Ooduah Orafidiya
Anthony Ayon

Case Study 4

Introduction

Northern California's water supply primarily comes from the Sierra Nevada Mountains. The Forest Service of the USDA (United States Department of Agriculture) monitors the water supply by utilizing a gamma transmission snow gage located in the Central Sierra Nevada which helps determine a depth profile of snow density. Snow packs in the mountains can absorb rainfall up to a certain extent, hence less dense snow packs are able to absorb more water. Analysis of snow packs helps monitor water supplies and advise on flood management. Measurement don't directly measure snow's density; therefore gamma ray emissions are used to calculate it. Due to instrument wear and tear as well as radioactive source decay, density readings from measured calculations could become inaccurate. In order to adjust the conversion method, a calibration run is made each year at the beginning of the season. Here we will develop a procedure to calibrate the snow gauge.

Data

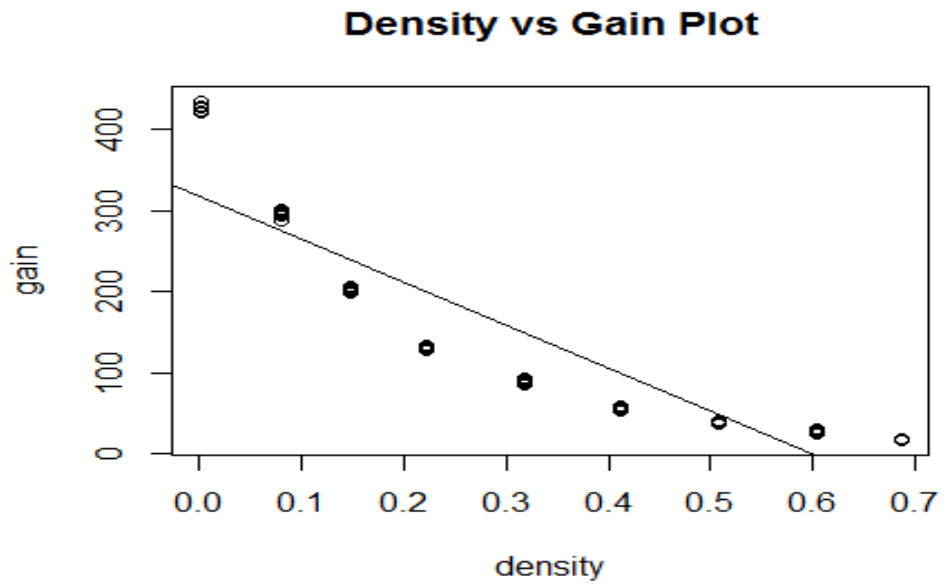
The data used in this study comes from a calibration run of the USDA Forest Service's snow gage located in the Central Sierra Nevada mountain range near Soda Springs. This calibration run consists of polyethylene blocks, whose purpose is to simulate snow. The block's density is measured by placing it between two snow gauge poles and then taking the readings of it. A total of 30 measurements are taken for each polyethylene block, except that in this dataset only the 10 middle measurements are reported. The reported measurement is an amplified version of the gamma photon count made by the detector which we call the gain. In the data, we will see 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene.

The goal of this case study is to develop a procedure to convert gain into density when the gauge is in operation. Experiment was done by changing density and then measuring the gain response, however when the gauge is in use, the snow- pack density is to be estimated from measured gain.

Scenario 1- Fitting

To start off and get a better visual of the calculated recording, we did a simple scatter plot of the Density with the Gain values (as seen in *figure 1*). It is important to point out that we have nine different density recordings that correspond to nine different polyethylene blocks, which explains the point clusters of our plot. Upon further analysis, we can see that the plot follows an exponential curved shape which may indicate that there is linearly present.

Figure 1: Density vs Gain Plot

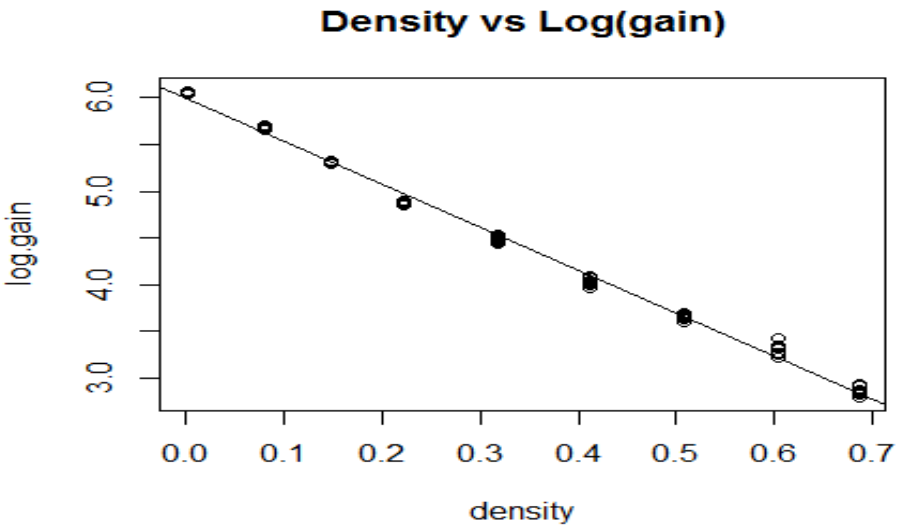


Furthermore, in order to satisfy linear regression we must make our plot linear, so this was achieved by taking the logarithm of all "gain" values. When this was executed and then replotted with the original density, we get a scatter plot with a clear negative linear association between points, illustrated in *figure 2*. Thus, we then proceeded by incorporating a Least Square Regression line so we could find the best fit for the scatterplot. Our results where:

***Slope*: -4.605937**
***y-intercept :* 5.997265**
***Formula*: Y = -4.605937xDensity + 5.997265**

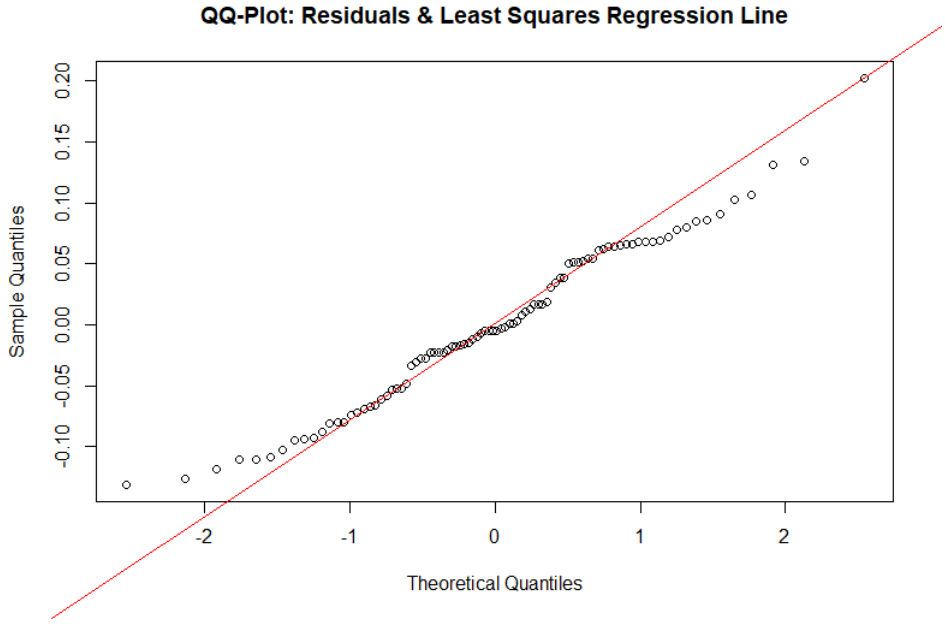
Judging from the results, we can hypothesize that there is linear association between Density and Log of Gain, but further test will be needed to confirm this statement.

Figure 2: Density vs Log of Gain Plot



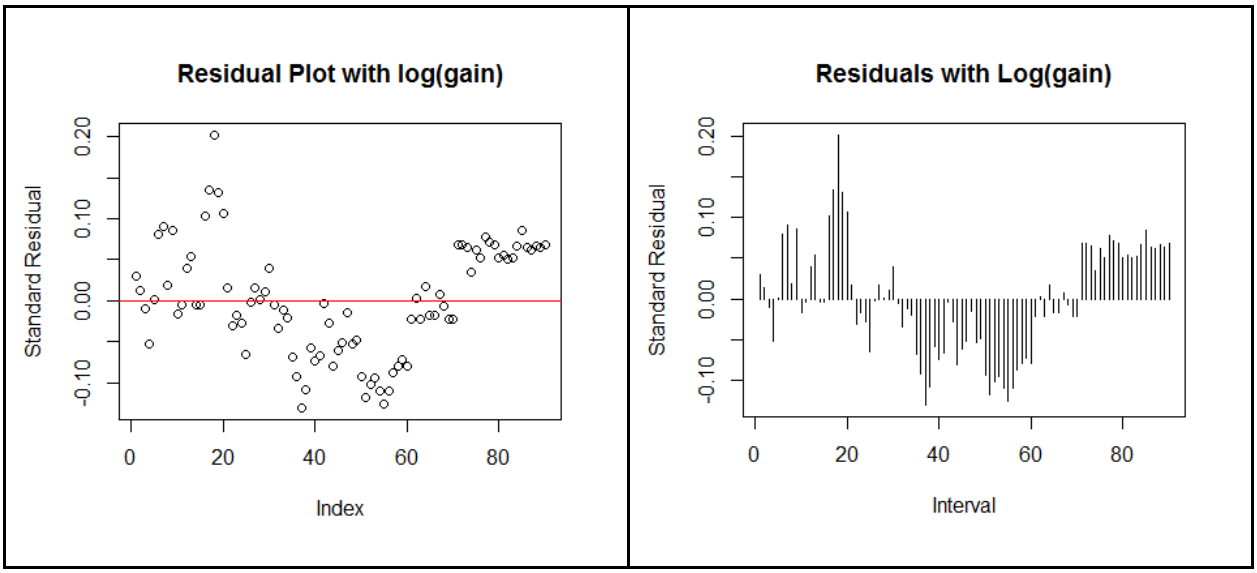
First, in order to check the normality of the residuals, we developed a QQ-plot that would give us a visual comparison between a normal and residual distribution. As we can see in *figure 3*, the residual points are close to following a normal distribution with some exceptions near the two tails of the plot. Nonetheless, we can say residuals do seem to roughly reflect a normal distribution.

Figure 3: Normal QQ-Plot with Residuals



Now in *figure 4* we will directly visualize the residuals and check for linearity. To do this we did two types of graphs that pretty much provide the same output, as both are essentially illustrating the behavior of residuals. For simplicity we will focus on the left one, the scatter residual plot. Even though the residuals seem to follow a slight oscillating path, the values are fairly close to zero indicating a good normality fit.

Figure 4: Same Residuals, Different Plot Types



One last aspect that we must address to establish linear association is checking for homoscedasticity. This check makes sure our fitted models are constant and not affected by any changes made to the x-values. Therefore, by referencing *figure 4*, we can conclude that there is in fact consistency in our model, thus homoscedasticity.

In all, we can say our model does follow the principles of linearity, where residual distributions are in fact normal. On the other hand, this can only be said about the data we were given, as any other discrepancies may cause drastic changes to these findings. One potential case is not reporting polyethylene densities accurately, which may lead to incongruities in the data and potential false rejections of linearity. Such impacts may directly influence the "gain" value, thus leading to inaccurate calculations and calibrations. Essentially, the more consistent a gauge measurement is the better.

Another potential scenario is the case where each block of polyethylene where to not be measured in a random order. For this instance, it is important we take into account that for most experiments done, random samples seem to provide more fruitful results since there are lower likelihoods of biases. This is a fundamental step in our procedure, as we try to encompass as much data as possible, so the machine is be able to adapt to different environments. If blocks where to be measured in a specific order, the machine would be exposed to outliers, thus provide imprecise results for the full range of 0.1 - 0.6 g/cm^3.

Scenario 2− Predicting

To get a clear visual of the boundaries of each point in *figure 2*, both the 95% confidence interval (blue) and prediction interval(orange) where plotted, shown in *figure 5*. These interval lines will help us predict or approximate the "density" value when selecting a random "gain" value.

As we can see, the confidence intervals are close to the least squares regression line which indicates that the average value of those points lies somewhere in between the lower and upper confidence intervals. On the other hand, we have the prediction intervals, which are wider. The prediction interval must account for both the points' average and the outliers of the plot, hence it makes sense that is wider than the confidence intervals. It is worth to point out that, as shown in *figure 5*, prediction intervals do not entirely encompass all the points inside its upper and lower bounds. This is the case of "density" value 0.6 where there is a minimal distance between the upper bound and one of the points.

Once having these intervals in place, we can start making approximations for the density for any given "gain" random value. As we can see from *table1*, when calculating the density for "gain" value 38.6, all least-squares confidence intervals and prediction intervals provide an accurate estimation for the density. Specifically, both the confidence and prediction intervals which fall within the acceptable range of the desired density value. In the case of "gain" value 426.7 we reach a special circumstance where all least-squares confidence intervals and prediction intervals contain negative values, showed in *Table2*. This is inaccurate considering that density cannot be negative even though it follows the same calculating procedure as all other density calculations. As illustrated in *figure 5*, density value 426.7 when taken the natural log, becomes 6.056 g/cm3 which barely falls within the scope of our graph. This is an exception in our plot since our decision to take the natural log eliminated the limit bound in the y-axis where the density approximated to zero. Meaning that the density can no longer infinitely approach zero and because of the linear structure, negative values can appear.

Figure 5:

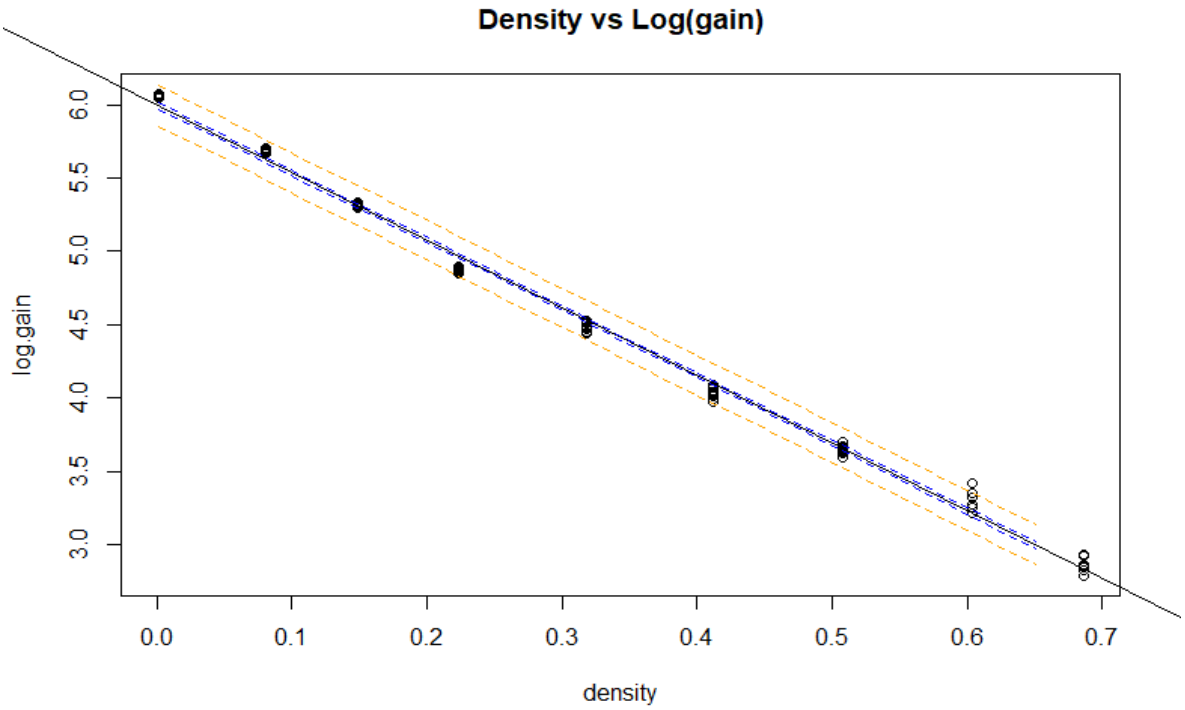


Table 1:

Density Prediction: gain value: 38.6

Original Density:	0.508 g/cm^3
Least-Square Regression:	0.5089113 g/cm^3
95% Confidence Intervals for Density:	(0.5049744, 0.5128482)g/cm^3
95% Prediction Intervals for Density:	(0.4793438, 0.5384787)g/cm^3

Table2:

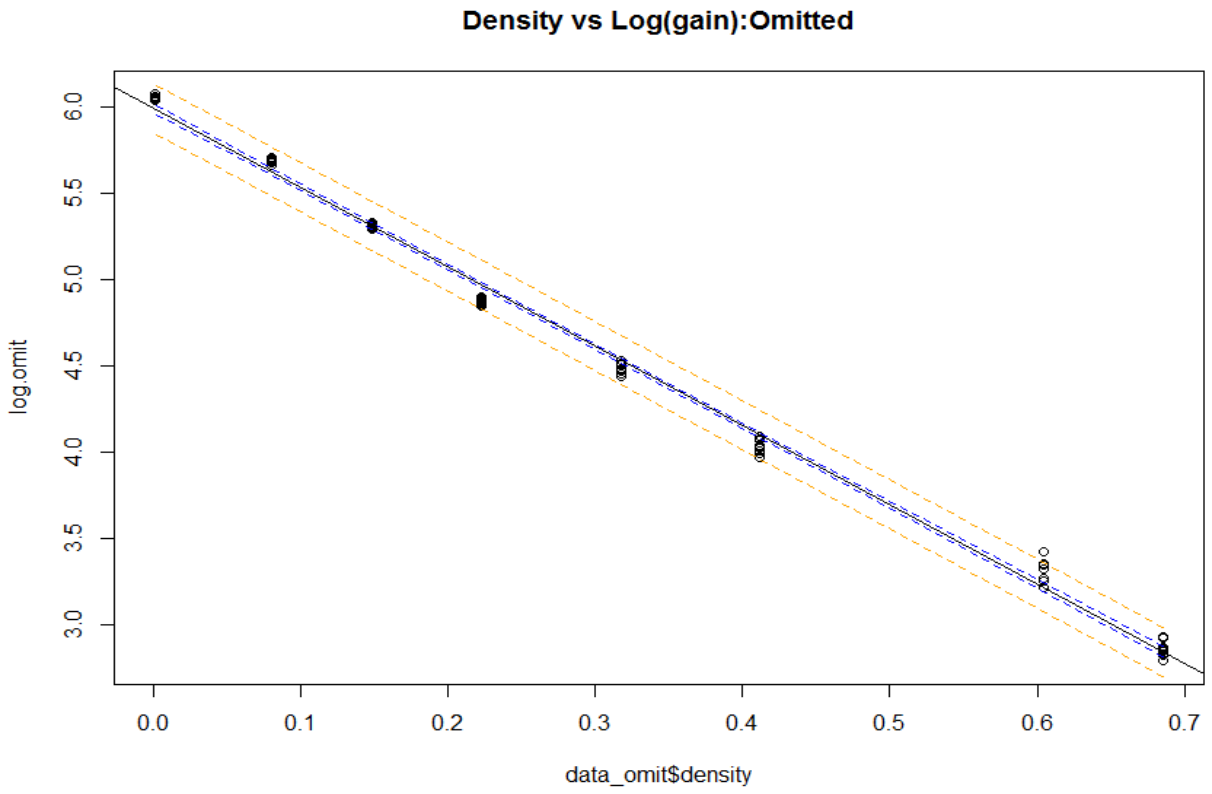
Density Prediction: gain value: 426.7

Original Density:	0.001g/cm^3
Least-Square Regression:	-0.01276954g/cm^3
95% Confidence Intervals for Density:	(-0.01841111, -0.007127971)g/cm^3
95% Prediction Intervals for Density:	(-0.04261185, 0.01707277)g/cm^3

Scenario 3 - Cross Validation

In this part, we focused on performing the same procedure done in scenario 2, but all the density values of 0.508 would be removed leaving the dataset with 80 samples rather than 90. Hence, we will use the remaining 80 as our training data and the 10 omitted as our test data. *Figure 6* shows the scatter plot and linear regression line of our training data with their corresponding confidence and predicting intervals. We can clearly see the points are gone but, there is not much change to the regression line, confidence intervals and prediction interval.

Figure 6:



Furthermore, we performed a density approximation using our training data. First, we approximated the density for the "gain" value of 38.6, results can be seen in *table 3*. We got fairly similar results to scenario one resulting in an actual 0.509 for the least-square regression, hence, making our density estimation more accurate. Regarding the 95% confidence and prediction intervals for the training data, there is a difference present in the ten-thousandths decimal spot, which doesn't really imply much change to our interval estimates.

In the case of the estimation of the 0.001 density for the training data we took a different approach from what was done in scenario two. For this case, we calculated the corresponding

gain of that specific density which ends up resulting in 400.43. This when used in the same application to calculate the least-square regression and intervals led to different results (*Table 4)* when compared to *Table 2.* In this case we have a confident interval spanning from a negative value to positive, which still encompasses the actual density calculation. Same can be said about the prediction interval where the intervals are actually wider than the ones found in scenario two.

Table 3

Density Prediction: gain value: 38.6

Original Density:	0.509 g/cm^3
Least-Square Regression:	0.5091927 g/cm^3
95% Confidence Intervals for Density:	(0.5046106, 0.5137747)g/cm^3
95% Prediction Intervals for Density:	(0.4779843,0.540401)g/cm^3

Table4:

Density Prediction: gain value: 400.43

Original Density:	0.001g/cm^3
Least-Square Regression:	-0.001g/cm^3
95% Confidence Intervals for Density:	(-0.004780354, 0.006780355)g/cm^3
95% Prediction Intervals for Density:	(-0.03040664, 0.03240664)g/cm^3

Theory

Linear Regression

In this lab we utilize linear regression in order to reveal the relationship between two variables. We seek to understand the strength of linear association between said variables. In doing this we wish to find the correlation between gain and density. We can also use a linear model to predict value by inserting x values into the equation itself. This does however come with a degree of uncertainty. This is called extrapolation as we are applying values from outside the data itself. The specific linear regression that we will be using is the Least Squares Regression which makes the vertical distance from the data points to the regression line as small as possible. The unique

piece of the Least Squares Regression is that it minimizes the variance, which is the sum of squares of the error which in other words means fitting a line to the points as closely as possible.

Residual Plots

Based on our Linear Regression plot we can also output a residual plot which shows the distance of every sample value from our linear regression line. A randomly dispersed residual plot indicates that it is a good fit for a linear model while on the other hand a non-randomly dispersed plot indicates a better fit for a non-linear model.

QQ-Plot

The purpose of our QQ plot is to check the similarity of two distributions by plotting their quantiles against each other. The similarities of the two distributions towards each other can be measured by how closely the plot follows the y=x line. A flatter, smoother like indicated more similarity between the two distributions.

Cross-Validation

We will be utilizing K-fold cross validation. The purpose of using cross validation is to estimate the ability of a machine learning model data points outside the current data set. In order to do this we will use a limited sample in order to estimate how the model is expected to perform on data not included in the sample. The way this process works is that we will shuffle the data set, divide all the values into K groups, split each group into training/testing data, fit a model to the training data and evaluate with the test set and finally collect the evaluation scores and summarize the accuracy of the model.

Conclusion

Due to gauges being subjected to wear and tear over time, there must be a yearly calibration run to continue functioning normally. In order to properly do so, various statistical methods need to be used to ensure accuracy. Throughout this paper we focused on using some of those data recordings in order to ensure each machine is well calibrated and ready to use in an actual scenario.

First, we focused on taking both data columns and plotting them in a scatterplot where we altered some of the gain recordings in order perform linear testing. We then used that plot to calculate least-squares regression line giving us a better idea of what was happening with the data while also performing some additional analyses in order to ensure the regression line was linear. This is an important element as fitting allows any user to visualize the behavior of the data and based on that, make various assumptions that can later be tested.

From there we moved on to making predictions with our model, where we focused on calculating the densities of random gain values. We centered around two specific densities for

these calculations, 38.6 and 426.7. With the help of confidence and prediction intervals we were able to see where those densities lied and what where their approximate values, thus providing a window of possible acceptable calculations.

One way of testing our finding is performing cross-validation where the data is split between training and testing data. In this case we omitted all densities with 0.508 and used it as our testing data while the remaining 80 density values where used as training data. We performed several calculated cross-validations and the data seemed to align with our original results, hence demonstrating the model's success.

In all we can say that the model used is appropriate for the dataset and a reliable source to calibrate the gauges. This case study, even though limited to 90 data points, can serve as one of many samples that could eventually help develop a model for the population data. Even though more calibrations will be needed to calculate more densities, we can conclude that our intervals provide high certainty that a block's density will lie within the given range.