

# Optimization of Retro-Reflectometry Measurements

Amal Latif  
Zhiguo Pan  
Aparna Rajgopal

May 5, 2004

**Abstract:** The Michigan Department of Transportation (MDOT) employs contractors to gather reflectivity readings from paint lines on roadways in Michigan using mobile retro-reflectometry vans. The contractors currently measure 2-mile subsections within every 10 miles of each roadway. We analyzed the reflectivity data to determine the minimum length of each subsection to be measured that would provide a reliable estimate of paint reflectivity within that section. The results of the analysis will be used to assist MDOT in minimizing the costs associated with obtaining the readings, while still preserving the accuracy of the readings.

# Table of Contents

1. Introduction.....	2
2. Reflectivity Data.....	2
3. Theoretical Framework.....	3
3.1. <i>Probability Distributions</i> .....	3
3.2. <i>Normal Distribution</i> .....	3
3.3. <i>Confidence Intervals</i> .....	5
3.4. <i>Sample Sizes</i> .....	6
4. Analysis of Reflectivity Data.....	6
4.1. <i>Assumptions</i> .....	6
4.2. <i>Minimum Length of Roadway Subsections not given <math>\sigma</math></i> .....	7
5. Results and Recommendations.....	8
6. Suggestions for future work.....	8
References.....	9
Appendix A.....	10
Appendix B.....	11
Appendix C.....	12

# 1. Introduction

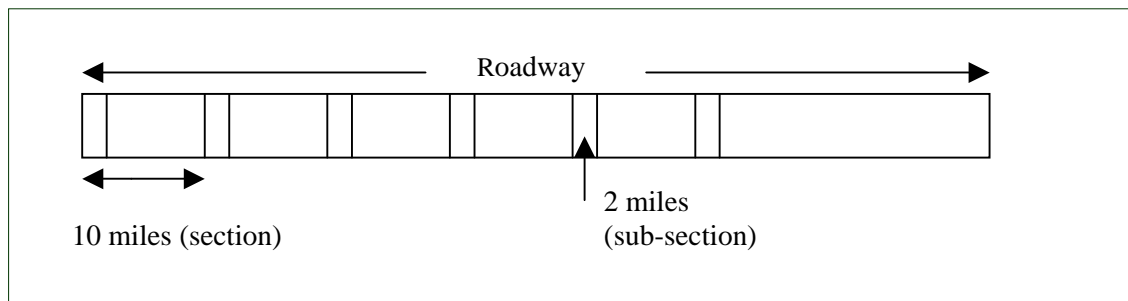
The Michigan Department of Transportation (MDOT) engages the use of mobile retro-reflectometry vans, driven by contractors, to measure the reflectivity from paint lines on freeways and state roadways in Michigan. The vans currently measure subsections within a roadway to gather data on reflectivity. The number of measurements taken, in turn determines the cost of hiring the contractors.

The goal of this project is to determine the minimum length of subsections to be measured, by identifying the smallest sample size that would be a reliable representation of the average reflectivity of paint for the entire section. These results will then be used to establish a cost-effective strategy for MDOT, while meeting traffic safety standards.

The report will focus on the analysis of the reflectivity data obtained from the mobile retro-reflectometry vans for the year 2003 using confidence intervals and sampling studies. Based on the results of our analysis, we will then provide recommendations for selecting a strategy that would reduce the cost of employing the contractors, without compromising the reliability of the readings.

## 2. Reflectivity Data

MDOT divides the state of Michigan into twelve geographic regions and employs mobile retro-reflectometry vans to measure the reflectivity from the paint lines on roadways in each region. These readings are reported in average MCD units ( $\text{mcd}/\text{m}^2/\text{lux}$ ) where, a lower reading represents a lower reflectivity of paint (see reference [1]). The vans take these measurements in October of every year to establish whether the paint lines meet the stipulated reflectivity requirements for traffic safety. This takes into account wear and tear due to traffic and weather conditions. Reflectivity or MCD data from these vans has been collected from 1998-2003.



**Figure 1.** Division of a roadway for current measurement procedure.

The vans currently measure paint reflectivity from 2-mile subsections within every 10-mile section of a roadway, as illustrated in Figure 1. This process is implemented for white and yellow lines, both edges and center. For every section, these readings are measured in both directions (north and south or east and west). Each subsection consists of a number of individual readings (raw data), which are collected,

and the average and standard deviation are determined for the entire 10-mile section. The statistics are then used by MDOT to identify the lines that require repainting and the quality of paint to be used so as to meet traffic standards.

### 3. Theoretical Framework

#### 3.1. Probability Distributions

In order to render a comprehensive examination of the given data, it is imperative to first identify the probability distribution that the data follows. The two main functions considered when studying probability distributions are the probability density functions (*pdf*) and cumulative density functions (*cdf*) [2]. A *pdf*,  $f$ , has the following properties:

$$f(x) \geq 0 \text{ for all } x,$$

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

$$P(a < x < b) = \int_a^b f(x)dx,$$

where  $P(a < x < b)$  is the probability that  $x$  lies between  $a$  and  $b$ .

On the other hand, a *cdf*,  $F$ , is a function defined over the real number line with the following properties:

$$F(x) = P(t \leq x) = \int_{-\infty}^x f(t)dt,$$

$$x_1 < x_2 \text{ implies } F(x_1) \leq F(x_2),$$

$$\lim_{x \rightarrow -\infty} F(x) = 0,$$

$$\lim_{x \rightarrow \infty} F(x) = 1.$$

Using this theoretical background, we identified that the MCD data follows the Normal Distribution.

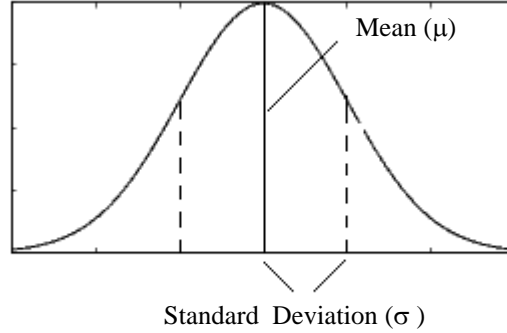
#### 3.2. Normal Distribution

The *pdf* of the normal distribution denoted as  $N(\mu, \sigma)$ , is a continuous, symmetric function, also called the *bell* or *normal curve* with two parameters: the mean  $\mu$  and the standard deviation  $\sigma$ . The justification for using the normal distribution for modeling is the central limit theorem, which states that the sum of independent samples from any distribution with finite mean and variance converges to the normal distribution as the sample size goes to infinity [3].

The normal distribution has pdf, given by equation (1):

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (1)$$

The plot in Figure 2. shows the bell curve of the normal pdf, with mean  $\mu$  and standard deviation  $\sigma$ .



**Figure 2.** The pdf of the Normal Distribution.

Given a population with unknown mean  $\mu$  and a random sample of size  $n$  from this distribution, say,  $X_1, X_2, \dots, X_n$ , the estimate of  $\mu$  is the sample average  $\bar{X}$  where  $\bar{X}$  is defined as

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Since the mean of a random sample of size  $n$  from the normal distribution  $N(\mu, \sigma^2)$  also follows the normal distribution  $N(\mu, \sigma^2)$ , the standardized variable

$$Y = (\bar{X} - \mu) / (\sigma / \sqrt{n}) \quad (2)$$

is  $N(0, 1)$ , known as the *standard normal distribution* with  $\mu = 0$  and  $\sigma = 1$ . Moreover, if  $n$  is large enough, then by the central limit theorem, the variable given by (2) has the approximate normal distribution  $N(0, 1)$  even though the underlying distribution might not be normal [2].

If the distribution of some phenomenon approaches a normal curve, the curve can be used to describe the probabilities associated with the phenomenon. The given MCD data provided from the previous years represents the averages of raw data measured from 2-mile subsections, reported with their standard deviations. Based on the theoretical framework of the various distributions, it is fairly accurate to approximate the MCD readings to a normal distribution.

### 3.3. Confidence Intervals

The sampling problem is a study where one wishes to determine the optimum sample size from a population with a desired level of confidence that the point estimate is correct within a given margin of error. The sample size is calculated based on the distribution of a random variable  $X$ , which represents the quantity under study, and whether or not the standard deviation of the population is known.

The theoretical framework of the normal distribution can be used to construct confidence intervals for the unknown means of distributions. For a random sample from a distribution with unknown mean  $\mu$  but known variance  $\sigma^2$ , given a probability  $\gamma$  and  $n$  sufficiently large, we can find a number  $z_0$ , known as the *critical value*, from the normal distribution table such that

$$P\left(-z_0 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_0\right) \approx \gamma. \quad (3)$$

For example, if  $\gamma=95\%$ , then  $z_0 = 1.96$ , and if  $\gamma=99\%$ , then  $z_0 = 2.58$ , where  $\gamma$  represents the *level of confidence* that the statement is true. Using equation (3), we can then construct the confidence interval that contains the mean  $\mu$  for a given  $\gamma$ , and since  $\sigma \geq 0$ ,

$$P\left(-z_0\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X} - \mu \leq z_0\left(\frac{\sigma}{\sqrt{n}}\right)\right) \approx \gamma, \quad (4)$$

$$P\left(\bar{X} - z_0\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + z_0\left(\frac{\sigma}{\sqrt{n}}\right)\right) \approx \gamma, \quad (5)$$

where equation (5) represents the required confidence interval for  $\mu$  and the upper and lower bounds are known as the *confidence limits*. Equivalently, the probability that the interval

$$\left[\bar{X} - z_0\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_0\left(\frac{\sigma}{\sqrt{n}}\right)\right],$$

includes the unknown mean  $\mu$  is approximately  $\gamma$ .

The confidence limits in equation (5) represent the sample average factored in with a margin of error, for a given sample of size  $n$  and confidence level  $\gamma$ . This error is given by the equation

$$z_0\left(\frac{\sigma}{\sqrt{n}}\right) = \varepsilon, \quad (6)$$

where  $\varepsilon$  represents the acceptable margin of error between  $\mu$  and  $\bar{X}$  [2].

### 3.4. Sample Sizes

The problem of determining the sample size depends on the standard deviation of the population. For the case when  $\sigma$  is given, the confidence interval would be the sample average with an error bound as expressed in equation (5). We can then find the required sample size needed to ensure an error  $\varepsilon$  with a confidence level  $\gamma$  by the relation

$$n = \frac{z_0^2 \sigma^2}{\varepsilon^2}, \quad (7)$$

where  $z_0$  is determined by the confidence level  $\gamma$ .

Typically, the standard deviation is not explicitly prescribed but follows its own distribution which may be computed from the given data. In this case, rather than considering specific values for the standard deviation, we consider the probability that  $\sigma$  lies within a certain interval which is calculated from a given distribution of  $\sigma$ . The probability in equation (4) is then calculated as the conditional probability

$$P\left(\left|\bar{X} - \mu\right| \leq \varepsilon\right) = \sum_{x=1}^m P(\sigma \in \sigma_x) \cdot P\left(\left|\bar{X} - \mu\right| \leq \varepsilon \mid \sigma \in \sigma_x\right), \quad (8)$$

where  $P(\sigma \in \sigma_x)$  is the probability that the standard deviation of the population lies within a certain interval  $\sigma_x$ ,

$m$  is the number of disjoint intervals in which  $\sigma$  could lie in, and

$P\left(\left|\bar{X} - \mu\right| \leq \varepsilon \mid \sigma \in \sigma_x\right)$  is the conditional probability given  $\sigma \in \sigma_x$ .

## 4. Analysis of Reflectivity Data

In order to perform an analysis to determine the optimal length of subsections to be measured, we first identified the assumptions that formed the framework of the study, and then computed the minimum number of individual MCD readings within a subsection that would represent the average reflectivity within that section.

### 4.1. Assumptions

MDOT currently divides the state of Michigan into geographic regions such as southwest, metro, superior regions etc., each of which is then assigned to a contractor to gather the MCD readings. This enables MDOT to monitor the wear and tear of the paint lines due to weather conditions and traffic density on the road. For the purpose of this study, we assume that the geographic divisions are small enough such that the weather conditions within each region do not vary. This enables us to eliminate the weather as a parameter of our study.

We also assume that the traffic density (number of vehicles per 10 miles of a road) within a particular region is uniform. Therefore, the wear and tear of paint due to tire friction is uniform over one section on a certain road. In addition we assume that the distribution of the standard deviations in each geographic region does not vary over the

years (see appendix A, Figure 3). Furthermore, if the machinery for measuring the retro-reflectivity is changed significantly, then the distributions of  $\sigma$  would need to be retabulated.

4.2. Minimum Length of Roadway Subsections not given  $\sigma$

The reflectivity data for a 2-mile subsection consists of approximately 1000 data points, which are then averaged to give the mean and standard deviation of the readings within that section. Equivalently, every mile is 500 data points.

In order to determine the optimal length of the roadway subsections, we attempted to determine the minimum number of data points, or the sample size, that would represent each sectional average  $\mu$ , for a certain roadway. Based on the theoretical framework of sample sizes when  $\sigma$  is not given, as described in Section 3.4, we formulated the problem to compute the sample size satisfying

$$P\left(\left|\bar{X} - \mu\right| \leq \varepsilon\right) > \gamma, \tag{9}$$

where  $\gamma$  is the probability or confidence level that the difference between the sample mean and actual sectional mean is within an acceptable margin of error  $\varepsilon$ .

Considering that  $\sigma$  is not fixed, we first computed the probabilities that the standard deviation of a subsection lies within intervals of ten. This was done based on the distribution of  $\sigma$  for the twelve geographic regions (See appendix A, Table 1). We then computed the conditional probability, as expressed in equation (8) by taking  $\sigma$  to be the upper bound of the interval. Thus, for an acceptable level of error  $\varepsilon$ , we solved equation (9) in which

$$P\left(\left|\bar{X} - \mu\right| \leq \varepsilon\right) = P(\sigma \leq 10) \cdot P\left(\left|\bar{X} - \mu\right| \leq \varepsilon \mid \sigma = 10\right) + P(10 < \sigma \leq 20) \cdot P\left(\left|\bar{X} - \mu\right| \leq \varepsilon \mid \sigma = 20\right) + \dots$$

where

$$P\left(\left|\bar{X} - \mu\right| \leq \varepsilon \mid \sigma = 10\right) = \int_{-\frac{\varepsilon}{10/\sqrt{n}}}^{\frac{\varepsilon}{10/\sqrt{n}}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,$$

to yield a sample size  $n$  that has an estimate of  $\mu$  with a confidence level  $\gamma$  and marginal error  $\varepsilon$ . The number of data points obtained was then translated to length, in miles, by dividing  $n$  by 500 to yield the minimum subsection length. This approach was repeated for various errors using different confidence levels (see appendix B, Table 2).

In order to perform the required calculations, we used the tools in Excel and also developed a code using the mathematical software, MATLAB, which facilitated easy computations of the required sample sizes and generated graphs for the same (see appendix C).

## **5. Results and Recommendations**

Our analysis revealed that the minimum number of readings required to give a reliable representation of average reflectivity for the entire section is less than the current measurement of 2 miles for almost all sections, but the minimum length varies from one region to the other (see appendix B, Table 2).

In general, the smaller the margin of error, the greater the length of the subsection to be measured. For example in the Bay region, for an error of 20 MCD, the required length while maintaining a 99% confidence level is 0.09 miles, while for an error of 5 MCD, the required length is 1.428 miles. As can be seen from the results, even with the smallest error, the minimum length that would still provide a reliable estimate of the actual readings is less than 2 miles for most sections. Furthermore, in some regions, for the smallest error of 5 MCD, the minimum length proved to be greater than 2 miles. For example, in the Southwest region, 1228 readings are required within each subsection to obtain a margin of error no greater than 5 MCD with confidence 99%.

Overall, the cost associated with hiring contractors to gather MCD readings throughout Michigan could be significantly reduced without compromising the accuracy of the readings. An illustration of the results can be seen in appendix B, Figure 4.

## **6. Suggestions for future work**

This study can be extended to determine the optimal length of each section or, equivalently, the number of sections a road must be divided into. This could eliminate redundancy in the data and also reduce the costs associated with gathering the MCD readings. However, the analysis needs an understanding of the correlation length, which in turn would require the full 500 point/mile data set over a distance of 10-50 miles for several representative roads.

A more detailed study of optimizing the measurement procedure in gathering MCD data may involve determining how the state of Michigan could be divided into regions. Currently, roadways are divided based on geographic regions, however, by conducting a correlation analysis between traffic density and MCD readings, one can determine the relation between the two factors and demarcate roads based on traffic density rather than geography. Furthermore, other factors such as weather conditions could also be factored into the analysis by taking the readings more than once a year. While the approach in this project focused on analyzing the data for one year and generalizing the results, a more detailed examination may be performed using data for several years.

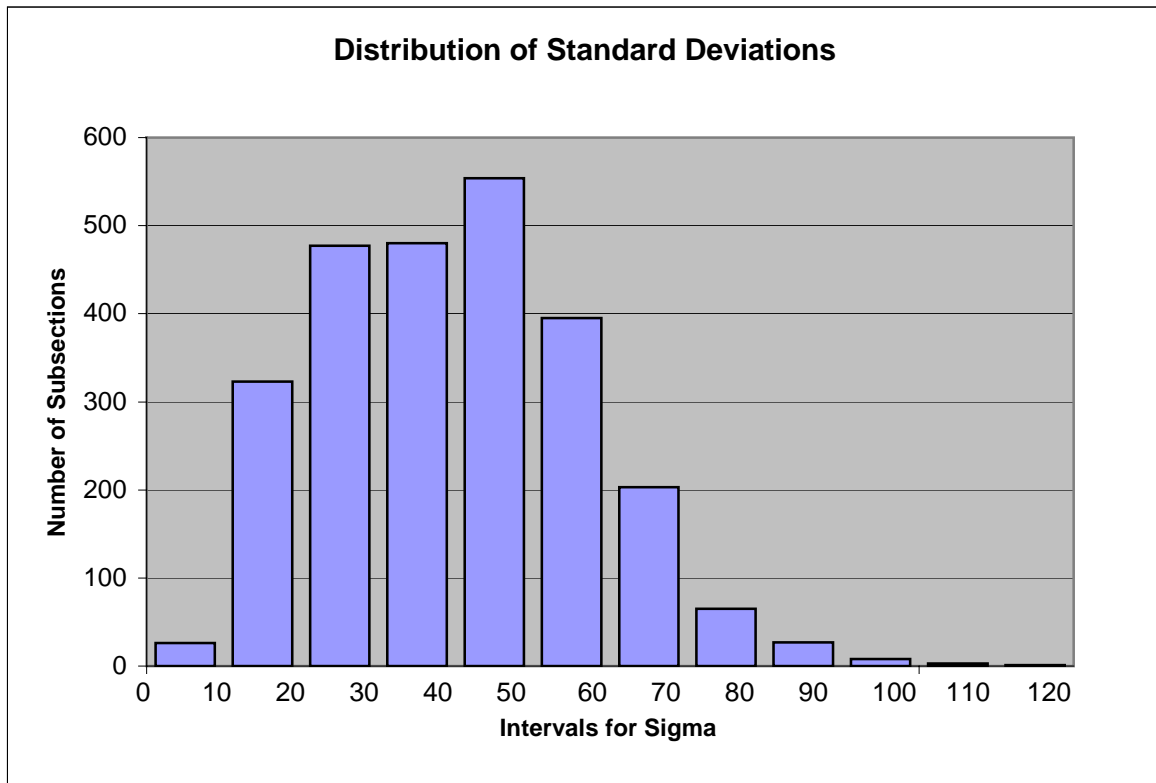
## References

- [1] B.C. Traffic Engineering, Inc., Prepared for MDOT, "Clark Highway Services-  
R<sub>L</sub> Summary", 2003, [12 regions].
- [2] Hogg, R.V. and E.A. Tanis, *Probability and Statistical Inference*, MacMillan  
Publishing Co., Inc., New York, 1977.
- [3] Evans, Merran, Nicholas Hastings and Brian Peacock, *Statistical Distributions*,  
John Wiley & Sons, New York, 2000.

## Appendix A

**Table 1.** Distribution and probabilities of standard deviations in Michigan

<b>Region</b>	(0, 10]	(10, 20]	(20, 30]	(30, 40]	(40, 50]	(50, 60]	(60, 70]	(70, 80]	(80, 90]	(90, 100]	(100, 110]	(110, 120]	<b>Total</b>
Bay	21	106	80	73	70	57	19	9	1	0	1	0	437
Grand	0	1	10	34	28	27	20	7	2	0	0	1	130
Metro	0	25	73	66	91	56	39	9	2	1	0	0	362
Alpena	2	14	6	21	26	16	11	2	0	1	0	0	99
Cadillac	0	41	25	20	42	20	10	0	0	0	0	0	158
Grayling	2	15	44	37	34	13	8	3	0	0	0	0	156
Traverse	0	15	48	27	29	33	7	0	0	0	0	0	159
S. West	0	1	15	29	43	39	25	10	11	3	2	0	178
Superior	1	81	75	83	107	85	43	22	11	2	0	0	510
Brighton	0	10	19	21	15	17	7	0	0	0	0	0	89
Jackson	0	10	45	18	27	14	6	2	0	0	0	0	122
Lansing	0	4	37	51	42	18	8	1	0	1	0	0	162
<b>Total</b>	<b>26</b>	<b>323</b>	<b>477</b>	<b>480</b>	<b>554</b>	<b>395</b>	<b>203</b>	<b>65</b>	<b>27</b>	<b>8</b>	<b>3</b>	<b>1</b>	<b>2562</b>
<b>Prob (%)</b>	<b>1.01</b>	<b>12.61</b>	<b>18.62</b>	<b>18.74</b>	<b>21.62</b>	<b>15.42</b>	<b>7.92</b>	<b>2.54</b>	<b>1.05</b>	<b>0.31</b>	<b>0.12</b>	<b>0.04</b>	

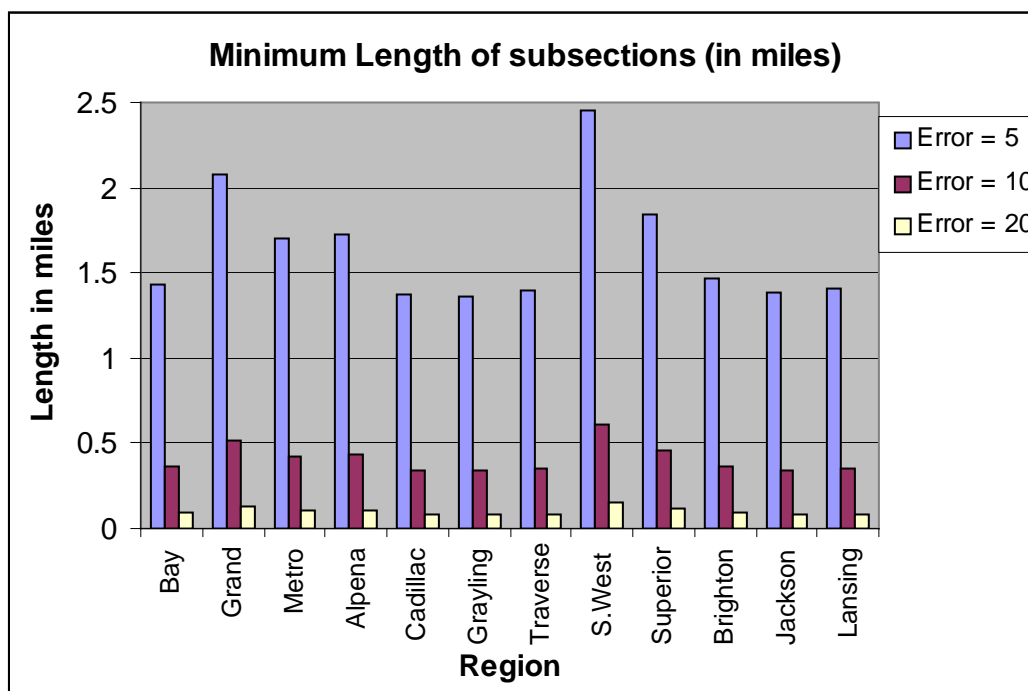


**Figure 3.** Distribution of standard deviations.

## Appendix B

**Table 2.** Minimum length of subsections for 99% confidence level.

Region	Error = 5		Error = 10		Error = 20	
	N	Length in miles	N	Length in miles	N	Length in miles
Bay	714	1.428	179	0.358	45	0.09
Grand	1037	2.074	260	0.52	65	0.13
Metro	850	1.7	213	0.426	54	0.108
North Alpena	864	1.728	216	0.432	54	0.108
North Cadillac	688	1.376	172	0.344	43	0.086
North Grayling	679	1.358	170	0.34	43	0.086
North Traverse	699	1.398	175	0.35	44	0.088
Southwest	1228	2.456	307	0.614	77	0.154
Superior	920	1.84	230	0.46	58	0.116
Univ Brighton	736	1.472	184	0.368	46	0.092
Univ Jackson	690	1.38	173	0.346	44	0.088
Univ Lansing	704	1.408	176	0.352	44	0.088
<b>Average</b>	<b>841</b>	<b>1.682</b>	<b>211</b>	<b>0.422</b>	<b>53</b>	<b>0.106</b>



**Figure 4.** Minimum length of subsections by region.

## Appendix C

**Program 1.** Source code to compute the probability of  $|X-\mu|<\text{Error}$ .

```
function [Pf]=PF(Error,N,Region)
%PF(Error,N,Region) returns the probability of  $|X-\mu|<\text{Error}$  for a sample
size N
%The Region Code is defined as
%0    Whole State
%1    Bay
%2    Grand
%3    Metro
%4    North Alpena
%5    North Cadillac
%6    North Grayling
%7    North Traverse
%8    Southwest
%9    Superior
%10   Univ Brighton
%11   Univ Jackson
%12   Univ Lansing
%Examples:
%If the sample size N is 100 and we need to calculate the probability
for the
%error to be less than 20 in Metro Region
% >> PF(10,100,3)
% ans =
%    0.9472
%which means that the probability for error to be less than 10 is
94.72%.

sig=(10:10:120);    % sig is the upper bound on sig for each interval

%psig(Region) is probability of selected region that sig is between 10*i
and
10*(i+1) for j=0..11
psig(1,:)= [26 323 477 480 554 395 203 65 27 8 3 1]/2562;%Whole
State
psig(2,:)= [21 106 80 73 70 57 19 9 1 0 1 0]/437;%Bay
psig(3,:)= [0 1 10 34 28 27 20 7 2 0 0 1]/130;%Grand
psig(4,:)= [0 25 73 66 91 56 39 9 2 1 0 0]/362;%Metro
psig(5,:)= [2 14 6 21 26 16 11 2 0 1 0 0]/99;%North
Alpena
psig(6,:)= [0 41 25 20 42 20 10 0 0 0 0 0]/158;%North
Cadillac
psig(7,:)= [2 15 44 37 34 13 8 3 0 0 0 0]/156;%North
Grayling
psig(8,:)= [0 15 48 27 29 33 7 0 0 0 0 0]/159;%North
Traverse
psig(9,:)= [0 1 15 29 43 39 25 10 11 3 2
0]/178;%Southwest
psig(10,:)= [1 81 75 83 107 85 43 22 11 2 0 0]/510;%Superior
psig(11,:)= [0 10 19 21 15 17 7 0 0 0 0 0]/89;%Univ
Brighton
psig(12,:)= [0 10 45 18 27 14 6 2 0 0 0 0]/122;%Univ
Jackson
psig(13,:)= [0 4 37 51 42 18 8 1 0 1 0 0]/162;%Univ
Lansing

z=Error*sqrt(N)/sqrt(2) ./sig;
Q=erf(z);
Pf=Q*psig(Region+1,:);
%P(|X-mu|<E|sigma=sigma_i and N=N)
%Sum the P's above times the psig's of the
%desired region.
```

**Program 2.** Source Code to obtain N for each region, for different levels of confidence.

```
%GetN(Error,P,Region) returns the minmum sample size N that can makes
the
%confidence interval of  $|X-\mu|<Error$  is greater than P.
%The funtion is implemented by using bineay search to get the numerical
%solution for N in equation  $pf(Error, N, Region) = P$ ;
%The Region Code is defined as
%0      Whole State
%1      Bay
%2      Grand
%3      Metro
%4      North Alpena
%5      North Cadillac
%6      North Grayling
%7      North Traverse
%8      Southwest
%9      Superior
%10     Univ Brighton
%11     Univ Jackson
%12     Univ Lansing

%Examples:
%If we need to calculate the minmum sample size N for Superior Region
%that can makes the confidence interval of  $|X-\mu|<10$  is greater than
99%.
% >> GetN(10,.99,9)
% ans =
%      230
%which means that N has to be greater than 230 to make sure the
probability of
 $|X-\mu|<10$  is greater than 99%.

function N=GetN(Error,P,Region)
N0=1;
N1=10000;
G1=pf(Error,N1,Region);
G0=pf(Error,N0,Region);
if ((P-G1)*(P-G0)<0)
    while (G1-G0>10^(-10))
        Nn=(N0+N1)/2;
        Gn=pf(Error,Nn,Region);
        if (Gn<P)
            N0=Nn;
            G0=Gn;
        else
            N1=Nn;
            G1=Gn;
        end % if (GN<P)
    end % while
    N=ceil(N1);
    pf(Error,N,Region)-P;
else
    output('Bad initial guess')
end
```

**Program 3.** Source code to calculate required N for each region.

```
% Calculate required N for each region.
%Ni(1)      Whole State
%Ni(2)      Bay
%Ni(3)      Grand
%Ni(4)      Metro
%Ni(5)      North Alpena
%Ni(6)      North Cadillac
%Ni(7)      North Grayling
%Ni(8)      North Traverse
%Ni(9)      Southwest
%Ni(10)     Superior
%Ni(11)     Univ Brighton
%Ni(12)     Univ Jackson
%Ni(13)     Univ Lansing
%where i=1,2,3

%Error = 5
for i=0:12
    N1(i+1)=GetN(5,0.99,i);
end

%Error = 10
for i=0:12
    N2(i+1)=GetN(10,0.99,i);
end

%Error = 20
for i=0:12
    N3(i+1)=GetN(20,0.99,i);
end
```