An importance sampling procedure for estimating crop yield

J. Benjamin Cook
jbenjamincook@gmail.com

1 Introduction

Unprecedented amounts of data are available on modern farms. From yield monitors to weather sensors to infrared imaging, farmers are able to keep track of every detail on their farms. However, most farmers are not taking advantage of this data. Much of the data is never reviewed after being collected. The data that is reviewed remains inaccessible, trapped in complicated legacy software, agronomist reports, and countless pages of spreadsheets.

Agricultural production lags far behind other industries in using data to make decisions, but data-based decision-making will be essential in the future. Farming is a challenging business. Reduced availability of water and increasing input costs are squeezing farmers' bottom lines during a time of extreme volatility in the commodities markets. New regulations on irrigation, pesticide use, and fertilizer application mean farmers must find a new way to boost yield. Soon, using data to optimize decision-making will no longer be a novel luxury – it will be essential for farmers to stay in business.

The purpose of this project is to develop and assess a stochastic procedure for estimating crop yield in a maize field with an eye toward making in-season forecasts. Giving farmers insight about how much yield to expect in their fields will arm them to effectively compete against professional traders in the futures market.

2 Background

Consider a corn field Ω and a yield function $f(x)$ that returns bushels per acre at any location in the field x. Total yield, in bushels can be evaluated by multiplying the area of the field, $|\Omega|$, by the integral, $I = \int_\Omega f(x)dx$. Assuming we can evaluate the function f at any location in the field, we can estimate the integral with $\hat{I}_{mc} = \frac{1}{N}\sum_i f(x_i)$, where the x_i are drawn uniformly from the area Ω. However, since yield is not distributed evenly throughout the field, this "vanilla" Monte Carlo approach results in unnecessarily high variance. Instead, it is possible to draw samples from a (possibly unnormalized) proposal distribution that is somehow "close" to f and then correct for the fact that the samples are no longer uniform. With importance sampling, the integral is estimated as:

$$\hat{I}_{is} = \int f(x)dx = \int g(x)\frac{f(x)}{g(x)}dx \approx \frac{1}{N}\sum_{x_i \sim g(.)} \frac{f(x_i)}{g(x_i)}$$

This approach assumes that $f(x)$ and $g(x)$ are normalized probability densities. Alternatively, we can use unnormalized functions if we correct for them as follows:

$$\hat{I}_{is} = \sum_i w_i f(x_i)$$

where $w_i = \frac{\widetilde{w}_i}{\sum_i \widetilde{w}_i}$ and $\widetilde{w}_i = \frac{f(x_i)}{g(x_i)}$ [2].

In order to build a procedure that is robust and effective, this paper addresses the following questions:

1. After the first pass, how should one derive an importance sampling function?

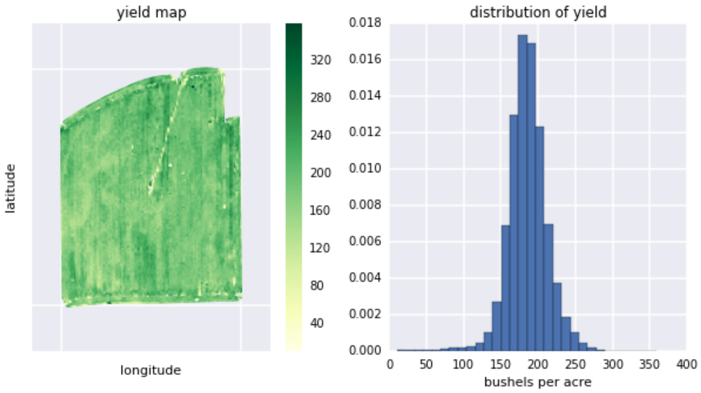

Figure 1: A yield map with the raw data and the distribution of yield at all measured locations

2. What is the best method for drawing samples from the importance sampling function?
3. On the second pass, how many points do we need to sample to achieve an acceptably low level of variance?

3 Data

The data come from a corn field in Butler County, Nebraska that was planted and harvested in 2009. Measurements are logged by a grain yield monitor which is connected to sensors on the arms of the tractor as it harvests the corn. The yield monitor records an estimate of yield in bushels per acre and geolocation approximately six times per second. The full dataset contains $16,898$ points. Figure 1 displays the raw data as a yield map in the left panel and a histogram of yield in the right panel.

Although the data coverage for this field is extensive, the sampling methods require that we can evaluate a yield function $f(.)$ at any continuous point $x_i \in \Omega$. Because of this requirement, $f(x_i)$ is defined as the average of the five nearest neighbors to point i. This is essnetially a pre-processing step and does not have a major impact on final results.

Because total number of bushels in the field is the quantity of interest, but data are in units of bushels per acre, we need an estimate of the total number of acres in the field. One convenient way to estimate the area is to do rejection sampling: draw points from a rectangle of known area and multiply the area of the rectangle by the fraction of points drawn that fall within the field boundary. Using this approach, the number of acres is found to be approximately 63.7. Figure 2 shows Monte Carlo samples where locations inside the field, Ω, are blue.

4 Method

The importance sampling procedure for estimating yield consists of four steps:

1. Collect a small number of samples from $f(.)$
2. Construct the proposal function $g(.)$
3. Sample from $g(.)$
4. Use the samples to estimate yield

This section enumerates each of the four steps.

The first step is to collect N_1 yield samples from the field. This is constrained to be a small number because it corresponds to someone actually walking the field and collecting stalks at N_1 randomly selected locations throughout the field.

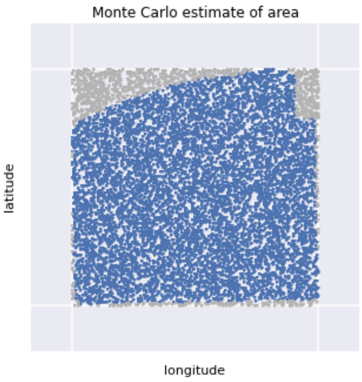

Figure 2: Monte Carlo estimate of field area

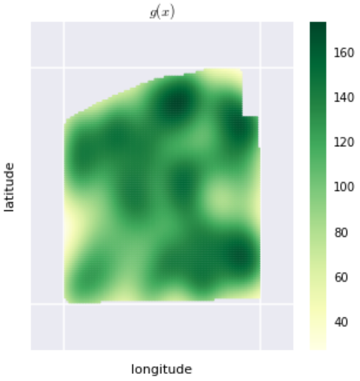

Figure 3: A Gaussian Process fit to the first N_1 points

Currently there is a strong simplifying assumption here: we can evaluate the actual yield function at these locations. That is, we can use the N_1 corn stalks to evaluate yield in bushels per acre at the sampled locations. In reality, inferring yield in bushels per acre based on volume of the grain from these N_1 stalks or based on height and circumference of the stalks will require a separate step.

The next step is to define the proposal function $g(.)$. Here, it is important to use a function that is somehow close to the true yield function so samples are concentrated in high yield areas. One simple way to build a function that is close to the true yield function is to fit a Gaussian Process (GP) to the N_1 samples. A GP is a distribution over functions and, assuming the mean is set to zero, is fully specified by a covariance kernel, $K(x_i, x_j)$ [4]. One common form for the covariance function is a squared exponential:

$$\sigma^2 \exp\left(-\|x_i - x_j\|^2/\phi\right)$$

Assuming we know the hyper-parameters, σ^2 and ϕ, we now have a function that we can use for importance sampling:

$$g(x_{new}) = K(x_{new}, X)K^{-1}(X, X)y$$

which corresponds to the mean of the GP. Here, X is the vecotor of locations for the N_1 samples and x_{new} is the location of any arbitrary new point. In theory, the hyper-parameters σ^2 and ϕ should not matter since we do not actually need a perfect model for the N_1 samples from the first pass. In practice, the GP is only "close" to the yield function for a relatively small range of the hyper-parameter values. Figure 3 shows $g(x)$ for $x \in \Omega$.

The third step is to draw samples from $g(x)$ to find good candidate points for evaluating crop yield. Two Markov Chain Monte Carlo algorithms were used to sample from $g(x)$: Slice Sampling [3]

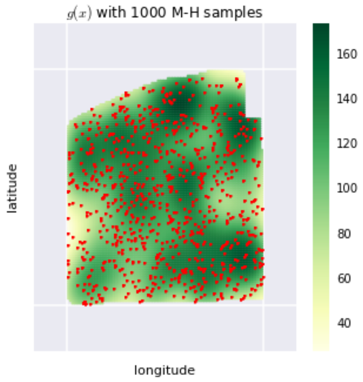

Figure 4: Samples drawn from the importance sampling function

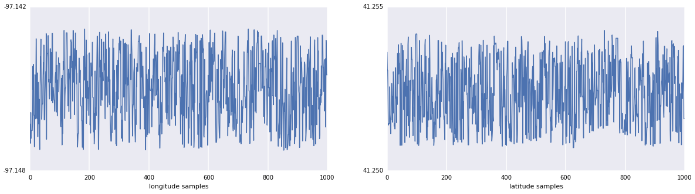

Figure 5: Trace plots of longitude and latitude samples from the M-H algorithm

and Metropolis-Hastings [1]. The M-H algorithm is preferred in this case because it does a good job of sampling from $g(x)$ and is faster than Slice Sampling. The proposal for the M-H samples is $q(x^*|x) \sim \mathcal{N}(x, \gamma)$, where γ is set to 0.002. 100 burn-in samples are discarded and thinning parameter is set to three, so that every third sample is saved. Points drawn outside of Ω are defined to be $-\infty$ and are therefore never accepted. The M-H algorithm performs well because even though $g(x)$ is multi-modal, the modes are relatively flat. Figure 4 shows 1,000 M-H samples from $g(x)$ and Figure 5 shows trace plots for longitude and latitude.

5 Results

The importance sampling procedure is an effective way to estimate the number of bushels the field yields. Unsurprisingly, as the number of M-H draws increases, the the variance of the estimate decreases. Table 1 show the average estimate and the standard deviation of those estimates for $N_mc = 100, 500$, and $1,000$. Each experiment was performed 100 times. The left panel of Figure 6 shows histograms of the three levels of M-H draws and the right panel shows a box plot of the estimates and variance.

Table 1: Results

N_{mc}	Bushels	sd($\hat{I}_{is}$)
100	11,987	160
500	11,979	79
1,000	11,988	58

In practice, how many M-H samples need to be drawn would have to be determined by the farmers and agronomists who use this procedure. Samples are relatively expensive in the sense that each one needs to be collected by hand. Fortunately, even with only $N_{mc} = 100$ samples, the standard

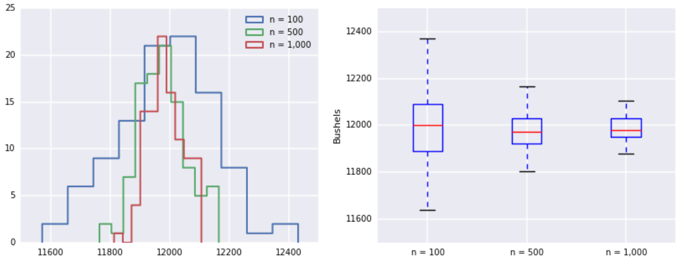

Figure 6: Distribution of estimates of $|\Omega|\hat{I}_{is}$ and variance of estimate as the number of Metropolis samples increases

deviation of 160 is less than 2 % of the number of bushels. This low variance makes the importance sampling method practical for estimating yield in real situations.

6 Conclusion

Stochastic optimization provides an important set of tools for estimating the crop yield in irregularly shaped fields. Fitting a GP to a few samples and then drawing several more importance samples from this proposal function decreases variance of estimated crop yield. This procedure can be used in a prediction setting with one additional step: by modeling how physical characteristics of the plant throughout the growing season drive yield at harvest.

References

[1] Chib, S., and Greenberg, E. Understanding the metropolis-hastings algorithm. *The American Statistician 49*, 4 (1995), 327–335.

[2] Murphy, K. P. *Machine Learning: a Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012.

[3] Neal, R. M. Slice sampling. *Annals of statistics* (2003), 705–741.

[4] Rasmussen, C. E. *Gaussian processes for machine learning*. Citeseer, 2006.