MINNESOTA DEPARTMENT OF TRANSPORTATION

RESEARCH

2009-41

Statistical Methods for Materials Testing

LRRB
LOCAL
ROAD RESEARCH
BOARD

Take the steps... Research...Knowledge...Innovative Solutions!

Transportation Research

Technical Report Documentation Page

1. Report No.
MN/RC 2009-41

2.

3. Recipients Accession No.

4. Title and Subtitle
Statistical Methods for Materials Testing

5. Report Date
December 2009

6.

7. Author(s)
Diwakar Gupta, Amy Peterson

8. Performing Organization Report No.

9. Performing Organization Name and Address
Industrial & Systems Engineering Program
University of Minnesota
111 Church Street SE
Minneapolis, MN 55455

10. Project/Task/Work Unit No.

11. Contract (C) or Grant (G) No.
(c) 89261 (wo) 51

12. Sponsoring Organization Name and Address
Minnesota Department of Transportation
Research Services Section
395 John Ireland Boulevard, MS 330
St. Paul, MN 55155-1899

13. Type of Report and Period Covered
Final Report

14. Sponsoring Agency Code

15. Supplementary Notes (including report URL, if available)
http://www.lrrb.org/pdf/200941.pdf

16. Abstract (Limit: 200 words)

Mn/DOT provides incentives to contractors who achieve high relative density via a pay factor applied to each unit of work. To determine the pay factor, Mn/DOT divides each day of a contractor’s work into a small number of lots. Then, core samples are taken from two locations within each lot and the relative densities of the cores are calculated by performing standardized tests in materials testing laboratories. The average of these two values is used as an estimate of the lot's relative density, which determines the pay factor.

This research develops two Bayesian procedures (encapsulated in computer programs) for determining the required number of samples that should be tested based on user-specified reliability metrices. The first procedure works in an offline environment where the number of tests must be known before any samples are obtained. The second procedure works in the field where the decision to continue testing is made after knowing the result of each test. The report also provides guidelines for estimating key parameters needed to implement our protocol.

A comparison of the current and proposed sampling procedures showed that the recommended procedure resulted in more accurate pay factor calculations. Specifically, in an example based on historical data, the accuracy increased from 47.0% to 70.6%, where accuracy is measured by the proportion of times that the correct pay factor is identified. In monetary terms, this amounted to a change from average over and under payment of $109.60 and $287.33 per lot, to $44.50 and $90.74 per lot, respectively.

17. Document Analysis/Descriptors
Hot mix paving mixtures, Specific gravity, Baye’s theorem, Bayesian methods, Expected value, Bids, Factor analysis, Testing, Representative Samples (Testing), HMA relative density testing protocol, Off-line and in-field testing, Sample size determination, Incentive/ disincentive calculations, Parameter estimation and updating

18. Availability Statement
No restrictions. Document available from:
National Technical Information Services,
Springfield, Virginia 22161

19. Security Class (this report)
Unclassified

20. Security Class (this page)
Unclassified

21. No. of Pages
74

22. Price

Statistical Methods for Materials Testing

Final Report

Prepared by:

Diwaker Gupta
Amy Peterson

Industrial & Systems Engineering Program
Department of Mechanical Engineering
University of Minnesota

December 2009

Published by:

Minnesota Department of Transportation
Research Services Section
395 John Ireland Boulevard, MS 330
St. Paul, Minnesota 55155-1899

This report represents the results of research conducted by the authors and does not necessarily represent the views or policies of the Minnesota Department of Transportation, the University of Minnesota, or the Industrial & Systems Engineering Program. This report does not contain a standard or specified technique.

The authors, the Minnesota Department of Transportation, the University of Minnesota, and the Industrial & Systems Enginering Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to this report.

Acknowledgements

The research team is grateful (in alphabetical order) to Minnesota Department of Transportation staff Debra Evans, John Garrity, Tom Ravn, Dr. Alan Rindels, Mike Robinson, and Curtis Turgeon for providing project data and feedback on the research described in this report.

Table of Contents

Chapter 1 Introduction **1**

Chapter 2 Understanding Current Protocols **4**
2.1 Data Description 4
2.2 Methodology 5
2.3 Results 6
2.4 Conclusions 13

Chapter 3 Procedures for Determining Sample Size **14**
3.1 Introduction 14
3.1.1 Offline Sample-Size Determination 14
3.1.2 In-field Sample-Size Determination 16
3.2 Results 17
3.3 Spatial Analysis 18
3.4 A Bayesian Approach for Pavement Density Sampling 22
3.4.1 Background 22
3.4.2 Testing Methods 24
3.4.3 Relevant Literature 25
3.4.4 Methodology 26
3.4.5 Analysis and Results 30
3.5 Conclusions 32

Chapter 4 Updating Estimated Variance **34**
4.1 Introduction 34
4.2 Methodology 37
4.2.1 Hypothesis-Test-Based Updates 39
4.2.2 Smoothing-Approach-Based Updates 40
4.3 Comparative Analysis 41
4.3.1 Experimental Setup 41
4.3.2 Performance Metrics 42
4.3.3 Results 44
4.4 Conclusions 49

Chapter 5 Conclusions **51**

References 53

Appendix A Background – HMA Mixing Process & Tests

A.1 Hot Mix Asphalt Mixing Process

A.2 Purpose of Tests

A.3 Testing Procedures

List of Tables

2.1 Daily averages for wearing mix, required density = 92%, bid price = $40.82. 11
2.2 Daily averages for wearing mix, required density = 92%, bid price = $40.82. 12

3.1 Recommended sample sizes for $\sigma^2 = 70 \times 10^{-6}$. 17
3.2 Recommended sample sizes for $\sigma^2 = 131 \times 10^{-6}$. 17
3.3 Recommended sample sizes for $\sigma^2 = 149 \times 10^{-6}$. 17
3.4 The *p*-value results from the various ANOVA tests. 20
3.5 Payment schedule for mat relative density (SP wearing mix, 4% void). 22
3.6 Reliability of pay factor calculations after two samples. 23
3.7 Economic impact of higher pay factor reliability. 32

4.1 Correlation analysis. 35
4.2 Bartlett hypothesis test results. 37
4.3 Specified variance values and an indicator of whether that variance is significantly different ($> 15\%$) than the previous day's variance in each scenario. 43
4.4 Proportion of type I and type II errors in each scenario with $s^2(0) = .0003$. 44
4.5 Mean absolute deviation per day for each scenario with $s^2(0) = .0003$. 45
4.6 Economic impact of performing too many tests when $s^2(0) = .0003$. 46
4.7 Economic impact of performing too many tests when $s^2(0) = .0002$. 46
4.8 Economic impact of performing too many tests when $s^2(0) = .0004$. 47
4.9 Economic impact of performing too few tests when $s^2(0) = .0003$. 47
4.10 Economic impact of performing too few tests when $s^2(0) = .0002$. 48
4.11 Economic impact of performing too few tests when $s^2(0) = .0004$. 48

List of Figures

2.1 Agency's value ± 0.03 followed by the contractor's value for each core. There were no unacceptable values. 6

2.2 Agency's value ± 0.03 followed by the contractor's value for each core. Unacceptable values are circled. 6

2.3 Agency's value ± 0.03 followed by the contractor's value for each core. There were no unacceptable values. 7

2.4 Agency's value $\pm .03/\sqrt{n}$ followed by the contractor's value for each core. Unacceptable values are circled. 7

2.5 Agency's value $\pm .03/\sqrt{n}$ followed by the contractor's value for each core. There were no unacceptable values. 7

2.6 A scatter plot of percent density for each lot with reference lines at 100% pay-factor limits for the wearing mix. 8

2.7 A scatter plot of percent density for each lot with reference lines at 100% pay-factor limits for the non-wearing mix. 9

2.8 Control chart for average bulk SpG for the wearing mix. The reference lines are the mean and the mean $\pm 2 \times$ standard deviation$/\sqrt{n}$. 9

3.1 Steps for using the offline procedure. 15

3.2 Steps for using the in-field procedure. 16

3.3 Number of cores in each offset level for SP 2180-94 shoulder. 19

3.4 Number of cores in each day of the project level for SP 2180-94 shoulder. 20

3.5 Scatterplot of average density by levels corresponding to day of the project. 21

3.6 Relative frequency of over- and underpayments under the current procedure. 31

3.7 Relative frequency of over- and underpayments under the proposed procedure. . . . 32

4.1 Updating timeline. 36

4.2 Confidence intervals. 38

4.3 Confidence intervals – continued. 39

4.4 Total benefit with different values of $s^2(0)$. 50

4.5 Total cost with different values of $s^2(0)$. 50

Executive Summary

Departments of Transportation (DOTs) hire contractors to perform pavement construction and rehabilitation work and use a variety of protocols to determine the quality of contractors' work. This report concerns hot-mix asphalt pavement construction and focuses on Minnesota DOT's (Mn/DOT's) problem of determining the number of samples that contractors should be required to test to determine pay factors. Although our approach is validated with Mn/DOT data, our procedure is applicable to all DOTs that use a similar approach for determining contractor payments.

Mn/DOT divides each day's work by contractors into a small number of lots. Then, core samples are taken from two locations within each lot and the relative densities of the cores are calculated by performing standardized tests in materials testing laboratories. The average of these two values is used as an estimate of the lot's relative density, which in turn determines the pay factor. This procedure is not sensitive to desired reliability in pay factor calculations. Unreliable calculations can result in excessive over- and under-payments, which affects project cost and the DOT's ability to affect contractor performance over time.

We describe a procedure, based on Bayesian statistics, that determines the requisite number of test samples to achieve user-specified reliability in pay factor estimation. We perform a variety of calculations to compare our approach with the current approach and describe the development of two decision support tools that can be used by DOT project engineers to specify test protocols. The first procedure works in an offline environment where the number of tests must be known before any samples are obtained. The second procedure works in the field where the decision to continue testing is taken after knowing the result of each test. We also provide guidelines for estimating key parameters needed to implement our protocol.

Chapter 1

Introduction

The Minnesota Department of Transportation (Mn/DOT) uses contractors to build/resurface pavements. A standard testing protocol is used to determine the quality of materials and contractors' compaction effort. This report is concerned with the goodness of the procedure used by Mn/DOT to determine contractors' compaction effort. According to this protocol, the daily amount of pavement built by each contractor is divided into a small number of lots. Then, core samples are taken from two locations within each lot and the relative densities of the cores are calculated by performing standardized tests in materials testing laboratories. The average of the two values is used as an estimate of the lot's relative density, which in turn determines the pay factor. [Mn/DOT obtains two cores per location, one core is tested by the contractor and the other by Mn/DOT. If the observed densities from paired cores are within a pre-specified range, then the contractors' values are used to calculate the average, otherwise Mn/DOT's values are used.]

Pay factors depend on ranges within which estimated relative density lies, with the pay factor remaining unchanged within each range. Pay factors translate into either incentive or disincentive (I/D) payments to the contractor for each piece of work (lot). For example, a pay factor of 102% results in a 2% incentive per ton, and a 96% pay factor penalizes the contractor by 4% per ton.

How many samples should Mn/DOT require the contractor to test per lot to determine pay factors? The trade offs that Mn/DOT must consider are as follows — too few samples increase the probability of assessing an incorrect pay factor, whereas too many samples increase testing cost. Too many cored samples also weaken the pavement. This is clearly an important question as illustrated by the following example. Suppose a testing regime determines the lot density to be in one range (say 100% pay factor range), but the true density lies in an adjacent range (say 98% range). For a typical material cost of $40 per ton and lot size of 500 tons, this amounts to a payment calculation error of $400 per lot. Furthermore, when the correlation between compaction effort and incentive payment is low on account of inaccurate testing procotol, I/D payment schemes are less likely to have their intended effect on contractor effort.

This project addresses the problem of determining the number of test samples needed to determine pay factors that are accurate up to a user-specified reliability level. We do not propose changing either the method used to detect if contractor's or Mn/DOT's density values should be used, or the I/D schedule (i.e. the relative density ranges that result in different pay factors), although both are worthy topics for future research. The approach presented in this report will remain valid for any I/D schedule.

We capture the reliability of the testing protocol by two metrics, which we call critical ratio and critical number. We first divide the range of relative densities in equal-sized intervals, called bins, such that each bin lies within a single pay-factor range. This step is necessary because in Mn/DOT's pay factor calculation method, the sizes of intervals over which pay factors remain invariant are not constant. Our procedure then simulates the result of taking each test sample, one at a time. After each simulation, a Bayesian approach is utilized to update (1) the current most likely bin within which the true relative density lies, (2) the critical ratio for each bin, and (3) the critical number. For each bin, the denominator of the critical ratio is the maximum likelihood among all bins that the true relative density lies in a bin, whereas the numerator is the likelihood that the true density lies in the chosen bin. The critical number is the number of bins for which the critical ratio is above a certain threshold specified by the user. The latter is referred to as the cutoff ratio. The process terminates when the most recent value of the critical number is less than or equal to the cutoff number.

For a given set of values of the cutoff number and the cutoff ratio, an iteration refers to the series of updates needed to obtain a single estimate of the required number of tests. For the same reliability metrics, we repeat these iterations many times to obtain an average number of tests needed. The number of iterations is determined by the desired level of accuracy in the estimation of the average number of required tests.

The first half of this report describes two variations of our procedure. The first is used when sample size must be determined in advance. This approach is intended to be used when cored samples are used to determine pavement density, because in that case, the number of cores to be cut must be known in advance. These cores are later tested in a lab. This approach is implemented using a computer program. The user enters certain parameters derived from historical data and a pair of reliability measures, i.e. the cutoff ratio and the cutoff number. The program then calculates the number of samples that should be tested for each lot.

The second variation of our procedure can be used when density observations are available immediately. This approach is implemented in an Excel worksheet for use in the field. As before, the user enters certain parameters derived from historical data and a pair of reliability measures up front. The difference is that after each sample is tested, the user also enters observed relative density value from the current test. The worksheet calculates the current most likely bin within which the true relative density lies and also recommends whether at least one more sample is needed to achieve the desired reliability or the procedure should terminate.

A key input for the aforementioned procedures is the variance of density values for each contractor. We denote this variance by σ^2. Our procedure for determining sample sizes requires fewer samples (cores) to be tested for contractors who have a smaller value of σ^2. This is likely to create an incentive for contractors to lower variance over time. Although such a change will be beneficial for Mn/DOT (because of more consistent pavement density), this creates the need to estimate σ^2 at regular intervals and to update its value when changes are detected. In particular, the problem involves the following steps: (1) determine the initial estimate of σ^2 when the new protocol is first implemented or each time a new contractor wins the bid ("new" means Mn/DOT does not have data on this contractor's performance from past projects), (2) determine for each contractor if σ^2 has changed during the course of a project or between projects, and (3) update the estimate of σ^2.

This second half of this report concerns how Mn/DOT would estimate and update σ^2 if our suggested protocol were to be implemented. Note that in our approach, variance values are expected to be different for different contractors over time depending on their performance. We develop and report a methodology for carrying out each of the three steps.

For setting up the initial estimate, we recommend using a value that is based on all available data for a particular mix or a set of similar mixes for a period of two or three years. This estimate would be the same for all contractors. It will ensure equal treatment of all contractors and allow Mn/DOT to assign initial values to new contractors as well. As contractors reduce their performance variability, industry-wide statistics would change. Because of this, the baseline inputs should be recalculated yearly (using data from the most recent two or three years) so that new contractors will continue to be held to the updated industry norms.

Our proposed methodology for carrying out steps 2 and 3 depends on the nature of change in σ^2. We conceptualize two types of changes. The first type of change occurs when the compaction process (and hence the variance of lot densities) is generally stable but experiences jumps every once in a while. The process returns to its stable state after the contractor detects and fixes the process anomaly. For this type of change, we develop an updating procedure based on hypothesis testing because the latter can detect jumps and allow Mn/DOT to update variance estimate. The second type of change occurs when the variance is not stable and changes frequently. For this type of change, we recommend an updating procedure based on a smoothing approach because that avoids large swings in estimated values.

The organization of the remainder of this report is as follows. The second chapter contains an overview and analysis of the current testing protocols employed by Mn/DOT. Our sample size determination approach is described and analyzed in the third chapter. The fourth chapter contains a description and analysis of the methodology for estimating and updating σ^2. The fifth chapter concludes the report.

Chapter 2

Understanding Current Protocols

2.1 Data Description

From the Mn/DOT Materials Lab in Maplewood, we received data from a project labeled SP 1013-80, including Test Summary Sheets and Density Incentive/Disincentive Worksheets. The project took place between Cologne and Chaska, MN and used two gyratory mix designs - SPWEB440 and SPNWB430. The contractor was Knife River and paving took place between 6/4/07 – 6/19/07.

The data consisted of Test Summary Sheets and Density Incentive/Disincentive Worksheets. The Test Summary Sheets include results from materials tests performed on the loose mix. These sheets also include, among other results, max specific gravity (SpG), air voids, and sample ton number which is the tonnage at which the sample was taken.

The Density Incentive/Disincentive Worksheets display reference information from the Test Summary Sheets, the date paved, total tons paved each day, sample number, the tonnage at which the sample was taken, how many tons since the last sample was taken, the individual voids of that sample, and the individual max SpG of that sample. Also recorded on these sheets is a moving average of the max SpG (used to find pavement density).

These sheets also include test results from both *mat* and *longitudinal joint* density samples. The mat is the main part of the road and the longitudinal joint is either the curb side of the mat or the end that joins another mat segment longitudinally. The longitudinal joint density is tested separately because there are geometric limitation of achieving density with an unsupported edge and/or when a roller bridges onto an existing mat, curb, or gutter. Results concerning the longitudinal joint density include sample number, bulk SpG, bulk SpG used (after consistency testing), average bulk SpG, pavement density, air voids (lots with low air voids do not receive incentive), tons represented, pay factor, and incentive/disincentive in dollars.

The pay factor for the mat is determined by the mean density of cores taken from the mat. For lots that are not tested for longitudinal joint density, this is the total pay factor. For longitudinal joint lots (20% of lots are tested for longitudinal joint density), two cores are sampled for measuring longitudinal density and each is assigned a pay factor according to a schedule that is specific to longitudinal cores. Then, the total pay factor for a lot for which longitudinal joint density is measured is the product of the three pay factors (one from the mat and two from the longitudinal joints).

2.2 Methodology

We analyzed the data to shine light on consistency, variability, quality, and correlation. All of these analyses were performed on the SP 1013-80 data, but our approach can be replicated for any similar data set. We compared the consistency of the contractor's bulk SpG values with that of the agency using two methods that Mn/DOT currently employs. First, the bulk SpG value of each contractor core must be within ± 0.03 of the agency's companion core value. Second, excluding any cores that failed the first method, the contractor's average daily bulk SpG value must be within $\pm 0.03/\sqrt{n}$ of the agency's average daily value, where n is the number of samples averaged. In addition to performing this analysis numerically, we also analyzed the data graphically to visualize if there were any patterns.

We tested variability using two methods. The first method highlights the variability of the bulk SpG measurements over time against the natural variability that is estimated from the first few days of data. The second method compares the variability of the contractor's bulk SpG measurements with that of the agency's. In the first method, we began by calculating the sample average and standard deviation of the first 32 samples (16 lots) of the SPWEB440 mix. We then graphed the average of the two bulk SpG values in each of the remaining 16 SPWEB440 lots along with previously calculated average and the average $\pm\, 2\, \times$ standard deviation $/\sqrt{2}$ as reference lines. Such time-series charts are also called X-bar charts. Processes are deemed to be in control if observations remain within the two standard deviation limits. In our experiment, the upper limit is not relevant because greater SpG is generally more desirable, except when it implies low voids.

To compare the variability in bulk SpG measurements by the contractor with that of the agency, we performed a hypothesis test. We performed this test in two different ways first testing loose mix and core sample data together and second, testing them separately. In each instance, our null hypothesis was that the variance of the contractor's measurements was no different from that of the agency's values.

Next we examined the pavement quality by graphing the percent density. Included in these graphs were 100% pay factor reference lines of 92 and 93 percent for the wearing mix and 93 and 94 percent for the non-wearing mix. Having percent density within these limits earns the contractor 100% pay for the corresponding lot. Assuming that actual density within the 100% pay factor reference lines represents a norm, graphing this data helped us visualize how frequently and how regularly the contractor performed quality work. Analogous to the X-bar chart, here too we paid more attention to density values that were below the 100% pay-factor level because higher density values are generally deemed to be desirable.

Lastly we performed a test of correlation between the Bulk SpG of the loose mix and the Bulk SpG of the core samples for approximately the same mix. To do this, we paired the core samples with the loose-mix sample taken from the nearest location (in terms of tons represented) and then performed an analysis of these pairings. However, we ran into some difficulty in pairing because some data were missing. Therefore, we performed the test in two different ways. In the first approach we treated any missing data as missing, whereas in the second approach, we replaced any missing data by daily averages.

2.3 Results

We found that by and large, the contractor was highly consistent with the agency; see Figures 2.1 – 2.5. In these plots, Figures 2.1 – 2.3 show data by core, whereas Figures 2.4 – 2.5 show data that is aggregated by day. Furthermore, Figures 2.1 – 2.2 show data for cores corresponding to the wearing mix and Figures 2.3 shows similar data for non-wearing mix cores. For the entire duration of the project, only two samples were not within 0.03 of the agency (see circled data in Figure 2.2) and only one day's average was not within $\pm 0.03/\sqrt{n}$ of the agency (see circled data in Figure 2.4). An interesting aspect of this analysis was that even at times when test outcomes were highly variable (for example, density values at the beginning of the project were more variable) the contractor was still consistent with the agency.

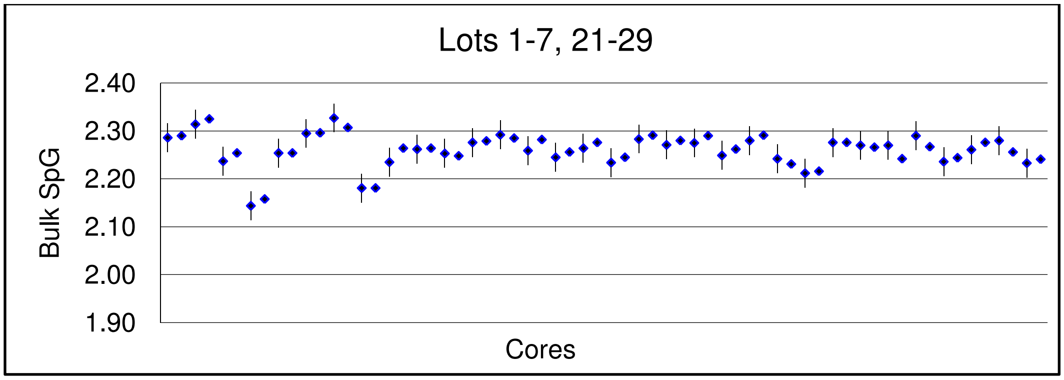


Figure 2.1: Agency's value ± 0.03 followed by the contractor's value for each core. There were no unacceptable values.

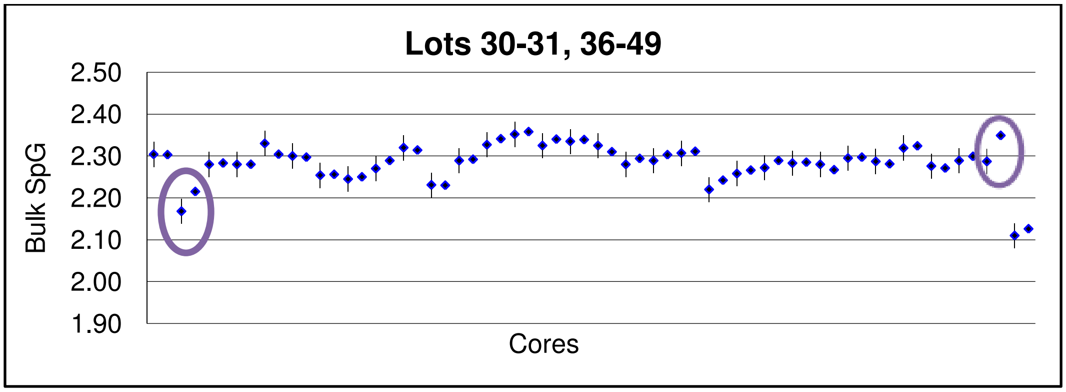


Figure 2.2: Agency's value ± 0.03 followed by the contractor's value for each core. Unacceptable values are circled.

Figures 2.6 and 2.7 contain plots of lot percent densities and the target density ranges. A significant number of lot densities are below the 100% pay-factor limits. In particular, about a fourth of the wearing mix lots and half of the non-wearing mix lots are below the 100% pay-factor limits. This suggests, at least in our limited analysis, that consistency testing encourages

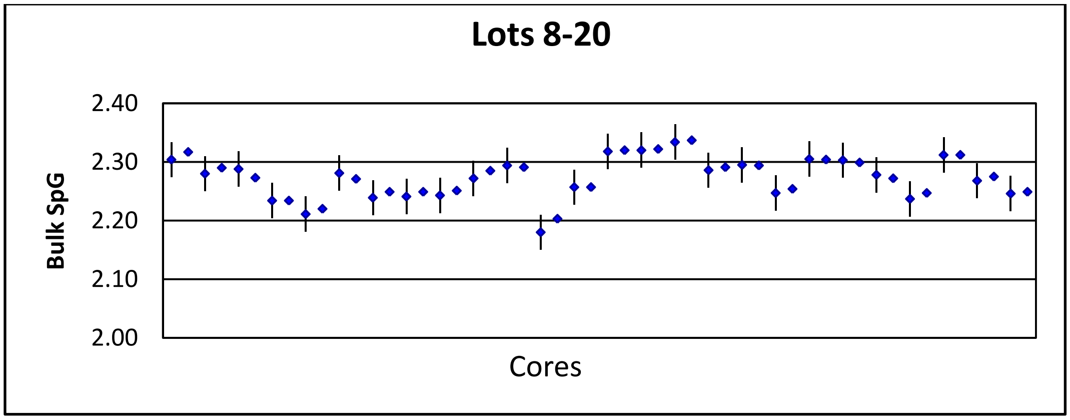


Figure 2.3: Agency's value ± 0.03 followed by the contractor's value for each core. There were no unacceptable values.

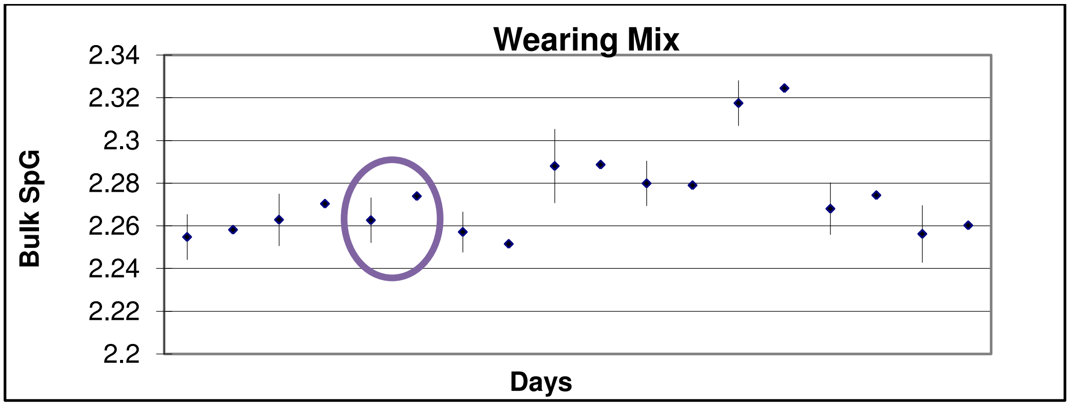


Figure 2.4: Agency's value $\pm .03/\sqrt{n}$ followed by the contractor's value for each core. Unacceptable values are circled.

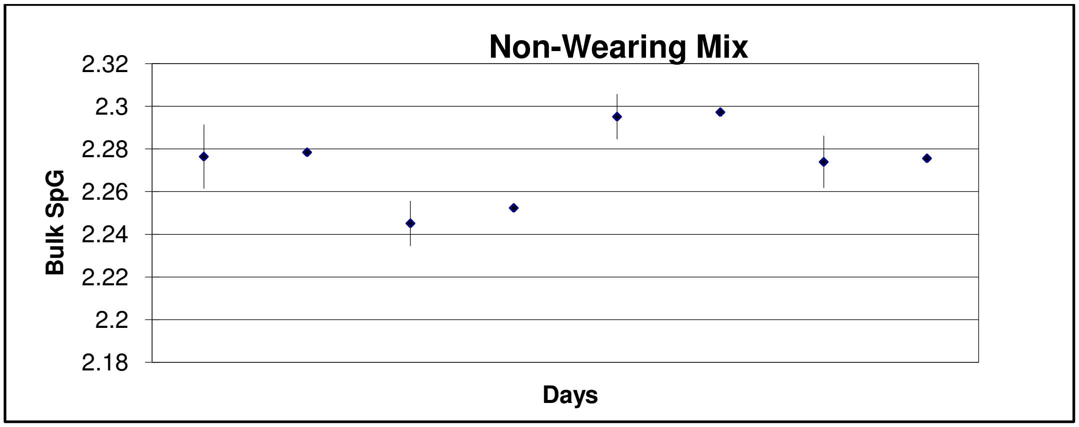


Figure 2.5: Agency's value $\pm .03/\sqrt{n}$ followed by the contractor's value for each core. There were no unacceptable values.

the contractor to be consistent (which is good), but it does not necessarily result in uniformly high quality (by which we mean percent densities at or above 100% pay factor values). We also hypothesize that variability in bulk SpG values over time may lead to poor road quality. Therefore, in addition to consistency, variability should also be taken into account when developing testing protocols.

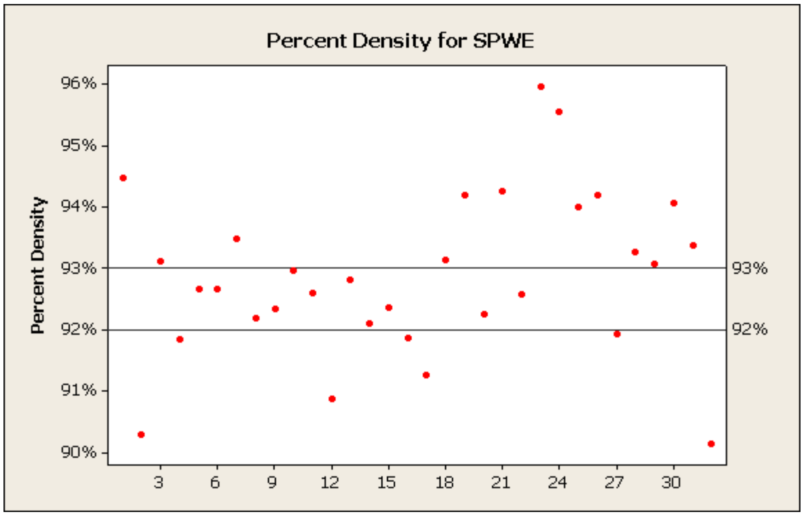


Figure 2.6: A scatter plot of percent density for each lot with reference lines at 100% pay-factor limits for the wearing mix.

Figure 2.8 is an attempt to display the natural variability in the Bulk Specific Gravity testing process. The points on the graph represent the averages of the two bulk specific gravities found by the contractor in the corresponding lot. The lot numbers represented are 30, 31, 36-49 (the last 16 lots of the SPWE mix). The center reference line is the average of the first 32 samples (16 lots) of the SPWE mix. The other reference lines are the average ± the standard deviation of the 32 samples divided by the square root of 2 (our sample size per lot).

Normally, if data lies outside the two standard deviation bounds, then the process is considered to be out of control. However, in this test, only data points that lie below two standard deviations of the mean would be considered undesirable because higher bulk SpG values are typically better. In Figure 2.8 we see that many data points lie above the upper control limit but none lie below the lower control limit. This implies that bulk specific gravity measurements by the contractor do not suggest an out of control process. Moreover, the variability in bulk SpG values appears to be higher during the first few days of the project.

Because a significant amount of variability exists in pavement densities from one lot to the next, we demonstrate one example protocol in this section that may encourage uniform higher density. This scheme is proposed as an example of what Mn/DOT may consider when revising I/D payment schemes. The determination of I/D schedule is not within the scope of this research project. The researchers do, however, recommend additional research effort on quantifying the impact of different schedules that reward uniformly greater compaction effort.

Under the current I/D payment scheme, the contractor is rewarded or penalized for each lot independently. Because the I/D scheme is set up this way, uniform good performance is not specifically rewarded and conversely, variability in performance is not specifically penalized. We demon-

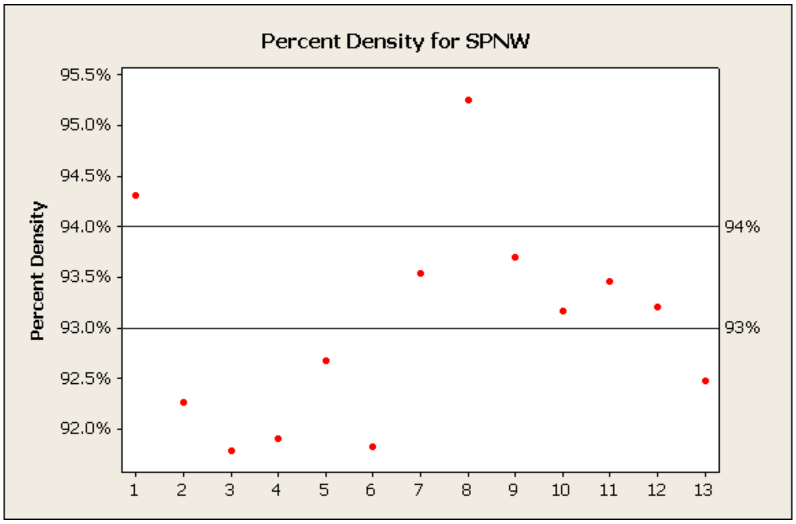


Figure 2.7: A scatter plot of percent density for each lot with reference lines at 100% pay-factor limits for the non-wearing mix.

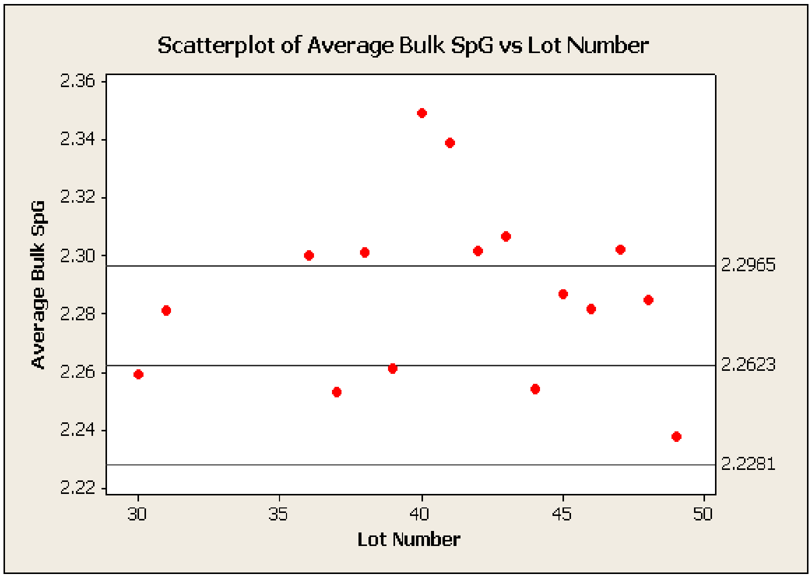


Figure 2.8: Control chart for average bulk SpG for the wearing mix. The reference lines are the mean and the mean ± 2 × standard deviation/$\sqrt{n}$.

strate an example I/D payment scheme that modifies the current I/D protocol to encourage uniform good performance. We modeled this scheme after the consistency testing protocol that Mn/DOT currently uses by not only testing individual cores but also daily averages. Analogously, the example scheme not only uses each lot's density (mat or longitudinal joint) but also daily average densities and the average density from the start of the project to date to calculate pay factors.

We calculate I/D payments according to the current method (see the column titled original under I/D payments in Tables 2.1 and 2.2), based on daily averages, and based on averages to date. The minimum of either the first two or of the first and third I/D payments is what the contractor would receive. To calculate I/D based on daily averages, first calculate the average density for each day (not including any lots that had low voids because lots with low voids will automatically receive no incentive, see Table 2.1). Then calculate I/D payments that each lot would receive based on this new average density. In Table 2.1, the final column entitled proposed I/D payments, shows the minimum of the original payment and the payment based on daily averages. For example, lots 30 and 31 in Table 2.1 have densities 91.3 and 93.1 (lot 31 also has longitudinal joint densities 92.9 confined and 87.7 unconfined) and original pay factors of 98% and $102\% \times 102\% \times 100\%$ The average of these mat densities is 92.2 and this average corresponds to a 100% pay factor. Lot 30 receives the minimum of 98% and 100% pay factor and so the pay factor remains at 98%. On the other hand, lot 31 receives the minimum of 102% and 100% pay factor and so changes to $100\% \times 102\% \times 100\%$.

To calculate I/D based on average density to date, first calculate the average of densities realized from the start of the project to date (again, not including any lots with low voids) and then calculate the corresponding I/D that each lot would receive (see Table 2.2). Finally, for each lot, the minimum of the original I/D and the new I/D based on average density to date is the pay factor that the contractor would receive for that lot. The final column in Table 2.2, entitled proposed I/D payments, is the minimum of these two payment schemes. For example, consider lots 30 and 31 in Table 2.2. The average mat density for lots 1-7 and 21-31 is 91.864, which translates to a 98% pay factor. Since the original pay factor for lot 30 was 98% this lot's pay factor remains unchanged. However, the pay factor for lot 31 was $102\% \times 102\% \times 100\%$ and now it is replaced with $98\% \times 102\% \times 100\%$. As expected, the modified procedure results in lower I/D payments because the contractor's performance is not uniform in this sample.

Because we observed variability in SpG values during the first few days of the project, we compared the SpG values from the first few days of production to those taken over the remaining days of the project. We did this in two ways, comparing the mean bulk SpG value and comparing the variance of the bulk SpG value. When we compared the means we found some statistically significant results which imply that the mean bulk SpG value of the first few days of production may be different from the rest of the days. For example, when we compared the mean SpG of the first two days of the project with the remaining days for the wearing data, our results were as follows: t-statistic = -1.88, and p-value = 0.064. Similarly, when we compared the variances, this time treating the first three days as the first sample and the remaining days as the second sample, we obtained F-statistic = 2.39 and p-value = 0.024, which also suggest that the mean and the variance of bulk SpG values may be different at the beginning of a project. It is in part for this reason that we recommend updating estimated σ^2 periodically (details in Chapter 4).

We performed a correlation analysis between loose mix and core bulk specific gravity values. When we treated the missing data as missing, the wearing mix results for $n = 32$ gave an estimate of correlation coefficient = -.006 and p-value = 0.976. The non-wearing results for $n = 13$ gave

Table 2.1: Daily averages for wearing mix, required density = 92%, bid price = $40.82.

				Pay Factor		I/D Payments	
Lot	% Densities (Low Voids)	Average % Densities per Day	Tons Paved Per Day	Original	Proposed	Original	Proposed
1	94.5 (low)	91.1	1899	1	1	$0.00	$0.00
2	90.3, 91u, 92.2c			0.94676	0.94676	-$1031.65	-$1031.65
3	93.1 (low)			1	1	$0.00	$0.00
4	91.9			0.98	0.98	-$387.58	-$387.58
5	92.7 (low)	92	1447	1	1	$0.00	$0.00
6	92.7 (low)			1	1	$0.00	$0.00
7	93.5, 93.2c, 91.5u (low)			1	1	$0.00	$0.00
21	92.2 (low)	92	3007	1	1	$0.00	$0.00
22	92.4 (low)			1	1	$0.00	$0.00
23	93.0 (low)			1	1	$0.00	$0.00
24	92.6, 91.1c, 89.0u (low)			1	1	$0.00	$0.00
25	90.9	92.02	4106	0.95	0.95	-$1676.05	-$1676.05
26	92.8			1	1	$0.00	$0.00
27	92.1, 92.2c, 91.1u			1.0404	1.0404	$1354.25	$1345.25
28	92.4			1	1	$0.00	$0.00
29	91.9			0.98	0.98	-$670.42	-$670.42
30	91.3	92.2	895	0.98	0.98	-$365.33	-$365.33
31	93.1, 92.9c, 87.7u			1.0404	1.02	-$670.42	-$670.42
36	94.2	93.35	3505	1.04	1.02	$1430.72	$715.37
37	92.3			1	1	$0.00	$0.00
38	94.3, 93.3c, 85.2u			1.00776	1.00776	$277.56	$277.56
39	92.6			1	1	$0.00	$0.00
40	96.0 (low)	92	2012	1	1	$0.00	$0.00
41	95.5, 93.6c, 93.8u (low)			1	1	$0.00	$0.00
42	94.0 (low)			1	1	$0.00	$0.00
43	94.2 (low)			1	1	$0.00	$0.00
44	91.9	92.77	1577	0.98	0.98	-$429.15	-$429.15
45	93.3			1.02	1	$429.15	$0.00
46	93.2, 91.4c, 88.2u			1.0302	1.01	$648.01	$214.58
47	94.1	92	3007	1.04	1	$868.09	$0.00
48	93.4, 93.0c, 92.4u			1.06121	1.0404	$1328.35	$876.79
49	90.2			0.91	0.91	-$1953.20	-$1953.20
Totals						$819.22	-$3402.69

Table 2.2: Daily averages for wearing mix, required density = 92%, bid price = $40.82.

				Pay Factor		I/D Payments	
Lot	% Densities (Low Voids)	Average % Densities per Day	Tons Paved Per Day	Original	Proposed	Original	Proposed
1	94.5 (low)	91.1	1899	1	1	$0.00	$0.00
2	90.3, 91u, 92.2c			0.94676	0.94676	-$1031.65	-$1031.65
3	93.1 (low)			1	1	$0.00	$0.00
4	91.9			0.98	0.98	-$387.58	-$387.59
5	92.7 (low)	92	1447	1	1	$0.00	$0.00
6	92.7 (low)			1	1	$0.00	$0.00
7	93.5, 93.2c, 91.5u (low)			1	1	$0.00	$0.00
21	92.2 (low)	92	3007	1	1	$0.00	$0.00
22	92.4 (low)			1	1	$0.00	$0.00
23	93.0 (low)			1	1	$0.00	$0.00
24	92.6, 91.1c, 89.0u (low)			1	1	$0.00	$0.00
25	90.9	92.02	4106	0.95	0.95	-$1676.07	-$1676.07
26	92.8			1	0.98	$0.00	-$670.43
27	92.1, 92.2c, 91.1u			1.0404	1.01959	$1354.25	$656.75
28	92.4			1	0.98	$0.00	-$670.43
29	91.9			0.98	0.98	-$670.42	-$670.43
30	91.3	92.2	895	0.98	0.98	-$365.34	-$365.34
31	93.1, 92.9c, 87.7u			1.0404	0.9996	$996.47	-$7.31
36	94.2	93.35	3505	1.04	1	$1430.72	$0.00
37	92.3			1	1	$0.00	$0.00
38	94.3, 93.3c, 85.2u			1.00776	0.969	$277.56	-$1108.82
39	92.6			1	1	$0.00	$0.00
40	96.0 (low)	92	2012	1	1	$0.00	$0.00
41	95.5, 93.6c, 93.8u (low)			1	1	$0.00	$0.00
42	94.0 (low)			1	1	$0.00	$0.00
43	94.2 (low)			1	1	$0.00	$0.00
44	91.9	92.77	1577	0.98	0.98	-$429.15	-$429.15
45	93.3			1.02	1	$429.15	$0.00
46	93.2, 91.4c, 88.2u			1.0302	1.01	$648.01	$214.58
47	94.1	92	3007	1.04	1	$868.09	$0.00
48	93.4, 93.0c, 92.4u			1.06121	1.0404	$1328.35	$876.79
49	90.2			0.91	0.91	-$1953.20	-$1953.20
Totals						$819.22	-$7222.36

an estimate of correlation coefficient = 0.1381 and p-value = 0.1278. In the analysis that replaced missing data with daily averages the SPWE mix results with $n = 32$ gave an estimate of correlation coefficient = 0.350 and p-value = 0.050. The non-wearing results with $n = 13$ gave an estimate of correlation coefficient = - 0.646 and p-value = 0.017. The negative correlation between these values can be explained in a number of ways. For example, upon observing relatively low loose-mix sample results, the contractor may increase compaction effort. Alternatively, low loose-mix bulk specific gravity values may be correlated with low loose-mix max specific gravity values. The latter would result in higher densities of cored samples, all else being equal. Overall, researchers did not pursue this analysis further because of the lack of practical significance when correlation can be negative for some data and positive for some other data.

2.4 Conclusions

Current testing protocol includes tests for consistency and quality for each lot. However because incentives/disincentives are awarded independently for each lot, this protocol does not pay attention to the variability in core densities over time. In other words, there is no memory of past performance in current testing protocols that could encourage uniformly good work over time. In most analyses that we performed we found variability in the bulk SpG values over time, which we conjecture may lead to poor road quality overall. For this reason, the proposed testing protocols presented in Chapter 3 reward both higher relative density and uniformity of density values across different lots.

One alternative that could be considered without adding any extra testing effort is to base incentives/disincentives not only on each lot separately but also on daily averages or averages from the start of the project to date. This approach would encourage contractors to aim for lower variability. However, it may not provide significant incentive for achieving high density in subsequent lots if the initial lots were of poor quality. Design of an appropriate I/D schedule is a worthy topic for future research, but not within the scope of the current project.

Chapter 3

Procedures for Determining Sample Size

3.1 Introduction

Mn/DOT determines pay factors based on an estimate of relative density for each lot. This chapter focuses on a method for specifying how many samples should be tested from each lot. If too few samples are tested, then this may result in a greater error in contractor payment calculations. In contrast, taking too many samples consumes time and money without significantly improving the accuracy of payment calculations. Throughout this report, we use density to refer to relative density of the paved road. In all cases, it is the relative density that determines the contractors' pay factors.

There are two variations to our procedure. The first is used offline to predetermine sample sizes. In contrast, the second is used in the field — as samples are observed it dynamically determines whether to take another sample or terminate sampling. Both variations use the same inputs. Based on past data, the user estimates σ^2 (the variance of density observations), and σ_0^2 (the variance of the observed mean density). The programs allow the user to also input an estimate of mean density, m_0. However, the determination of sample sizes is, for the most part, insensitive to the choice of m_0. Therefore, our approach will select a default mean density of 0.925, if nothing is specified by the user. This default lies at the center of 100% pay factor range for a commonly used wearing mix. Details of how to estimate these parameters from historical data can be found in Section 3.4. The user also specifies two reliability metrics — a cutoff ratio above which a bin remains a candidate and a cutoff number of bins that are permitted to remain candidates. An explanation of how to choose the reliability metrics can also be found in Section 3.4. In this section, we focus on describing the steps necessary to implement our approach in practice.

3.1.1 Offline Sample-Size Determination

This approach will be used when sample sizes need to be determined in advance. For example, when core sampling is used to determine pavement density, sample are taken in advance and tested in an offsite materials lab and so the sample size must be predetermined.

Figure 3.1 displays the steps a user would take to implement this variation of the procedure and the following paragraph gives detail to these steps. After the inputs are entered into a computer program developed for this purpose (the program is coded in Matlab), the procedure randomly generates a density sample from a normal distribution with a known mean and variance. Based on this new information, the procedure then updates the distribution of the mean density using a Bayesian approach. Because in practice, one would not know what mean density value to use to generate samples, our procedure is repeated with different assumed values of the true mean. Our program reports the range of sample sizes needed for different assumed pavement mean density values. As we show in Section 3.4, the assumed value of density does not affect our calculations in a significant way. In nearly all cases, the required sample sizes vary by at most 1.

Using the updated distribution obtained above, the procedure determines the likelihood that mean density lies in each of several intervals of density values. Each interval is called a bin. Finally, the procedure takes the ratio of the likelihood that the mean lies in each of the bins to the maximum likelihood among all bins. The procedure continues taking samples until the reliability criteria are met and then records how many samples were needed to meet the reliability criteria. This is counted as one iteration of the simulation. The procedure described above is repeatedly simulated (details of how many simulations should be performed are also found in Section 3.4). Finally, the largest average sample size needed to meet the reliability criteria from the simulations, irrespective of the value of the assumed mean, is the recommended number of samples.

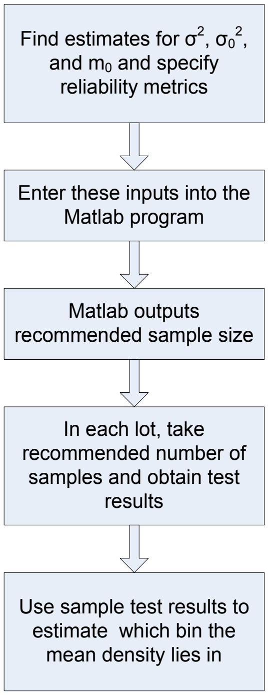

Find estimates for σ^2, σ_0^2, and m_0 and specify reliability metrics
Enter these inputs into the Matlab program
Matlab outputs recommended sample size
In each lot, take recommended number of samples and obtain test results
Use sample test results to estimate which bin the mean density lies in

Figure 3.1: Steps for using the offline procedure.

The user will then take the recommended number of samples and obtain test results from those samples. The test results will be entered into a program that uses Bayesian updating to determine which bin is most likely to contain the true mean (available as an Excel Worksheet), on which the pay factor will be based.

3.1.2 In-field Sample-Size Determination

This variation will be used when results from density observations are available immediately, as is the case with a nuclear density gauge. Figure 3.2 displays the steps a user would take to implement this variation of the procedure and the following paragraph gives detail to these steps.

First determine the inputs and reliability metrics described in Section 3.1.1. Next, observe one density sample and enter this data into a computer program developed for this purpose (the program is imbedded in a Microsoft Excel Worksheet, which provides a familiar user interface). The program then determines if another sample is needed or if the user can stop taking samples. If another sample is needed, then the user will continue observing density samples one at a time and entering them into the program until the program determines that enough samples have been taken. Once the program determines that enough samples have been taken, it will recommend that the user should stop taking samples. The program will also display the current best estimate of which bin contains the true mean.

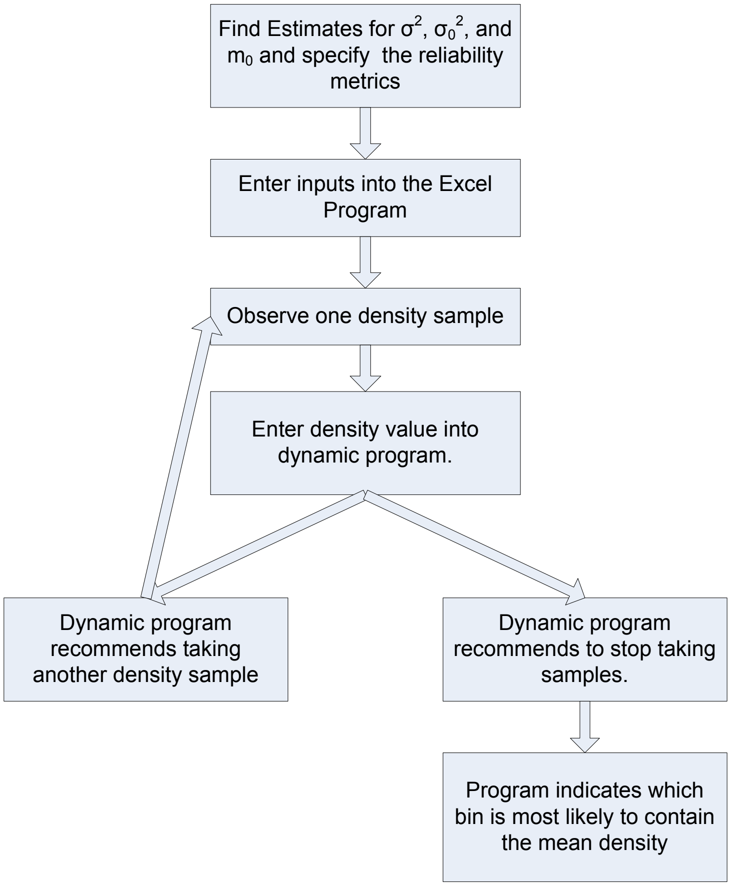

Find Estimates for σ^2, σ_0^2, and m_0 and specify the reliability metrics
Enter inputs into the Excel Program
Observe one density sample
Enter density value into dynamic program.
Dynamic program recommends taking another density sample
Dynamic program recommends to stop taking samples.
Program indicates which bin is most likely to contain the mean density

Figure 3.2: Steps for using the in-field procedure.

3.2 Results

The section displays sample outputs from the offline variation of our procedure that determines sample sizes in advance. The following tables display the output of the program for a particular estimate of the density variance, σ^2, and for various reliability criteria. In these tables, r is the cutoff ratio above which bins remain candidates and b is the cutoff number of bins that are allowed to remain candidates.

Table 3.1: Recommended sample sizes for $\sigma^2 = 70 \times 10^{-6}$.

	r = 0.5	r = 0.6	r = 0.7	r = 0.8	r = 0.9
b = 1	7	5	3	2	2
b = 2	3	1	1	1	1
b = 3	1	1	1	1	1

Table 3.2: Recommended sample sizes for $\sigma^2 = 131 \times 10^{-6}$.

	r = 0.5	r = 0.6	r = 0.7	r = 0.8	r = 0.9
b = 1	13	9	6	3	2
b = 2	4	3	2	1	1
b = 3	2	1	1	1	1

Table 3.3: Recommended sample sizes for $\sigma^2 = 149 \times 10^{-6}$.

	r = 0.5	r = 0.6	r = 0.7	r = 0.8	r = 0.9
b = 1	15	9	6	4	2
b = 2	5	3	2	1	1
b = 3	2	2	1	1	1

Observe that for each pair of reliability metrics, the sample sizes are non-decreasing in the value of σ^2. This means that our procedure will prescribe more testing for contractors with greater historical variability in observed density values. In the long run, our procedure will provide an incentive for contractors to focus attention on achieving consistent density values across all lots in a project. In contrast, the current testing procedure is insensitive to a contractor's historical performance as well as Mn/DOT's desire to achieve different levels of pay factor reliability in different projects depending on the importance of the work being performed, the unit cost of materials, and the dollar amount affected by pay factors.

3.3 Spatial Analysis

Core locations are defined by their *offset*, which is generally the distance from the center of lane, and *station*, which is the distance from the start of the lot. Both the offset and station are randomly generated in accordance with the Mn/DOT Bituminous Manual Section 5-693.7 or ASTM D3665 Section 5 Table A. To find the offset, a random two-decimal number between 0 and 1 is generated and then multiplied by the width of the mat. The resulting number is the offset of the core location. Similarly, another randomly generated two-decimal number is multiplied by the total length of the lot which results in the station of the core location.

Mats are thicker near the center of lane. Hughes (1989) reports that thicker mats cool more slowly. Consequently, thicker mats have more time for adequate compaction to occur before cooling beyond the cessation limit — the cessation limit is the temperature below which compaction will not have a significant effect on pavement density. That is, if mats are thicker nearer the center of lane, then it would be reasonable to hypothesize that the pavement closer to the center of lane is likely to have higher density.

If the offset is found to be a statistically significant factor influencing density, then one option would be to adjust payment schedule to account for the spatial distribution of mat density. This is because a core with a randomly generated location that is closer to the center of lane will be more likely to result in a higher pay factor than a core whose location is further from the center of lane. So, the payment schedule should potentially take into account both the location and the observed density of a core.

A spatial analysis of core locations was performed in order to determine if the offset of a core sample has a statistically significant effect on the density of that sample. The researchers had no evidence that suggested that the station might influence the pavement density. Consequently the spatial analysis was limited to analyzing the influence of the offset. To perform this analysis, an Analysis of Variance (ANOVA) was performed on the data. ANOVA is a method for testing the impact levels of one or more *factors*, in this case the levels of offset, on a *response variable*, in this case pavement density (Hayter 2007).

To begin, a One-Way ANOVA was performed with density as the response variable and offset as the factor. Another One-way ANOVA was performed with density as the response variable and the day of the project on which the sample was cored as the factor. Also, a General Linear Model ANOVA was performed with the density as the response variable and both offset and the day of the project as factors.

An advantage of the General Linear Model is that it can test more than one factor for significance. In addition, the General Linear Model can be unbalanced, as was the case in this analysis. A balanced design has the same number of observations for each combination of factors. In contrast, an unbalanced design has missing or unequal number of observations for certain combinations of factors. The General Linear Model reports whether each factor and/or any interaction between the factors had a significant influence on the response variable.

The researchers had access to core location and density data for eight projects. The offset data for each project was grouped into ranges of offset levels. Similarly, days of the project were grouped into several ranges of days. Without the grouping, ANOVA would treat each offset value and each date as a different level. That would cause problems because each day there is a different number of lots tested and also the number of offset values realized is different. We need ranges of offset and project days as distinct levels to ensure that the ANOVA would be meaningful.

For each project, the groupings were chosen so as to make the number of density samples in each grouping as close to equal as possible. Consequently, each project had its own unique groupings, corresponding to the data available for that particular project. For example, in project SP 2180-94, the offset values were grouped into five levels:

Level 1 : 0 to 0.2

Level 2 : 0.2 to 0.4

Level 3 : 0.4 to 0.5

Level 4 : 0.5 to 0.6

Level 5 : 0.6 to 1

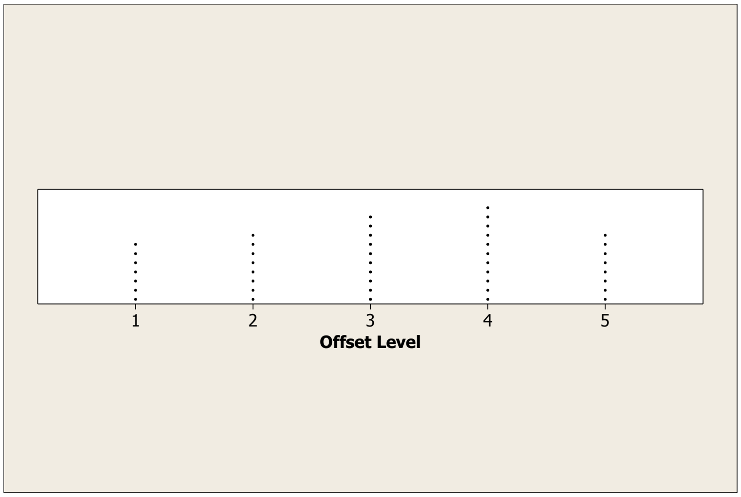


Figure 3.3: Number of cores in each offset level for SP 2180-94 shoulder.

Figure 3.3 displays a dotplot of these four levels which demonstrates that the levels have approximately the same number of cores. For the same project, the day of the project was grouped into three levels:

Level 1 : Days 1 to 2

Level 2 : Days 3 to 5

Level 3 : Days 6 to 8

Figure 3.4 displays a dotplot of how many core samples were in each level of the day of the project in order to demonstrate that the levels contain approximately the same number of core samples.

Table 3.4 displays the resulting p values, corresponding to each combination of project and run of ANOVA. A p-value of 0.05 or smaller implies that the factor does have statistically significant influence on the response variable. In Table 3.4, statistically significant data is shown in bold font.

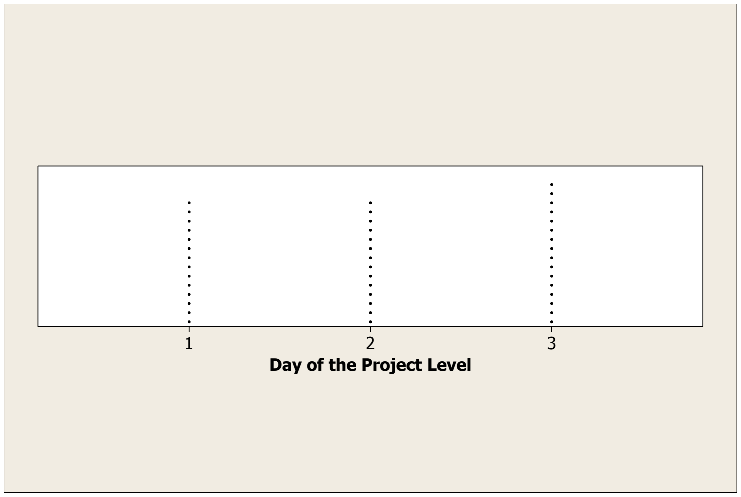


Figure 3.4: Number of cores in each day of the project level for SP 2180-94 shoulder.

Table 3.4: The *p*-value results from the various ANOVA tests.

			General Linear Model		
SP	One-Factor ANOVA for Offset	One-Factor ANOVA for Day of Project	Offset	Day of Project	Interaction
2180-94 - Shoulder	**0.047**	**0.011**	**0.009**	**0.004**	0.299
2180-94 - SPWE	0.298	0.646	0.391	0.670	0.974
2903-10	0.065	0.796	0.066	0.785	0.264
2509-19	0.475	**0.008**	0.487	**0.006**	0.122
3108-63	0.322	**0.023**	0.192	**0.023**	0.588
7602-15	**0.012**	0.579	**0.039**	0.507	0.737
8003-29 - Shoulder	**0.048**	**0.000**	0.423	**0.000**	0.299
8003-29 - SPWE	0.062	**0.001**	0.340	**0.000**	0.487

The results were that when a One-Way ANOVA was performed with offset as a factor, three out of eight *p* values were statistically significant. When a One-Way ANOVA was performed with day of the project as a factor, five out of eight *p* values were statistically significant. With the General Linear Model of ANOVA with both offset and day of the project as factors, the results for the offset had two out of eight significant *p* values, the results for the day of the project had five out of eight significant *p*-values, and the results for the interaction between the day of the project and the offset had zero out of eight significant *p* values.

Because there are so few significant *p*-values corresponding to offset as a factor, the conclusion of this analysis is that the offset of a core location does not have a consistent statistically significant influence on the density of that core sample. This implies that applying a payment schedule that adjusts for the spatial distribution would be erroneous in the majority of instances.

In contrast, more *p* values are significant when we consider the date of the project as a factor. This means that the day of a project has some effect on the density in a reasonably consistent manner. Because of this, scatterplots were made which show the average density corresponding to each level of days for a project and are displayed in Figure 3.5. In analyzing these scatterplots, about half seemed to show density increasing as time went on and the other half showed density decreasing over time. Because no pattern was observed, we cannot conclude that density values consistently increase or decrease with time. Rather, it appears that there is variability in production from day to day which influences the density differently on different days. In short, the day of the project is not correlated with realized densities according to a pattern that would be useful for materials testing.

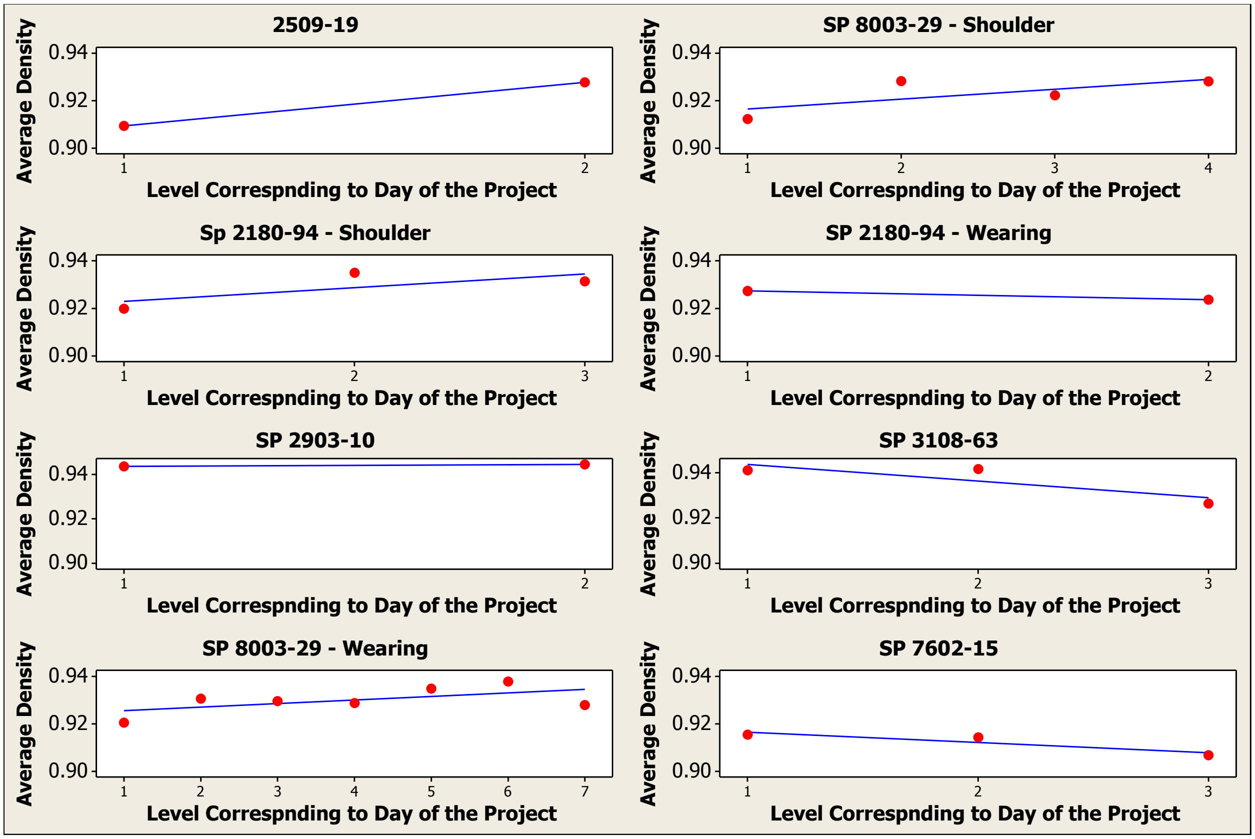


Figure 3.5: Scatterplot of average density by levels corresponding to day of the project.

There may be several explanations for why the offset is not a statistically significant factor influencing pavement density. In addition to the offset, there are other influences on pavement density. For example, as mentioned above thicker mats have more time available for adequate compaction before it cools below the cessation point. A carefully planned compaction effort can overcome the limited time constraint for thinner mats. Moreover, factors such as mix design, air temperature, base (material below the mat) temperature, moisture content in the base, and human error all influence pavement density (Hughes 1989). These factors introduce substantial variability in observed density values.

3.4 A Bayesian Approach for Pavement Density Sampling

3.4.1 Background

Mn/DOT determines pay factors according to a published schedule Mn/DOT (2007). The schedule for determining the pay factor for a particular type of mix and target level of air voids is shown in Table 3.5 below. Air voids are “the pockets of air between the asphalt-coated aggregate particles in a compacted asphalt mixture” (AASHTO T 269-97). A lab technician calculates the air voids (AASHTO T269-97) of core sample by measuring the bulk SpG (A) and the maximum SpG (B). Then, the air voids are equal to $100(1 - A/B)$ (for further details, see Section 3.4.2 and Appendix A). No incentive payments are given if the realized air voids exceed a pre-specified threshold. All relative density observations are rounded to three decimal digits accuracy. Thus, the range “0.936 and above” in Table 3.5 equals $(0.935, 1]$, and the next range equals $(0.93, 0.935]$, and so on. Similar pay factor schedules are available for other types of mix.

Table 3.5: Payment schedule for mat relative density (SP wearing mix, 4% void).

Density Value	Pay Factor
0.936 and above	1.04
0.931 - 0.935	1.02
0.92 - 0.93	1.00
0.91 - 0.919	0.98
0.905 - 0.909	0.95
0.900 - 0.904	0.91
0.895 - 0.899	0.85
0.890 - 0.894	0.7
less than 0.89	0.7

This research focuses on more reliable pay factor calculations. For this purpose, it suffices to identify the correct pay factor range for each unit of contractor’s work. Our approach is designed to achieve precisely this outcome. Before presenting the technical details of this approach, we illustrate its key ideas with the help of an example. Consider project number SP 8055-19, which was completed during 9/5/07 to 9/25/07. Specifically, in this example, we focus on the sixth lot that was a part of the contractor’s second day of work. For the sixth lot, the contractor’s bulk SpG values were 2.334 and 2.347 and the max SpG value was 2.484. Thus, the average relative density for this lot was 0.942 which corresponded to a 4% pay factor (see Table 3.5).

How reliable is the pay factor calculation described above? In order to answer this question, we need to determine the relative likelihood that the estimated pay factor is the correct pay factor. For this purpose, we divide the range of relative density values in equal-length intervals, called bins, and assess the likelihood that the true density lies in each interval — see Table 3.6. Calculations based on our procedure, which is described in more detail later in this section, show that the bin most likely to contain the lot’s relative density is $(0.93, 0.935]$ with a pay factor of 1.02. However, bins $(0.925, 0.93]$ and $(0.935, 0.94]$ cannot be ruled out as possible candidates because the likelihood that the true density lies in these bins is not sufficiently low relative to the most likely bin

$(0.93, 0.935]$. That is, pay factors 1.0 and 1.04 cannot be ruled out.

Table 3.6: Reliability of pay factor calculations after two samples.

bin	0.91–.915	0.915–.92	0.92–.925	0.925–.93	0.93–.935	0.935–.94	0.94–.945	0.945–.95	0.95–.955
likelihood	0.001	0.011	0.064	0.194	0.311	0.266	0.120	0.029	0.004
critical ratio	0.003	0.036	0.205	**0.622**	**1.000**	**0.854**	0.387	0.093	0.012

With the background developed above, we are now ready to provide more details of our approach. Note that most of the technical details are presented in Section 3.4.4, which may be skipped by those not interested in mathematical details without affecting the readability of the rest of this chapter. We first divide the range of relative densities in equal-sized intervals (see Table 3.6), such that each bin lies within a single pay-factor range. This step is necessary because in Mn/DOT's pay factor calculation method, the sizes of intervals over which pay factors remain invariant are not equal. In Table 3.6, double vertical lines demarcate pay factor ranges. Our procedure, which utilizes a Bayesian approach, then simulates the result of taking each test sample, one at a time. At each iteration, the outcomes of this procedure are (1) the current most likely bin within which the true relative density lies, (2) the critical ratio for each bin, and (3) the critical number. The critical ratio is a fraction. For each bin, its denominator is the maximum likelihood, among all bins, that the true relative density lies in a bin, and the numerator is the likelihood that the true density lies in the chosen bin.

The user (Mn/DOT) in our protocol will specify two numbers to indicate desired reliability — a cutoff ratio and a cutoff number. The critical number is the number of bins for which the critical ratio is above a threshold specified by the user, which we refer to as the cutoff ratio. The cutoff ratio is used to determine the critical number after each sample is observed and the process terminates when the critical number is less than or equal to the cutoff number. In one iteration of the procedure, we repeatedly simulate taking test samples and determine the average number of tests needed to achieve a desired combination of cutoff ratio and cutoff number. The procedure is run several times using different parameters to generate sample data. We also test our procedure against the current procedure used by Mn/DOT. Our procedure results in significantly more accurate pay factor determination and provides a direct link between a user-specified reliability of pay factor calculations and the number of tests required.

Returning to the calculations underlying Table 3.6, we assume that relative density is not constant at different spots within a lot. In particular, density is normally distributed with parameters μ and σ^2. The aim of our procedure is to determine μ, and we estimate σ^2 from historical data. That is, the variance of relative density values observed in historical samples is assumed to reflect the natural variability underlying density measurements. However, the mean density may vary from one lot to another. Because there may exist some prior information about μ, e.g. a particular contractor's prior record or performance on previous days of the same project, we assume that the uncertainty about the true value of μ is captured by a normal distribution with initial parameters m_0 and σ_0^2. We set $m_0 = 0.925$ because it corresponds to center of the 100% pay factor range and therefore gives the contractor the benefit of the doubt, and variance $\sigma_0^2 = \sigma^2/2$ because two samples are taken from each lot in the current procedure used by Mn/DOT. We will show that the value of m_0 does not significantly affect the number of test samples needed. However, the estimated value of σ_0^2 does affect the testing protocol.

The estimated, σ_0^2, based on the variance of three previous 2007 projects performed by the same contractor who performed project SP 8055-19 is 0.00015. After observing each sample, we update the estimated parameters of the distribution of μ. This process, after two samples gives an estimated mean 0.934, and variance 0.0000374, knowing which the likelihoods of the mean relative density lying in each bin can be determined. Table 3.6 displays 9 density-value bins (note that more bins were considered but are not shown in this table). The first row shows the likelihood that the mean density lies in each bin. This likelihood was found using the updated parameter estimates of the distribution of the mean density. Next a ratio of each of these likelihoods to the maximum among all likelihoods is found and these ratios are shown in the second row of Table 3.6.

Observe that after two samples are taken and measured, the bin most likely to contain the mean density is the bin which correspond to the range 0.93 to 0.935. Moreover, two adjacent bins also have relatively high likelihoods. In particular, the bin that corresponds to a range of 0.935 to 0.94 has a 26.5% likelihood of containing the mean, relative to a 31.1% likelihood for range 0.93 to 0.935. That is, after observing two samples, Mn/DOT cannot conclude with confidence that pay factor 1.02 is the right pay factor. This results contrasts with the current method in two ways. First, we make use of historical information and obtain a different estimate than what one would obtain by simply averaging the densities of the two samples. Second, our procedure provides relative likelihood estimate that the chosen bin is the correct bin to contain the mean density. If we were to apply our procedure with cutoff ratio of 0.5 and cutoff number equal to 2, two samples would be inadequate for this level of reliability because as seen in Table 3.6 there are three bins that have a critical ratio of at least 0.5.

We used our procedure and the current Mn/DOT procedure (Mn/DOT 2007) to compare the magnitude of payment errors. Assuming that a lot equals 500 tons of material and that the unit price is $40 per ton, the procedure currently used by Mn/DOT resulted in an average overpayment of $109.60 and underpayment of $287.33 per lot based on representative historical data. Also, the relative frequency of accurate payment, where accurate implies within $100 of the correct payment, was 47%. When we used the number of tests recommended by our method, it resulted in an average overpayment of $44.50 and underpayment of $90.74 per lot. The relative frequency of accurate payment was 71%. Details of experiments that led to these results are described in Section 3.4.5.

3.4.2 Testing Methods

The most common method for obtaining pavement density observations is core sampling in which cylindrical cores are cut from the pavement and tested in an offsite materials lab. A lab technician calculates the bulk specific gravity, bulk SpG (AASHTO T166-07), of each core by measuring its dry mass in air (A), its mass in water in water (B), and its saturated surface dry mass (C). The bulk SpG is calculated as

$$\text{Bulk SpG } = \frac{A}{C-B}.$$

The calculation of the pavement density also requires the maximum specific gravity (max SpG) of the mix. The max SPG is an intrinsic property of the mix which is determined by the mix composition and is used to determine air voids and a target level of compaction (AASHTO T209-05). The max SpG calculation measures loose mix because this gives a theoretical maximum specific gravity of mix that has no air voids. This is obtained by performing the Rice Test (AASHTO T209-05) on a loose-mix sample. The test requires measurements of the mass of dry (heated) loose mix in air (A), in water (B), and its saturated surface dry weight(C). The max SpG is calculated as

$$\text{Max SpG } = \frac{A}{A-(C-B)}.$$

Mn/DOTs then determines the relative density of each core as the ratio of the bulk SpG of that core sample and the max SpG. The average of core densities for each lot is treated as that lot's relative density.

Yet another method for testing pavement density uses a nuclear density gauge WSDOT (2008). This gauge records how gamma radiation interacts with the electrons in the pavement to determine the pavement density. The key advantages of this method are that it is non-destructive, it gives results in about five minutes, and the entire procedure can be performed at the worksite. The disadvantages to this method are that the use of nuclear density gauges requires a special license because of the radiation hazard and that it requires an initial calibration, as well as periodic recalibration, to ensure that its readings are correctly converted into pavement density values. Usually, each calibration exercise may require up to ten core samples to be tested. In addition, the nuclear density gauge may add additional noise to the density readings. The effect of additional noise will be addressed in Section 3.4.4.

Irrespective of the method used, a key question facing DOT project engineers is how to determine the number of density observations that should be obtained in any given situation. Too many observations are expensive. Too many cores can also compromise the integrity of the pavement and increase contractors' effort of filling the holes left behind after coring. Thus, additional testing requirements would be factored into the contractors' bid as higher bid prices. Too few observations may lead to errors in pay factor determination and uncertain pavement quality.

3.4.3 Relevant Literature

Hughes (1989) describes the effect of compaction on pavement quality. This article concludes that good compaction effort is necessary to achieve high pavement quality. This report also describes many factors that influence the amount of compaction effort needed to achieve quality pavement density, such as base temperature, moisture in the base, lay down temperature, and lift thickness.

Lenth (2001) describes the importance of adequate sample sizes. The author states that "sample size is important for economic reasons: an under-sized study can be a waste of resources for not having the capability to produce useful results, while an over-sized one uses more resources than are necessary." The article lists several methods for determining sample size including specifying a confidence interval and a Bayesian approach, both of which are used in the procedure described in this report.

Rilett (1998) explores the use of end product (paved road) specification. This is consistent with the approach used by Mn/DOT that allows the contractor freedom in choosing how to compact the roads using a subset of standard engineering practices so long as the end product meets certain specifications. This article also seeks to determine an optimal sampling strategy for density testing. One difference between Rilett (1998) and our research is that the former is concerned with a percent within limits specification for density whereas our procedure considers each lot separately. Rilett

(1998) first estimates the variance of density values, say σ, from historical data, and then argues that the sample size, n, can be determined by considering a confidence interval on the average density, e.g. $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$ is the 95% confidence interval. The sample size is chosen to achieve a desired precision in the size of the confidence interval. In contrast, our approach for sample size determination is a Bayesian approach that depends on the user-specified reliability measures.

McCabe et al. (1999) also utilizes the precision of the confidence interval to determine sample size. This article suggests that agencies may incorporate variance (in addition to mean) to their end product specifications. While our procedure does not add a specification for variance, it does reward contractors who are more consistent and thus provides an incentive to contractors to have a lower variance. Another benefit gained from our methodology is that it takes into account both historical data and sample observations from the current project.

Buttlar and Hausman (2000) describe a computer simulation program that calculates the risk to the agency and to the contractor of using a particular sampling method, measurement method, spec limits, and pay scales. Similar to our procedure, the user inputs the mean density and standard deviation of the density. The user also inputs the standard deviation of the testing method, number of samples to be taken, the sampling procedure, the spec limits, the pay factor schedule, and any cap on pay factors. The program, ILLISIM, then calculates the risk to the contractor (underpayment) and the risk to the agency (overpayment). The goal of this program is to help the user determine the optimal value for the inputs. In contrast with this paper, our research mainly focuses on determining the number of samples to be tested. In addition, given reliability criteria specified by the user, our procedure gives a specific method for determining the sample size.

Chouband et al. (1999) compare the use of nuclear density gauges to core sampling. Our procedure can be adapted for use with a nuclear density gauge, although this may add some additional variability to the density test results. Chouband et al. (1999) show that core sampling is a more reliable method of sampling. In addition, the article concludes that nuclear density gauges should be used with caution and after careful calibration. Romero and Kuhnow (2002) compare nonnuclear density gauges to nuclear density gauges and conclude that nuclear density gauges are more reliable than nonnuclear gauges.

3.4.4 Methodology

Inputs

Our procedure assumes that density observations are normally distributed with parameters μ and σ, where μ, the mean density, is unknown and σ^2, the variance, is estimated from historical data and denoted by s^2. To analyze what distributions fit the data, we took data from one project and tried to fit it to several known distributions. The normal distribution was among several distributions that fit the data. Others in the literature have also assumed that pavement density is normally distributed (see Rilett 1998).

Our procedure further assumes μ is normally distributed, with mean, m_k, and variance, σ_k^2 (similar assumption was also made in Rilett 1998). Again, the estimate of σ_k^2 is denoted by s_k^2. The parameters (m_k, s_k^2) are estimates of the distribution of μ after the k-th sample results have been observed. In this notation, $k = 0$ corresponds to the state of knowledge before sampling begins.

In reality the true mean is just a number and not a random variable. The reason for assigning a prior distribution to μ and developing a procedure for updating the parameters of this distribution after each sample observation, is to incorporate both the prior information and the sample information to estimate its most likely value. In addition to computational simplicity, another reason for using a normal distribution is that it has two parameters that can be modified independently. For example, if the user is less confident in the initial estimate of the mean of μ, then a larger σ_0^2 can be chosen to reflect this lack of confidence. The initial estimates of the density variance σ^2, mean of the mean density m_0, and variance of the mean density σ_0^2 are the inputs for this procedure.

Assuming that both the density and the mean density are normally distributed greatly simplifies the Bayesian updating process because the normal distribution is self-conjugate. This means that if the prior distribution is normal then the posterior distribution will also be normal. After each observation, the Bayesian updating method updates the prior distribution using a combination of new data and prior information. The result is the posterior distribution (Carlin and Louis 2000). The posterior distribution obtained after the k-th sample becomes the prior before the $(k+1)$-th sample.

There are various ways to estimate σ^2, m_0, and σ_0^2, some of which are described next. For example, the parameter σ^2 may be estimated by the sample variance of density values from similar past projects because sample variance is an unbiased estimator of population variance. To overcome any objection to this estimation, each contractor's inputs may be based on their individual past performance. In the event that there is not enough data for a particular contractor, an industry average may be used. The parameter σ_0^2 may be estimated as $s^2/2$ because in the current testing protocol, two samples are taken per lot. However, in situations where a Mn/DOT engineer prefers to use a more conservative approach, we recommend setting $s_0^2 = s^2$ as the default relationship.

In addition, one may nominally set $m_0 = 0.925$ because this is the center of the 100% pay factor bin and provides a neutral starting disposition towards the contractor's quality. In fact, while m_0 is a necessary input for this procedure, its value can be chosen arbitrarily and thus setting it at the neutral value of .925 provides an unbiased prior. An explanation of this is presented later in this section, after additional notation has been introduced. In addition to these inputs, the user must specify two reliability metrics — a cutoff ratio r, and a cutoff number b. Smaller values of r and b imply stricter reliability requirements. Details on how to choose reliability metrics are presented in Section 3.4.5.

Updating

After s^2, m_0, s_0^2, and the reliability criteria are entered, the procedure randomly generates a sample core density value from a normal distribution, with a known mean and variance, in order to simulate sample data. Given the initial estimate s_0^2 of σ_0^2, define $l_0 = \sigma^2/s_0^2$. Then, after obtaining the kth sample, using a Bayesian approach we obtain the following updated distributional parameters of the unknown mean density (see Clemen and Reilly 2001 for details).

$$m_k = \frac{l_0 m_0 + (x_1 + \cdots + x_k)}{l_0 + k} \tag{3.1}$$

and

$$s_k^2 = \frac{\sigma^2}{l_0 + k}. \tag{3.2}$$

Consistent with the notation introduced above, we also set $l_k = l_0 + k$.

Let $(a_{-\infty}, b_{-\infty})$, ... , (a_∞, b_∞) denote the bins. Note that bins are chosen such that $b_j - a_j$

is a constant that is independent of j. Specifically in our procedure, $b_j - a_j = .005$. Then using the updated prior distribution, the procedure determines the likelihood of the mean density lying in each bin and the ratio of the likelihood that the mean density lies in each bin to the maximum likelihood among all bins. Specifically, the probability that the mean density, μ, lies in bin i after k samples is

$$P(a_i \leq \mu \leq b_i) = \Phi((b_i - m_k)/s_k) - \Phi((a_i - m_k)/s_k) \tag{3.3}$$

The ratio associated with the ith bin is

$$r_i = \frac{P(a_i \leq \mu \leq b_i)}{\max_j P(a_j \leq \mu \leq b_j)} = \frac{\Phi((b_i - m_k)/s_k) - \Phi((a_i - m_k)/s_k)}{\max_j[\Phi((b_j - m_k)/s_k) - \Phi((a_j - m_k)/s_k)]} \tag{3.4}$$

Let c^* and r^* denote the cutoff number and the cutoff ratio which are the reliability criteria for this procedure. Our procedure continues to generate sample data and ultimately terminates when there are no more than c^* bins for which the above ratio is greater than r^*.

Effect of m_0

In order to show that a change in the value of m_0 does not significantly affect the outcome of the procedure, consider two runs of the procedure which generate sample observations from normal distributions with the same mean and variance but each run has a different assumed value for m_0. In particular, the first run assumes m_0 and the second run assumes $m_0 + \delta$, for some $\delta > 0$. Assume both runs generate data from the exact same sample density values from a normal distribution, say $x_k \sim N(\mu_s, \sigma^2)$. Also, assume that both runs have the same estimates for s_0^2, which implies that both runs assume the same value of l_0.

After each sample is generated, the estimated distribution of the mean density μ is updated and denoted by N(m_k, s_k^2) and N(m'_k, s'^2_k), for the first and second run respectively. The area under these normal curves and within the ranges of a certain bin, represents the likelihood that the true mean lies in that particular bin. Because ($b_j - a_j$) is a constant independent of j and k, in each iteration (that is, after an additional sample is observed) of the two runs, the bin containing the center of the normal curve will also have the most area under the curve. Moreover, in each iteration of the two runs, the center of the normal curve will be either m_k or m'_k because this is the current best estimate of the mean of μ. In the event that m_k lies on the edge of two bins, the two bins will have an equal likelihood for containing the true mean because the normal distribution is symmetrical about the mean and bin size remains constant. So, the bin(s) most likely to contain the true mean density also contain(s) m_k. In the first run, let the most likely bin be denoted by B_k which represents (a_{i*}, b_{i*}) and similarly in the second run the most likely bin is and B'_k which represents (a'_{i*}, b'_{i*}).

Equation 3.2 shows clearly that as k increases, s_k^2 will decrease. This implies that as more samples are observed, the estimated distribution of the true mean, μ, will be less spread out and more concentrated around m_k. Because s_k^2 decreases in k, the denominator in Equation 3.4 will increase in k. However, for a particular bin (a_i, b_i), the likelihood that μ lies in that bin (the numerator in Equation 3.4) depends both on m_k and s_k^2. This implies that the likelihood may increase or decrease as we observe more samples. So r_i may increase or decrease over time.

Also, equation 3.2 shows that, for a given k, $s_k^2 = s'^2_k$ because both runs of the procedure begin with the same values for σ^2 and l_0. Thus, termination of the procedure is determined entirely by the current estimate of m_k. From Equation 3.1 we see that if the two runs of the procedure drew the same set of sample data and for a given k, the difference between m_k and m'_k will always be $\frac{\delta}{1+k/l_0}$. This is because

$$\begin{aligned} m_k - m'_k &= \frac{l_0 m_0 + \sum x_i}{l_0 + k} - \frac{l_0(m_0 + \delta) + \sum x_i}{l_0 + k} \\ &= \frac{l_0 \delta}{l_0 + k} \\ &= \frac{\delta}{1 + k/l_0} \end{aligned} \tag{3.5}$$

This difference quickly goes to zero as k increases.

Number of Simulations

Since the data used in this procedure was randomly generated, each simulation of this procedure can result in a different recommended number of samples. Therefore the procedure is repeatedly simulated. The number of simulations, say m, must be large enough to overcome the variation between simulations. If $n_1, n_2, \ldots, n_m$ is a series of the sample sizes from simulations $1, 2, \ldots, m$, then $\bar{n} = \sum_{i=1}^{m} n_i/m$ is the recommended sample size.

We use a two-step procedure to determine m. The first step performs 1000 simulations to determine $s_n^2 = \sum_{i=1}^{1000}(n_i - \bar{n})^2/999$, an estimate of the variance of recommended sample sizes. Next, we choose a number m that limits the size of the confidence interval of $\bar{n}$ to a user specified limit. For example, the half length of a 95% confidence interval equals $\frac{1.96 \times s_n}{\sqrt{m}}$. If the user wants this quantity to be at most $0.01 \times \overline{n}$, then we can obtain the minimum value of m by solving $\frac{1.96 \times s_n}{\sqrt{m}} = 0.01 \times \bar{n}$, which gives $m = (\frac{1.96 \times s_n}{0.01 \times \bar{n}})^2$. If m is larger than 1000, then we perform additional $m - 1000$ simulations and the average of the recommended sample sizes from the total m simulations is the recommended sample size. If m is less than 1000, then the average of recommended sample sizes from the original 1000 simulations is the recommended sample size.

Recall that the simulation randomly generates sample data from a normal distribution with a known mean and variance. The variance used to generate the sample data is the input s^2. The entire procedure is repeated for several values of the mean (μ_s), ranging from .88 to .96 in increments of .005. To be conservative, the maximum recommended samples size from each of these repetitions is the final recommended sample size.

Implementation

Knowing the recommended sample size, the contractor takes the appropriate number of samples and obtains test results. The test data is entered into a separate program that uses Bayesian updating to determine which bin is most likely to contain the true mean density. If a potential user were implementing a method of density testing that requires the sample size to be predetermined, as is the case when core samples are observed, this procedure would be run offline and in advance. In contrast, if the user were implementing a non-destructive method that can test pavement density quickly in the field, then a dynamic variation of this procedure could be used. For example, if a nuclear density gauge were being used then the contractor could enter density observations as they are observed. The updating procedure would be the same as described above. The contractor would continue taking samples until the number of bins for which the ratio of likelihoods is above a certain threshold, based on the specified performance measure. In this variation, the number of samples taken for each lot is not necessarily equal to the expected number of tests.

Devises used to test pavement density in the field may add to the variation in test readings because such test devices are usually less accurate than offsite testing of cores. In such cases, the variation of the testing devise should be ascertained by taking pairs of density readings and core samples from the same locations. Then, the difference between the variation of the density readings and the variation of the core sample test results would be the added variation, say s_d^2, due to the testing devise. Finally, this variation would be added to the initial estimate of the variance (based on historical data) to find a total variance, $s_{total}^2 = s^2 + s_d^2$, which will be the input for density variance in the procedure.

Generally cores taken from the mat are evaluated apart of those taken from the longitudinal joint. Current Mn/DOT protocol requires contractors to evaluate the longitudinal joint density on 20% of the lots each day, with a minimum of 1 lot designated as a longitudinal joint density lot. Currently, in a longitudinal joint density lot, the contractor takes 2 longitudinal joint samples on each side of the lane that is being paved. Since the longitudinal joint has a relatively small area compared to the area of the mat, taking too many samples would compromise the integrity of the longitudinal joint. For these reasons, our procedure may not be appropriate to the longitudinal joint unless a non-destructive test is being used.

3.4.5 Analysis and Results

The procedure was analyzed to determine if it improved the accuracy of I/D payments. To do so a random mean was generated. Then, based on this mean, random samples were generated from a triangular distribution, which provided a good fit with data from one historical project. To simulate the current procedure two samples were generated to simulate one lot. To simulate the proposed procedure, a performance measure of one bin remaining above a cutoff ratio of 0.7 was used and this performance measure required 13 samples per lot. For each lot, the generated samples were averaged to estimate the mean density. Finally, the payments based on the estimate of the mean density were compared with payments based on the true mean density, which had been used to generate the samples.

This procedure was simulated $1,500$ times and the relative frequency of under- and overpayments were recorded. Lots where the payment based on the estimated mean density was higher than the payment based on the true mean density counted as an overpayment. Conversely, lots where the payment based on the true mean was higher than the payment based on the estimated mean density counted as an underpayment. The results from the current procedure are displayed in Figure 3.6 and the results from the proposed procedure are displayed in Figure 3.7. All values on the horizontal axis correspond to under- or overpayments within $100 of that value. For example, payments between -$100 and $100 were considered accurate and are displayed above the 0 in Figures 3.6 and 3.7.

In Figure 3.6 accurate payment was given 47.0% of the time, the average overpayment was $109.60, and the average underpayment is $287.33. In figure 3.7 accurate payment was given 70.6% of the time, the average overpayment was $44.50, and the average underpayment is $90.74.

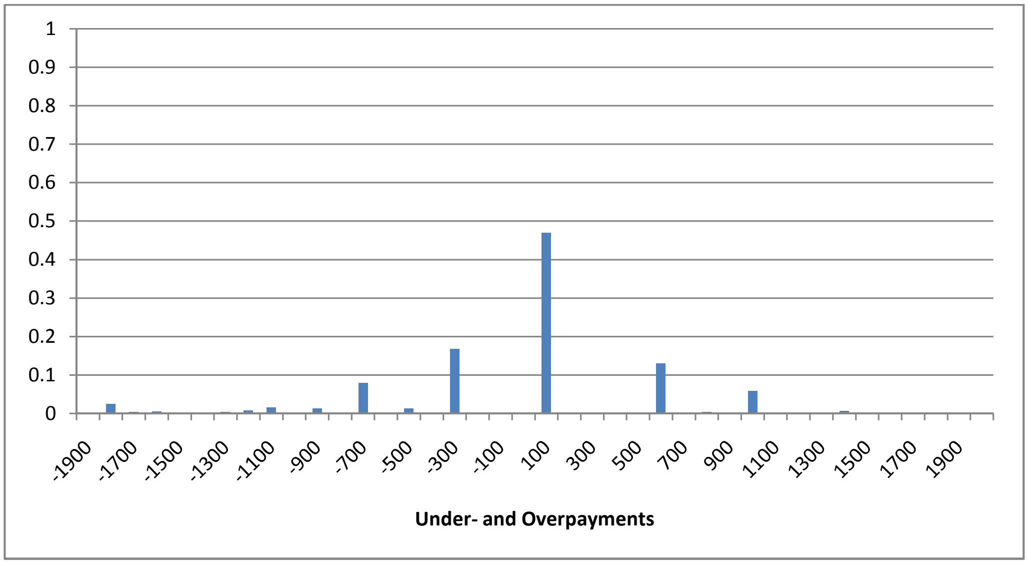


Figure 3.6: Relative frequency of over- and underpayments under the current procedure.

According to this, using the proposed procedure increased accurate payments by over 20% as well as significantly reduced under- and overpayments.

To help the user understand how to choose a performance measure, Table 3.7 displays the required sample size and incremental benefit per test for various performance measures. The incremental benefit per test is the increase in accuracy of pay factors that the agency would realize for each additional sample observation. To determine the incremental benefit, an analysis was performed which randomly generated density samples from a normal distribution with a known mean. The number of density samples generated was based on the various required sample sizes, which are recorded in Table 3.7. Then the under or overpayment was calculated by comparing the payment based on the mean density of the generated samples to the payment based on the known mean used to generate these samples. Finally, this was simulated 1500 times so that an average under- or overpayment for our procedure could be estimated. In addition, to simulate Mn/DOT's current procedure the expected under- or overpayment based only on the first two samples was also calculated. This was calculated as follows:

$$\text{Incremental Benefit per Test} = \frac{A - B}{C - 2}$$

where for each performance measure A was the expected under- or overpayment from Mn/DOT's procedure, B was the expected under-or overpayment from our procedure, and C was the required sample size based on our procedure.

As long as the expected benefit per test is greater than the cost per test, then choosing that performance measure would be beneficial to the agency. Moreover, choosing the strictest performance measure that still has a greater expected benefit per test than cost per test will give the agency the most benefit.

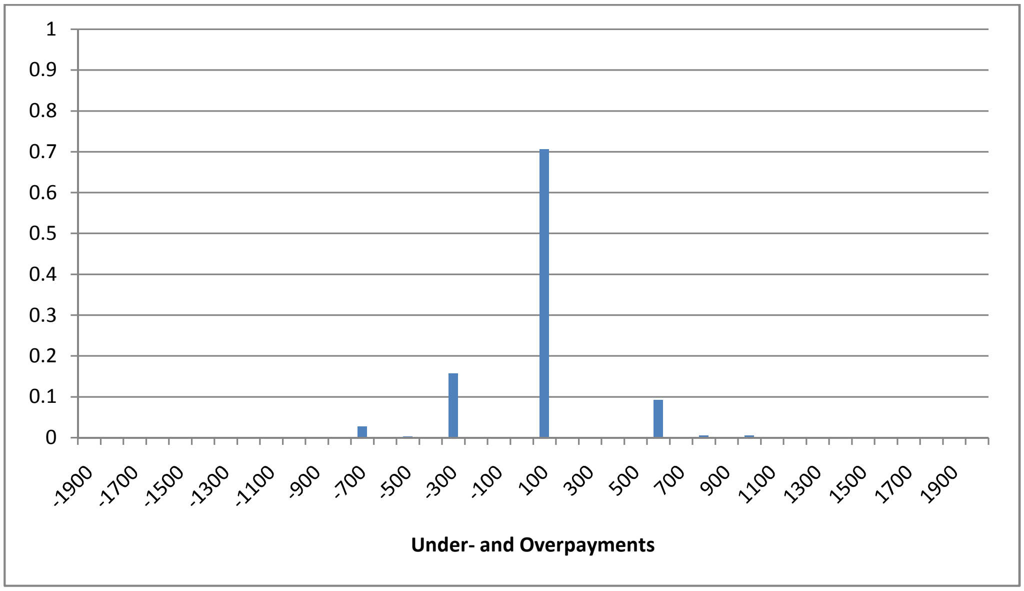


Figure 3.7: Relative frequency of over- and underpayments under the proposed procedure.

Table 3.7: Economic impact of higher pay factor reliability.

	$r = 0.5$		$r = 0.6$		$r = 0.7$		$r = 0.8$		$r = 0.9$	
	Sample Size	Extra Benefit	Sample Size	Extra Benefit	Sample Size	Extra Benefit	Sample Size	Extra Benefit	Sample Size	Extra Benefit
$b = 1$	37	$4.99	25	$7.04	16	$11.92	10	$21.45	5	$40.40
$b = 2$	14	$13.11	10	$21.45	7	$28.68	4	$33.33	2	$0.00
$b = 3$	6	$34.90	5	$40.40	3	$47.74	2	$0.00	1	$188.27

3.5 Conclusions

Materials testing activities are important steps in pavement construction process because they help improve quality. This chapter describes a procedure for determining the number of samples that should be taken to meet user-specified reliability criteria. In addition to the reliability measures, several inputs based on past data are entered prior to running the procedure. These inputs determine the prior distribution of the mean density, which represents current knowledge about the true mean density. Then the procedure generates random density samples. After each sample the prior distribution is updated. Next, using the updated prior distribution, the procedure determines the likelihood of the mean density lying in each of many equal-sized bins. Lastly, the procedure finds the ratio of the likelihood that the mean density lies in each bin to the maximum likelihood among all bins. The procedure terminates when the number of bins for which the ratio of likelihoods is above a cutoff ratio is equal to or less than a cutoff number. This procedure is repeatedly simulated to find the expected required sample size.

After analysis of the current procedure and the one described here, it was determined that using the sample size suggested by the procedure described here resulted in more accurate I/D payments than using the current sample size of two. One simulation increased accuracy from 47.0%, under the current method, to 70.6%, under the suggested method.

For smaller values of σ^2, our procedure results in smaller recommended sample sizes. In other words, contractors with consistent performance are rewarded because they are required to do fewer tests. In addition, contractors are still encouraged to perform high quality work because of the pay factor schedule. So implementing our procedure will encourage contractors to perform consistently high quality work.

Chapter 4

Updating Estimated Variance

4.1 Introduction

Mn/DOT's current testing protocol rewards a contractor for achieving high mean lot density. It does not consider variability in density observations as a measure of pavement quality. In reality, density variation also determines pavement quality. Our proposed procedure (described in Chapter 3) requires more samples to be tested when a contractor's quality (i.e. pavement relative density) is more variable. This gives the contractor an incentive to reduce variance.

Suppose σ^2 denotes the variance of density observations that is used to determine the number of required samples for a particular contractor. Before implementing our proposed procedure, Mn/DOT needs to address three issues:

1. How to calculate an initial estimate of σ^2 when the new protocol is first implemented or when a new contractor wins the bid to do the work? [Note, "new" in our terminology means that Mn/DOT does not have data on this contractor's performance from past projects.]

2. How to determine for each contractor if σ^2 has changed during the course of a project or between projects?

3. How to update the estimate of σ^2 if a change is detected in step 2?

In this chapter, we develop and report a methodology for carrying out each of the three steps. But first, we analyze the correlation between the mean and variance of density samples from thirteen projects. For each project, the mean and variance were calculated for each day of the project. Then a correlation analysis was performed. The results are shown in Table 4.1. The purpose of this analysis is to demonstrate that a contractor can achieve high mean density even as its day-to-day performance varies significantly.

Table 4.1: Correlation analysis.

Project SP	Number of Days	p-value	Correlation Coefficient
0206-116	21	0.174	-0.308
0402-05	10	0.930	0.017
0405-12	18	0.272	-0.273
0413-30	18	0.927	0.023
0415-18	5	0.089	0.820
0506-10	11	0.505	0.226
0603-11	8	0.301	-0.419
0710-29	20	**0.000**	-0.725
1013-80	8	0.530	-0.263
1805-71	16	0.755	0.085
4009-14	11	0.206	-0.501
8055-19	14	0.730	-0.101
8611-18	12	0.330	-0.308

Only one out of the thirteen projects, SP 0710-29, had a significant correlation between mean and variance of daily pavement densities (i.e. p-value ≤ 0.05). This implies that generally the mean and variance of a project are not correlated. The correlation coefficient across the thirteen projects is more often negative, suggesting that higher mean density is weakly associated with lower variance, but this effect is not statistically significant. A possible inference from this analysis is that a change in reward structure is needed to incentivize contractors to place greater effort on improving the consistency of their compaction effort.

In the remainder of this section, we set up a framework for thinking about the three questions posed above. This differs from the presentation in Chapter 3 because we now allow the possibility that σ^2 may change over time. We define an update epoch to be a moment when new information is available to potentially consider the question whether the variance of pavement density has changed. From a practical viewpoint, update epochs will arise no more frequently than once each day after that day's cores are tested and their results are tabulated. However, our formulation is general and allows an arbitrary definition of an update epoch.

Because we allow both the true variance σ^2 and its estimate to change at each update epoch, we index both these quantities by the epoch index $t = 0, \cdots$. Let $\sigma^2(t)$ and $\sigma_m^2(t)$ denote the population and mean lot density variances at update epoch t. We use $s^2(t)$ and $s_m^2(t)$ to denote their estimates. In the implementation of our approach, we will work primarily with $s^2(t)$ and $s_m^2(t)$ because $\sigma^2(t)$, $\sigma_m^2(t)$ cannot be ascertained. In this terminology, the initial variances are denoted by $\sigma^2(0)$, $\sigma_m^2(0)$ and their estimates by $s^2(0)$ and $s_m^2(0)$.

In general, we expect the relationship between $s^2(t)$ and $s_m^2(t)$ to be independent of t. That is, at the start of each new cycle of samples, we will have some relative uncertainty about the mean lot density, which will be captured by $s_m^2(t)$. We expect this to be a known function of $s^2(t)$. It is reasonable to expect that the assumed functional relationship will remain unchanged because $s_m^2(t)$ captures the variance of mean density of a subgroup, given a population variance of $s^2(t)$.

Put differently, updating primarily affects $s^2(t)$. We obtain $s_m^2(t)$ from a hypothesized relationship between $s^2(t)$ and $s_m^2(t)$. In situations where a Mn/DOT engineer does not have a strong intuition about what relationship to use, we recommend setting $s_m^2(t) = s^2(t)$ as the default relationship.

Figure 4.1 displays a timeline for the updating procedure. At update epoch t, there is some estimate of the variance $\hat{s}^2(t)$ which is based on a sample size of $n(t-1)$. This estimate of the variance is used as an input to the procedure described in Chapter 3, which determines the sample size for lots observed between the t and $(t+1)$-th update epochs. So, $n(t)$ total samples are observed between the t and $(t+1)$-th update epochs and these samples have a variance of $s^2(t)$. Then at the $(t+1)$-th update epoch, a new estimate of the variance is found (using one of the two methodologies described in this chapter), $\hat{s}^2(t+1)$, which is based on a sample size of $n(t)$. Then, this estimate of the variance is used to determine sample size for lots observed between the $(t+1)$ and $(t+2)$-th update epochs, according to the procedure described in Chapter 3.

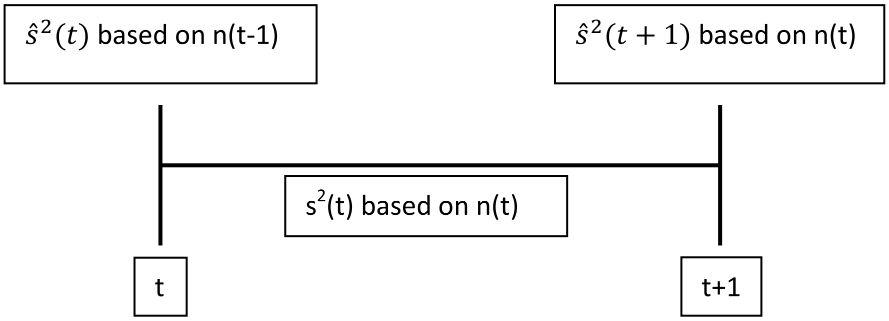

$\hat{s}^2(t)$ based on n(t-1)
$\hat{s}^2(t+1)$ based on n(t)
s^2(t) based on n(t)
t
t+1

Figure 4.1: Updating timeline.

The Bayesian updating procedure describes the process by which we refine our estimate of $s_m^2(t)$ for each lot by taking more samples. In particular, starting with $s_{m,0}^2(t) = s_m^2(t)$, the new estimate $s_{m,k}^2(t)$ after taking k samples is

$$s_{m,k}^2(t) = \frac{s^2(t)}{l_0(t) + k}, \tag{4.1}$$

where $l_0(t) = s^2(t)/s_m^2(t)$ is the equivalent number of initial samples corresponding to the assumed functional relationship between $s^2(t)$ and $s_m^2(t)$. We caution the readers not to be confused by the two levels of updating – one that updates $s^2(t)$ at each update epoch t and the other that updates $s_m^2(t)$ by taking the required number of samples from each lot being tested between update epochs t and $(t+1)$. Chapter 3 concerned the latter, whereas Chapter 4 (this chapter) concerns the former.

We describe two methodologies for updating $s^2(t)$. The first methodology is based on applying a hypothesis testing procedure at each update epoch to check if the variance has changed. This methodology is appropriate in situations where the variance is generally stable but periodically jumps to another value. An advantage of this procedure is that it can be quite sensitive to changes in the variance. Conversely, a disadvantage is that it may change $s^2(t)$ too frequently, leading to unnecessary fluctuations in sample size.

The second methodology uses a smoothing approach to update the value of $s^2(t)$ using current and historical data. This methodology is appropriate in situations where the variance changes frequently but by a small amount each time. An advantage of this procedure is that it avoids large swings in values of $s^2(t)$. A disadvantage is that it always lags the observed changes in $s^2(t)$, which can make it relatively unresponsive.

We provide both methodologies because Mn/DOT project engineers can best decide which methodology will work best in particular situations. We do report an extensive comparative analysis of both methodologies to help project engineers in this task.

4.2 Methodology

To illustrate how the variance of density values changes over time, we analyzed data from nine projects. For each project, the data was clustered into groups of 4 days. For example, clusters consisted of days 1-4, days 5-8, and days 9-12. Groups of 4 days were chosen because generally a day had 4 lots with 2 samples per lot, which gives an approximate sample size of 32 for each group. This is just enough to carry out meaningful comparisons of changes across groups.

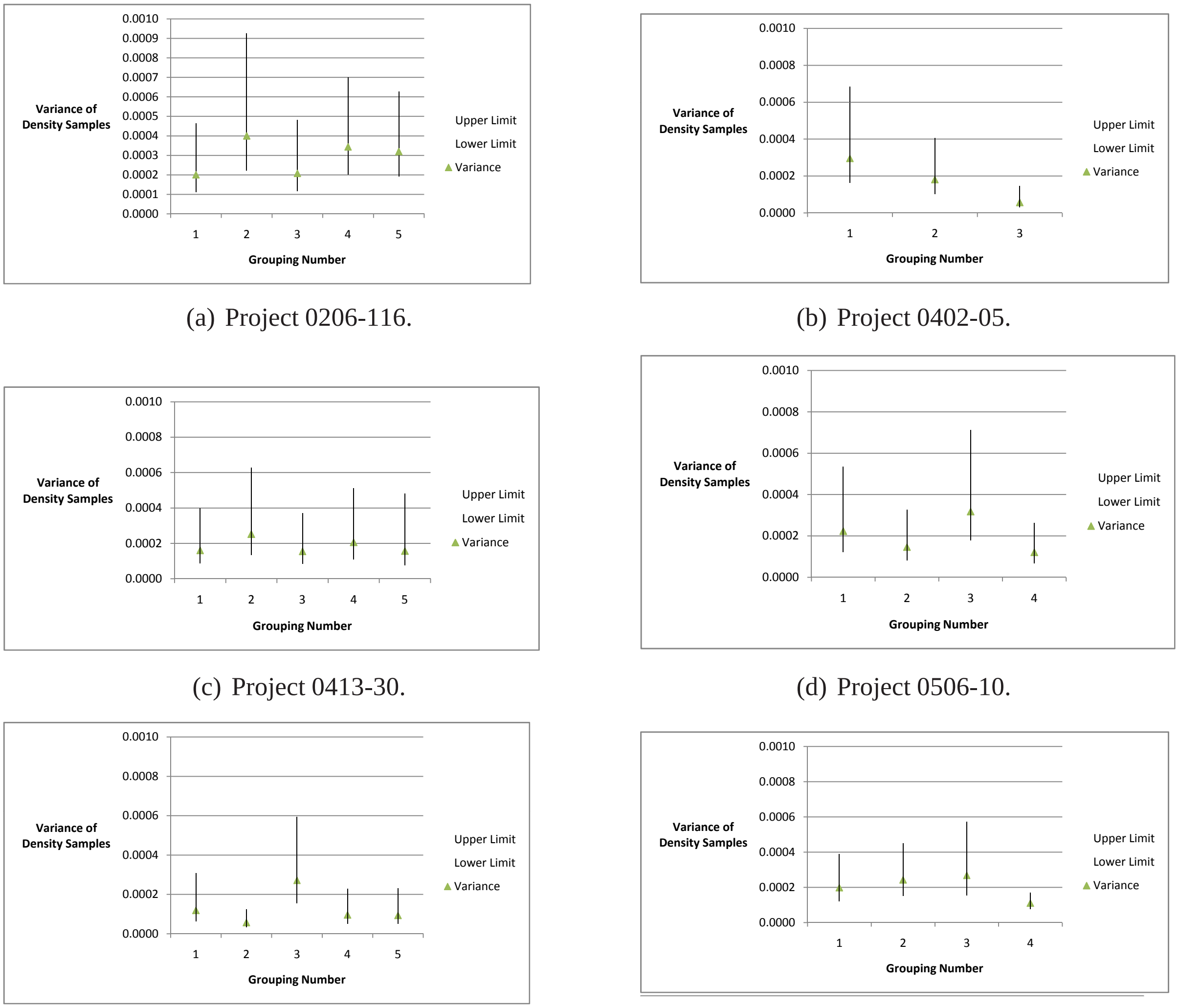
We performed a hypothesis test, known as the Bartlett's test, for each project. The null hypothesis for this test was that all groups have equal variances and the alternative hypothesis was that at least one of the groups has a different variance. The results of these tests are shown in Table 4.2. The significant (less than 0.05) p-values are shown in bold font.

Table 4.2: Bartlett hypothesis test results.

Project SP	p-value	Number of Groups	Number of Adjacent Groups	Number of F-tests that failed
0206-116	0.556	5	4	0
0402-05	**0.016**	3	2	0
1413-30	0.864	5	4	0
0506-10	0.184	4	3	1
0710-29	**0.016**	5	4	2
1805-71	**0.044**	4	3	1
4009-14	**0.013**	2	1	1
8055-19	0.941	4	3	0
8611-18	0.387	6	5	0

In addition, for each project, a hypothesis test was performed over all pairs of adjacent groups. This hypothesis test was an F-test with the null hypothesis being that the two groups have the same variance and the alternative hypothesis being that the two groups have different variances. For example, if a project had 4 groupings, there were 3 hypothesis tests with the following three null hypotheses: $\sigma^2(1) = \sigma^2(2)$, $\sigma^2(2) = \sigma^2(3)$, and $\sigma^2(3) = \sigma^2(4)$. Table 4.2 also displays, for each project, the total number of groups, the number of adjacent groups, and the number of F-tests that failed (had significant p-values). In the terminology introduced in the previous section, each group represents a potential update epoch. For each project, we also plotted the variance of each group and a 95% confidence interval on the variance. These results are shown in Figures $4.2(a)$ through $4.3(c)$.

Table 4.2 shows that in four out of nine projects, at least one group had a significantly different variance than the others. Table 4.2 also displays that five out of nine projects had zero rejected F-tests, three out of nine projects had one rejected F-test, and one out of nine projects had two rejected F-tests. This means that most projects had no change in variance from one group to the next, a few projects had one change between adjacent groupings, and one project had two changes.



(a) Project 0206-116.

(b) Project 0402-05.

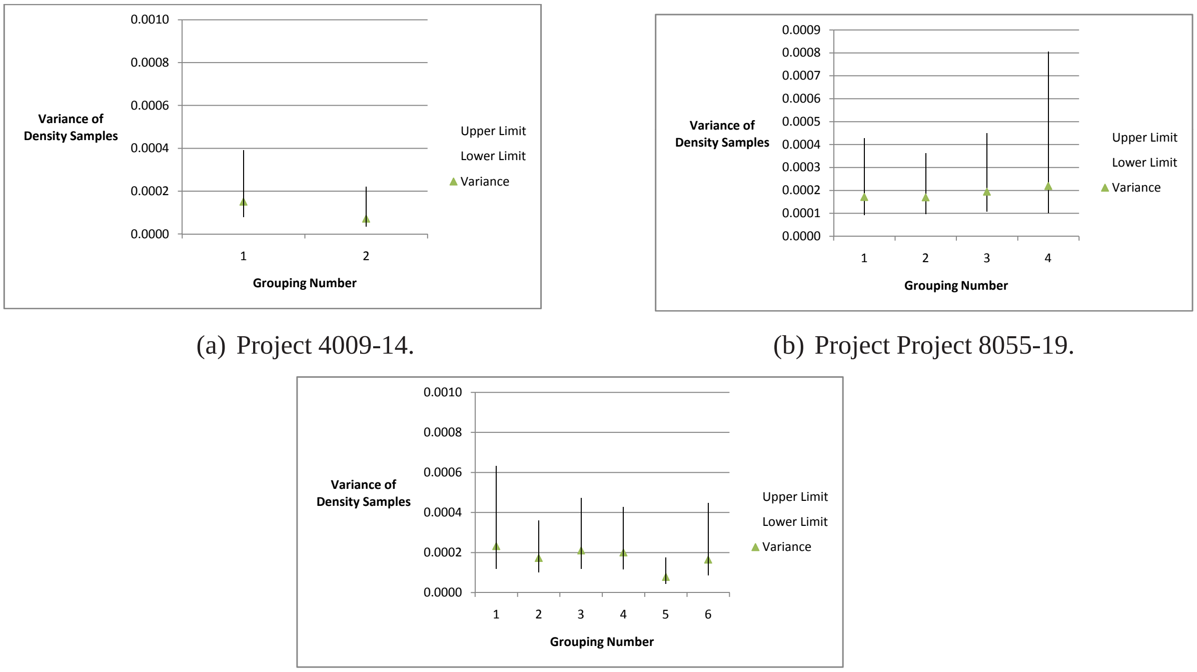
(c) Project 0413-30.

(d) Project 0506-10.

(e) Project 0710-29.

(f) Project Project 1805-71.

Figure 4.2: Confidence intervals.



(a) Project 4009-14.

(b) Project Project 8055-19.

(c) Project 8611-18.

Figure 4.3: Confidence intervals – continued.

From Figures $4.2(a)$ through $4.3(c)$, it appears that the variance changes smoothly and does not jump much. This is particularly noticeable for projects 0206-116, 0413-30, 8055-19, and 8611-18, which are shown in Figures $4.2(a)$, $4.2(c)$, and $4.3(b)$, respectively. The graph for project 0710-29, Figure $4.2(e)$, seems to have jumps between the 2nd and 3rd and the 3rd and 4th groupings. Also, project 1805-71, Figure $4.2(f)$, might have a jump between the 3rd and 4th groupings. These are visually the most noticeable jumps. Overall, it is not possible to say conclusively whether the changes in variance occur gradually or via a jump process. Therefore, in what follows, both procedures are presented and the choice of which methodology to implement is left to the user.

For both methods described below, the initial estimate $s^2(0)$ is obtained from the most recent 2-3 years of data from projects involving the same or similar mix. We propose to update this initial estimate each year on a rolling basis to allow for changing industry standard of performance.

4.2.1 Hypothesis-Test-Based Updates

We propose performing a hypothesis test at each update epoch to determine whether the variance of the most recent data, $s^2(t)$, is significantly different from the most recent estimate of the variance. The latter is denoted by $s^2(t)$. If $n(t)$ (the number of samples that have been observed between the t and the $(t+1)$-th update epoch) is very small, then a test of hypothesis would not be able to detect a change even when a change has actually occurred. Therefore, it is important to define the update epoch carefully. We propose that update epochs occur daily (if the number of samples observed each day is at least 30) or over the minimum number of days that contain just over 30 samples. A more precise argument can be constructed to determine the appropriate update interval. However, this argument cannot be operationalized because it requires the user to specify certain costs that cannot be estimated from available data.

There are two possible results of the hypothesis test – either the null hypothesis is not rejected, or it is rejected. In each instance, the proposed updating procedure will find a slightly different estimate of the variance, $\hat{s}^2(t+1)$, that will be used to determine sample size for lots taking place between the $(t+1)$ and the $(t+2)$-th update epoch. A failed hypothesis test implies that the variance of observed density values since the most recent update epoch is significantly different from the most recent estimate of the variance, $\hat{s}^2(t)$. In this case, we drop historical data and set $\hat{s}^2(t+1) = s^2(t)$, which is the current best estimate of variance and is based on a sample of size $n(t)$.

In contrast, if the null hypothesis is not rejected, we utilize the new information in conjunction with the historical information. Specifically, if $\hat{s}^2(t)$ were the previous best estimate of variance of densities and this estimate were based on a sample size of $n(t-1)$, we would compute $\hat{s}^2(t+1) = \frac{n(t)s^2(t)+n(t-1)\hat{s}^2(t)}{n(t)+n(t-1)}$ as the updated estimate of variance and set $n(t) = [n(t) + n(t-1)]$ as the size of the sample on which this estimate is based.

Regardless of how the updated estimate is obtained, the new $\hat{s}^2(t+1)$ is used to determine the required sample sizes for lots completed between the $(t+1)$ and the $(t+2)$-th update epoch.

4.2.2 Smoothing-Approach-Based Updates

Rather than attempting to detect changes that have already occurred in the data, this procedure attempts to predict future changes based on past behavior. In a typical smoothing procedure, we have both the forecast of the unknown quantity and its realized value at each update epoch. We denote the forecast of variance for period t by $\hat{s}^2(t)$ and the actual observed variance by the previously introduced notation $s^2(t)$. The procedure is initialized by setting $\hat{s}^2(1) = s^2(0)$. For update epoch $t \geq 1$, we perform the following operation.

$$\hat{s}^2(t+1) = \alpha s^2(t) + (1-\alpha)\hat{s}^2(t) \tag{4.2}$$

where α is a smoothing constant chosen by the user.

Suppose $d_i(t)$ is the observed density from the i-th sample in period t and $\overline{d}(t)$ = average density for that period. Then, using the data observed in epoch t, we obtain an estimate of variance at the end of period t as follows:

$$s^2(t) = \frac{\sum_{i=1}^{n(t)}(d_i(t) - \overline{d}(t))^2}{n(t)-1} \tag{4.3}$$

Ideally, it is desirable to weight $s^2(t)$ values differently for each t because we do not expect the number of samples taken to be invariant between two consecutive update epochs. However, weighting of $s^2(t)$ values introduces significant complexity and is not likely to have much of an impact on the quality of this procedure because α values are chosen optimally and periodically updated, as explained next

We obtain the best α by applying smoothing procedure to a set of data that we call the training sample. At each update epoch, we keep track of $s^2(t)$ and $\hat{s}^2(t)$. This allows us to calculate the sum of squares of errors $\sum_{\tau=1}^{t}[s^2(t) - \hat{s}^2(t)]^2$, where t in this expression is the number of observations in the training sample. We then choose a value of α that minimizes this sum. Note that the sum of squares of errors is one measure of the accuracy of $\hat{s}^2(t)$ to the actual variance $s^2(t)$. There are other criteria that could be used to measure accuracy and to obtain the best α with respect to those criteria.

4.3 Comparative Analysis

In this section, we compare and contrast the two methodologies discussed above. This section is divided into three parts. We describe the experimental set up first, followed by the key performance metrics we calculated to compare the two methodologies, and finally, we present the results. A discussion of the implications of these results and our results are presented in the next section.

4.3.1 Experimental Setup

A full evaluation of the two procedures required the ability to adjust the number of samples taken each day according to the previous day's test results. Since historical data contained at most 2 samples per lot, historical data could not be used to test our procedures. Because of this, we simulated density observations in the experiments described in this section.

Historical projects typically lasted between 10 and 30 days and most days had 4 lots. On the basis of this observation, in the experiments we conducted, each scenario was simulated for 20 days and each day had 4 lots. Also, in each scenario the starting value of true process variance was $\sigma^2(0) = 0.0003$. The number of samples were determined according to a target cutoff ratio of .6 and a cutoff number of 2 (see Chapter 3 for details on choosing these performance measures).

The two methodologies described in the previous sections were analyzed under three different scenarios. The first scenario represented a situation in which the true process variance changed randomly each day from the base-case value of 0.0003 during the course of the project. In the second scenario, the variance also changed randomly, but an up or down change could last for up to three consecutive days. The third scenario contained an initial period of instability after which the variance stabilized, representing contractor's efforts to gain control over the production process. Unstable values at the start of a project are a common occurrence in historical data because contractors try to vary mix composition and production processes to realize higher density. Each scenario was simulated three times using a different initial value for $s^2(0)$ each time – the correct value of .0003 and two incorrect values which were .0002 and .0004. In each case, we assumed that $s^2(0)$ was based on an initial sample of size $n(0) = 100$.

Under each scenario, data was randomly generated to represent density samples. Consider, for example the case when the initial estimate of the variance of was .0003. We used the static program from Chapter 3 to find that 8 samples should be taken per lot. So each scenario began by generating data samples for 4 lots with 8 samples each (a total of 32 samples) for day 1. In each scenario, the density samples were generated from a normal distribution with a mean .925 and the specified variance for that particular day and scenario. Under the first scenario, the variance values were either a big or small jump from .0003, where a small jump is 5%-20% above or below .0003 and a big jump is 20% to 50% above or below .0003. The second scenario had the same type of jumps, except that big jumps could remain for up to 3 days. Under the third scenario, the first 5 days of the project consisted of big jumps from .0003 and days 6 through 20 consist of small jumps from .0003. The specified values for the variance can be found in Table 4.3. For example, the first day of the third scenario generates 32 random samples from a normal distribution with mean .925 and variance .000412412.

Once the samples were generated, the hypothesis testing methodology found the value of the test statistic, $f = \hat{s}^2(t)/s^2(t)$, where $\hat{s}^2(t)$ was the variance used to determine sample size for day t and and $s^2(t)$ was the observed variance on day t. Then, the critical points, $f_{\alpha/2,n(t-1),n(t)}$ and $f_{1-\alpha/2,n(t-1),n(t)}$, were found, where $\alpha = 0.05$ was the significance level, $n(t-1)$ was the sample size used to find $\hat{s}^2(t)$ and $n(t)$ was the sample size used to find $s^2(t)$.

If we observed that $f > f_{\alpha/2,n(t-1),n(t)}$ or $f < f_{1-\alpha/2,n(t-1),n(t)}$ then the null hypothesis was rejected and $\hat{s}^2(t+1)$ was set equal to $s^2(t)$. Otherwise, the hypothesis test was deemed inconclusive and we updated $\hat{s}^2(t+1)$ as follows: $\hat{s}^2(t+1) = \frac{n(t)s^2(t)+n(t-1)\hat{s}^2(t)}{n(t)+n(t-1)}$ and $n(t) = [n(t) + n(t-1)]$. The new value of $\hat{s}(t+1)$ was then used to determine the sample size for the next day, using the static program developed in Chapter 3. This process was continued throughout the 20 simulated days in each scenario.

The smoothing methodology required an extra step in which we used a training sample to determine optimal α. For each scenario, we used a unique training sample of 120 days, in which the variance values were randomly generated in the same fashion that the variance values were generated for the analysis. The training sample for each scenario resulted in a unique optimal value for alpha. For the first scenario, $\alpha = 0$, for the second scenario $\alpha = 0.14634$, and for the third scenario, $\alpha = 0.00053$.

However, the agency will not always be confident that the available training sample was an accurate representation of the project on which the smoothing methodology is implemented. So, rather than using a training sample to determine the optimal value of α, some literature suggests a nominal value of $\alpha = .10$ can be used when the process is somewhat stable (not too variable) (Silver et al., 1998). In addition, the performance of the smoothing procedure is normally quite insensitive to the chosen value of α. Due to this, the smoothing methodology was simulated three times for each scenario with $\alpha = .10$ and the initial value of $s^2(0) = .0003$, .0002, and .0004. Analyses that used the customized training samples to determine α will be referred to as Ideal Smoothing and the analysis using the nominal value of $\alpha = .10$ will be referred to as Nominal Smoothing.

After the α values were found, the analysis could be performed. To begin, sample density values were randomly generated from a normal distribution with a mean .925 and the corresponding variance for each day and scenario. Once the samples were generated, the smoothing methodology estimated the variance for day t, denoted by $\hat{s}^2(t)$. Then the smoothing procedure was applied in order to estimate the variance used in the next epoch. The smoothing procedure is as follows:

$$\hat{s}^2(t+1) = \alpha s^2(t) + (1-\alpha)\hat{s}^2(t) \tag{4.4}$$

The value of $\hat{s}^2(t+1)$ was then used to determine the sample size for each lot on the next test epoch. This procedure was repeated for each day for each test scenario.

4.3.2 Performance Metrics

The simulation results were analyzed in terms of three key performance metrics to test the effectiveness of each procedure. The first metric calculated the proportion of Type I and Type II errors.

A Type I error would occur if it was deemed that the variance had significantly changed when it actually had not changed and a Type II error would occur if it was deemed that the variance had not significantly changed when it actually had. This analysis considered a change to be significant if the variance changed by more than 15% from one day to the next (note that this level of significance could be set to another percentage). Table 4.3 displays the variance used on each day in each scenario as well as an indicator of whether the variance on a particular day is significantly different than the previous day. This performance metric is relevant only for the hypothesis-test-based updating method.

The second metric measured the difference between the number of samples that were obtained on a particular day (based on the estimate of the variance) and the ideal number of samples that should have been obtained that day (based on the actual variance). Note that each day had 4 lots. So this metric equaled the average absolute difference in actual and the ideal number of samples per lot over the 20 days multiplied by four. A metric such as this is also called mean absolute deviation (MAD) in statistics and operations research literature.

Table 4.3: Specified variance values and an indicator of whether that variance is significantly different ($> 15\%$) than the previous day's variance in each scenario.

Day	Variance Jumps		Variance Jumps for up to 3 Days		Initial Chatter in Variance	
	Variance	Change	Variance	Change	Variance	Change
1	0.000232764	change	0.000224222	change	0.000412412	change
2	0.00027594	change	0.000184061	change	0.000429679	no change
3	0.000346777	change	0.000268627	change	0.000226105	change
4	0.000396462	no change	0.000316477	change	0.000212513	no change
5	0.0002468	change	0.000320343	no change	0.000412803	change
6	0.000280703	no change	0.000353463	no change	0.000268127	change
7	0.000356188	change	0.000246757	change	0.000326252	change
8	0.000166403	change	0.000240402	change	0.000335311	no change
9	0.000359444	change	0.000262746	no change	0.000279845	change
10	0.000320562	no change	0.000266211	no change	0.000348596	change
11	0.000356113	no change	0.00035386	change	0.000335519	no change
12	0.000328325	no change	0.000265233	change	0.000345423	no change
13	0.000270069	change	0.000284182	no change	0.000337954	no change
14	0.000270707	no change	0.000405208	change	0.000274099	change
15	0.000221313	change	0.000450505	no change	0.000340227	change
16	0.000201703	no change	0.000320809	change	0.000333308	no change
17	0.000331298	change	0.000267149	change	0.000318237	change
18	0.000338313	no change	0.000207188	change	0.000356318	change
19	0.000248293	change	0.000229312	no change	0.000344064	no change
20	0.000352914	change	0.00017083	change	0.000357063	no change

Finally, the third metric measured the economic impact of using each updating methodology. For each method, on any given day we might take either too few or too many samples per lot relative to the ideal number. The latter can only be known in a simulation exercise such as the one we constructed. If too many samples were taken, this resulted in both an extra cost and an extra benefit. Clearly, the extra cost would come from performing more than the ideal number of tests. The benefit would come from increased accuracy in Incentive and Disincentive payment calculations.

Similarly, if fewer samples were taken, then this would also have two opposing effects. In this case, the cost would be caused by the decreased accuracy in Incentive and Disincentive payment calculations, whereas the benefit would be smaller testing cost.

4.3.3 Results

The proportion of Type I and Type II errors for hypothesis-testing-based updating procedure are shown in Table 4.4 below. Recall a Type I error would occur if it was deemed that the variance had significantly changed when it actually had not changed and a Type II error would occur if it was deemed that the variance had not significantly changed when it actually had. Also, recall that here a change in variance is deemed significant if only the variance changed by more than 15% from one day to the next (note that this level of significance could be set to another percentage).

Table 4.4: Proportion of type I and type II errors in each scenario with $s^2(0) = .0003$.

Initial Value of $s^2(0)$	Scenario	Proportion of Type I Errors	Proportion of Type II Errors
.0003	Variance Jumps	.1	.5
	Variance Jumps for Up to 3 Days	.05	.5
	Initial Chatter in Variance	.1	.5
.0002	Variance Jumps	0	.5
	Variance Jumps for Up to 3 Days	.1	.65
	Initial Chatter in Variance	0	.3
.0004	Variance Jumps	0	.45
	Variance Jumps for Up to 3 Days	0	.5
	Initial Chatter in Variance	0	.5

Clearly, Table 4.4 shows that the proportion of Type I errors was at most $.1$ in our simulations which confirms that the proposed procedure is biased toward protecting against Type I errors. This is exactly as one would expect. However, protecting against Type I errors can result in an increased frequency of Type II errors, which we see in this analysis. Type II errors are more likely when the change in variance is relatively small, as is the case in this our simulations. Using an incorrect estimate of $\sigma^2(0)$ that errs on the side of being larger produces better outcomes, as one would expect.

Similarly, Table 4.5 shows calculations of mean absolute deviation for each scenario. A smaller value for the MAD is desirable and this implies greater accuracy of updating protocol. Consider the hypothesis testing and ideal smoothing methodologies. When $s^2(0) = .0003$, the hypothesis testing methodology has a smaller MAD in one scenario and the ideal smoothing methodology has smaller values in the other two scenarios. When $s^2(0) = .0002$, the hypothesis testing methodology has a smaller MAD in two scenarios and the ideal smoothing methodology has a smaller MAD in one scenario. When $s^2(0) = .0004$, the hypothesis testing methodology has a smaller MAD in one scenario and the two methodologies have the same MAD in the remaining scenarios. So, while neither methodology clearly dominates, the hypothesis testing methodology appears to be slightly more accurate.

Table 4.5: Mean absolute deviation per day for each scenario with $s^2(0) = .0003$.

Initial Value for $s^2(0)$	Scenario	Hypothesis Testing Methodology	Ideal Smoothing Methodology	Combination Smoothing Methodology
.0003	Variance Jumps	2.45	1.35	1.35
	Variance Jumps for Up to 3 Days	1.55	1.75	1.45
	Initial Chatter in Variance	1.7	1.5	1.5
.0002	Variance Jumps	1.75	2.45	2.45
	Variance Jumps for Up to 3 Days	2.2	1.7	2.15
	Initial Chatter in Variance	2	3.1	3.1
.0004	Variance Jumps	2.15	2.65	2.5
	Variance Jumps for Up to 3 Days	2.1	2.1	3.15
	Initial Chatter in Variance	2.0	2.0	2.05

Also, Table 4.5 shows that when the ideal smoothing methodology is used the MAD is significantly higher when $s^2(0) = .0002$ and $.0004$ than when $s^2(0) = .0003$ in two of the three scenarios. In addition, when the combination smoothing methodology is used the MAD is significantly higher when $s^2(0) = .0002$ and $.0004$ than when $s^2(0) = .0003$ in all three scenarios. Because of this, we can conclude that having the wrong initial estimate of $s^2(0)$ results in an increased MAD and underscores the importance of having accurate initial estimates.

Tables 4.6, 4.7, and 4.8 display the cost and benefit for performing too many tests. Recall that each simulation has 20 days with 4 lots each. Tables 4.6, 4.7, and 4.8 show for each simulation, how many days took too many tests, the total number of extra tests that were taken during the course of the simulation, the average number of extra tests over all 20 days, and the average number of extra tests over just the days that had extra tests. These extra tests are considered a cost. In addition, Tables 4.6, 4.7, and 4.8 record the total savings from taking the additional tests, the average savings over all 20 days, and the average savings over just the days that took extra tests.

Table 4.6: Economic impact of performing too many tests when $s^2(0) = .0003$.

	Scenario	A	B	C	D	E	F	G
Hypothesis Testing	Variance Jumps	10	104	1.3	2.6	$1,429.39	$71.47	$142.94
	Variance Jumps for Up to 3 Days	11	84	1.05	1.91	$1,214.24	$60.71	$110.39
	Initial Chatter in Variance	4	64	0.8	4	$886.03	$44.30	$221.51
Ideal Smoothing	Variance Jumps	6	40	0.5	1.67	$759.89	$37.99	$126.65
	Variance Jumps for Up to 3 Days	6	56	0.7	2.33	$981.63	$49.08	$163.61
	Initial Chatter in Variance	2	16	0.2	2	$295.62	$14.78	$147.81
Nominal Smoothing	Variance Jumps	6	40	0.5	1.67	$905.36	$45.27	$129.34
	Variance Jumps for Up to 3 Days	8	56	0.7	1.75	$1,022.73	$51.14	$127.84
	Initial Chatter in Variance	3	32	0.4	2.67	$556.70	$27.84	$185.57

[A: number of occurrences, B: total number of extra tests, C: average number of extra tests per lot for all days, D: average number of extra tests per lot for days with too many tests, E: total savings, F: average savings per day for all days, G: average savings per day for days with too many tests.]

Table 4.7: Economic impact of performing too many tests when $s^2(0) = .0002$.

	Scenario	A	B	C	D	E	F	G
Hypothesis Testing	Variance Jumps	7	64	0.8	2.29	$1,330.73	$66.54	$147.86
	Variance Jumps for Up to 3 Days	12	120	1.5	2.5	$1,710.13	$85.51	$142.51
	Initial Chatter in Variance	10	96	1.2	2.4	$1,223.37	$61.17	$122.34
Ideal Smoothing	Variance Jumps	1	4	0.05	1	$78.76	$3.94	$78.76
	Variance Jumps for Up to 3 Days	4	32	0.4	2	$646.73	$32.34	$161.68
	Initial Chatter in Variance	0	0	0	0	$0.00	$0.00	$0.00
Nominal Smoothing	Variance Jumps	4	32	0.4	2	$568.22	$28.41	$142.06
	Variance Jumps for Up to 3 Days	7	56	0.7	2	$1,008.78	$50.44	$144.11
	Initial Chatter in Variance	0	0	0	0	$0.00	$0.00	$0.00

[A: number of occurrences, B: total number of extra tests, C: average number of extra tests per lot for all days, D: average number of extra tests per lot for days with too many tests, E: total savings, F: average savings per day for all days, G: average savings per day for days with too many tests.]

Table 4.8: Economic impact of performing too many tests when $s^2(0) = .0004$.

	Scenario	A	B	C	D	E	F	G
Hypothesis Testing	Variance Jumps	8	72	.9	2.25	$1,163.26	$58.16	$145.41
	Variance Jumps for Up to 3 Days	6	76	0.95	3.17	$1,194.97	$59.75	$199.16
	Initial Chatter in Variance	1	20	0.05	5	$295.74	$14.79	$295.74
Ideal Smoothing	Variance Jumps	20	212	2.65	2.65	$3,031.92	$151.60	$151.60
	Variance Jumps for Up to 3 Days	9	104	1.3	2.89	$1,736.62	$86.83	$192.96
	Initial Chatter in Variance	17	156	1.95	2.29	$2,176.80	$108.84	$128.05
Nominal Smoothing	Variance Jumps	10	108	1.35	2.7	$1,620.88	$81.04	$162.09
	Variance Jumps for Up to 3 Days	16	148	1.85	2.31	$2,349.40	$117.47	$146.84
	Initial Chatter in Variance	9	76	0.95	2.11	$1,080.18	$54.01	$120.02

[A: number of occurrences, B: total number of extra tests, C: average number of extra tests per lot for all days, D: average number of extra tests per lot for days with too many tests, E: total savings, F: average savings per day for all days, G: average savings per day for days with too many tests.]

Table 4.9: Economic impact of performing too few tests when $s^2(0) = .0003$.

	Scenario	A	B	C	D	E	F	G
Hypothesis Testing	Variance Jumps	7	92	1.15	3.29	$1,982.14	$99.11	$220.24
	Variance Jumps for Up to 3 Days	5	40	0.5	2	$528.12	$26.41	$105.62
	Initial Chatter in Variance	7	72	0.9	2.57	$1,358.69	$67.93	$169.84
Ideal Smoothing	Variance Jumps	10	68	0.85	1.70	$1,102.04	$55.10	$110.20
	Variance Jumps for Up to 3 Days	10	84	1.05	2.1	$1,381.37	$276.27	$138.14
	Initial Chatter in Variance	15	104	1.3	1.73	$1,888.89	$94.44	$125.93
Nominal Smoothing	Variance Jumps	10	68	0.85	1.70	$1,025.16	$51.26	$102.52
	Variance Jumps for Up to 3 Days	7	56	0.7	2	$782.33	$39.12	$111.76
	Initial Chatter in Variance	10	76	0.95	1.9	$1,193.24	$59.66	$119.32

[A: number of occurrences, B: total number fewer tests, C: average number of fewer tests per lot for all days, D: average number of fewer tests per lot for days with too few tests, E: total cost, F: average cost per day for all days, G: average cost per day for days with too few tests.]

Table 4.10: Economic impact of performing too few tests when $s^2(0) = .0002$.

Hypothesis Testing	Variance Jumps	9	64	0.8	1.78	$981.64	$49.08	$109.07
	Variance Jumps for Up to 3 Days	5	56	0.7	2.8	$1,064.03	$53.20	$177.34
	Initial Chatter in Variance	6	64	0.8	2.67	$1,346.14	$67.31	$224.36
Ideal Smoothing	Variance Jumps	17	192	2.4	2.82	$3,846.40	$192.32	$226.26
	Variance Jumps for Up to 3 Days	10	100	1.25	2.5	$2,015.53	$100.78	$201.55
	Initial Chatter in Variance	18	248	3.1	3.44	$4,952.03	$247.60	$275.11
Nominal Smoothing	Variance Jumps	13	116	1.45	2.23	$2,048.03	$102.40	$157.54
	Variance Jumps for Up to 3 Days	7	72	0.9	2.57	$1,258.04	$62.90	$179.72
	Initial Chatter in Variance	17	164	2.05	2.41	$2,925.48	$146.27	$172.09

[A: number of occurrences, B: total number fewer tests, C: average number of fewer tests per lot for all days, D: average number of fewer tests per lot for days with too few tests, E: total cost, F: average cost per day for all days, G: average cost per day for days with too few tests.]

Table 4.11: Economic impact of performing too few tests when $s^2(0) = .0004$.

Hypothesis Testing	Variance Jumps	11	108	1.35	2.45	$1,889.12	$94.46	$171.74
	Variance Jumps for Up to 3 Days	5	56	0.7	2.8	$1,850.19	$92.51	$185.02
	Initial Chatter in Variance	16	140	1.75	2.19	$2,511.41	$125.57	$139.52
Ideal Smoothing	Variance Jumps	0	0	0	0	$0.00	$0.00	$0.00
	Variance Jumps for Up to 3 Days	7	64	.8	2.29	$1,042.47	$52.12	$148.92
	Initial Chatter in Variance	1	4	0.05	1	$39.38	$1.97	$39.38
Nominal Smoothing	Variance Jumps	5	20	0.25	1	$247.57	$12.38	$49.51
	Variance Jumps for Up to 3 Days	4	24	0.3	1.5	$330.86	$16.54	$82.71
	Initial Chatter in Variance	3	12	0.15	1	$196.00	$9.80	$65.33

[A: number of occurrences, B: total number fewer tests, C: average number of fewer tests per lot for all days, D: average number of fewer tests per lot for days with too few tests, E: total cost, F: average cost per day for all days, G: average cost per day for days with too few tests.]

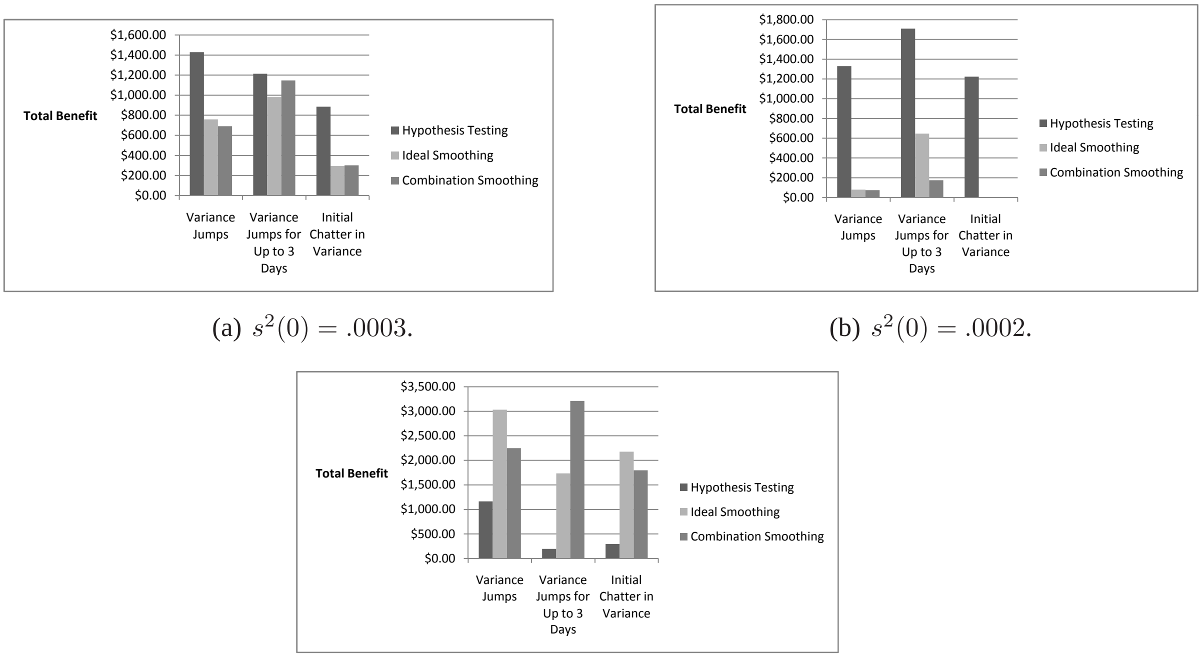
Similarly, Tables 4.9, 4.10, and 4.11 display the cost and benefit for performing too few tests. Tables 4.9, 4.10, and 4.11 show for each simulation, how many days took too few tests, the total number of fewer tests that were taken during the course of the simulation, the average number of fewer tests over all 20 days, and the average number of fewer tests over just the days that had fewer tests. These fewer tests are considered a savings. In addition, Table Tables 4.9, 4.10, and 4.11 record the total cost from taking fewer tests, the average cost over all 20 days, and the average cost over just the days that took fewer tests.

For calculating economic impact of incorrect I/D payments, it was necessary to carry out simulations with the true mean density and the best estimate of density that was obtained from using either the ideal number of tests or the number of tests that our update procedure recommended. Consider for example, the situation when the updating procedure leads to too few tests. In each simulation when this happened, lots where the payment based on the estimated mean density was higher than the payment based on the true mean density counted as an overpayment. Conversely, lots where the payment based on the true mean was higher than the payment based on the estimated mean density counted as an underpayment. This procedure was simulated 1,500 times and the average absolute error for all 1,500 simulations (either an over- or underpayment) was found. Lastly, the difference between average absolute error in payments based on the proposed procedure and the ideal number of samples was found. This is the cost of taking too few samples in one lot. This number was multiplied by 4 to represent the cost of taking too few samples for that day. A similar procedure was repeated to find the benefit from taking more than the ideal number of samples.

Tables 4.6, 4.7, and 4.8 show the economic impact of simulations that performed too many tests. When $s^2(0) = .0002$ and $.0003$, the hypothesis-test-based updates tended to take extra tests more frequently, take a greater number of extra tests, and have a higher average number of extra tests over all days than the smoothing methodology. In addition, the average number of extra tests over days that had extra tests were also higher when $s^2(0) = .0002$, $.0003$, and $.0004$. Tables 4.9, 4.10, and 4.11 display the economic impact of simulations that performed too few tests. These results correspond to times when our procedure recommended taking fewer tests than should have been observed based on the true variance of a particular day of a simulation. Here, the smoothing methodology seems to take fewer tests more frequently as well as take more fewer tests than the hypothesis testing methodology when $s^2(0) = .0002$ and $.0003$.

4.4 Conclusions

Neither method dominates the other. This is not surprising because each method is designed to perform better in different situations. Figures $4.4(a)$, $4.4(b)$, and $4.4(c)$ show the Total benefit for each scenario and methodology when $s^2(0) = .0003$, .0002 and .0004. Figures $4.5(a)$, $4.5(b)$, and $4.5(c)$ show the Total benefit for each scenario and methodology when $s^2(0) = .0003$, .0002 and .0004. In our simulations, there is a higher total benefit for the hypothesis test methodology than the smoothing methodology when $s^2(0) = .0003$ and .0002 than when $s^2(0) = .0004$. Also, the total cost for the hypothesis testing methodology is lower than the total cost for the smoothing methodology when $s^2(0) = .0002$ and higher when $s^2(0) = .0004$. So, in general, it seems that if the estimate for $s^2(0)$ is low, then the hypothesis test methodology is better in terms of economic impact. In contrast, it seems that if the estimate for $s^2(0)$ is high, then the smoothing methodology is better in terms of economic impact.

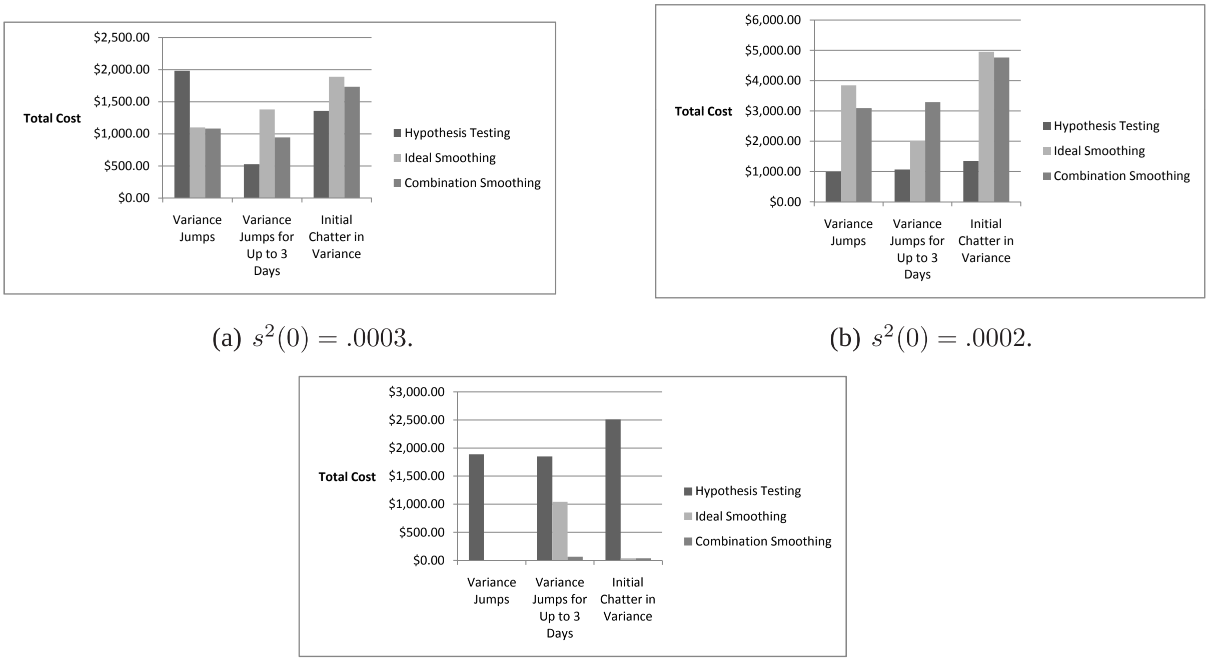


(a) $s^2(0) = .0003$.

(b) $s^2(0) = .0002$.

(c) $s^2(0) = .0004$.

Figure 4.4: Total benefit with different values of $s^2(0)$.



(a) $s^2(0) = .0003$.

(b) $s^2(0) = .0002$.

(c) $s^2(0) = .0004$.

Figure 4.5: Total cost with different values of $s^2(0)$.

Chapter 5

Conclusions

This report describes a procedure to determine how many samples should be taken to achieve a user-specified level of accuracy in identifying mean lot density. Inputs to this procedure include reliability measures (called the cutoff ratio and cutoff number) and several estimates based on past data, which represent the current knowledge about the true mean density. Then the procedure randomly generates sample density values and updates the prior distribution based on this new sample data. Using the updated prior distribution, the procedure determines the likelihood of the mean density lying in each of several ranges of density values that are equal in length. Lastly, the procedure finds the ratio of the likelihood that the mean density lies in each bin to the maximum likelihood among all bins. Termination of sampling occurs when the number of bins for which the ratio of likelihoods is above a cutoff ratio is equal to or less than a cutoff number. This procedure is repeatedly simulated to determine the recommended sample size.

Analysis of the current sampling procedure and the one we describe in this report determined that using the sample size recommended by our procedure resulted in more accurate incentive and disincentive payments than using the current sample size of two. In a simulated scenario based on representative data, accuracy increased from 47.0% under the current sampling procedure, to 70.6% under the recommended procedure.

In addition, in our procedure, smaller estimates of the variance, s^2, result in smaller recommended sample sizes. Put differently, contractors whose performance is more consistent are rewarded because they are required to test fewer samples. Moreover, the pay factor schedule still encourages contractors to perform quality work. Because of this, our procedure will encourage contractors to perform both consistent and high quality work.

As mentioned above, s^2 is a measure of the consistency of a contractor's compaction effort. Because s^2 is a key input to our procedure, we developed a method for estimating it. Furthermore, since a contractors level of consistency may change over time, we also developed methods for updating the estimate of s^2 if a change is detected. Due to the nature of changes in the variance, two methods were developed - one based on hypothesis testing and one cased on smoothing.

A comparative analysis of the two methods revealed that neither method dominates the other. This is not surprising because each method is designed to perform better in different situations. Moreover, both methodologies result in fairly accurate incentive and disincentive payments. Figures 4.5(a) to 4.5(c) show that the total cost in inaccurate incentive and disincentive payments was at most \$5000 for a project lasting 20 days. Assuming a bid price of \$40 per ton and production quantity of 2000 tons per day, a project lasting 20 days would cost \$1.6 million. So inaccuracy of \$5000 is relatively low as a fraction of total project cost. Moreover, even when an inaccurate estimate of s^2 is used, the two methodologies still result in fairly accurate incentive and disincentive payments. In summary, implementation of the procedure presented in Chapter 3 after choosing rational reliability measures, and estimating and updating the variance according to methods presented in Chapter 4, will together result in more accurate incentive and disincentive payments. This approach will also help achieve the original intent of instituting I/D schemes by strengthening the coupling between contractor performance and its rewards/penalties.

References

Buttlar, W. G. and Hausman, J. J., 2000, "ILLISIM program for end-result specification development", *Transportation Research Record* (1712), 125–136.

Carlin, B. P. and Louis, T. A., 2000, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, New York, Second Edition.

Chouband, B., Upshaw, P. B., Sholar, G. A., Page, G. C. and Musselman, J. A., 1999, "Nuclear density readings and core densities, A comparative study", *Transportation Research Record* (1654), 70–78.

Clemen, R. T. and Reilly, T., 2001, *Making Hard Decisions*, Duxbury, Thomson Learning, Pacific Grove, California.

Hayter, A., 2007, *Probability and Statistics for Engineers and Scientists*, Third Edition, Thomson Books/Cole, Belmont, California.

Hughes, C. S., 1989, *Compaction of asphalt pavement*, NCHRP Synthesis of Highway Practice 152, Transportation Research Board, Washington, DC.

Lenth, R. V., 2001, "Some practical guidelines for effective sample-size determination", *The American Statistician* **55**(3), 187–193.

McCabe, B., AbouRizk, S. and Gavin, J., 1999, "Sample size analysis for asphalt pavement quality control", *Journal of Infrastructure Systems* **5**(4), 118–123.

Mn/DOT, 2007, *Minnesota Department of Transportation Standard Specifications for Construction*. Available at http://www.dot.state.mn.us/pre-letting/prov/order/2360-2350-combined.pdf, Modified by Special Provisions, December 11, accessed August 25, 2009.

Rilett, L. R., 1998, "Effect of variability on end product specifications", *Journal of Construction Engineering and Management* **124**(2), 139–145.

Romero, P. and Kuhnow, F., 2002, "Evaluation of new nonnuclear pavement density gauges with data from field projects", *Transportation Research Record* (1813), 47–54.

Silver, E. A., Pyke, D. F. and Peterson, R., 1998, *Inventory Management and Production Planning and Scheduling*, Third Edition, John Wiley and Sons, New York.

WSDOT, 2008, *Washington State Department of Transportation Pavement Guide, Nuclear Density Gauge*. Available at http://training.ce.washington.edu/wsdot/modules/07_construction/nuclear_gauge.htm, accessed August 25, 2009.

Appendix A

Background – HMA Mixing Process & Tests

Hot Mix Asphalt Mixing Process

- Over 90% of roads are paved with HMA (Hot Mix Asphalt) because it is long lasting, cost effective, and recyclable. Asphalt is the same as Blacktop and comes from Petroleum Oil but it should not be confused with Coal Tar which comes from coal. Asphalt comes from refined crude oil and Asphalt Cement, abbreviated as AC, is a type of asphalt) that is used in Hot Mix Asphalt (HMA). To make HMA, crushed stone, gravel, and sand are mixed with AC that performs as a glue to hold the aggregates together. The mix is paved while hot which is why it is called Hot Mix Asphalt. The compaction temperature depends on both the air temperature and mat thickness and can range from 225-270 °F. HMA is designed to increase fuel efficiency of automobiles and reduce traffic build up and wear on roads due to traffic. Reclaimed Asphalt Pavement (RAP) can be reused in future road mixtures, which reduces landfill space and is just as effective as roads paved only with new materials.

- The properties of the asphalt cement and the aggregate used in a mix are important factors that determine how well the pavement will perform. The properties of asphalt cement are found using one of three procedures: penetration and viscosity grading systems, Superpave performance grading (pg) system, and temperature-viscosity characteristics. The viscosity grading system measures the viscosity of the AC at 140 degrees Fahrenheit. The Superpave performance grading system tests the performance of the AC in the climate in which the mix will be paved. The temperature-viscosity method measures how the viscosity in the AC changes as the temperature changes.

- The most important property of aggregate is surface texture because this determines its frictional resistance. Greater friction increases rutting resistance. The surface texture is measured in terms of the coarse aggregate angularity (CAA) and fine aggregate angularity (FAA) tests (see description below). Another important quality of the aggregate is particle size distribution, also known as gradation.

- Two types of facilities make HMA, called batch and drum facilities. The key difference between the two is that batch facilities make mixes in batches, whereas drum facilities produce mix in a continuous fashion.

- The following mixing procedures are common to both types of facilities:

  - The aggregates are stored in stock piles
  - Front end loaders fill a series of bins with aggregate
  - Adjustable openings at the bottom of these bins allow aggregate to fall onto a conveyor belt at varying rates which are monitored by computer and can be adjusted by plant employees. In addition, plant employees also control which aggregates are added to the mix, how much of each aggregate is added, the speed of the conveyor belt, how much oil is added to the mix later, and the temperature of the mix. Computerized monitoring allows precise amount of each ingredient to be used to create the HMA ensuring consistency with the Mix Design Report (MDR) and the production of the right quantity needed for the project.

  - From the conveyor belt, the aggregate is moved to a dryer where aggregate mix moisture is removed by heating the mix in a long cylindrical drum
  - A burner is located at one end of the dryer, where the aggregate tumbles in hot air and becomes dry and heated

- Batch facilities make mixes in batches
  - From the dryer, the aggregate is transported in an elevator to the mixing tower
  - The hot aggregate is separated by size and placed into different weigh buckets
  - From the weigh bucket, the aggregate is transported to a mixing bucket which mixes the aggregate and asphalt cement together until the mixture is properly coated
  - From the mixer, the mix is either loaded directly into trucks or stored in a holding silo until needed

- Drum facilities make mixes in a continuous stream
  - In Drum facilities the drying occurs at the beginning of a long cylindrical drum.
  - After drying, the mix passes to the other end of the drum where the asphalt and RAP (if needed) are added.
  - The mix cannot be placed directly into the trucks because the mix is made continuously. So the mix is stored in a silo from where trucks pick up the mix as needed.

- The holding silo stores and keeps the mix heated until it is picked up. The truck driver drives the truck under the silo and a plant employee loads the truck with the desired weight of HMA.

- Trucks should not have to travel too far between the plant and the paving site because otherwise the mix may cool too much in transit and no longer be at the necessary paving temperature.

Purpose of Tests

- Maximum Specific Gravity – Max specific gravity is the ratio of the mass of loose material to the mass of water with the same volume as the material and at the same temperature. This test measures the volume of the mix without any air voids. The max specific gravity of HMA depends on the ingredients used in making the mix. The results of this test are used in calculating air voids (Pa).

- Bulk Specific Gravity – Bulk specific gravity is the ratio of the mass of compacted material to the mass of water with the same volume as the material and at the same temperature. Bulk specific gravity is used in calculating air voids of compacted samples.

- Asphalt Content (AC) – Asphalt content of HMA has a direct effect on the overall life and performance of the pavement. Pavement that has too much asphalt is unstable and pavement with too little asphalt is not durable.

- Gradation – Gradation testing determines the percentage of aggregate that falls within certain ranges of size and whether these percentages meet the specifications. There are three different gradation tests one for material that does not pass the # 4 (4.75 mm) sieve, one for material that passes the #4 sieve but not the #200 (75 μm), and one for material that passes through the #200 sieve. The last test verifies whether the pavement will be susceptible to frost and/or permeability.

- Voids in Mineral Aggregate (VMA) – This test verifies whether there is a sufficient amount space between the aggregate so that the asphalt can adequately coat the aggregate. Roads that do not have enough asphalt coating the aggregate tend to be less durable.

- Adjusted Asphalt Film Thickness (Adjusted AFT) – This test is similar to the VMA test but recently has been deemed more accurate. This calculation is a ratio between effective asphalt volume and aggregate surface area.

- Air Voids – This is the percentage of the mix that is made up of air pockets. Pavement that has high air voids is susceptible to stripping, premature oxidation, premature deterioration, and rutting. On the other hand, pavement that has low air voids is susceptible to bleeding and shear flow.

- Fines/Effective Asphalt Content (AC) – This is a ratio between the percentage of aggregate that passes the #200 (75 μm) sieve and the effective asphalt content.

- Superpave consensus properties

  – Coarse Aggregate Angularity (CAA) – This test measures the percentage of coarse aggregate that is angular (not round). Coarse aggregate is aggregate that does not pass the #4 (4.75 mm) sieve. Mixes with sufficiently angular aggregate will have increased internal friction and be resistant to rutting.

  – Fine Aggregate Angularity (FAA) – This test has the same purpose as the CAA test except it is performed on aggregate that does pass through the #4 (4.75 mm) sieve.

  – Flat and Elongated Particles – This test determines the percentage of aggregate that have flat surfaces because aggregate with flat faces can crack and break during construction.

  – Clay Content – This test measures how much clay is in the mix. Water causes clay to expand and so too much clay in the mix will lead to poor road quality.

Testing Procedures

- Take sample (about 6000 grams) out of oven once it reaches 160-230 degrees Fahrenheit

- Blend the mix by hand

- For the Marshall design, quarter the mix for three Marshall specimens and one rice specimen

  - Load the three Marshall quarters with an equal weight (ideally 2.5 inches tall and 1200-1225g of mix) into three Marshall molds that were also heated in oven to 275 10 degrees Fahrenheit
  - Rice sample is 2000-2050 grams
  - Place rice sample in a larger pan and spread it out so that the mix is no larger than the largest aggregate or as small as of an inch
  - Re-blend the leftover material and take an extraction sample of 2000-2100g which will be used for the % Asphalt Cement and Gradation tests

- For the gyratory design (which require larger samples), divide the mix in half
  - Load two Gyratory specimens that are each 4800 g and 115mm 5 in height. Also include a thermometer in the bucket to indicate when the mix reaches compaction temperature (based on pg grade of the sample).
  - Replace the specimens in the oven.
  - Re-blend the leftover material
  - Take out 2000-2050g for rice test (same as above)
  - Take out 2000-2100g for extraction (same as above)

- Gyratory Compaction
  - Remove thermometer once the mix reaches compaction temperature
  - Place paper at the bottom of mold
  - Place mix in mold in one pour so as not to cause segregation
  - Place top paper and top plate on top of mix
  - Place mold in machine, set the machine for the desired number of gyrations and angle. Push start.
  - A ram places pressure on the mix from above while the machine spins the mold and tilts the mold at a desired angle.
  - If needed, leave material to set and cool for a few minutes
  - Extract sample, remove top plate, top paper, and bottom paper
  - Set sample in front of fan to cool completely

- Marshall Compaction
  - Once mix is at compaction temperature remove from oven and remove thermometer
  - Set all three molds on machine
  - Turn machine on and spade the material 15 times on the outside, 10 in the middle, and then bring the material to a cone
  - Turn the machine off
  - Put paper on the top of each sample

  - Load hammers, which have been heated on a hot plate (200-300 degrees Fahrenheit), on machine and engage safety counter-weight
  - Set the machine to the required number of hammer blows (determined by mix type) and turn machine on
  - The hammers each give a certain number of blows to the mix inside the mold while the molds spin
  - Make sure all hammers had the exact same number of blows
  - Turn samples over, remove any mix that might have been loosened, rotate hammers, and repeat the same number of blows in the same manner on the other side
  - Remove hammers and replace on hot plate
  - Remove samples, remove paper from both sides of sample, mark the last side pounded on each sample with the sample number
  - Set samples in front of fan to cool to room temperature

- Maximum Specific Gravity (Rice) Test

  - Chop up mix for rice test so that it is smaller than its largest aggregate or smaller than 1/4 inch
  - Transfer all of the mix into a calibrated container along with a screen that allows the water to flow freely through the mix
  - Weigh the sample in air. The scale used in weighing the samples should be sufficiently accurate so that the max SpG can be calculated to four decimal places.
  - Fill the container with 77 ± 1.8 degree Fahrenheit water at least a half inch higher than the mix
  - Make sure there are no floating particles. If there are some that cannot be knocked down by hand, use up to 15 drops of aerosol OT. Aerosol OT is a wetting agent that solubilizes, emulsifies, and reduces interfacial tension.
  - Place sample in rice apparatus, put top on to seal apparatus, turn vacuum on until the sample reaches a vacuum of 30mm of mercury pressure, and then turn on vibrator and set timer for 15 minutes
  - After 15 minutes turn off vibrator, slowly remove vacuum, remove sample from apparatus, and check for floating particles, knocking them down with finger or using any remaining drops of aerosol OT (recall that the maximum is 15 drops).
  - Carefully place sample on scale in 77 ± 1.8 degree Fahrenheit water bath
  - Set timer for 10 minutes
  - After 10 minutes, weigh sample and remove sample from scale
  - The max SpG is calculated by:

$$\text{Max SpG } = \frac{A}{A-(C-B)}$$

where:
A = weight of dry sample in air in grams
B = weight of container in water in grams
C = weight of container and sample in water in grams

- Bulk Specific Gravity
  - Marshall specimens
    - Bulk all three samples at once
    - The test requires a scale that reads to 1/10g, a water bath that is 77 ± 1.8 degrees Fahrenheit, timer for 3-5 minutes, and a damp towel
    - Weigh each dry sample in air
    - Place samples in water bath for 3-5 minutes, turn off bath so there is no motion, and weigh each sample under water
    - Take less than 15 seconds to roll the samples on the damp towel (removes excess water) and weigh each saturated sample in air. This is the surface saturated dry weight (SSD).
  - Gyratory specimens
    - Only bulk one sample at a time
    - Measure dry weight in air
    - Place sample in water bath for 3-5 minutes and weigh sample in water
    - In less than 15 seconds, take sample out, roll on damp towel, and weigh surface saturated weight
  - The bulk Specific Gravity calculation is:

$$\text{Bulk SpG } = \frac{A}{B-C}$$

    where:
    A = weight in grams of the specimen in air
    B = weigh in grams, surface saturated dry
    C = weight in grams, in water
  - The bulk SpG should be recorded to the nearest 0.001 grams.
- Asphalt Content (%AC)
  - Wear a face shield, gloves, long sleeve jacket or long sleeves, and apron that are all appropriate for high temperatures
  - Preheat the ignition oven to $1000\,^{\circ}\mathrm{F}$
  - Weigh the container and lid to the nearest 0.1 gram at $300\,^{\circ}\mathrm{F}$ or more.
  - Pour sample tray and spread it out with a hot spatula. Return tray to preheat oven until it reaches a constant mass. Constant mass is achieved when the change in mass of the

sample and tray does not exceed 0.01% of the original mass of the specimen. For example, in step 2 check if:

$$\frac{W1 - W2}{W2} \times 100 \leq .01\% \times \text{ original mass}$$

If this is not true, repeat until in Step 3:

$$\frac{W2 - W3}{W3} \times 100 \leq .01\% \times \text{ original mass}$$

Keep repeating until the change in mass is less than 0.01% of the original mass. After constant mass is achieved, weigh the tray and material to the nearest 0.1 gram. This is the initial weight of the specimen.

  - Place the tray in the ignition oven (recall that it is at $1000\,^{\circ}\mathrm{F}$).
  - After the burn cycle is complete, remove the tray from the oven using safety equipment. Place the sample in an oven heated to $300\,^{\circ}\mathrm{F}$. Once the specimen reaches $300\,^{\circ}\mathrm{F}$, record the weight. The return the tray to the ignition oven ($1000\,^{\circ}\mathrm{F}$) for an additional 15 minutes to ensure constant mass has been achieved. If constant mass has not been achieved, repeat this step until it is achieved
  - Finally, once the tray and sample have achieved constant mass at or above $300\,^{\circ}\mathrm{F}$, weigh the sample and tray together to the nearest 0.1 gram.
  - Deduct tray and lid weight from all measurements.
  - AC is calculated as:

$$\%\text{AC } = \frac{W_s - W_a}{W_s} \times 100 - C_f$$

    where:
    W_a = the total weight of aggregate remaining after ignition
    W_s = the total weight of the HMA sample prior to ignition
    C_f = calibration factor, percent by weight of HMA sample

  - C_f depends both on the asphalt calibration factor and the aggregate correction factor.
  - %AC should be recorded to the nearest 0.01%

- Gradation
  - The sieves used in this test are 1.06 in. (26.5 mm), 3/4 in. (19.0 mm), 1/2 in. (12.5 mm), 3/8 in. (9.5 mm), #4 (4.75 mm), #8 (2.36 mm), # 16 (1.18 mm), #30 (600 μ m), #50 (300 μm), #100 (150 μm), and #200 (75 μm).
  - Fine and Coarse Aggregate
    - For fine aggregate, this test is performed on aggregate that passes the #4 sieve but not the #200 sieve. For coarse aggregate, this test is performed on aggregate that does not pass the #4 sieve.

        - ∗ Weigh the sample to the nearest 0.1g. This measurement will be used to calculate the percentage in each sieve as well as to check for any lost material.
        - ∗ Nest the sieves in ascending order with the smallest sieve at the bottom. Load the aggregate into the top of the nested sieves, taking care not to overload the sieves.
        - ∗ Place nested sieves in a shaking mechanism
        - ∗ After sieving, remove each tray and record each corresponding weight to the nearest 0.1g, taking care to remove all material from each tray, including any material stuck to the sides or in the sieve openings.
        - ∗ The total weight of the aggregate in each sieve should be within 0.3% of the original weight (i.e., no more than 0.3% should be lost in this testing process).
    - – Finer than #200 sieve aggregate
        - ∗ Take a sample of mix (entire mix, not just mix that passes the #200 sieve). Dry the sample to a constant weight. Weigh the sample to the nearest 0.1 gram, this is the dry weight.
        - ∗ Place the sample into a large container and cover the material with water (if necessary, add wetting agent).
        - ∗ Agitate the container so that particles finer than the #200 sieve are suspended in the water
        - ∗ Pour the water (not the mix) through a #200 (75gm) sieve.
        - ∗ Add more water to the container and re-pour through the #200 sieve until the water is clear. Any aggregate remaining in the #200 sieve should be re-placed in the original container.
        - ∗ Place the aggregate remaining in the container into a tray and heat in an oven, on an electric skillet, or over an open flame until the sample reaches a constant weight. Record this as the new dry weight.
        - ∗ The calculation for percentage passing the #200 sieve is

$$\text{finer than \#200 sieve } = \frac{B-C}{B} \times 100$$

          where:
          B = Original dry weight of sample in grams
          C = Dry weight of sample after washing and drying to constant weight in grams
        - ∗ The percentage passing the #200 sieve should be recorded to the nearest 0.1%.

- Voids in Mineral Aggregate (VMA)
    - – This test measures the percentage of this mix that is not made up of air voids or effective AC
    - – VMA is calculated by:

$$\text{VMA } = 100 - \frac{Ps \times Gmb}{Gsb}$$

      where:
      Ps = Aggregate content, percent by total mass of mixture

Gsb = bulk specific gravity of total aggregate
Gmb = bulk specific gravity of compacted mixture

- Adjusted AFT
  - This test is similar to the test for VMA and will eventually replace the VMA test because it is more accurate
  - The purpose of this test is to calculate the surface area of the aggregate so that an estimate can be found for the amount of asphalt coating the aggregate
  - First calculate the aggregate surface area (SA) by:

    $$Sa = 2 + 0.02a + 0.04b + 0.08c + 0.14d + 0.30e + 0.60f + 1.60g$$

    where: a, b, c, d, e, f, g are the percentage of aggregate passing through the #4, #8, #16, #30, #50, #100, and #200 sieves, respectively. The percentage passing through the #200 sieve should be rounded to the nearest 0.1% and the rest to the nearest 1%. Recall that sieve #4 is 4.75 mm, #8 is 2.36 mm, #16 is 1.18 mm, #30 is 600 μm, #50 is 300 μm, #100 is 150 μm, and #200 is 75 μm.
  - Next AFT is calculated by:

    $$\text{AFT } = \frac{Pbe \times 4870}{100 \times Ps \times SA}$$

    where:
    Pbe = Effective asphalt content as a percent of the total mixture 4870 = Constant Conversion Factor
    Ps = Percent Aggregate in Mixture/100 or equivalently $\frac{100-Pb}{100}$
    Pb = Percent Total Asphalt Cement in Mixture
    SA = Calculated Aggregate Surface Area in SF/lb.
  - Finally, Adjusted AFT is calculated by:

    $$\text{Adjusted AFT } = AFT + 0.06 \times (SA - 28)$$

- Air Voids
  - A ratio of how many air voids are present in the mix
  - Air voids are calculated as:

    $$\text{Air Voids } = 100(1 - \frac{A}{B})$$

    where:
    A = bulk specific gravity
    B = maximum specific gravity

- Fines / Effective AC
  - This test is a ratio of the percentage of aggregate that passes the #200 (75 μm) sieve divided by the effective AC
  - The fines/effective AC is determined by:

    $$\text{fines/effective AC } = \frac{A}{Pbe}$$

    where:
    A = percentage of aggregate passing the #200 sieve
    Pbe = effective asphalt

    $$\text{Pbe} = Pb - \frac{Pba \times Ps}{100}$$

    Pb = percent asphalt in mix
    Pba = percent absorbed asphalt
    Ps = percent stone in mix (100-%AC)

    $$\text{Pba} = 100 \times \frac{Gse - Gsb}{Gse - Gsb} \times Gb$$

    Gse = effective specific gravity of the aggregate
    Gsb = bulk specific gravity of the aggregate
    Gb = specific gravity of the asphalt

    $$\text{Gse } = \frac{100 - Pb}{\frac{100}{Gmm} - \frac{Pb}{Gb}}$$

    Gmm = max gravity (rice test) of the mix
    Pb = percent asphalt in mix

- Superpave – Superior Performing Asphalt Pavement. Four consensus properties are critical for testing the quality of Superpave
  - Coarse Aggregate Angularity (CAA)
    - Ensures a high degree of aggregate internal friction and rutting resistance
    - It is defined as the percent by weight of particles retained on the #4 (4.75 mm) sieve with one or more fractured faces
    - The procedure to determine the percentage varies from state to state but usually involves manually counting the stones and visually determining whether they have been crushed adequately
    - Crushed particles are preferred because they will interlock with each other
  - Fine Aggregate Angularity (FAA)
    - This also ensures a high degree of aggregate internal friction and rutting resistance
    - It is defined as the percent of air voids present when the aggregates that do pass the #4 (4.75 mm) sieve are loosely packed in a container

    - The sample of fine aggregate is blended and weighed and poured into a funnel and held in place. Then the mixture is allowed to flow into a cylindrical container of known volume. The cylinder is then weighed to determine the amount of air voids.
    - The calculation is: (V-W/Gsb)/V x 100% where V is the known volume of the cylinder, W is the weight of the aggregate, and Gsb is the fine aggregate bulk specific gravity. Superpave requirements for FAA are 40-45%

- Flat and Elongated Particles
    - Flat surfaced aggregate is undesirable because it can crack and break during construction
    - This test is performed on materials retained on the #4 (4.75 mm) sieve
    - A proportional caliper device is used to measure the dimensional ratio of a sample of aggregate. The ratio of the longest dimension to the shortest dimension should not exceed 5:1, otherwise the aggregate is considered flat. No more than 10% of the aggregate should have more than a 5:1 ratio. If a mix fails, the aggregate content would need to be adjusted
- Clay Content
    - Water causes clay to expand. So this test is used to limit the amount of clay in the mix because in excessive amounts clay can harm the quality of the pavement.
    - Mix any sand present in the MDR with a sample of aggregate that passes the #4 (4.75 mm) sieve in order to keep the clay content in proportion to the MDR
    - Place the mix in a graduated cylinder. Agitate the cylinder until the ingredients settle
    - Measure the clay and sedimented sand. Then the height of the sand is divided by the height of the clay and sand (the sand settles to the bottom and the clay settles in the middle) and then that quotient is multiplied by 100. The clay content has a maximum of 40-50%.