STA 232B Final Project

Qian Li, Yu Pei, Jean Du,
Yifei Wang, Yu Pei, Heather Darnell

March 20, 2014

1 Paper

The goal of the source paper is to analyse the effectiveness of LANDSAT satellite data. The data consists of 36 Iowa farmland segments for which pixels in satellite images were counted to estimate the area of corn and soybeans in the segment. The satellite data is to be compared to actual reports of hectarage for the two crops in the same segments, obtained through less experimental means.

The 36 segments in the data came from 12 different counties, with most counties reporting on more than one segment. Thus, a county-based effect was considered to account for this clustering.

The problem was addressed through what the author calls an Components-Of-Variance Model. The model foremostly fits a linear relationship between the predictive satellite data and the responsive reported hectarage. The error is modeled to come from two sources, the county-formed clustering and the background variation.

The crux of the model is that while background variation has an independent effect on all observations, cluster-based variation only affects observations from different counties independently.

The Components-Of-Variance Model is essentially what we knew in modern days as the Linear Mixed Model, without generalizing to multiple sources of clustering. The estimation methods were also similar, with details provided in the Proofs Section.

The numerical results of this model stated that both corn and soybean satellite data had a significant association with reported corn hectarage. However, only soybean satellite data had a significant association with soybean hectarage.

The source of variation for corn hectarage was roughly evenly distributed between cluster and background, while for soybeans the cluster-based variation took a 60% majority.

For prediction, predominantly two methods were compared - the less error-prone BLUP, and the computationally simpler Survery Regression Predictor. It was observed that while the BLUP produced much lower standard errors than the survery regression predictor for lower sample size per cluster, the difference became negligible when sample sizes per cluster increased.

2 Proofs

2.1 II.I Form and Distribution of the Best Predictor for Realisation of the Random Effect

The following proof relates to the claims surrounding Equations 3.1, 3.2.

We start off understanding that The best predictor of random effect realisation v_i is its conditional distribution given $\bar{u}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} u_{ij}$. Using definition $u_{ij} = v_i + e_{ij}$, we produce $\bar{u}_{i\cdot} = v_i + n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Additionally, we know that each $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$ idependentally and identically.

With any arbitrary i, since v_i is normal and $\bar{u}_{i\cdot}$ is a sum of v_i and a set of other independent normal random variables - the e_{ij} - we may conclude that v_i and $\bar{u}_{i\cdot}$ are jointly normal. As for their parameters, we have.

$$v_i \sim N(0, \sigma_v^2) \Rightarrow E(v_i) = 0, Var(v_i) = \sigma_v^2.$$

$$\bar{u}_{i\cdot} = v_i + n_i^{-1}\sum_{j=1}^{n_i} e_{ij} \Rightarrow E(u_{i\cdot}) = 0 + n_i^{-1}[n_i \cdot 0] = 0,$$

$$Var(\bar{u}_{i\cdot}) = Var(v_i) + (n_i^{-1})^2[n_i \cdot Var(e_{ij})] = \sigma_v^2 + n_i^{-1}\sigma_e^2$$

$$Cov(v_i, \bar{u}_{i\cdot}) = Cov(v_i, v_i) + n_i^{-1}Cov(v_i, \textstyle\sum_{j=1}^{n_i} e_{ij}) = Var(v_i) + 0 = \sigma_v^2$$

Putting things together, we get

$$\begin{pmatrix} v_i \\ \bar{u}_{i\cdot} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + n_i^{-1}\sigma_e^2 \end{pmatrix} \right)$$

Using properties of joint normality, we have, for $g_i = (\sigma_v^2 + n_i^{-1}\sigma_e^2)^{-1}\sigma_v^2 = Var(\bar{u}_{i\cdot})^{-1}\sigma_v^2$

$$E(v_i|\bar{u}_{i\cdot}) = 0 + \sigma_v^2(\sigma_v^2 + n_i^{-1}\sigma_e^2)^{-1}(\bar{u}_{i\cdot} - 0) = \bar{u}_{i\cdot}g_i$$

$$\begin{aligned} E((v_i - \bar{u}_{i\cdot}g_i)^2) &= Var(v_i - \bar{u}_{i\cdot}g_i) + E(v_i - \bar{u}_{i\cdot}g_i)^2 \\ &= Var(v_i(1-g_i) - g_i n_i^{-1}\textstyle\sum_{j=1}^{n_i} e_{ij}) + 0 = (1-g_i)^2\sigma_v^2 + (g_i n_i^{-1})^2[n_i\sigma_e^2] \\ &= [Var(\bar{u}_{i\cdot})^{-1}n_i^{-1}\sigma_e^2]^2\sigma_v^2 + [Var(\bar{u}_{i\cdot})^{-1}n_i^{-1}\sigma_v^2]^2 n_i\sigma_e^2 \\ &= Var(\bar{u}_{i\cdot})^{-2}[\sigma_v^2\sigma_e^2 n_i^{-1}](n^{-1}\sigma_e^2 + \sigma_v^2) = \sigma_v^2[Var(\bar{u}_{i\cdot})^{-1}\sigma_e^2 n_i^{-1}] \\ &= \sigma_v^2(1-g_i) \end{aligned}$$

Which are exactly 3.1 and 3.2, respectively.

2.2 Unbiasedness and Distribution of the Fitting-Of-Constants Estimator for Residual Variance

The following proof relates to the claims surrounding Equation 3.8.

The term $\hat{e}'\hat{e}$ in the σ_e^2 estimator (3.8) is meant to be the residual sum of squares for the regression of $y_{ij} - \bar{y}_{i\cdot}$ on $x_{ij} - \bar{x}_{i\cdot}$, for counties that have $n_i > 1$.

Without loss of generality, assume all counties have more than 1 segment observed (remove those that don't from the dataset before proceeding). We start this proof by describing the matrix that transforms our original data $Y = \{y_{ij}\}_{i\in 1,\ldots,T, j\in 1,\ldots,n_i}$ into the deviation data $\{y_{ij} - \bar{y}_{i\cdot}\}$

Let $\mathbb{1}_i$ denote a row vector of length n with 0s in all entries except those that correspond with segments in county i, which are instead 1. Note that $n_i^{-1}\mathbb{1}_i Y = \bar{y}_{i\cdot}$ $n_i^{-1}\mathbb{1}_i X = \bar{x}_{i\cdot}$.

Let $\mathbb{1}_F$ be a matrix whose rows are $n_i^{-1}\mathbb{1}_i$, with each $n_i^{-1}\mathbb{1}_i$ repeated n_i times.

Setting Up the Problem

It follows that $\mathbb{1}_F Y$ provides a vector of county-restricted means, $(I - \mathbb{1}_F)Y$ provides the deviation data, and $(I - \mathbb{1}_F)(Y - X\beta - V)$ - where V is a vector of realised county effects - provides the associated residuals, $\hat{e}$.

To speak further of the properties of $(I-\mathbb{1}_F)$, note that for any i, $\mathbb{1}_i\mathbb{1}_i' = n_i$, making $(n_i^{-1}\mathbb{1}_i)(n_i^{-1}\mathbb{1}_i') = n_i^{-1}$, which is exactly the entry in $\mathbb{1}_{F,pq}$ for any p, q belonging to the same county. When p, q are in different counties $i \neq j$, $(n_i^{-1}\mathbb{1}_i)(n_j^{-1}\mathbb{1}_j') = 0$, which would also be the value of $\mathbb{1}_{F,pq}$. This lets us conclude that $\mathbb{1}_F$, and therefore $I - \mathbb{1}_F$, is idempotent.

We lastly obtain $\text{rank}(I - \mathbb{1}_F)$. Note that $I - \mathbb{1}_F$ is a block diagonal matrix where each block is a $n_i \times n_i$ matrix with all entires n_i^{-1}, except the diagonals which are $1 - n_i^{-1}$. It is clear that such a block has rank $n_i - 1$ when $n_i > 1$.

Since $I - \mathbb{1}_F$ is being used to fit a linear model, we should take the parameters thereof into account. The deviation data has no intercept term (they were cancelled out in the subtraction process) so the remaining slopes only take away 2 ranks, giving $I - \mathbb{1}_F$ a total rank of $\sum_{i=1}^{T}(n_i - 1) - 2$.

The Proof Itself

The proof is split into two parts, the unbiasedness of 3.8 and the distribution of 3.8. Both can be addressed by taking the expectation. Note that $\sigma_e^{-1}\hat{e} = \sigma_e^{-1}(Y - X\beta - V)$ is trivially multivariate normal with mean 0 and identity variance under the assumed model and using C.2 from the text (*Linear and Generalized Linear Mixed Models and Their Applications*).

$$\begin{aligned} E((\sigma_e^{-1}\hat{e})'(\sigma_e^{-1}\hat{e})) &= E((\sigma_e^{-1}(Y - X\beta - V))'(I - \mathbb{1}_F)'(I - \mathbb{1}_F)(\sigma_e^{-1}(Y - X\beta - V))) \\ &= 0 + \text{tr}((I - \mathbb{1}_F)) = \text{rank}(I - \mathbb{1}_F) = \textstyle\sum_{i=1}^{T}(n_i - 1) - 2 \quad \text{(tr=rank by idempotence)} \end{aligned}$$

We can conclude the unbiasedness of $\hat{\sigma}_e^2$ in (3.8) by noting from here that $(\sigma_e^2)^{-1}E(\hat{e}'\hat{e}(\sum_{i=1}^{T}(n_i - 1) - 2)^{-1}) = 1$. In other words, $E(\hat{\sigma}_e^2)/\sigma_e^2 = 1$.

The distribution can be seen by idempotence of $I - \mathbb{1}_F$, and 0-mean normality of $\sigma_e^{-1}\hat{e}$. These lead us to conclude that $(\sigma_e^2)^{-1}\hat{e}'\hat{e}$ has a $\chi^2_{\text{rank}(I-\mathbb{1}_F)}$ distribution.

$$\Rightarrow d_e\hat{\sigma}_e^2/\sigma_e^2 \text{ has } \chi^2_{d_e} \text{ distribution, where } d_e = \sum_{i=1}^{T}(n_i - 1) - 2$$

2.3 Expectation of the Squared Average of the Ordinary Least-Squares Residuals of a Given County, for Purposes of Estimating County-Based Variance

The following is a proof of Equation 3.10.

We use the same $\mathbb{1}_i$ as in the previous proof. Recall $n_i^{-1}\mathbb{1}_i Y = \bar{y}_{i\cdot}$, $n_i^{-1}\mathbb{1}_i X = \bar{x}_{i\cdot}$, and $\mathbb{1}_i\mathbb{1}_i' = n_i$.

Since $\overset{\triangle}{u}$ is the average of residuals, we can expect it to be 0 on average. This makes $E(\overset{\triangle}{u}^2) = Var(\overset{\triangle}{u})$. Combining with the characterisation of $\mathbb{1}_i$, we may rewrite $\overset{\triangle}{u}$

$$= n_i^{-1}\mathbb{1}_i Y - n_i^{-1}\mathbb{1}_i X(X'X)^{-1}X'Y = n_i^{-1}\mathbb{1}_i(I - X(X'X)^{-1}X')Y$$

$$\Rightarrow E(\overset{\triangle}{u}^2) = Var(\overset{\triangle}{u}) = n_i^{-1}\mathbb{1}_i(I - X(X'X)^{-1}X')Var(Y)(I - X(X'X)^{-1}X')\mathbb{1}_i' n_i^{-1}$$

$Var(Y) = \sigma_v^2 ZZ' + \sigma_e^2 I$ has multiple components that add up to it. Using the distributive property of non-random values, we can evaluate the expected value of $\overset{\triangle}{u}$ by separately evaluating it for different components of $Var(Y)$, then adding them back up.

Component One

The first component is $\sigma_v^2 ZZ'$. Z is a $n \times T$ matrix mapping county-based effects column matrix to the segments belonging to them. In other words, Z is a matrix with columns $\mathbb{1}_j'$ for $j \in 1, ..., T$. This makes XZ a matrix where each column is $n_j\bar{x}_{j\cdot}'$. Similarly, $Z'X$ is a matrix where each row is $n_j\bar{x}_{j\cdot}$. Thus, $X'ZZ'X = \sum_{j=1}^{T} n_j^2 \bar{x}_{j\cdot}\bar{x}_{j\cdot}'$

Additionally, note that $\mathbb{1}_i Z$ is a row vector with n_i in the i^{th} entry and 0 in all others. This makes $\mathbb{1}_i ZZ'X = n_i^2\bar{x}_{i\cdot}$. We can now put everything together

$$\begin{aligned}
&\sigma_v^2 n_i^{-1}\mathbb{1}_i(I - X(X'X)^{-1}X')ZZ'(I - X(X'X)^{-1}X')\mathbb{1}_i' n_i^{-1} \\
&\quad = \sigma_v^2[n_i^{-1}\mathbb{1}_i ZZ'\mathbb{1}_i' n_i^{-1} - 2n_i^{-1}\mathbb{1}_i ZZ'X(X'X)^{-1}X'\mathbb{1}_i' n_i^{-1} \\
&\qquad\qquad + n_i^{-1}\mathbb{1}_i X(X'X)^{-1}X'ZZ'X(X'X)^{-1}X'\mathbb{1}_i' n_i^{-1}] \\
&\quad = \sigma_v^2[n_i^{-1}n_i^2 n_i^{-1} - 2n_i^{-1}(n_i^2\bar{x}_{i\cdot})(X'X)(\bar{x}_{i\cdot}') + \bar{x}_{i\cdot}(X'X)^{-1}(\textstyle\sum_{j=1}^{T} n_j^2\bar{x}_{j\cdot}\bar{x}_{j\cdot}')(X'X)^{-1}\bar{x}_{i\cdot}'] \\
&\quad = \sigma_v^2[1 - 2n_i\bar{x}_{i\cdot}(X'X)\bar{x}_{i\cdot}' + \bar{x}_{i\cdot}(X'X)^{-1}(\textstyle\sum_{j=1}^{T} n_j^2\bar{x}_{j\cdot}\bar{x}_{j\cdot}')(X'X)^{-1}\bar{x}_{i\cdot}']
\end{aligned}$$

Take the factor multiplying σ_v^2 and call it b_i. Note it is exactly the b_i described in Equation 3.10.

Component Two

The second component is $\sigma_e^2 I$. Replacing YY' with it and taking the expected value gets us.

$$\begin{aligned}
&\sigma_e^2 n_i^{-1}\mathbb{1}_i(I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X')\mathbb{1}_i' n_i^{-1} \\
&\quad = \sigma_e^2 n_i^{-2}\mathbb{1}_i(I - X(X'X)^{-1}X')\mathbb{1}_i'
\end{aligned}$$

$$= \sigma_e^2 n_i^{-2}(\mathbb{1}_i\mathbb{1}_i' - \mathbb{1}_i X(X'X)^{-1}X'\mathbb{1}_i) = (n_i - n_i\bar{x}_{i\cdot}(X'X)^{-1}\bar{x}_{i\cdot}' n_i)$$
$$= \sigma_e^2 n_i^{-1}(1 - \bar{x}_{i\cdot}(X'X)^{-1}\bar{x}_{i\cdot}')$$

The second line comes from the weak assumption that X is of full rank, making the projection $(I - X(X'X)^{-1}X')$ idempotent.

Take the factor multiplying σ_e^2 and call it d_i. Note it is exactly the d_i described in Equation 3.10.

Putting everything together, we have $E(\overset{\triangle}{u}^2) = \sigma_v^2 b_i + \sigma_e^2 d_i$, as we wanted.

3 Data in Detail

The data is available from Battese et al. (1988).

The data is comprised of information drawn from satellite images of corn and soybean crops in 12 counties throughout central Iowa. The data frame consists of 37 observations of 10 variables. These variables include the number of segments in each county, sample segment ID per county, true hectarage of corn and soybeans for each sample segment, the number of pixels classified as corn or soybeans for each sample segment, county mean number of pixels classified as corn or soybean, whether an observation is an outlier and the names of the county to which each sample segment belongs. Both classes of pixels are used as predictors in two separate models. The first model estimates hectares of corn, while the second estimates hectares of soybeans. It should be noted that the data is unbalanced, containing different sample sizes among the counties.

A nested error regression model is proposed for the data, with county as the random effect. The model is of the form

$$y_{ij} = \beta_0 + \beta_1 x_{1,ij} + v_i + e_{ij}$$

where i represents county, j represents segment in county, $x_{1,ij}$ and $x_{2,ij}$ represent the number of pixels classified as corn or soybeans respectively, v_i is the area specific random effect and e_{ij}s are sampling errors. The model assumptions are $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$ where σ_v^2 and σ_e^2 are unknown. The response, y_{ij} can be taken as hectares of either corn or soybeans.

There are 5 candidate variables for the model. They are $x_{1ij}, x_{2ij}, x_{1ij}^2, x_{2ij}^2, x_{1ij}x_{2ij}$. In the following parts, we use BIC, Adaptive Fence with minimum dimension criteria, and Adaptive Fence with MSPE to perform model selection.

4 Model Selection via BIC

The initial goal is to select a low dimensional model that can be used to interpret the relationship that corn and soybean pixel counts from satellite photographs may have with the true hectorage of either crop. We first consider model selected by the Bayesian Information Criterion (BIC). Here, BIC $= \hat{Q}(M) + \log(n)|M|$, where $\hat{Q}(M)$ is the negative log-likelihood of the model. Whichever

model minimizes this criterion will be chosen. BIC penalizes models with a higher number of terms, thus this method favors simple models. The top 3 models with lowest BIC when either corn or soybean is the response are listed below:

Response=Corn Hectarage	Response= Soybean Hectarage
1. corn^2	1. soybean
2. corn	2. soybean+soybean^2
3. corn^2+soybean^2	3. soybean+corn×soybean

So when hectares of corn is the response, the model with only the $corn^2$ has the lowest BIC. When the response is hectares of soybeans, the chosen model is the one with predictor variable *soybean*. The result for corn indicate that $corn^2$ can give us the best fitting result and has the lowest possible model complexity. The chosen model when response is hectares of soybeans is very straightforward. It is expected that pixel count of soybeans in satellite images will be predictive of true hectorage of soybeans.

However, the BIC method of model selection is not very flexible. And it puts a large penalty on complex models. Another problem when using BIC in the dataset is the effective sample size n is not well defined since the data from the same county might be correlated. The actual n should be smaller than the number of data points. Hence we consider an alternate algorithm for model selection, the adaptive fence (AF).

5 Adaptive Fence - Introduction

The fence method, proposed by Jiang et al., consists of a procedure to isolate a subgroup of what are considered correct models via the inequality

$$\hat{Q}(M) \leq \hat{Q}(M_f) + C$$

where $\hat{Q}(M)$ is the the measure of lack-of-fit for model M; we use negative log-likelihood in this project, but other measures may be considered. $\tilde{M}$ is the baseline model that has the minimum $\hat{Q}$. C is the cut-off. Any model satisfying this inequality is said to fall within the fence. The optimal model is then selected from the models within the fence according to a criterion of optimality, which can be flexible. This project considers two criterion: minimum dimension, and mean squared prediction error (MSPE).

6 Adaptive Fence with Minimum Dimension

Finite-sample performance of the fence method depends heavily on the choice of the cut-off, or the tuning parameter C. This, in a way, creates a difficulty similar to that in the information criteria. So the adaptive fence is proposed to let the data speak on how to choose this cut-off. We assume

that the full model is the correct model, and among all the possible submodels, we wish to select C to maximize the probability of choosing the optimal model, defined as a correct model with minimum dimension among all the correct models. This means that we choose C that maximize

$$P = P(M_c = M_{opt})$$

Where M_{opt} represent the optimal model, and M_c is the model selected by the fence with the given C. Since we dont know the probability, or the optimal model, we proceed assuming the full model is correct. We generate bootstrap samples from the full model to approximate P on the right hand side of the above equation.

To find the appropriate cutoff, we use the idea of maximum likelihood. We let $p^*(M) = p^*(M_c = M)$ be the empirical probability of M obtained by the bootstrapping. This means $p^*(M)$ is the sample proportion of times out of the total number of bootstrap samples that model M is selected by the fence method with the given C. Let $p^* = \max_M p^*(M)$. And we choose the C that maximizes p^*. The model corresponds with p^* will be the model that we choose.

Here is a detailed outline of the Adaptive Fence algorithm using the minimum dimension criterion of optimality for model selection.

1. Fit the full model with all the linear and quadratic terms of the covariates:

   $$y \sim x_1 + x_2 + (x_1)^2 + (x_2)^2 + x_1 \times x_2 + v + e$$

   From the fitting of the full model, we obtain the parameter estimates of beta, county and error variations and the negative log-likelihood of the full model.

   Then we define the minimum model as M^*, where only the random effects terms of the county and the sampling error are involved. $M^* : y_{ij} \sim v_i + e_{ij}$ We fit this model and obtain the negative log-likelihood for the minimum mode.

   The value of C, the cut-off, lies within the interval: $[0, Q(M^*) - Q(M_f) + (m^*, mf)]$

2. Based on the parameter estimates we obtained from the fitting the full model, we generate 100 bootstrapped data sets.

3. Take all the different cut-off values, C_ns, equally spaced in the cut-off interval. For each C_n, we do the following:

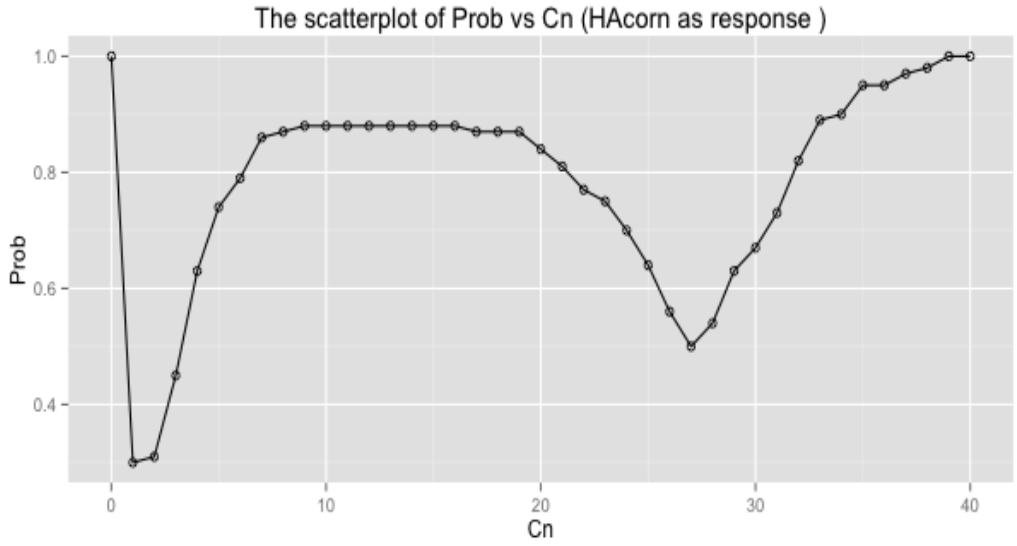
   - Generate 100 datasets using the bootstrap method. Within each bootstrapped data set, we have many possible models based on the dimension criterion. The dimensions range from dimension = 1, with only 1 variable, which is the simplest model, to dimension = 6, with all 6 variables, which is the most complex model.
   - We fit the model from the simplest to the most complex until one model falls within the fence, satisfying the cut-off inequality.

- If there are two or more models with the same dimension all within the fence, then we select the model with the smallest negative log-likelihood.

4. Based on this approach, we find the model that is selected most frequently, and denote the highest frequency as p_i.

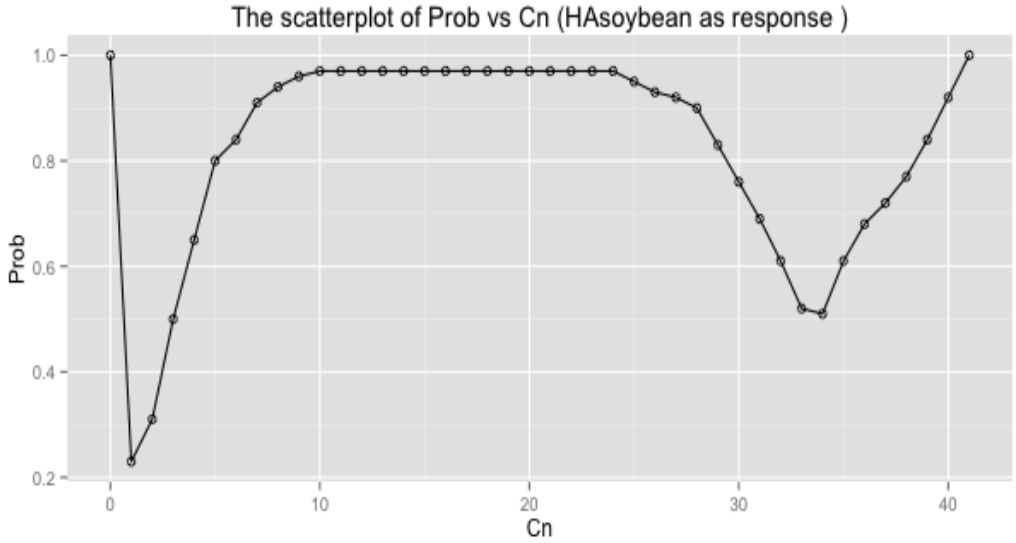
5. Finally, we plot the probabilities p_i versus the cutoff values C_n, identify the C_n with the largest probability, and choose the corresponding model.

This is the plot of probabilities p_i versus cutoff values C_n where corn area is the response, and C ranges from 0-40.



We ignore the extreme minimum and maximum values of C, because when $C = 0$, the full model is selected every time with probability = 1. And when C is large, potentially every model falls within the fence, so the model with the minimum dimension would always be selected. Therefore, we only focus on the middle part of the plot, and it appears that Cs ranging from 10 to 19 have approximately the same highest probabilities (prob = 0.84). It is likely that the Cs may depend heavily on the criterion, which in this case is the minimum dimension criterion. Importantly, we observe that values of C between 10 and 19 all correspond to the same model. Thus, any C in the flat interval leads to the same model, which indicates the value of C here is not very sensitive. This, in part, also reflects the robustness of the adaptive fence method. A non-robust method would give very different results over small intervals of C.

This is the plot of probabilities p_i versus cutoff values C_n where soybean area is the response, with C ranging from 0-41.



It is clear that this plot for the soybean area response has similar shape to the previous plot for the corn data. When we focus on the plateau in the center, it appears that Cs ranging from 10 to 24 have approximately the same highest probabilities (prob = 0.98). In addition, we also observe that values of C between 10 and 24 all correspond to the same model. Thus, the result for the soybean data is similar to the result for the corn data, in that any C in this interval leads to the same model. The value of C here is not very sensitive, again, in part, reflecting the robustness of the adaptive fence method.

The resulting models for corn and soybean are selected by the Ci with corresponding maximums p_i = 0.84 and 0.98 respectively. When corn area is the response variable, the model selected involves the (corn pixels)^2 variable, and when soybean is the response variable, the model selected involves the soybean pixels variable.

If we compare the model selection by BIC and AF (use the minimum-dimension criterion as the criterion of optimality), these two method chose with same models. This is largely due to the similarities in both algorithms. Both put heavy emphasis on models with few predictors.

7 Adaptive Fence with MSPE

One of the advantages of fence method is that the criterion of optimality can incorporates practical considerations. When the goal is to predict the hectares of corn (or soybeans) in each county, minimizing MSPE, which is a measure of prediction error, would be a more proper criterion than minimum dimension.

When adopt mpse as the criterion, we have to combine the data for first three counties because:

1. The terms used in estimating MSPE has $n_i - 1$ as denominator. The first three counties only have one sample. Thus we must combine them to avoid having 0 in the denominator.

2. Having more than one sample for each group enable us to estimate the random effect of the corresponding group. Otherwise, we can not discriminate the random effect from the error terms in the model.

The algorithm using MSPE is very similar to the algorithm using minimum dimension. The difference is the procedure used for choosing optimal model:

1. Steps I and II, fitting the full model and getting the bootstrapped data are just the same as in the previous section.

2. After generating Cs ($C_1, \ldots, C_n$) with equal space

   (a) Fit all 32 potential models (1 for dim=0, 5 for dim=1, 10 for dim=2, 10 for dim=3, 5 for dim=4, and 1 for dim=5), obtain their estimates $(\hat{\beta}, \hat{\sigma_v^2}, \hat{\sigma_e^2})$ and $\hat{Q}(M)$, which is still negative log-likelihood.

   (b) Find the models that fall in the fence. For these models, calculate the MSPE. The optimal model is the one with smallest MSPE.

3. Calculate P^* for each C_n.

4. Plot the P^* against C_n as in the previous section.

For the models that lie in the fence, the MSPE is calculated with following formula:

$$mspe(M) = \sum_{i=1}^{m} \{\tilde{\mu_i}^2(\hat{\phi}) + 2a_i(\hat{\sigma_v^2}, \hat{\sigma_e^2})\bar{X_i}'\hat{\beta}\bar{y_i} + b_i(\hat{\sigma_v^2}, \hat{\sigma_e^2})\hat{\mu_l}^2\},$$

where

$$\tilde{\mu_i}^2(\hat{\phi}) = \bar{X_i}'\hat{\beta} + \{\tfrac{n_i}{N_i} + (1 - \tfrac{n_i}{N_i})\tfrac{n_i\hat{\sigma_v^2}}{\hat{\sigma_e^2} + n_i\hat{\sigma_v^2}}\}(\bar{y_i} - \bar{x_i}'\hat{\beta})$$

$$a_i(\hat{\sigma_v^2}, \hat{\sigma_e^2}) = (1 - \tfrac{n_i}{N_i})\tfrac{\hat{\sigma_e^2}}{\hat{\sigma_e^2} + n_i\hat{\sigma_v^2}}$$

$$b_i(\hat{\sigma_v^2}, \hat{\sigma_e^2}) = 1 - 2\{\tfrac{n_i}{N_i} + (1 - \tfrac{n_i}{N_i})\tfrac{n_i\hat{\sigma_v^2}}{\hat{\sigma_e^2} + n_i\hat{\sigma_v^2}}\}$$

$$\hat{\mu_l}^2 = \tfrac{1}{n_i}\textstyle\sum_{j=1}^{n_i} y_{ij}^2 - \tfrac{N_i - 1}{N_i(n_i - 1)}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2,$$

and this term is a design unbiased estimator for $\bar{Y_i}^2$.

For the term $\tilde{\mu_i}^2$, we have to do further calculations.

$$\tilde{\mu_i}^2(\hat{\phi}) = \hat{\beta}'\bar{X_i}\bar{X_i}'\hat{\beta} + [\{\tfrac{n_i}{N_i} + (1 - \tfrac{n_i}{N_i})\tfrac{n_i\hat{\sigma_v^2}}{\hat{\sigma_e^2} + n_i\hat{\sigma_v^2}}\}(\bar{y_i} - \bar{x_i}'\hat{\beta})]^2 + 2\{\tfrac{n_i}{N_i} + (1 - \tfrac{n_i}{N_i})\tfrac{n_i\hat{\sigma_v^2}}{\hat{\sigma_e^2} + n_i\hat{\sigma_v^2}}\}\hat{\beta}'\bar{X_i}(\bar{y_i} - \bar{x_i}'\hat{\beta})$$

Note that term $\bar{X_i}\bar{y_i}$ exists in the formula of mspe and $\tilde{\mu_i}^2(\hat{\phi})$, of which the expected value is$\bar{X_i}\bar{Y_i}$. The terms $\bar{X_i}\bar{X_i}$ and $\bar{X_i}\bar{x_i}$ (the former is the expected value of the latter) exist in the formula of mspe. The value of these terms could be hard to estimate when there are squares or interaction of

variables in the model. Hence we have to use design unbiased estimators to estimate these terms or the expected value of these terms.

$$\bar{X^2_{1,i}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1ij}^2 - \frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(x_{1ij} - \bar{x_{1i\cdot}})^2$$

$$\bar{X^2_{2,i}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{2ij}^2 - \frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(x_{2ij} - \bar{x_{2i\cdot}})^2$$

$$\bar{X_{1,i}}\bar{X_{2,i}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1ij}x_{2ij} - \frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i} x_{1ij}x_{2ij} - \sum_{j=1}^{n_i} x_{1ij}x_{2ij}$$

$$\bar{X^2_{1,i}}\bar{y_{i\cdot}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1ij}^2 y_{ij} - \frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(x_{1ij}^2 - \sum_{j=1}^{n_i} x_{1ij}^2)(y_{ij} - \bar{y_{i\cdot}})$$

$$\bar{X^2_{2,i}}\bar{y_{i\cdot}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{2ij}^2 y_{ij} - \frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(x_{2ij}^2 - \sum_{j=1}^{n_i} x_{2ij}^2)(y_{ij} - \bar{y_{i\cdot}})$$

$$\bar{X_{1,i}}\bar{X_{2,i}}\bar{y_{i\cdot}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1ij}x_{2ij}y_{ij} - \frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(x_{1ij}x_{2ij} - \sum_{j=1}^{n_i} x_{1ij}x_{2ij})(y_{ij} - \bar{y_{i\cdot}})$$

When we use MSPE as the optimality criterion, we have the following properties.

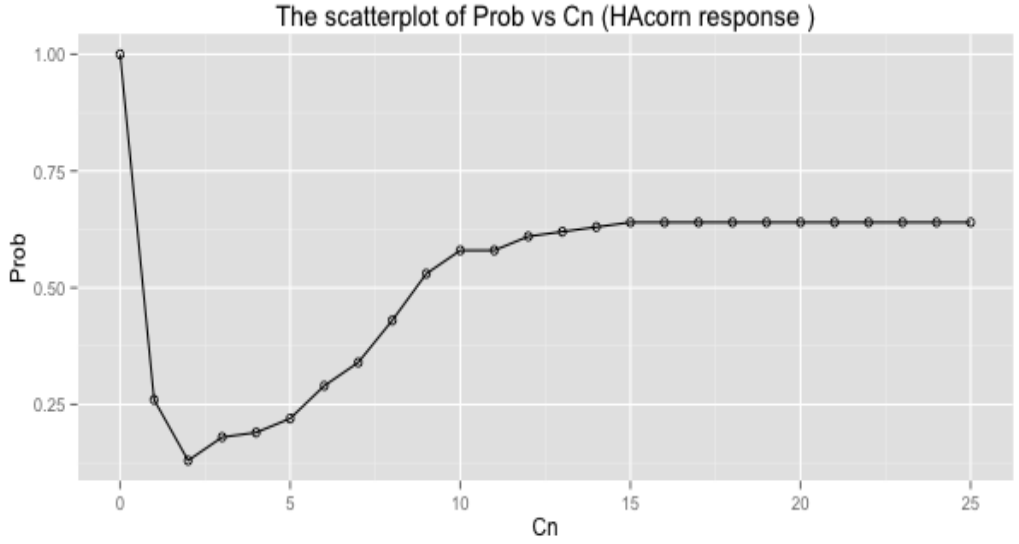
1. The curve will not be in W shape, instead, it will be flat in until the end. The argument for this property is as follows:

   When c is small, only the full model will be in the fence. However, the full model could have a comparatively large MSPE due to overfit. When c gets larger, the models with lower dimension will be in the fence. Once the model with the smallest MSPE is in the fence, it will always be chosen as optimal.

2. As compared to minimum dimension, MSPE optimality criterion choses models with higher dimension. This is mainly because overly simplistic models often do not predict well.

3. p^* does not reach 1 when C is large. Several models can all have small MSPE. When fitting bootstrapped data, these candidates all have a chance being chosen.

The AF algorithm with MSPE criterion, with corn as the response, results in the following plot.

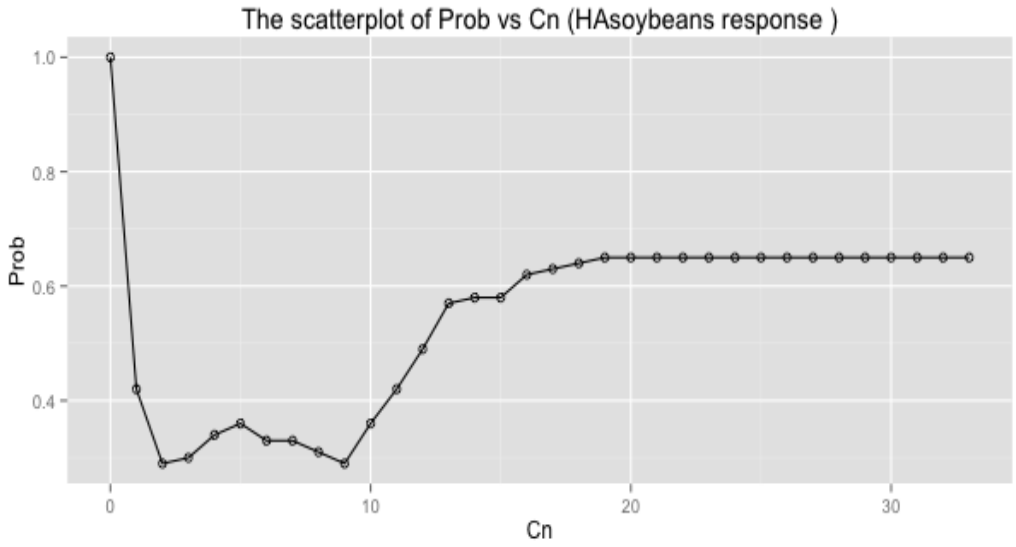


As in the previous section, we ignore low values of C. The portion of the graph that is of interest is the plateau, which occurs when Cn ranges from 15 to 25. This section is comprised of three different models.

1. $Soybean + Soybean^2 + Soybean \cdot Corn$
2. $Corn + Soybean^2 + Soybean \cdot Corn$
3. $Corn + Soybean + Soybean^2$

Although these three models have the highestl p^* (0.6, 0.13, 0.12), the first is chosen most frequently. Thus we will consider the first model as the model chosen by the fence algorithm. It seems odd that the predictors of hectares of corn in this model do not involve corn directly. However, our purpose here is to predict, not to interpret.

When using soybeans as response, the algorithm produces the following curve for P_i vs.C_i.



The shape of this plot is, not surprisingly, similar to that of the previous. Here we are interested in the values of C that range from 20 to 32. In this range, there is only one model, $Soybean^2$, with p^* roughly 0.65.

8 Computational Issues

The procedures for AF, using either minimum dimension criteria or MSPE criteria have been given in above sections. In terms of fitting each potential model, we use function *lmer* from *lme4* package in R, with county as the random effect, and fit the model parameters using the maximum likelihood method.

Translating the procedure into an R function is quite straightforward. The BIC values and log likelihood values can be accessed from the fitted model object.

Our function for minimum dimension AF takes less than one minute to run. And we can easily parallelize this process for each value of C_n. So a total of about one and a half minutes is all it takes to obtain the results. For the AF with MSPE criterion, it takes about 3 minutes to cycle through the algorithm for just one C_n . So running in parallel on 32 core machines is the fastest way to get the final results. If this is done, the total run time is just over 5 minutes.

Overall, with proper coding, these algorithms are not particularly computationally expensive. However, it should be noted that the sample size and number of predictors used for the LANDSAT data is relatively small. Also only bootstrap samples of size 100 were constructed. If one has a relatively large dataset, or would like to create more than 100 bootstrap samples, the AF algorithm can become quite labor intensive. Therefore, in some cases, it may be prudent to consider other versions of the adaptive fence such as fast AF, which is not covered in this paper. Less computationally-based measures of lack of fit or optimality criterion could also be used.

9 Summary and concluding remarks

The BIC method is quite straightforward, and our result is reasonable and in line with other groups results. The fence method offers much flexibility in the choice of criterion, whether it is for measure of lack-of-fit or optimality. For example, in problem two we switch to using prediction accuracy as criterion of optimality,which yields different results than minimum dimension. The two fence methods can be summarized in the following table.

	Measure of Lack of Fit: Q(M)	Criterion of Optimality
AF1	Negative log-likelihood	Minimum-dimension
AF2	Negative log-likelihood	Minimum mspe

In this dataset, the MSPE method will have some difficulty estimating the unknown variables, especially for the higher interaction terms, which have large variation. This seems to be the driving force behind the inconsistencies in the different groups results.

There are also some issues with the fence method. When the candidate variable number gets larger, the computational aspects of the algorithm will become significantly more difficult and expensive.

There has been much learned from using linear mixed models in small area estimation, various adaptive fence methods for model selection, and observed best prediction as a measure for model optimality. The more appropriate method of analysis depends largely on the investigators interests. If simplistic interpretation of the relationship between predictors and response is the goal, then the methods using BIC or minimum dimensionality AF should be used. Minimum dimensionality AF may be preferable to the researcher who wishes to tailor or compare measures of lack of fit. If the

intent is to create a robust model that predicts well, the adaptive fence using MSPE criterion is more appropriate. Overall, the adaptive fence method seems to be a better choice in both cases, as it allows for multiple considerations for both lack of fit and optimality, which can be adjusted based on the question at hand.