Sampling Strategies

Stephen Kyle Wilshusen

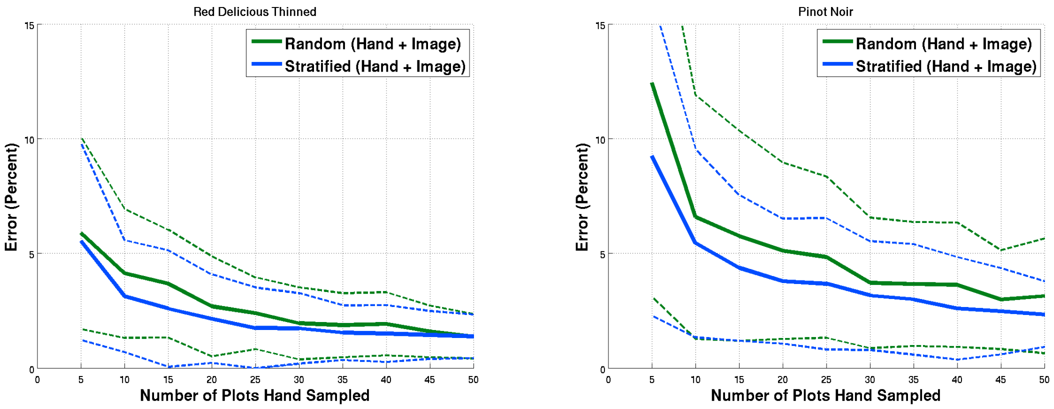
August 2014

1 Introduction/Abstract

This paper is a study of sampling methods to estimate yield in orchards in vineyards. There are five datasets that were analyzed: Two apple datasets and three grape datasets. The main takeaways are different for both fruit types. For apple datasets, it is seen that mapping image estimates to yield estimates can be done with a "scaling factor," or multiplying by a scalar value. For grape datasets, it is seen that a linear function needs to be estimated and used to map image estimates to yield estimates. In addition, the best performance of our method has the same performance as the classical technique of extrapolation of groundtruth measurements. Nonetheless, a groundtruth extrapolation does not yield dense fruit estimates, as our method does.

2 Stratified random sampling

Stratified sampling focuses on breaking a sampling space up into different groups (or strata). For our purposes, we created strata that divided up the yield distribution. Each strata encompasses one section of the yield distribution (from high, to medium, to low levels of counts). Samples are selected from each level, to ensure that a representative sample is selected.



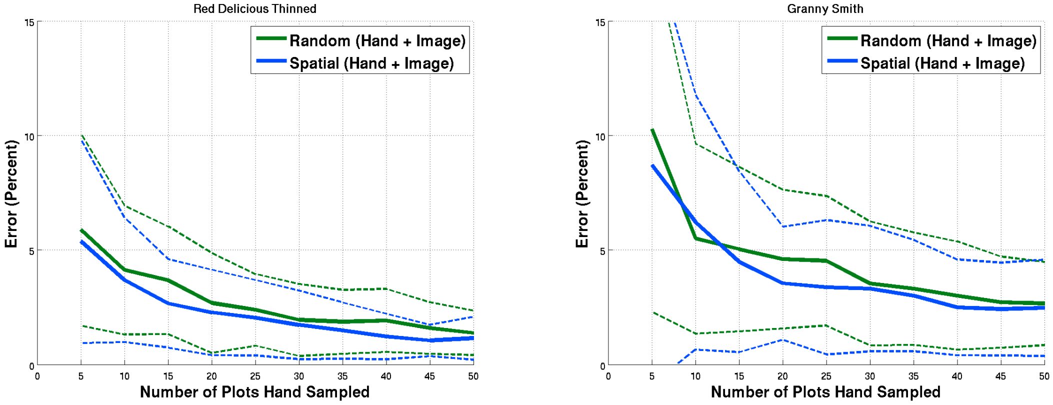
(a) Red Delicious

(b) Pinot Noir

Figure 1: Stratified Sampling is modestly effective for all types of datasets

3 Spatial sampling

The current spatial sampling method simply focuses on iteratively moving points to locations that are far away from all other points.



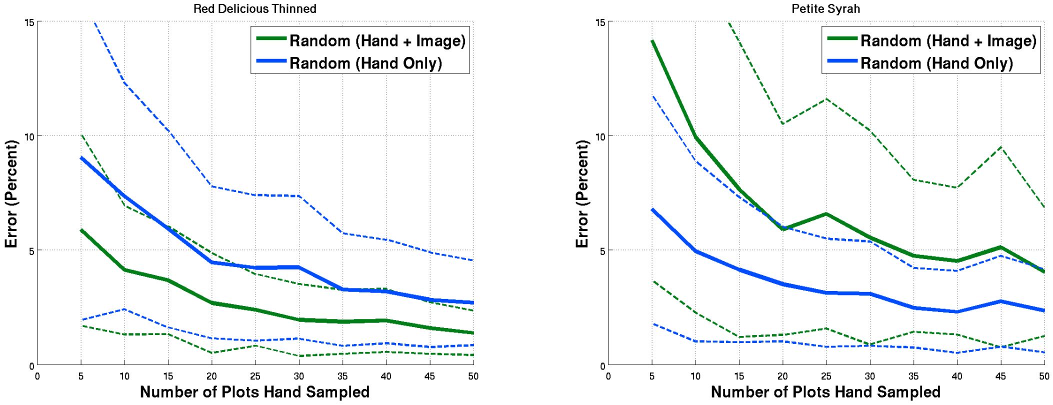
(a) Red Delicious

(b) Granny Smith

Figure 2: Spatial sampling yields modest improvements for apple datasets

3.1 Direct Extrapolation: No Sensor Data

All of our sampling methods are based on a process that starts with initially sampling groundtruth data. In this section, the results show that simply extrapolating this groundtruth subsample measurement can gain accuracy that is greater than scaling the sensor data. For the apple datasets, it is seen that direct extrapolation is not effective. For the grape datasets, direct extrapolation of the groundtruth sample is very effective. This phenomenon of direct extrapolation being more effective than any sampling method is explained by an analysis of using linear offset functions to map image estimates to yield estimates.



(a) Red Delicious Groundtruth Extrapolation

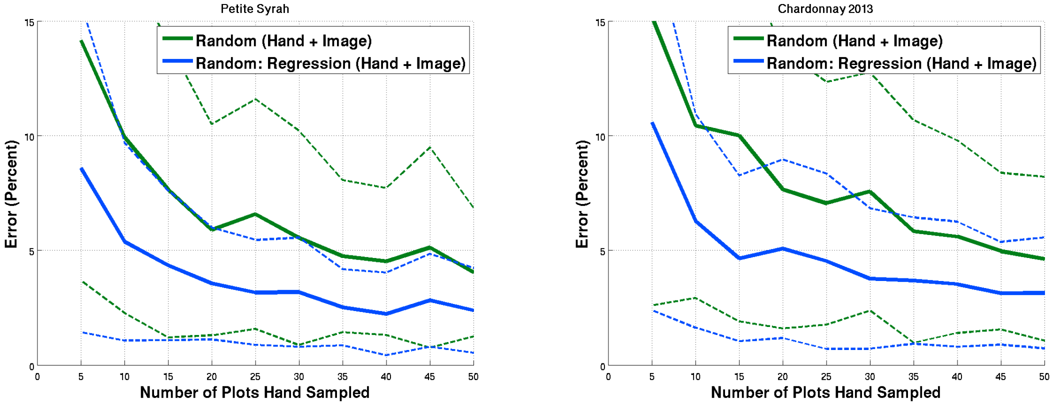
(b) Petite Syrah Groundtruth Extrapolation

Figure 3: Direct extrapolation is ineffective for apple datasets and effective for grape datasets

4 Sampling: Scaling Function from Linear Regression

Instead of scaling with a scalar value, a linear scaling function can be obtained by doing linear regression on the subsample data collected. With the resulting function, after regression, sensor data can be mapped to predicted yield more accurately.

A linear regression is seen to be effective for grape datasets.

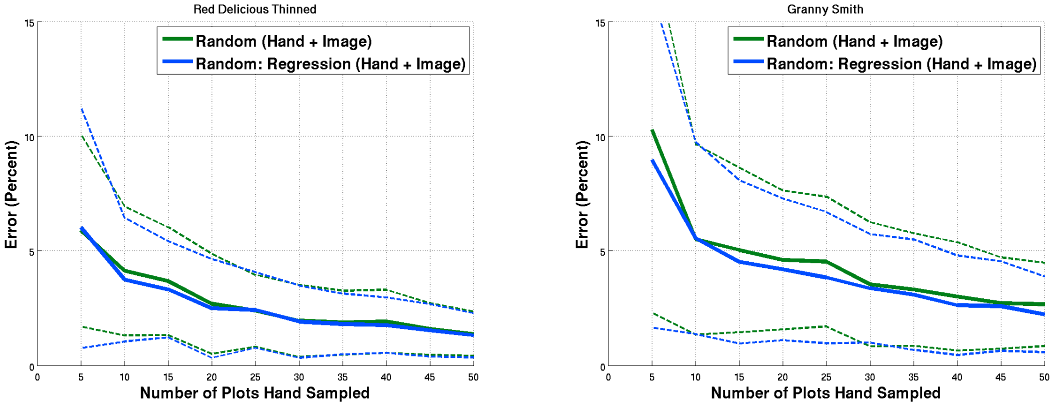


(a) Petite Syrah

(b) Chardonnay

Figure 4: A linear regression operation to find an offset linear function is the most important factor to transferring grape image measurements to yield estimates

A linear regression is seen to be less effective for apple datasets.



(a) Red Delicious

(b) Granny Smith

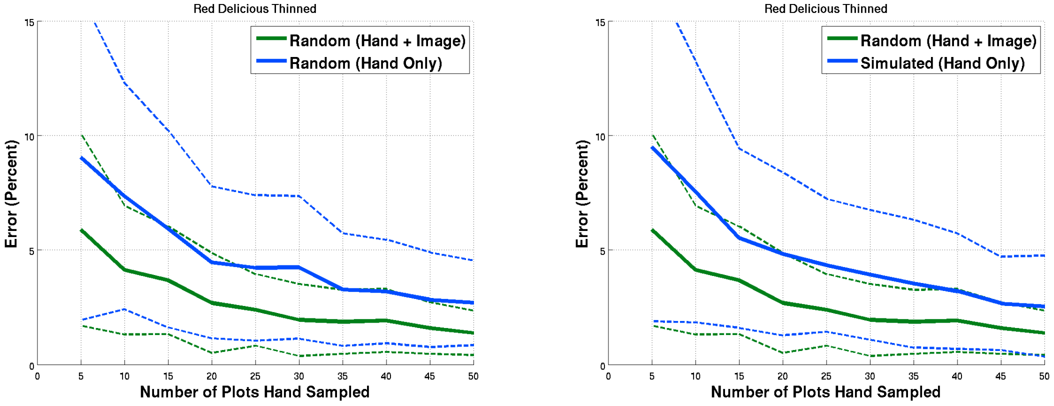
Figure 5: Linear regression to obtain offset linear function: Apple datasets

This suggests that there are some systematic differences between apple detection and grape detection. This is logical as the grape code includes features that the apple code does not, such as a clustering step in the grape algorithm.

5 Section: Simulating algorithm performance

The previous sections have displayed the results that we have seen by using an emphircal approach. In this emphirical approach, we assume that we have already collected all of the groundtruth data from all farm plots, as well as image data from all farm plots. We have noticed through this emphirical study that the groundtruth data is organized by a normal distribution. As well, our algorithm has consistent error rates, which can also be modelled by a normal distribution. This means that we can first simulate groundtruth data and then simulate our algorithm's expected performance by adding an error distribution to the simulated groundtruth data. Through this process of simulation, we can model our algorithm's expected performance to a high degree of accuracy.

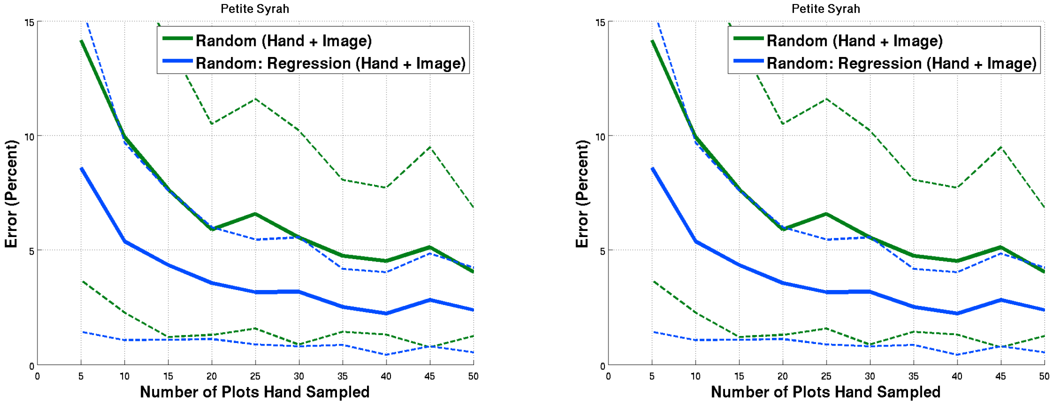
We can model the performance of only using groundtruth data:



(a) Red Delicious: Empirical: Groundtruth Extrapolation (b) Red Delicious: Simulated: Groundtruth Extrapolation

Figure 6: Simulation of groundtruth extrapolation

We can also model the performance of our algorithm over grape datasets. In this model a linear function is used to transfer from hand to algorithm counts.



(a) Petite Syrah: Empirical (b) Petite Syrah: Simulated

Figure 7: Simulated data yields almost identical results to empirical data.

We can also model the performance of our algorithm over apple datasets. In this model a scaling factor is used to transfer from hand to algorithm counts.

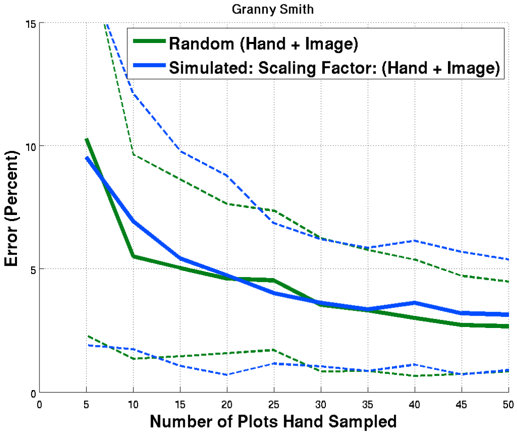
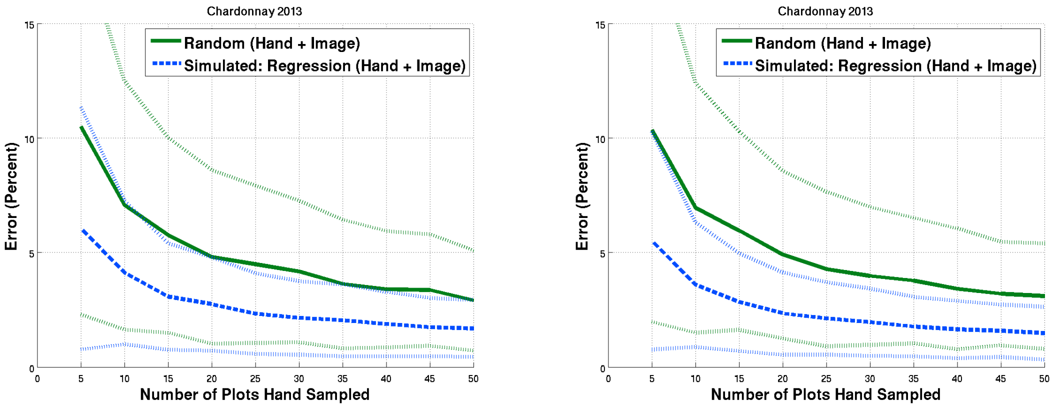


Figure 8: Simulated performance almost overlaps with empirical performance. Note: Empirical is labelled as ’random’

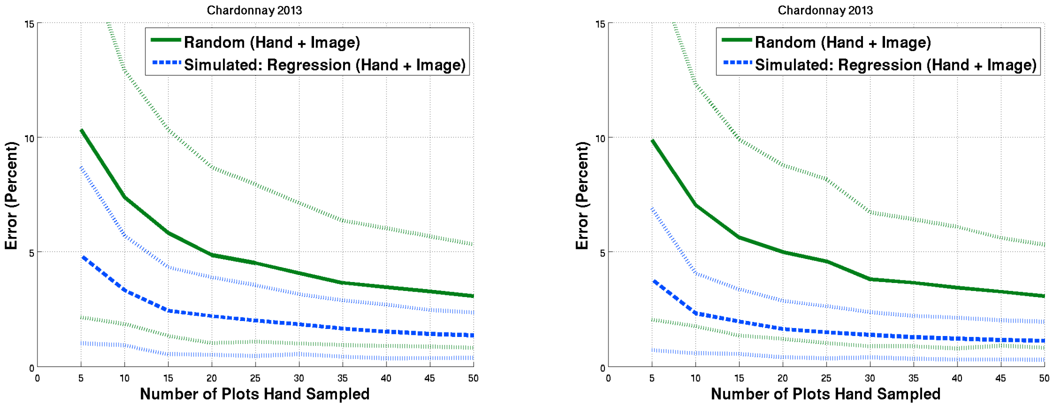
Decreasing R-Squared values does not seem to have much effect on changing the error plots that are generated. This is explicable. If the errors can be represented by a normal distribution, then the errors act to balance each other out in our simulations. This means that unless R-squared values are very high, a decrease in performance is not estimated. This occurs as even when section by section error is at 20 percent, the overall error is not high because these section errors balance each other out.



(a) Petite Syrah: R^2 : 0.50

(b) Petite Syrah: R^2 : 0.93

Figure 9: "Low" R-squared values yield graphs that are indistinguishable from one another

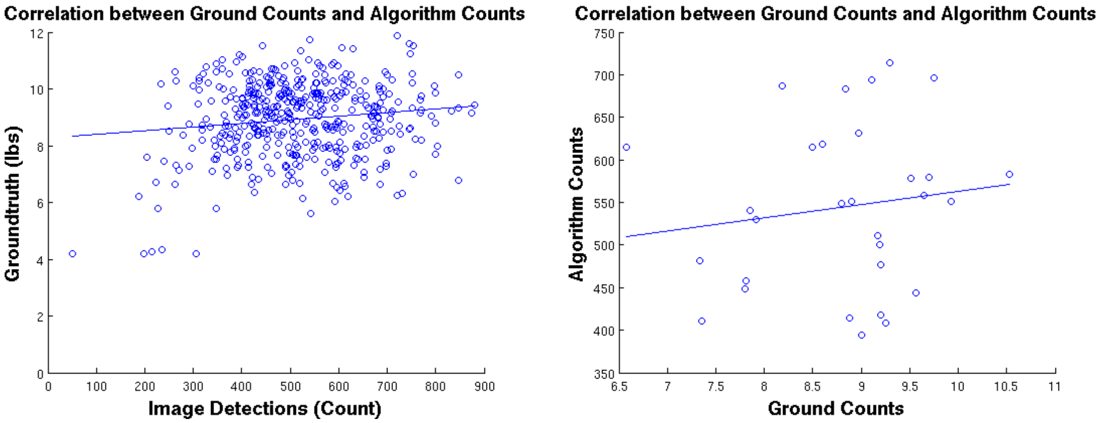


(a) Petite Syrah: R^2 : 0.97

(b) Petite Syrah: R^2 : 0.99

Figure 10: Only higher R-squared values yield graphs with distinguishable differences from one another

Analyzing the correlation between ground counts and algorithm counts shows that there is a low correlation between ground counts and algorithm counts. A lack of correlation between the harvest monitor data and our estimates could be due to errors in the harvest monitor data. To analyze this, we attempted to smooth the sensor data over individual rows. In this approach, we tried to not smooth sensor data accross adjoint rows. With this approach, a low correlation score between ground counts and algorithm counts is still apparent.



(a) Correlation: 2 Meter Step

(b) Correlation: 50 Meter Step only over rows

Figure 11: Correlation between Ground and Algorithm Counts

6 Conclusion

This document summarizes results found between July and August 2014, with regards to sampling methods in orchards and vineyards. This includes an analysis of sampling methods that revolve around a scaling factor: random sampling, stratified random sampling, and spatial sampling. In addition, this also includes an analysis of pure groundtruth extrapolation and the usage of a linear function instead of a scaling factor.

The main takeaway conclusions are different for apple and grape datasets.

For grape datasets, stratified and spatial sampling approaches only improve performance marginally. Improvements, instead, came from either using a linear function for scaling sensor data or directly extrapolating sensor data. In fact, directly extrapolating the groundtruth subsample yields similar accuracy as using a linear scaling function and any sampling method (stratified, spatial, random) for all 3 datasets. The message from these results is that 1.) a linear scaling function is necessary for mapping image estimates to yield estimates and 2.) the sensor data is not adding much signal currently, if an extrapolation of the groundtruth sample is as effective as using sensor measurements and the groundtruth sample.

For apple datasets, stratified and spatial sampling lead to small increases in performance that are marginal. A linear scaling function or direct extrapolation of groundtruth measurements did not lead to performance increases.