weekday Sample Design, Factor Group 1**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 2 | 1442.952 | 154.323 | 1437 | 1459.445 | 1305.292 |
| 3 | 4679.739 | 360.245 | 4759.667 | 4612.842 | 4271.694 |
| 7 | 690.498 | 79.11 | 706.867 | 683.463 | 701.118 |
| 8 | 521.994 | 55.156 | 578.2 | 532.259 | 527.458 |
| 9 | 1220.628 | 137.496 | 1261.444 | 1197.479 | 1173.395 |
| 10 | 1774.414 | 71.524 | 1818.692 | 1783.815 | 1731.343 |
| 12 | 1893.256 | 237.475 | 1886 | 1804.308 | 1549.953 |
| 14 | 1653.58 | 271.524 | 1541.6 | 1499.36 | 1357.901 |
| 50 | 2350.38 | 354.346 | 2426.333 | 2321.528 | 2582.077 |
| 54 | 200.928 | 22.256 | 205.267 | 195.617 | 217.075 |
| 56 | 399.926 | 26.704 | 409.467 | 390.489 | 394.269 |
| 100 | 34757.809 | 2822.516 | 36505.602 | 33988.645 | 33844.844 |
| 102 | 49045.395 | 8655.972 | 49552.801 | 48080.09 | 50315.621 |
| 103 | 34088.012 | 5972.301 | 34141.266 | 32952.961 | 36120.281 |
| 104 | 25207.348 | 4098.61 | 25297.668 | 24253.197 | 26806.203 |
| 110 | 19287.453 | 3405.92 | 19335.668 | 18559.986 | 19580.789 |
| 164 | 7051.922 | 646.971 | 7552.1 | 6840.228 | 7403.124 |
| 166 | 4262.319 | 715.591 | 3638.6 | 3912.32 | 4010.983 |
| 170 | 2897.82 | 130.077 | 3021.333 | 2895.396 | 2934.033 |
| 172 | 10739.834 | 760.702 | 11155.133 | 10509.165 | 10734.038 |
| 179 | 1383.419 | 246.034 | 1412.4 | 1415.47 | 1499.237 |
| 197 | 2980.281 | 223.196 | 3025.467 | 2879.718 | 3270.84 |
| 199 | 1072.482 | 162.954 | 1118.786 | 1067.795 | 1166.172 |
| 211 | 2718.858 | 311.072 | 2795.133 | 2739.186 | 2877.626 |
| 213 | 5632.057 | 421.719 | 5909 | 5764.28 | 6005.351 |
| 216 | 1250.74 | 144.523 | 1302.133 | 1255.407 | 1352.916 |
| 217 | 671.303 | 32.161 | 699.333 | 675.869 | 705.2 |
| 219 | 2616.678 | 443.961 | 2677.846 | 2693.857 | 2824.446 |
| 225 | 2211.242 | 268.147 | 1877.8 | 2012.186 | 1956.861 |
| 226 | 358.562 | 18.896 | 373.2 | 358.326 | 401.85 |

**Table A.8. Weekday Sample Design, Factor Group 3**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 52 | 357.751 | 19.223 | 381.8 | 352.974 | 405.58 |
| 175 | 12770.388 | 1738.113 | 13144.667 | 12816.264 | 14392.916 |
| 187 | 10870.194 | 1860.596 | 11863.231 | 10642.218 | 12590.803 |
| 200 | 20607.844 | 3360.305 | 21010.533 | 19950.887 | 24905.41 |
| 204 | 4637.32 | 544.111 | 4948.067 | 4618.396 | 6516.196 |
| 207 | 1228.075 | 74.897 | 1271.667 | 1212.373 | 1246.808 |
| 208 | 17583.816 | 656.586 | 17370.066 | 17690.607 | 18292.523 |
| 215 | 617.64 | 116.518 | 520.8 | 556.656 | 643.446 |
| 218 | 452.125 | 75.563 | 439.667 | 436.765 | 541.12 |
| 224 | 1644.561 | 221.717 | 1737.733 | 1676.322 | 1875.058 |

**Table A.9. Weekday Sample Design, Factor Group 4**

| ATR # | Bayes Estimate | Posterior S.D. | Simple Average | Factored Average | AADT |
|---|---|---|---|---|---|
| 1 | 1805.409 | 201.262 | 1886.267 | 1833.95 | 1734.453 |
| 51 | 397.997 | 52.368 | 486.2 | 411.857 | 466.764 |
| 55 | 601.778 | 37.797 | 641 | 606.574 | 621.495 |
| 57 | 1021.052 | 177.929 | 862.3 | 928.572 | 1277.032 |
| 214 | 466.778 | 23.963 | 517.067 | 471.846 | 502.598 |
| 220 | 1827.828 | 150.476 | 2042.2 | 1851.626 | 1896.391 |
| 221 | 1964.783 | 115.848 | 2139.6 | 1974.944 | 2003.977 |
| 223 | 2482.245 | 113.753 | 2912.5 | 2465.708 | 2651.259 |

# APPENDIX B

## MINITAB and MATLAB Subprograms

## Developed for this Research

# APPENDIX B

# MINITAB AND MATLAB SUBPROGRAMS DEVELOPED FOR THIS RESEARCH

## MINITAB MACROS

**IWLS** computes estimates of the monthly and day-of-week terms appearing in the lognormal regression model described in Chapter 2.

The columns of input are:

YDAT - a column containing the the natural logarithms of daily traffic counts. Missing values should be coded using Minitab's '*' missing value code. The time-series for successive stations should be stacked on top of each other, but this is not necessary;

STATFAC - column containing an ATR id number;

MONFAC - column containing a code for the month of each count;

DAYFAC - column containing a code for the day-of-week of each count.

The columns of output are:

RES - column containing the regression residuals;

BV - column containing the regression parameter estimates;

DIFF - column containing the squared difference between estimates from successive iterations.

Subcommands:

NSTAT - number of ATR stations in data set;

NDAY - number of days of observations for each stations;

NITER - number of re-weighted least-squares iterations.

```
MACRO
IWLS YDAT STATFAC MONFAC DAYFAC RES BV DIFF;
    NSTAT NN1;
    NDAY NN2;
```

```
    NITER NN3.
MCONSTANT NN1 NN2 NN3 NERF W1 I KNTR
MCOLUMN X.1-X.NN1 Y1 Y2 YDAT STATFAC MONFAC DAYFAC RES BV BV2 DIFF
DEFAULT NN1=2 NN2=365 NN3=1
GLM YDAT=STATFAC MONFAC DAYFAC;
    RESIDUALS RES;
    COEFFICIENTS BV.
LET I=1
WHILE I < NN3
 UNSTACK RES X.1-X.NN1;
  SUBSCRIPTS STATFAC.
 LET KNTR=1
 LET W1=STDEV(X.KNTR)
 LET W1=1/(W1*W1)
 ERASE Y1
 LET NERF=NN2
 SET Y1
 NERF(W1)
 END
 WHILE KNTR < NN1
  LET KNTR=KNTR+1
  LET W1=STDEV(X.KNTR)
  LET W1=1/(W1*W1)
  SET Y2
  NERF(W1)
  END
  STACK Y1 Y2, Y1
 ENDWHILE
 GLM YDAT=STATFAC MONFAC DAYFAC;
```

```
  WEIGHTS Y1;
  RESIDUALS RES;
  COEFFICIENTS BV2.
 LET DIFF(I)=SUM((BV-BV2)**2)
 LET BV=BV2
 PRINT DIFF BV
LET I=I+1
ENDWHILE
ENDMACRO
```

**MACF** computes autocorrelation coefficients when some of the data in the time series are missing, using Dunsmuir and Robinson's formula.

The input is:

XIN - column containing the time-series of interest. Missing data should be coded using Minitab's missing data code '*';

MAXLAG - the number of lags for which autocorrelation coefficients are sought.

The output is:

ARCOE - column containing the estimated autocorrelation coefficients;

ARSE - column containing estimates of each coefficients standard error.

```
MACRO
MACF XIN MAXLAG ARCOE ARSE
MCONSTANT SS K SUM1 S2 N2 MAXLAG
MCOLUMN XIN ARCOE ARSE X1 X2 X3
LET X1=XIN-AVER(XIN)
LET SS=STAN(X1)
LET S2=SS*SS
```

```
LET X2=X1
DO K=1:MAXLAG
LAG X2,X2
LET X3=X1*X2
LET SUM1=SUM(X3)
LET N2=N(X3)
LET ARCOE(K)=(SUM1/N2)/S2
LET ARSE(K)=1/SQRT(N2)
ENDDO
ENDMACRO
```

**MATLAB FUNCTIONS**

**ebclas** computes posterior factor group classification probabilities using the Empirical Bayes prior described in Chapter 3.

Input:

sd - 2-column array with first column containing daily traffic counts for one year, the second column contains a 0-1 code, with 0-indicating missing data for that day, 1-indicating good data forthat day

dd - 1 column array containing 0-1 codes, with 1-indicating that the day is included in the sample, 0-indicating that the day is excluded form sample

pripi- column with one element for each factor group, containing the prior classification probabilities

empri- 3-column array with rows of the form [$\phi 1$ $\phi 2$ $\sigma$], containing estimates of the covariance parameters for an ATR

Output:

piv - 3-column array; first column contains posterior classification probabilities, second column contains Sk terms as defined in equation (3.10), third column contains pripi input.

```
function piv=ebclas(sd,dd,pripi,empri)
% piv=ebclas(yy,dd,pripi,empri)
% piv=posterior classification probabilities
% sd=[daily count, good data code]
% dd=sampling plan
% pripi=prior classification probabilities
% empri=prior estimates of phi and sigma
% global mf wf month day
global mf wf month day
NS=sum(dd);
N=length(sd);
NP=length(empri);
[nm ng]=size(mf);
ns2=0;
Xm1=zeros(NS,12);
Xw1=zeros(NS,7);
x1=ones(NS,1);
for t=1:N
 if dd(t) > 0
   ns2=ns2+1;
   y1(ns2,1)=log(sd(t,1));
   Xm1(ns2,month(t))=1;
   Xw1(ns2,day(t))=1;
 end
end
pout=zeros(3,1);
for kk=1:NP
% kk
 V=vmats(empri(kk,:),dd);
```

```
 sig2=V(1,1);
 sig=sqrt(sig2);
 V(1,1)=1;
 dx=det(V);
 Vi=inv(V);
 ss=x1'*Vi*x1;
 for k=1:ng
  e1=y1-Xm1*mf(:,k)-Xw1*wf(:,k);
  aa=(x1'*Vi*e1)/ss;
  e2=e1-aa*x1;
  s2=(e2'*Vi*e2);
  mm=sqrt(dx*ss)*(sig^(NS-1))*exp(s2/(2*sig2));
  pvec(k,1)=1/mm;
 end
 pout=pout+pvec
% pause
```

**vmats** is used by ebclas to compute the sample correlation matrix assuming for the seasonal AR model described in Chapter 3.

```
function vm=vmats(vp,ds)
% vm=vmats(vp,ds)
% computes autocovariance matrix for sample plan in ds
% vp=[phi1,phi2,sig]
% phi1=AR(1) parameter
% phi2=AR(7) parameter
% sig=noise standard deviations
nn=length(ds);
phi1=vp(1);
```

```
phi2=vp(2);
phi3=phi1*phi2;
sig=vp(3);
sig2=sig*sig;
ns=0;
for i=1:nn
 if ds(i,1) > 0
  ns=ns+1;
  dp(ns)=i;
 end
end
mx=dp(ns)-dp(1);
FF=fmat([phi1 phi2]);
ff=zeros(9,1);
ff(1,1)=sig2;
gam=FF\ff;
for i=1:8
 rho(i)=gam(i+1)/gam(1);
end
for i=9:mx
 rho(i)=phi1*rho(i-1)+phi2*rho(i-7)-phi3*rho(i-8);
end
vm=eye(ns);
for i=1:ns
 for j=i+1:ns
  vm(i,j)=rho(dp(j)-dp(i));
  vm(j,i)=vm(i,j);
 end
end
```

```
vm(1,1)=gam(1);
```

**fmat** is used by vmats to initialize the computation of the correlation matrix

```
function fm=fmat(phiv)
% fm=fmat(phiv);
% fm=coefficient matrix
% phiv=[phi1 phi2]
% phi1=AR(1) parameter
% phi2=AR(7) parameter
phi1=phiv(1);
phi2=phiv(2);
phi3=phi1*phi2;
fm=[1 -phi1 0 0 0 0 0 -phi2 phi3;
-phi1 1 0 0 0 0 -phi2 phi3 0;
0 -phi1 1 0 0 -phi2 phi3 0 0;
0 0 -phi1 1 -phi2 phi3 0 0 0;
0 0 0 -(phi1 + phi2) 1+phi3 0 0 0 0;
0 0 -phi2 phi3 -phi1 1 0 0 0;
0 -phi2 phi3 0 0 -phi1 1 0 0;
-phi2 phi3 0 0 0 0 -phi1 1 0;
phi3 -phi2 0 0 0 0 0 -phi1 1];
```

**opsamp2** solves the nonlinear program described in Chapter 3 to approximate the optimal sampling design for discriminating between two factor groups

Input:

nin-12x7 array containing initial guess of sampling design

ss1-column index of mf and wf for first factor group

ss2-column index of mf and wf for second factor group

NS-number of days in sample

nits-number of iterations of algorithm (50 seems to work well)

```
function nout=opsamp2(nin,ss1,ss2,NS,nits)
% nout=opsamp2(nin,ss1,ss2,NS,nits)
% computes "optimal" sample design of a given size for
% discriminating two specified factor groups
% nout="optimal" sample design of size NS
% nin=initial sample design of size NS
% ss1=index of first factor group
% ss2=index of second factor group
% NS=sample size
% nits=number of iterations
% global mf wf nb
nn1=nin;
for i=1:nits
 i
 nn2=updaten(nn1,ss1,ss2,NS,i)
 max(max(abs(nn2-nn1)))
 nn1=nn2;
end
nout=nn1;
```

**updaten** is used by opsamp2

```
function nout=updaten(nin,ss1,ss2,NN,itr)
% nout=updaten(nin,ss1,ss2,NN,itr)
% nout=12x7 matrix containing updated sampling plan
% ss1=index for first group
% ss2=index for second group
% NN=sample size
% itr=iteration number
% global mf wf nb
md=mf(:,ss1)-mf(:,ss2);
wd=wf(:,ss1)-wf(:,ss2);
if itr == 0
 itr=1
 mbar=0
 wbar=0
else
 nj=sum(nin);
 ni=sum(nin');
 mbar=ni*md/NN;
 wbar=nj*wd/NN;
end
kntr=0;
for i=1:12
for j=1:7
 kntr=kntr+1;
 ij(kntr,:)=[i j];
 xx=md(i,1)+wd(j,1);
 dd(kntr)=xx*xx-2*xx*(mbar+wbar);
```

```
end
end
[dds id]=sort(dd);
nnt=zeros(12,7);
nsum=0;
kntr=85;
while nsum < NN
 kntr=kntr-1;
 ii=ij(id(kntr),1);
 jj=ij(id(kntr),2);
 nnt(ii,jj)=min([nb(ii,jj),NN-nsum]);
 nsum=nsum+nnt(ii,jj);
end
alpha=1/itr;
nout=nin+alpha*(nnt-nin);
```

**ebmdt** computes Empirical Bayes estimates of MDT as described in Chapter 4.

Input:

yy-column vecor containing logarithms of daily traffic counts

ds-column vector containing sampling plan

postpi-array output from ebclas

empri-same as in eblcas

Output:

ov-2-column array, one row for each factor group; first column contains posterior means of MDT, the second column contains posterior standard deviations of MDT

```
function ov=ebmdt(yy,dd,postpi,empri)
% ov=ebmdt(yy,dd,postpi,empri)
% ov=posterior mean and variance of mdt, conditional on group
```

```
% yy=log daily count vector
% dd=sampling plan
% postpi=output from ebclas
% empri=prior estimates of phi and sigma
% global mf wf month day
global mf wf month day
NS=sum(dd);
N=length(yy);
NP=length(empri);
[nm ng]=size(mf);
ns2=0;
Xm1=zeros(NS,12);
Xw1=zeros(NS,7);
x1=ones(NS,1);
for t=1:N
 if dd(t) > 0
  ns2=ns2+1;
  y1(ns2,1)=yy(t);
  Xm1(ns2,month(t))=1;
  Xw1(ns2,day(t))=1;
 end
end
mv=zeros(ng,1);
sv=zeros(ng,1);
for kk=1:NP
% kk
 V=vmats(empri(kk,:),dd);
 sig2=V(1,1);
 sig=sqrt(sig2);
```

```
  V(1,1)=1;
  dx=det(V);
  Vi=inv(V);
  ss=x1'*Vi*x1;
  for k=1:ng
    e1=y1-Xm1*mf(:,k)-Xw1*wf(:,k);
    aa=(x1'*Vi*e1)/ss;
    e2=e1-aa*x1;
    s2=(e2'*Vi*e2);
    mm=sqrt(dx*ss)*(sig^(NS-1))*exp(s2/(2*sig2));
    mvec(k,1)=exp(aa+(.5)*sig2*(1+(1/ss)))/mm;
    vvec(k,1)=exp(2*aa+sig2*(1+(2/ss)));
    vvec(k,1)=vvec(k,1)/mm;
  end
  mv=mv+mvec
  sv=sv+vvec
% pause
end
for i=1:ng
  ov(i,1)=mv(i,1)/postpi(i,2);
  ov(i,2)=(sv(i,1)/postpi(i,2))-ov(i,1)^2;
end
```