A Unit Level Model for Prediction of Categorical Data with Auxiliary Information

Gabriel Demuth, Emily Berg, Philip Dixon

Iowa State University

September 14, 2017
The NRI is administered by Natural Resources Conservation Service, Iowa State University provides statistical support.

- Uses rotating panel survey design.
- PSU is Segment (160 acres), each segment contains 3 points (SSU)
- Point classification done by human interpretation of aerial photos.
- Slow process: 2012 data released in 2015
- Interest in NRI based estimates/predictions ahead of release.

Goal for this talk: Incorporate auxiliary information to improve prediction.
In Iowa, about 0.2% of points changed into or out of cropland between 2008 and 2010

NRI based predictions will be no change, incorporate auxiliary information.
The CDL classifies land use in the United States using computer interpretation of satellite images.

- Offers complete coverage of the US.
- Resolution is $30 \times 30$ meter pixels (early releases were $56 \times 56$ meters).
- Trained using June Area Survey, other data sources.
- Prone to misclassification from an NRI point of view.

We use the CDL pixel closest to each NRI point.
Consider a population of individuals \( i = 1, \ldots, N \). At any time \( t \), an individual is in one of \( K \) discrete states.

- Suppose we can observe an individual’s state without error using slow/expensive instrument \( X \) (NRI)
- We assume that points’ states evolve according to a first order Markov process

Let

\[
\pi_{k|h} = P(X_{t+1} = k \mid X_t = h)
\]

- Want to predict \( X_1 \) for each point \( i \)
Suppose also that we have access to a fast/cheap instrument $Y$ (CDL) that measures $X$ with error.

- Assume missclassification probabilities for $Y$ depend only on $X$.

$$p_{m|k} = P(Y_t = m | X_t = k)$$
Under these assumptions and using 0-1 loss, the Bayes’ Rule for predicting $X_{t+1}$ is $P(X_{t+1} | Y_{t+1}, X_t)$, and by Bayes’ Rule:

$$
\tilde{p}_{k|m,h} \equiv P(X_1 = k | Y_1 = m, X_0 = h) = \frac{\pi_k | h p_m | k}{\sum_{j=1}^{K} \pi_j | h p_m | j}
$$

**Example:** Suppose $K = 2$, and we observe $X_0 = 1$, $Y_1 = 2$.

<table>
<thead>
<tr>
<th>Markov Process</th>
<th>Y Accuracy</th>
<th>Predictive Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1</td>
<td>1$</td>
<td>$p_1</td>
</tr>
<tr>
<td>.8</td>
<td>.5</td>
<td>.8</td>
</tr>
<tr>
<td>.8</td>
<td>.75</td>
<td>.57</td>
</tr>
<tr>
<td>.8</td>
<td>.95</td>
<td>.17</td>
</tr>
<tr>
<td>$\pi_2</td>
<td>1$</td>
<td>$p_2</td>
</tr>
<tr>
<td>.8</td>
<td>.5</td>
<td>.2</td>
</tr>
<tr>
<td>.8</td>
<td>.75</td>
<td>.43</td>
</tr>
<tr>
<td>.8</td>
<td>.95</td>
<td>.83</td>
</tr>
</tbody>
</table>
Under the model assumptions and 0-1 loss, the Bayes Rule is to predict \( X_1 = \max\{\tilde{p}_1|m,h, \ldots, \tilde{K}_k|m,h\} \)

- How does this predictor synthesize \( X, Y \) information in practice?

- We investigate this question by comparing the expected *correct* prediction rate using \( \tilde{p} \) and only \( X \) information.

**Correct Prediction Rate:** Condition on \( X_0 = h \), weight probability of correct prediction under \( \tilde{p}_o|m,h \) by \( P(Y_1 = m \mid X_0 = h) \)

\[
CR_\pi = \sum_{h=1}^{K} T_h \times \max\{\pi_1|h, \pi_2|h, \ldots, \pi_K|h\}/N
\]

\[
CR_{\tilde{p}} = \sum_{h=1}^{K} \sum_{m=1}^{K} T_h \times \max\{p_{m1}|\pi_1|h, p_{m2}|\pi_2|h, \ldots, p_{mK}|\pi_K|h\}/N
\]
Behavior of Predictor in "Stable" System

Assume that $\pi_{h|h} > \pi_{j|h}, j \neq h$ and $p_{m|m} > p_{j|m}, j \neq m$.

1: If $\pi_{h|h}p_{m|h} \geq \pi_{j|h}p_{m|j}$ $\forall j, h, m$, then

$$\min\{\pi_{1|1}, \ldots, \pi_{K|K}\} \leq CP_{\bar{p}} \leq \max\{\pi_{1|1}, \ldots, \pi_{K|K}\}$$

And $CR_{\bar{p}} = CR_{\pi}$.

2: If 1 holds except for at least one $m^*$ such that

$\pi_{m^*|h}p_{m^*|m^*} > \pi_{j|h}p_{m^*|j}$ $\forall j \neq m^*$, then

$$CP_{\bar{p}} > CP_{\pi}$$
Expected Correct Prediction Rate of $\tilde{\rho} \mid m, h$

$\pi_1 \mid 1 = \pi_2 \mid 2 = .8$
For this example, we consider the population of real core points in Iowa ($N = 2983$). We model at the level of Broad Uses. **NRI (X) Categories:**

1. Cropland (Broad Uses 1, 2)
2. Other (Broad Uses 3 - 12)

We estimate the $\pi$s using 2008 - 2009 and 2009 - 2010 transitions, saving 2010 - 2011, 2011 - 2012 for verification. **CDL (Y) Categories:**

1. Cropland
2. Other

$p$ parameters are estimated using 2009 and 2010.

- In this case $M = K$ and NRI (X), CDL (Y) categories have the same meaning.
Application 1 Results

At the level of broad use, cropland is very stable, and the CDL is fairly accurate.

Parameter Estimates

<table>
<thead>
<tr>
<th>( \pi )</th>
<th>Est</th>
<th>( p )</th>
<th>Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_{1</td>
<td>1} )</td>
<td>0.998</td>
<td>( p_{1</td>
</tr>
<tr>
<td>( \pi_{2</td>
<td>2} )</td>
<td>0.984</td>
<td>( p_{2</td>
</tr>
</tbody>
</table>

Predictive Distribution

<table>
<thead>
<tr>
<th>( \tilde{\pi} )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{\pi}<em>k \mid Y</em>{t+1} = 1, X_t = 1 )</td>
<td>0.9998</td>
<td>0.0002</td>
</tr>
<tr>
<td>( \tilde{\pi}<em>k \mid Y</em>{t+1} = 2, X_t = 1 )</td>
<td>0.9885</td>
<td>0.0115</td>
</tr>
<tr>
<td>( \tilde{\pi}<em>k \mid Y</em>{t+1} = 1, X_t = 2 )</td>
<td>0.1413</td>
<td>0.8587</td>
</tr>
<tr>
<td>( \tilde{\pi}<em>k \mid Y</em>{t+1} = 2, X_t = 2 )</td>
<td>0.0029</td>
<td>0.9971</td>
</tr>
</tbody>
</table>

Condition 1 applies \( \rightarrow \) \( \tilde{\pi} \) predicts no change, and misclassifies 33 points.
Correct Prediction Rate for Application 1
The General Joint Distribution

Assume that we have $Y$ data available for years $1, \ldots, T$, and that we have no interest in predicting out farther ahead than $T$. Then shifting notation slightly

$$P(X_1 = x_1, \ldots, X_T = x_T | X_0 = x_0, Y_1 = y_1, \ldots, Y_T = y_T) =$$

$$\prod_{b=1}^{T} \prod_{t=1}^{b-1} \frac{\pi_{x_t} | x_{t-1} p_{y_t} | x_t}{\sum_{s \in S_T} \prod_{j_t \in s} \pi_{j_t} | j_{t-1} p_{y_t} | j_t}$$

Where $S_T$ is the set of all sets of $X$ paths of length $T$.

- if $K = 2$, $T = 2$ then

$$S_T = \left\{ x_1, x_2 \text{ s.t. } x_1, x_2 \in 1, 2 \right\} = \left\{ \{1,1\}, \{1,2\}, \{2,1\}, \{2,2\} \right\}.$$ 

The denominator is the sum of probabilities of all possible $X$ paths, conditioned on $X_0, Y_1 \ldots Y_B$. 
Onwards from the Joint

Equation (1) is the joint distribution of a particular X path $T$ steps into the future. We can easily use it to calculate other distributions.

Example: We observe $Y_1$ through $Y_4$, but are only interested in predicting $X_2$. We simply

- Calculate

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4 | x_0, y_1, y_2, y_3, y_4)$$

- Marginalize out $X_1, X_3, X_4$.

Which gives

$$\hat{p}^* \equiv P(X_2 = x_2 | x_0, y_1, y_2, y_3, y_4)$$

$$= \frac{p_{y_2 | x_2} \sum_{j_1=1}^{K} \sum_{j_3=1}^{K} \sum_{j_4=1}^{K} \pi_{j_1 | x_0} \pi_{x_2 | j_1} \pi_{j_3 | x_2} \pi_{j_4 | j_3} p_{y_1 | j_1} p_{y_3 | j_3} p_{y_4 | j_4}}{\sum_{s \in S_4} \prod_{j_t \in s} \pi_{j_t | j_{t-1}} p_{y_t | j_t}}$$

(2)
Applying $\tilde{p}^*$

No change is a good predictor for cropland one year ahead.

- Two years ahead, predicting No Change (only using NRI information) misses 49 points (1.6%).

We can use the distribution $\tilde{p}^*$ found in (2) to incorporate CDL information from 2010 - 2014.

- There are 32 distinct predictive distributions (combinations of $X_0$ values and possible $Y$ histories)

Each of these generate a unique prediction.
$	ilde{p}^*$ No Change Probabilities
Two Years Ahead: Results

No Change misclassifies 49 points in 2012, the model misclassifies 329.

Why?
The model currently assumes that the $p$ probabilities do not depend on time.

- Estimating using data for 2011 gives $p_{1|1} = 0.8808$ and $p_{2|2} = 0.7720$. Values are similar for 2012.

Using 2011 estimates of $p$, the model only misclassifies 282 points in 2012, and only predicts change for two histories:

- $(X_0 = 1, Y_1 = 2, Y_2 = 2, Y_3 = 2, Y_4 = 2)$ (4/168 change)
- $(X_0 = 2, Y_1 = 1, Y_2 = 1, Y_3 = 1, Y_4 = 1)$ (12/97 change)

Suggesting that it is still overly optimistic about the accuracy of the CDL(Y).
Data Application 2: Crop Rotation

Cropland tends to remain cropland; however, Iowa cropland is dominated by corn and soybeans in an every-other-year rotation.

NRI (X) Categories:

1. Corn
2. Soybeans
3. Other Cropland
4. Other (non-cropland)

CDL (Y) Categories:

1. Corn
2. Soybeans
3. Other
Data Application 2: Parameter Estimates

Table: Estimates of $\pi_k | h$

<table>
<thead>
<tr>
<th></th>
<th>$h = 1$</th>
<th>$h = 2$</th>
<th>$h = 3$</th>
<th>$h = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Corn)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 1$ (Corn)</td>
<td>0.284</td>
<td><strong>0.944</strong></td>
<td>0.045</td>
<td>0.002</td>
</tr>
<tr>
<td>$k = 2$ (Soy)</td>
<td><strong>0.689</strong></td>
<td>0.036</td>
<td>0.016</td>
<td>0.002</td>
</tr>
<tr>
<td>(Crop)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 3$ (Crop)</td>
<td>0.018</td>
<td>0.012</td>
<td><strong>0.926</strong></td>
<td>0.001</td>
</tr>
<tr>
<td>(Other)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 4$ (Other)</td>
<td>0.008</td>
<td>0.009</td>
<td>0.013</td>
<td><strong>0.994</strong></td>
</tr>
</tbody>
</table>

Note the corn-soybean cycle.

Table: Estimates of $p_m | k$

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Corn)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 1$ (Corn)</td>
<td><strong>0.714</strong></td>
<td>0.186</td>
<td>0.096</td>
<td>0.262</td>
</tr>
<tr>
<td>(Soy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 2$ (Soy)</td>
<td>0.140</td>
<td><strong>0.672</strong></td>
<td>0.066</td>
<td>0.192</td>
</tr>
<tr>
<td>(Other)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 3$ (Other)</td>
<td>0.146</td>
<td>0.142</td>
<td><strong>0.837</strong></td>
<td>0.546</td>
</tr>
</tbody>
</table>
Results: One Year Ahead

There are 12 distinct predictive distributions (combinations of $(X_0, Y_1)$ values).
Accuracy of 1-Step Ahead

Using only NRI(X) data, the best predictions are corn $\rightarrow$ soy, soy $\rightarrow$ corn, and no change for the other categories.

- 2851 misclassified points using only NRI(X)
- 2193 of these are corn that remain in corn

<table>
<thead>
<tr>
<th>$x_0$</th>
<th>$y_1$</th>
<th>Prediction</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>corn</td>
<td>corn</td>
<td>1366</td>
<td>1100</td>
<td>31</td>
<td>15</td>
<td>1146</td>
</tr>
<tr>
<td>corn</td>
<td>soy</td>
<td>539</td>
<td>3065</td>
<td>11</td>
<td>16</td>
<td>566</td>
</tr>
<tr>
<td>corn</td>
<td>other</td>
<td>288</td>
<td>531</td>
<td>42</td>
<td>57</td>
<td>387</td>
</tr>
<tr>
<td>soy</td>
<td>corn</td>
<td>3628</td>
<td>33</td>
<td>16</td>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>soy</td>
<td>soy</td>
<td>573</td>
<td>39</td>
<td>15</td>
<td>12</td>
<td>66</td>
</tr>
<tr>
<td>soy</td>
<td>other</td>
<td>552</td>
<td>13</td>
<td>21</td>
<td>32</td>
<td>66</td>
</tr>
<tr>
<td>crop</td>
<td>corn</td>
<td>33</td>
<td>7</td>
<td>270</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>crop</td>
<td>soy</td>
<td>10</td>
<td>23</td>
<td>219</td>
<td>6</td>
<td>39</td>
</tr>
<tr>
<td>crop</td>
<td>other</td>
<td>30</td>
<td>29</td>
<td>2174</td>
<td>32</td>
<td>91</td>
</tr>
<tr>
<td>other</td>
<td>corn</td>
<td>28</td>
<td>6</td>
<td>2</td>
<td>4382</td>
<td>36</td>
</tr>
<tr>
<td>other</td>
<td>soy</td>
<td>7</td>
<td>22</td>
<td>2</td>
<td>2988</td>
<td>31</td>
</tr>
<tr>
<td>other</td>
<td>other</td>
<td>18</td>
<td>20</td>
<td>13</td>
<td>8665</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7072</td>
<td>4888</td>
<td>2816</td>
<td>16222</td>
<td>2585</td>
</tr>
</tbody>
</table>
We can make predictions using the corn, soy, etc categories, then remap them to Cropland/Not Cropland as before.

- Finer partition of the $X$, $Y$, spaces may allow for a more accurate predictor than the direct model.

At the Broad Use level, this misclassifies 305 points.

- Better than using $\tilde{p}^* = P(X_1 | x_0, y_1, y_2, y_3, y_4)$ and only Cropland/Not Cropland categories (329 misclassified), but not hugely.

- Problem may still be model overconfidence in the CDL (see corn/corn).
Incorporation of Weights

Up to now we have only provided predictions for specific points.

- Really want a prediction for, say, the amount of cropland in Iowa.
- The NRI is a survey, each point has a weight $W_i$

Let $\hat{A}_{k,t}$ be the estimated area in category $k$ $t$ years ahead. Two obvious ways to incorporate weights and calculate $\hat{A}_{k,t}$.

1. $\hat{A}_{k,t} = \sum_{i=1}^{N} 1\{\text{Predict } k\}$
2. $\hat{A}_{k,t} = \sum_{i=1}^{N} P(k | \circ)$

Note that if we assume each point is an observation of a multinomial of size $W_i$ (represents $W_i$ acres), 2 is the Bayes’ Rule.
Incorporation of Weights Example

We apply both schemes for incorporating weights to the crop rotation example.

- Only core segments, but does include pseudopoints.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Corn</th>
<th>Soy</th>
<th>Crop</th>
<th>Other</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Class.</td>
<td>44371</td>
<td>44371</td>
<td>44371</td>
<td>44371</td>
<td>39292</td>
</tr>
<tr>
<td>Model Alloc.</td>
<td>109277</td>
<td>82720</td>
<td>41393</td>
<td>58339</td>
<td>2697</td>
</tr>
<tr>
<td>NRI Only Class.</td>
<td>79310</td>
<td>112666</td>
<td>42249</td>
<td>57526</td>
<td>23775</td>
</tr>
<tr>
<td>NRI Only Alloc.</td>
<td>108385</td>
<td>82915</td>
<td>41974</td>
<td>58477</td>
<td>3119</td>
</tr>
<tr>
<td>2011 NRI</td>
<td>113435</td>
<td>79560</td>
<td>41725</td>
<td>57031</td>
<td></td>
</tr>
</tbody>
</table>

- Allocation outperforms classification.
- Model allocation is substantially the best by Root Mean Squared Error.
Adding Covariates.

- Let transition/misclassification probabilities depend on covariate information (distance from urban center, number of years in current state, etc, confidence measure in classification, etc)
- Easy to do via multinomial regression for each individual

Continuous $Y$ Data.

- Assume we still observe $Y$ on a discrete schedule, but the measurement is a continuous value (reflectivity of a light band at a pixel)
- Specify density of $Y_t | X_t$ (e.g. $(Y_t | X_t = k) \sim N(\mu_k, \sigma^2)$)
- Model then essentially performs classification for us.
- Extend to multivariate $Y$ observations, more frequent $Y$ observations...
Joint work with Emily Berg and Philip Dixon of Iowa State University.

References:

- Cropland Data Layer Website: https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php