Can BIGDATA Help in the Production of Reliable Local Area Statistics?

Partha Lahiri

JPSM, University of Maryland, College Park

plahiri@umd.edu

The Fourth Baltic-Nordic Conference on Survey Statistics - BaNoCoSS-2015, 24-28 August 2015 in Helsinki, Finland.

August 24, 2015

Three Examples of Local Area Statistics

- Estimation of crop acreage, crop production, crop yield for the purpose of local agricultural decision making, payments to farmers if crop yields are below certain levels.

- Estimation of transportation related variables such as purpose of the trip (work, shopping, social, etc.), means of transportation (car, walk, bus, subway, etc.), travel time of trip to assist transportation planners and policy makers who need comprehensive data on travel and transportation patterns.

- Estimation of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions.

Problem 1: Remote Sensing BIGDATA

- Can earth resources satellite data provide useful ancillary data source for county estimates of crop acreage?
- Satellite information is recorded for *pixels* (a term for *picture elements*). A pixel is about .45 hectares;
- Based on satellite readings in early Fall, it is possible to classify the crop cover all pixels. This generates big data.

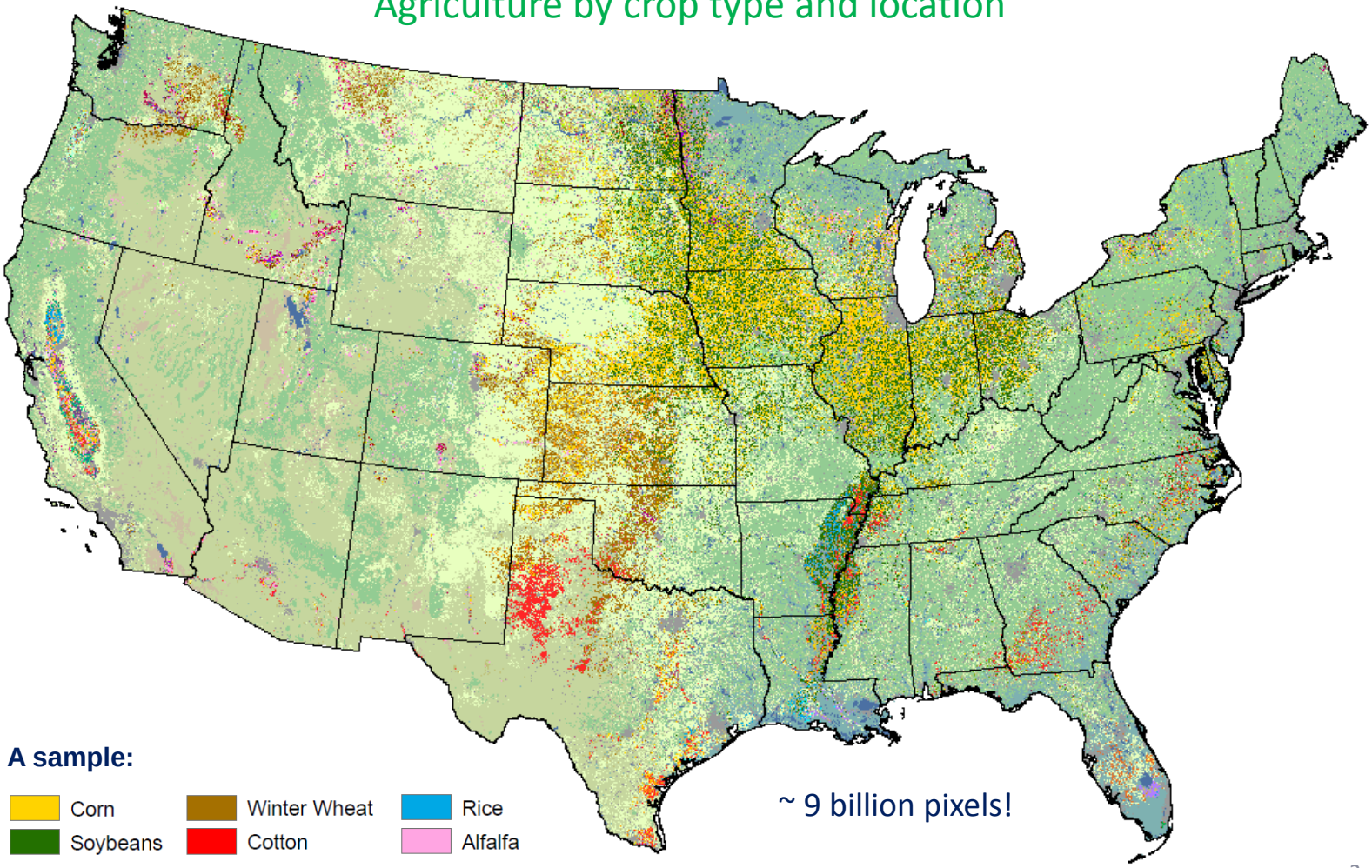
A Quote from Bellow et al.

The polar-orbiting Landsat satellites contain a multi-spectral scanner (MSS) that measures reflected energy in four bands of the electromagnetic spectrum for an area of just under one acre. The spectral bands were selected to be responsive to vegetation characteristics. In addition to the MSS sensor, Landsats IV and V have a Thematic Mapper (TM) sensor which measures seven energy bands and has increased spatial resolution. The large area (185 by 170 km) and repeat (16 day per satellite) coverage of these satellites opened new areas of remote sensing research: large area crop inventories, crop yields, land cover mapping, area frame stratification, and small area crop cover estimation.

Courtesy of Carol Crawford, NASS-USDA (6 slides)

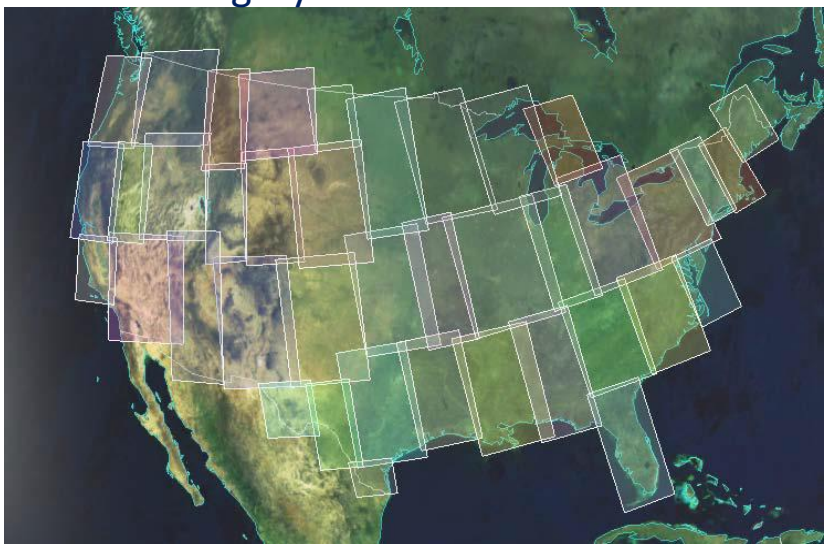
Cropland Data Layer

Agriculture by crop type and location

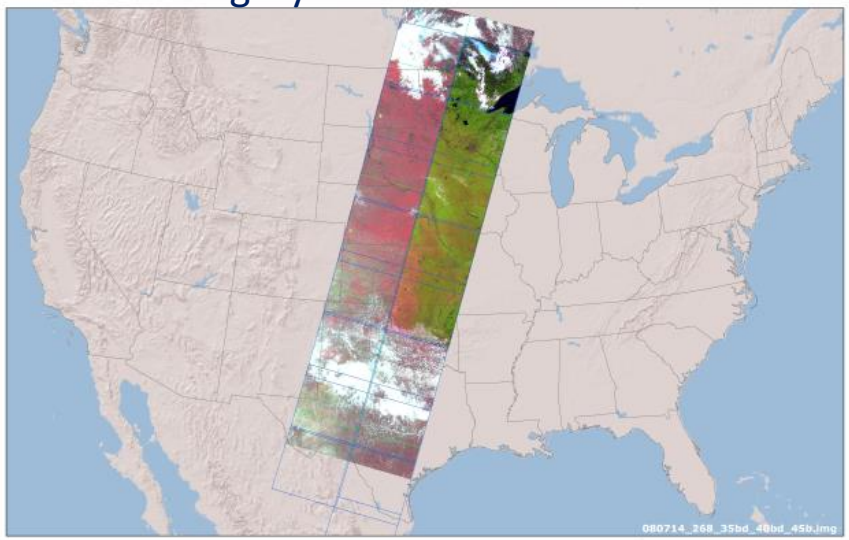


2014 Cropland Data Layer Inputs

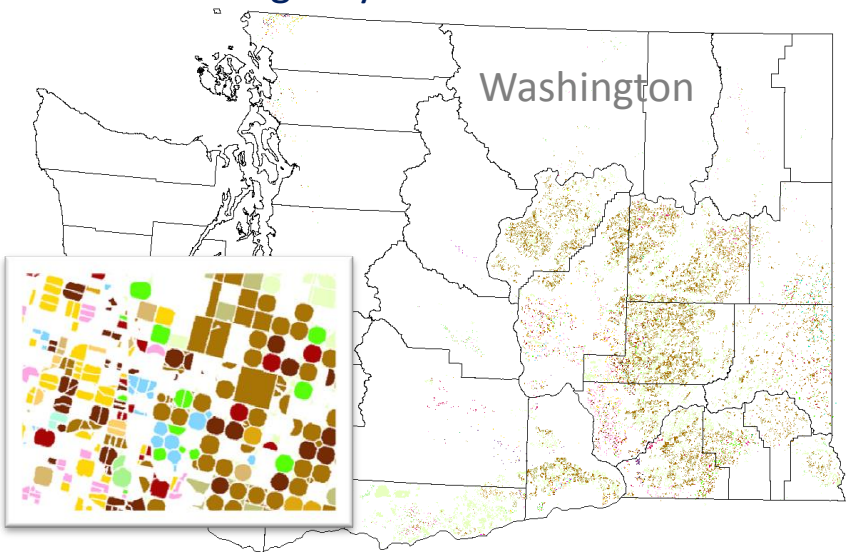
Satellite Imagery – Deimos & UK2



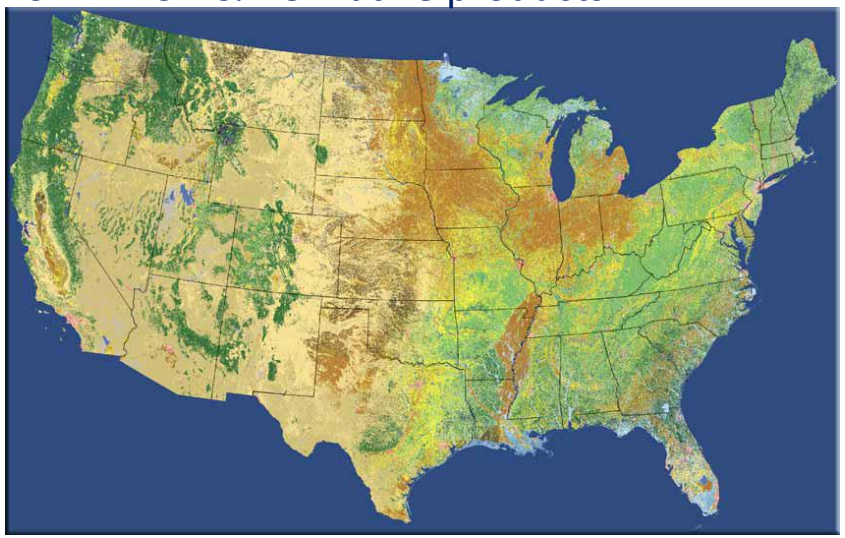
Satellite Imagery – Landsat 8



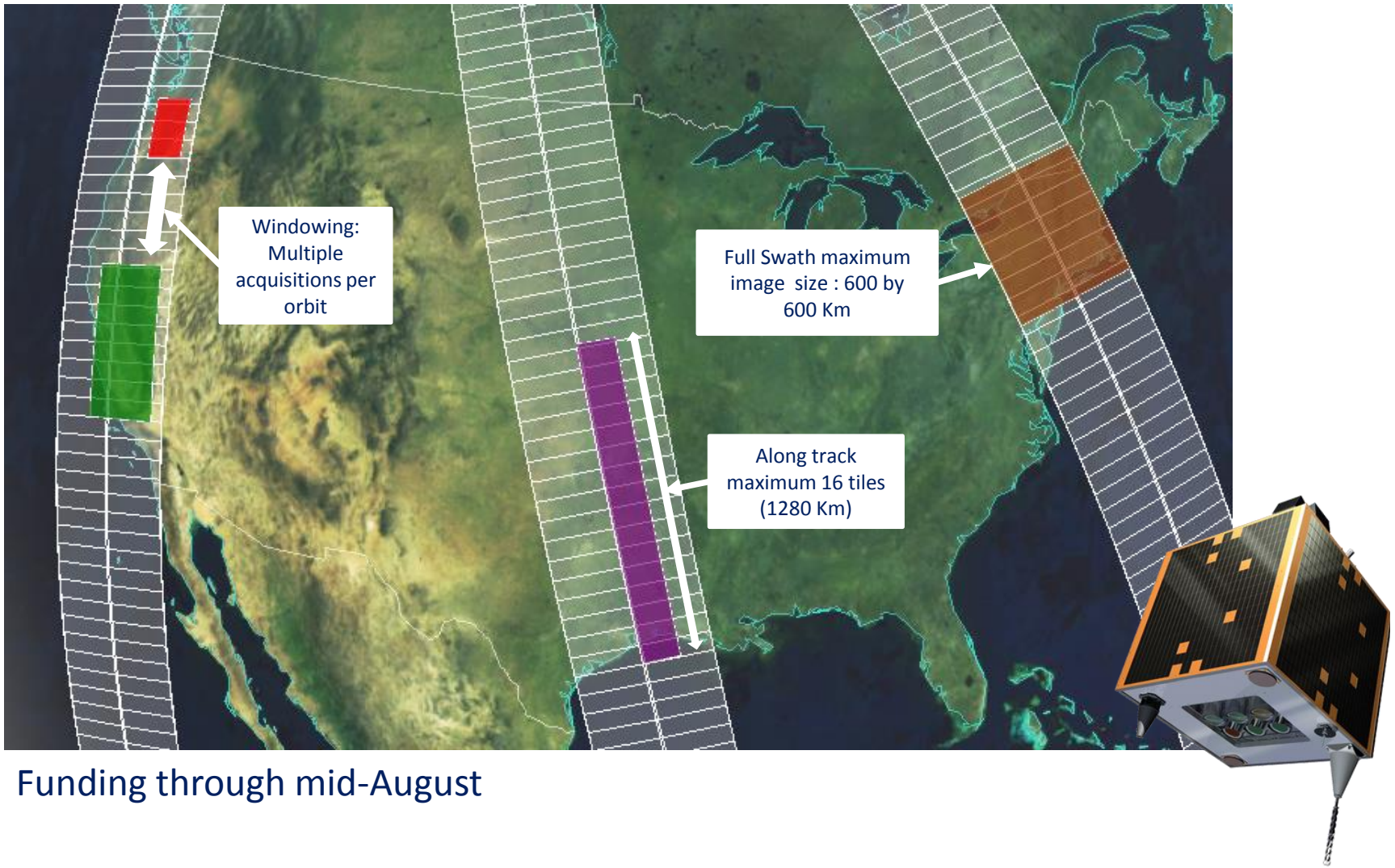
Farm Service Agency: Common Land Unit



2011 NLCD & Derivative products

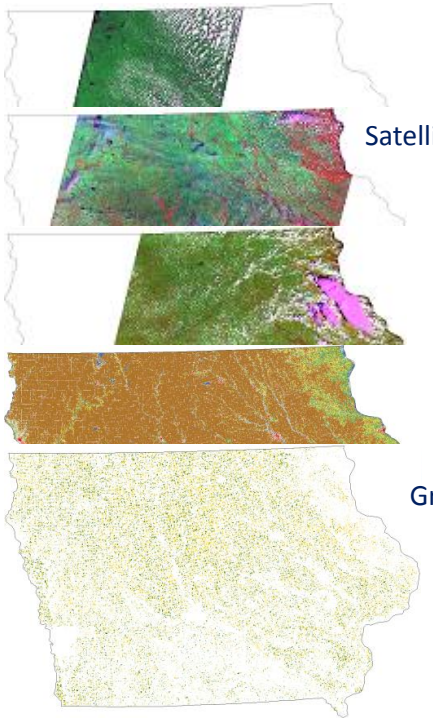


2014 Deimos-1/UK2 Satellite Tasking



Funding through mid-August

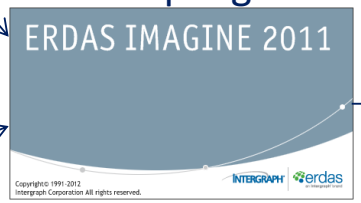
Processing a CDL



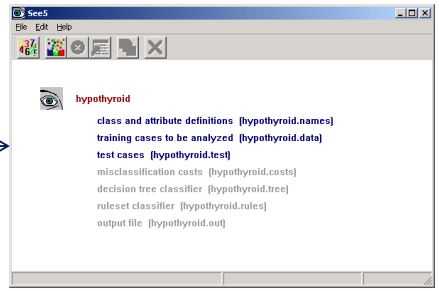
Satellite Imagery

Ground Truth

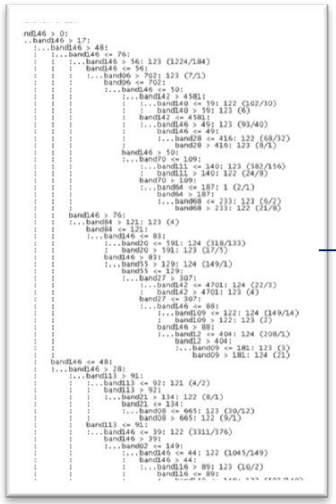
Sampling



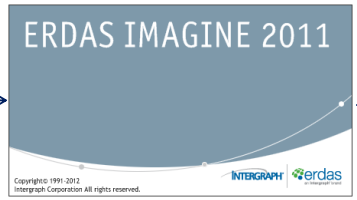
See5



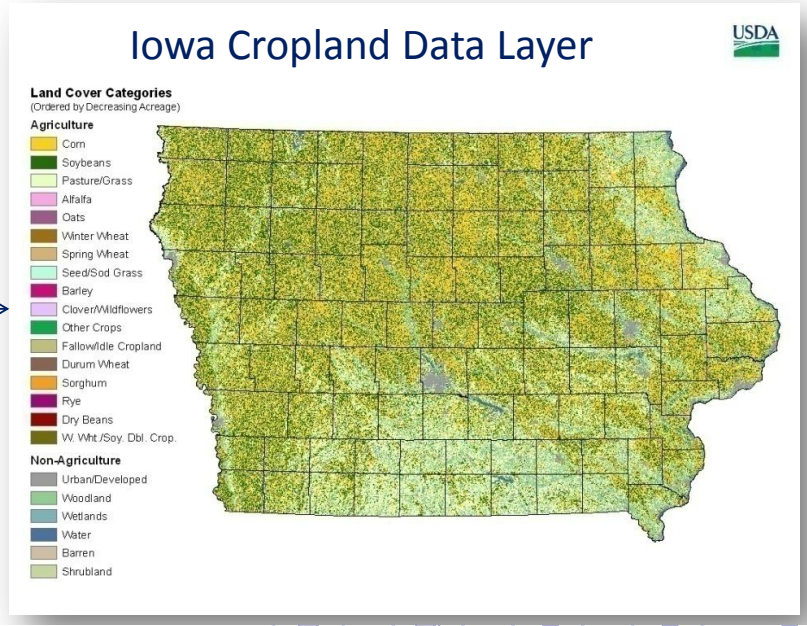
Decision Tree



Classification

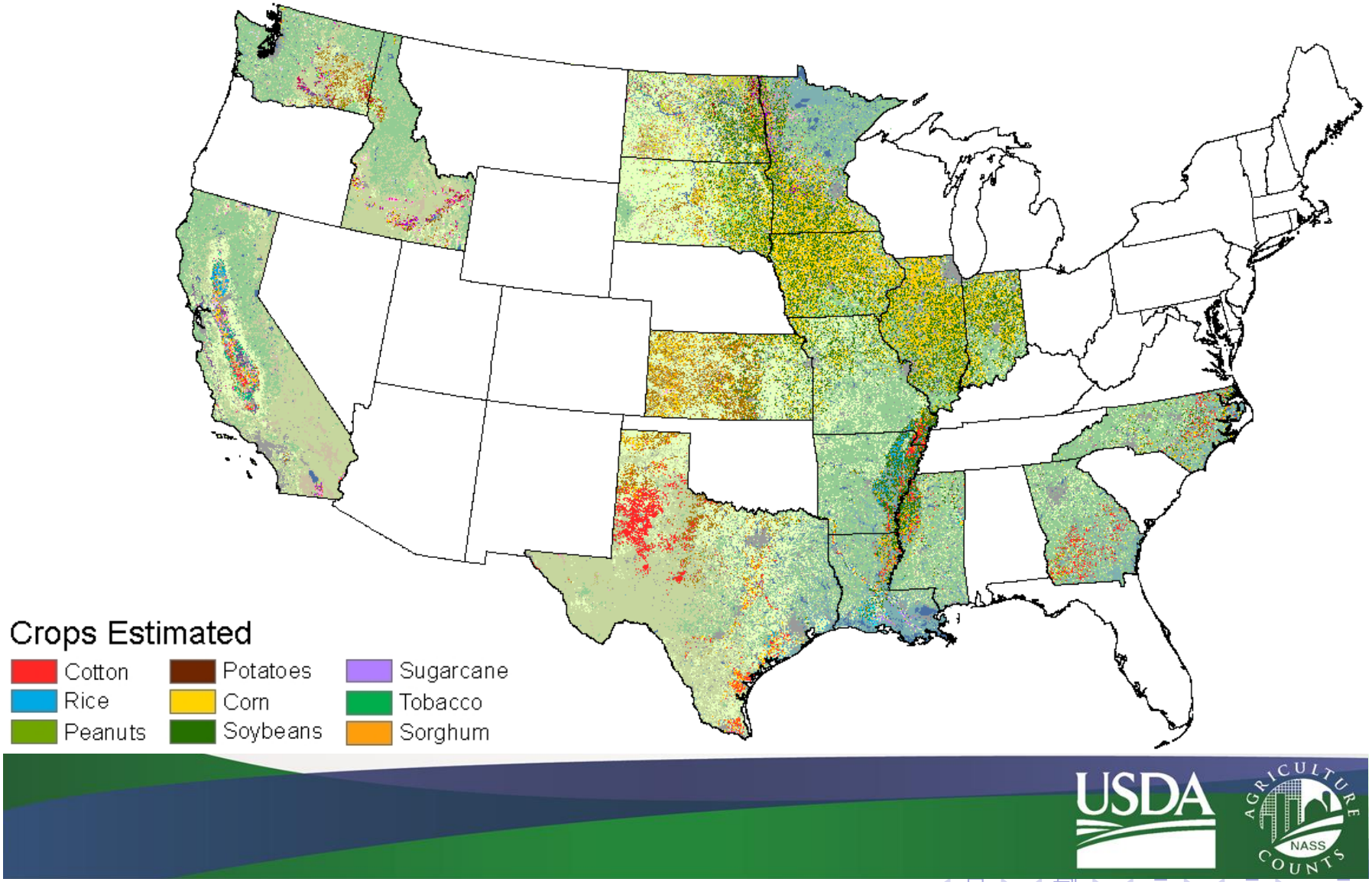


Iowa Cropland Data Layer



September

17 States Classified
9 Crops Estimated
Imagery from April - August



Problem 2: Vehicle Probe Project (VPP) BIGDATA

- Original goal: to enable a wide-variety of transportation operations and planning applications that require a high-quality data source.
- Applications include congestion management systems, traveler information systems, travel-time on changeable message signs.
- Data contains travel time, speed, historic speed, etc. for different road segments
- If data for a whole year, for all 12,295 TMC segments in Maryland were to be downloaded, the estimated number of records is 6.46 billion. The physical disk size of this data is estimated to be 375GB.

FIGURE: Location of NJ11-0009 segment in New Jersey, near Philadelphia.

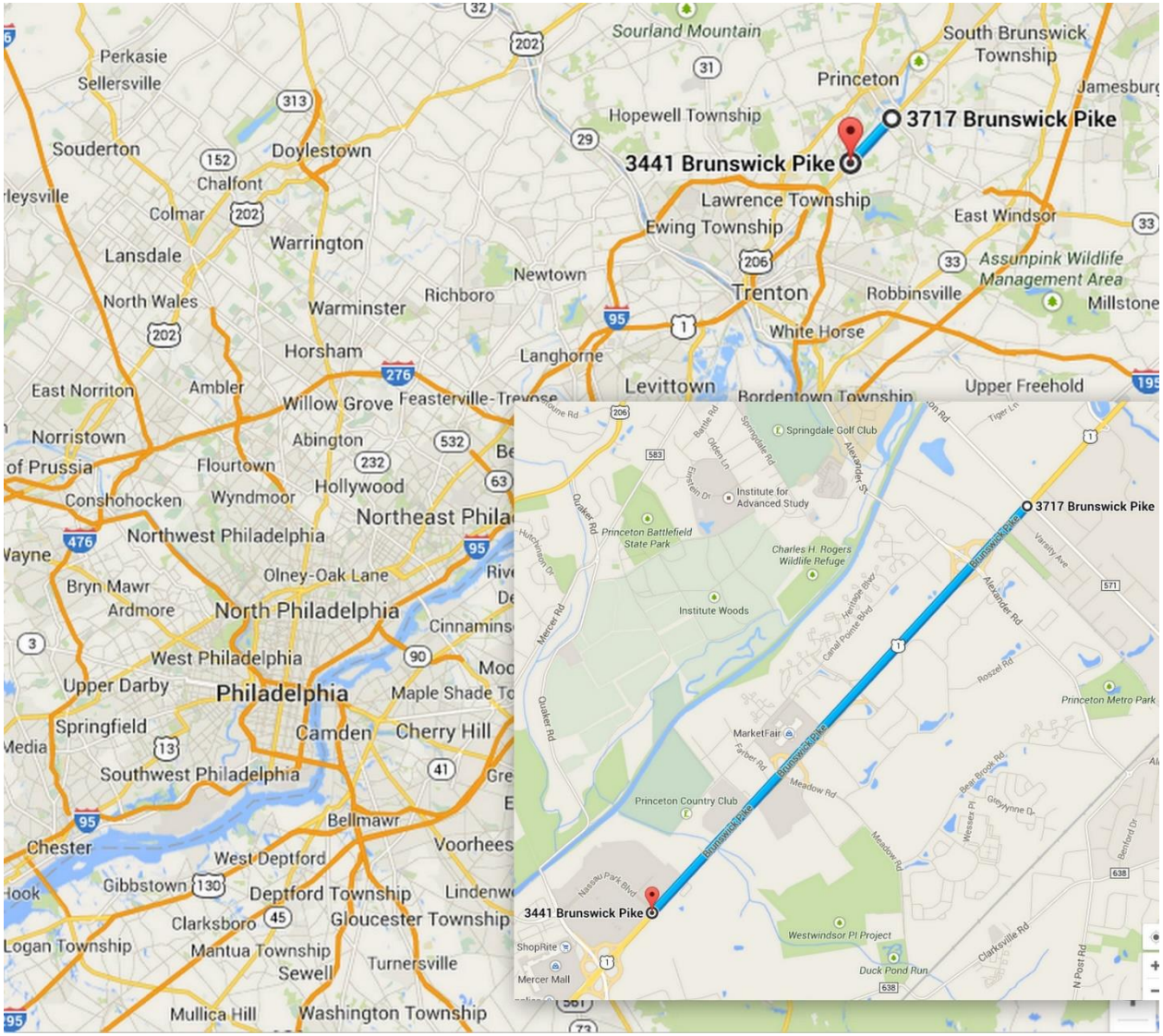
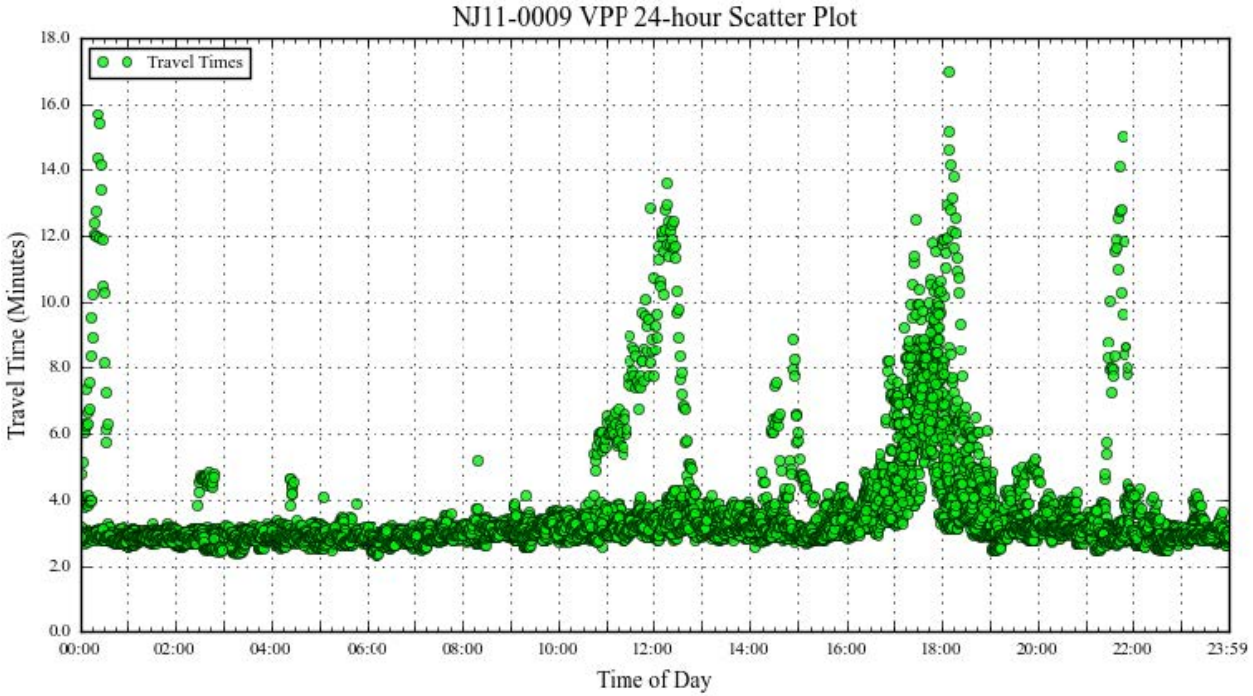


FIGURE VPP travel time data



Problem 3: BIGDATA from Administrative Records

- Internal Revenue Service Data
- Supplemental Nutrition Assistance Program (SNAP) data

Some features of BIGDATA

- May not contain the variable(s) of interest
- errors due to measurement, classification, self selection, etc.
- massive complex data for local area
- computational issue

How do we correct Big Data?

Look for existing sample data or conduct a new survey

Some features of sample surveys

- Finite populations
- Representativeness
- Large samples for large areas, but small or no sample for small areas
- Variable(s) of interest can be included
- Chance selection: equal/epsem
- Stratification to improve precision and administrative control

Sample Survey Data

- Problem 1: June Enumerative Survey
- Problem 2: National Household Travel Survey (NHTS) and American Community Survey (ACS)
- Problem 3: ACS

How do we combine Big Data with Sample Survey Data?

Two Cases:

- **Case 1:** No or little overlap between the two data sources
- **Case 2:** Most of the survey data can be linked with Big Data

Case 1: Statistical Matching

Small Area Level Model

Ref: Fay and Herriot (JASA 1979)

For $i = 1, \cdots, m$,

$$\textit{Level 1: (Sampling Distribution)}: \; y_i|\theta_i \sim N(\theta_i, \psi_i);$$
$$\textit{Level 2: (Prior Distribution)}: \; \theta_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, A)$$

where

- m : number of small area;
- y_i : direct survey estimate of θ_i;
- θ_i : true mean for area i;
- $\boldsymbol{x}_i$: $p \times 1$ vector of known auxiliary variables;
- ψ_i: known sampling variance of the direct estimate;
- The $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ and model variance A are unknown.

Estimation Method

Parameter of Interest: θ_i

Inferences based on the posterior distribution of θ_i:

$$\theta_i|y;\beta,A \overset{ind}{\sim} N(\hat{\theta}_i^B,\ \sigma_i^2(A)),$$

where

- $\hat{\theta}_i^B = (1-B_i)y_i + B_i\boldsymbol{x}_i'\boldsymbol{\beta}$
- $B_i = \frac{\psi_i}{A+\psi_i}$
- $\sigma_i^2(A) = (1-B_i)\psi_i$

EB: Treat β and A fixed and estimate them by consistent estimators (e.g., ANOVA, ML, REML, adjusted ML)

HB: Put priors, possible non-informative flat priors, on β and A. The inference is based on the posterior distribution of the target parameter.

The James-Stein Estimator

$$\hat{\theta}_i^{JS} = (1 - \hat{B}_{JS})y_i, \text{ where } \hat{B}_{JS} = \frac{m-2}{\sum_{j=1}^{m} y_j^2}.$$

Results:

- Total MSE (TMSE) of direct estimator: $\sum_{j=1}^{m} E[(y_i - \theta_i)^2|\theta] = m$
- TMSE of JS estimator: $\sum_{j=1}^{m} E[(\hat{\theta}_i^{JS} - \theta_i)^2|\theta] \leq m - \frac{(m-2)^2}{m-2+\sum_i \theta_i^2}$. (Efron)

Remarks:

- If $\theta_i = 0,\ (i = 1, \cdots, m)$, then TMSE of JS$\leq [m - (m-2)] = 2$. Thus, the largest reduction is obtained when $\theta_i = 0\ (i = 1, \cdots, m)$ and m large.
- If any $|y_j| \to \infty$, the JS converges to the direct.

Measurement Error Issue in Big Data

Two Situations:

- **Situation 1:** The sources of measurement error can be reasonably identified and we have enough data to explain them.
- **Situation 2:** The sources cannot be easily detected or we do not have data to explain the measurement error even if the sources of error are identified.

Situation 1: An Example

$$\text{Level 1 (Sampling model):} \begin{pmatrix} y_i \\ \boldsymbol{x}_i \end{pmatrix} | \theta_i, \boldsymbol{X}_i \overset{ind}{\sim} N\left(\begin{pmatrix} \theta_i \\ \boldsymbol{X}_i \end{pmatrix}, \left(\begin{array}{c|c} \psi_{iy} & 0 \\ \hline 0 & \boldsymbol{\Psi}_{ix} \end{array} \right) \right)$$

$$\text{Level 2 (Linking model):} \quad \theta_i | \boldsymbol{X}_i \overset{ind}{\sim} N(\boldsymbol{X}_i' \boldsymbol{\beta}, A)$$

Remark: The above model reduces to the FH model when $\Psi = \mathbf{0}$.

The Bayes estimator of θ_i under FH:

$$\hat{\theta}_i^B = (1 - B_i) y_i + B_i \boldsymbol{x}_i' \boldsymbol{\beta},$$

where

$$B_i = \frac{\psi_{iy}}{A + \psi_{iy}}$$

The Bayes estimator of θ_i under FH with ME:

$$\hat{\theta}_i^{B*} = (1 - B_i^*) y_i + B_i^* \boldsymbol{x}_i' \boldsymbol{\beta},$$

where

$$B_i^* = \frac{\psi_i}{A + \psi_i + \boldsymbol{\beta}' \boldsymbol{\Psi}_{ix} \boldsymbol{\beta}}$$

Remarks:

Under the FH-ME,

$$\text{MSE}(\hat{\theta}_i^B) = (1 - B_i)\psi_{iy} + B_i^2 \boldsymbol{\beta}' \boldsymbol{\Psi}_{ix} \boldsymbol{\beta},$$

which is greater than ψ_{iy} if $\boldsymbol{\beta}' \boldsymbol{\Psi}_{ix} \boldsymbol{\beta} > A + \psi_{iy}$ but

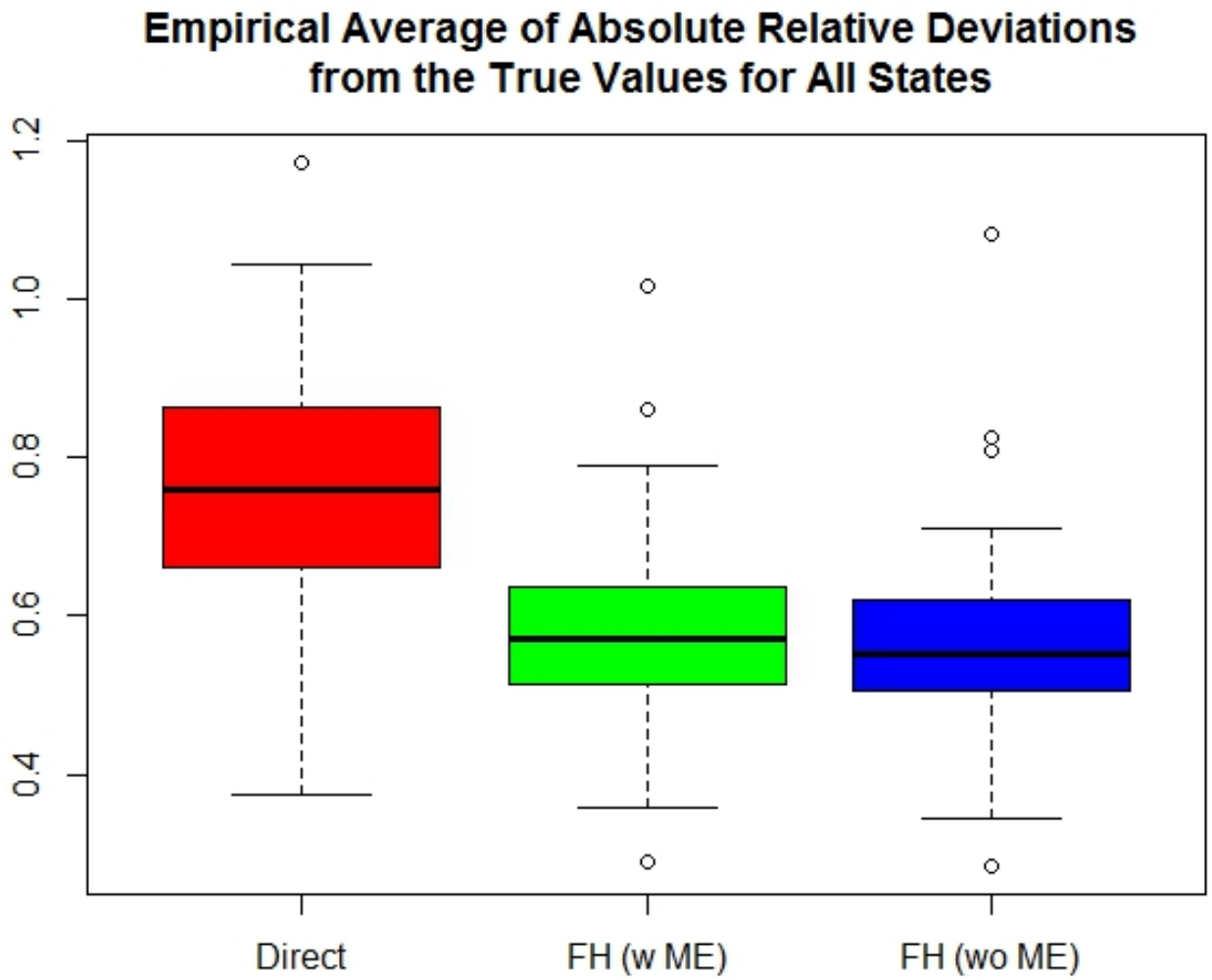
$$\text{MSE}(\hat{\theta}_i^B) = (1 - B_i^*)\psi_{iy} < \psi_{iy}$$

An Illustration

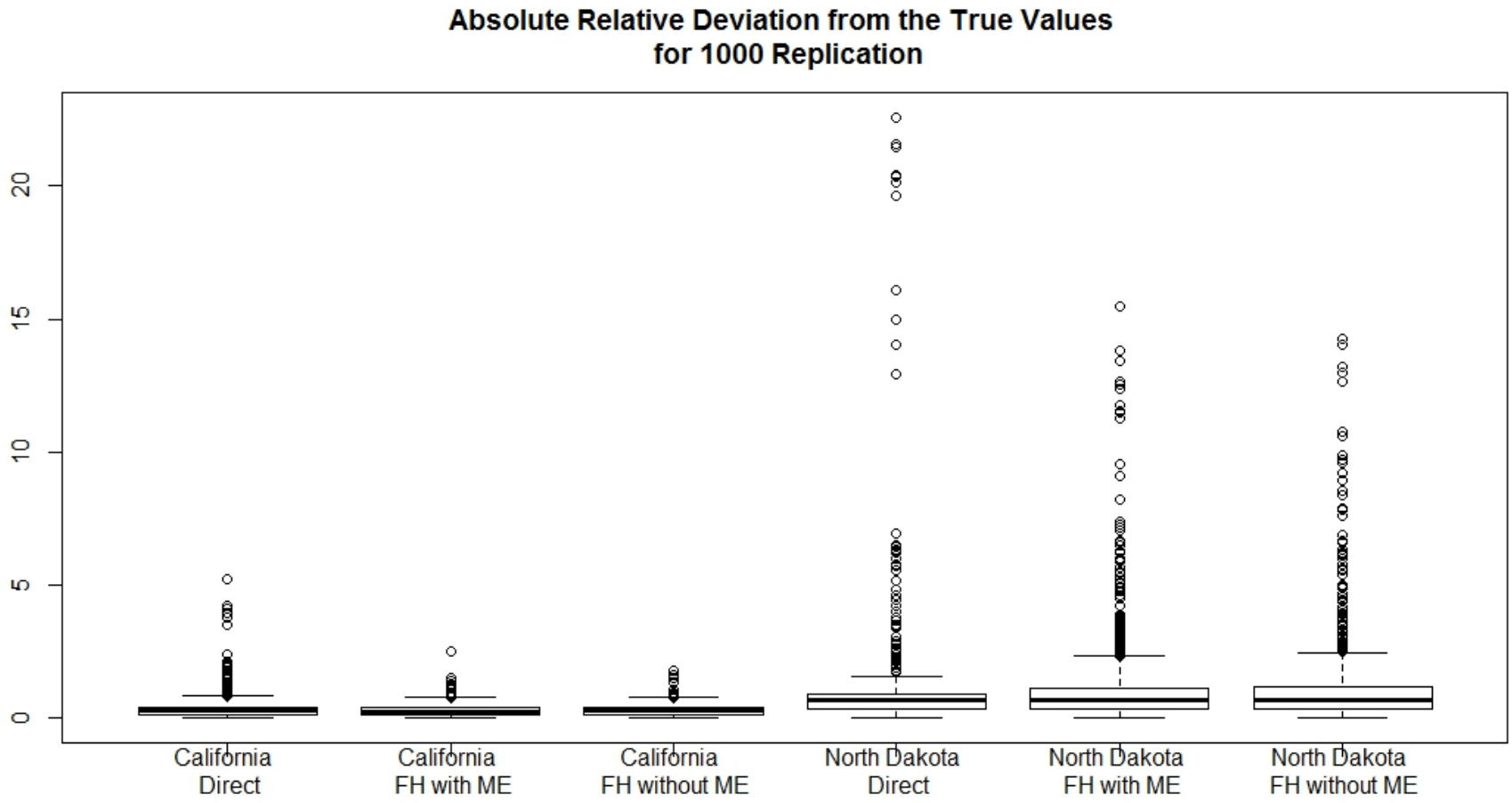
Mosaferi, S. (2015)

- Parameter of interest: Average number of full-time employment for 49 states of the U.S.
- Big Data: American Community Survey (PUMS micro data used)
- Survey Data: Annual Survey of Employment and Payroll (ASPEP)
- FH-ME model in the logarithmic scale.
- Design-based evaluation using Census of Governments

Mosaferi 2015



Mosaferi 2015



Situation 2: A partial Solution

Ref: Datta and Lahiri (1995, JMVA)

An Outlier Resistent Model For $i = 1, \cdots, m$,

$$\textit{Level 1: (Sampling Distribution)}: \; y_i|\theta \overset{ind}{\sim} N(\theta_i, \; \psi_i);$$

$$\textit{Level 2: (Prior Distribution)}: \; \theta_i|\boldsymbol{\beta}, \; A \overset{ind}{\sim} \frac{1}{\sqrt{A}} p_i \left(\frac{\theta_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sqrt{A}} \right)$$

where $p_i(x) = \int_0^\infty r^{1/2} \psi(x r^{1/2}) g_i(r) dr$, $\phi(x)$ being the pdf of a standard normal distribution.

To retain shrinking in presence of an outlier in residual, use a heavy tail distribution (e.g., Cauchy) for the mixing distribution $g_i()$

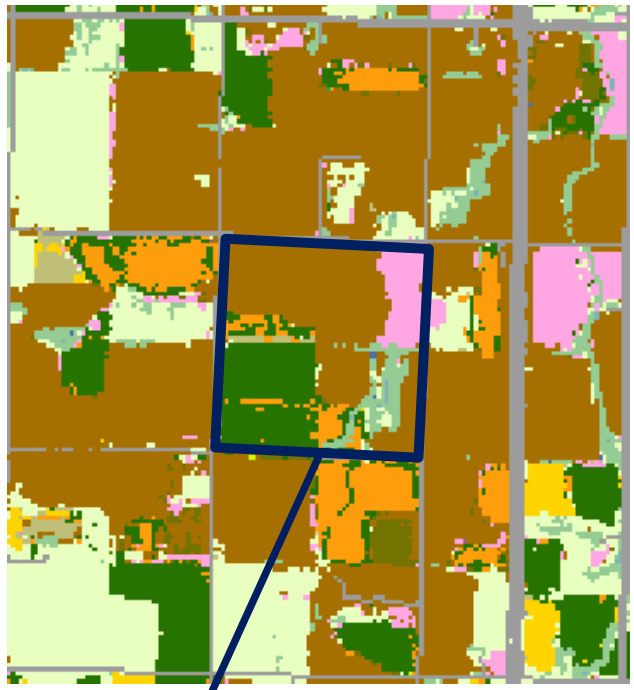
Case 2: Record Linkage

PAGE 2 **SECTION D - CROPS AND LAND USE ON TRACT** 17

How many acres are inside this blue tract boundary drawn on the photo (map)? .

Now I would like to ask about each field inside this blue tract boundary and its use during 2000.

FIELD NUMBER		01	02	03	04	05
1. Total acres in field		828	828	828	828	828
2. Crop or land use. [Specify]						
3. Occupied farmstead or dwelling		843				
4. Waste, unoccupied dwellings, buildings and structures, roads, ditches, etc.		---	---	---	---	---
5. Woodland		831	831	831	831	831
6. Pasture	Permanent (not in crop rotation)	842	842	842	842	842
	Cropland (used only for pasture)	856	856	856	856	856
8. Idle cropland - Idle all during 2000		857	857	857	857	857
9. Two crops planted in this field or two uses of the same crop.		☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No
	[Specify second crop or use]					
	Acres	844	844	844	844	844
10. Acres left to be planted		610	610	610	610	610
11. Acres irrigated and to be irrigated [If double cropped, include acreage of each crop irrigated]		620	620	620	620	620
16. **Winter Wheat** (include cover crop)	Planted	540	540	540	540	540
17.	For grain or seed	541	541	541	541	541
18. **Rye** (include cover crop) [Exclude ryegrass]	Planted	547	547	547	547	547
19.	For grain or seed	548	548	548	548	548



REGRESSION VARIABLES:

Dependent Y

Independent X

	Enumerated JAS Segments	CDL Classified Acres
Soybeans	227	273
Wheat	337	541

Battese, Harter and Fuller (1988 JASA)

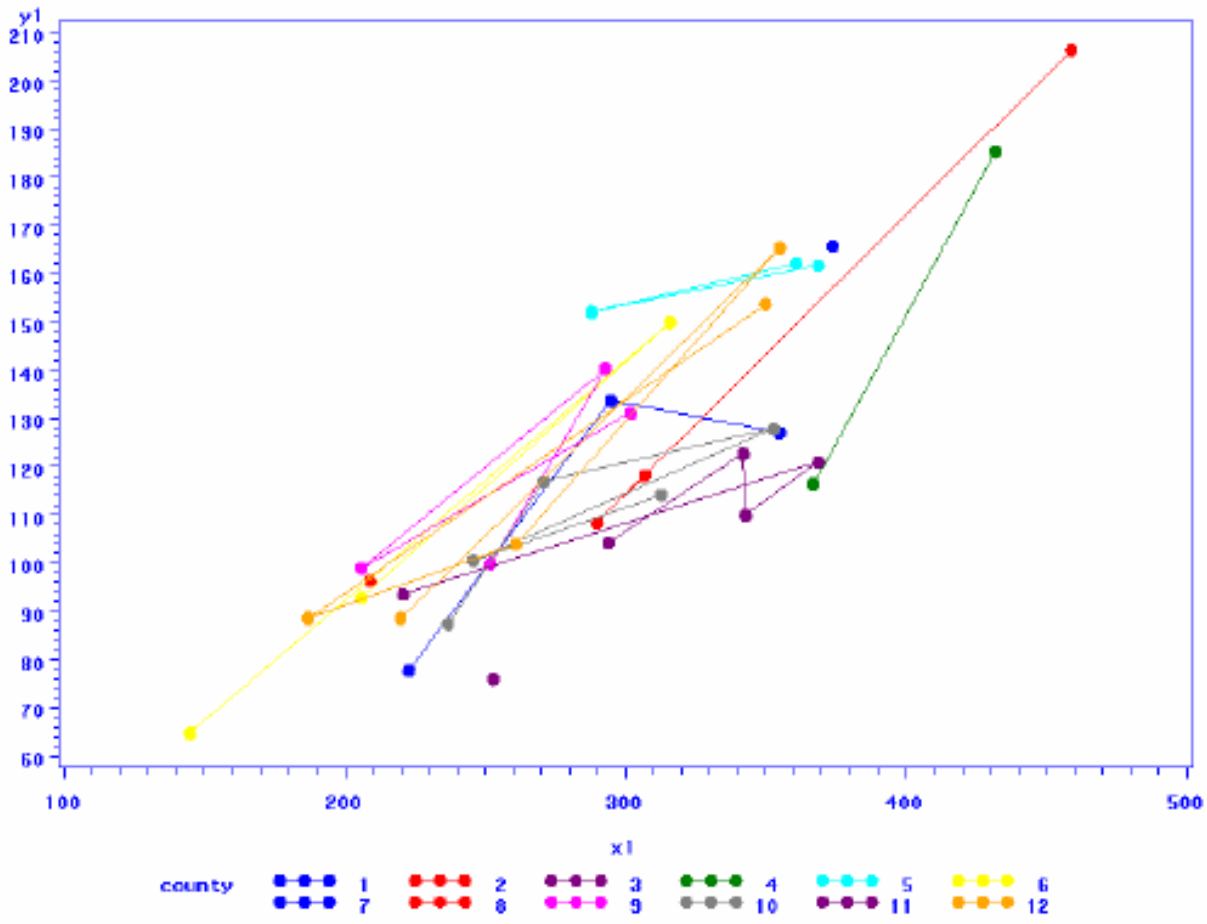
Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
County	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	169.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	216	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			106.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

* The mean number of pixels of a given crop per segment in a county is the total number of pixels classified as that crop, divided by the number of segments in that county.

How to make BIGDATA useful?

Fig 2: Plot of Corn Hectares versus Corn Pixels by County



This plot also reflects the strong relationship between the reported hectares of corn and the number of pixels of corn for counties separately. But the slopes and/or intercepts

How do we combine information?

- y_{ij} : value of the study variable for the jth unit of the i small area population $(i = 1, \cdots, m; \;\; j = 1, \cdots, N_i)$
- We are interested in estimating the finite population means:

$$\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}.$$

Nested Error Regression Model

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij},$$

where x_{ij} is a $p \times 1$ column vector of known auxiliary variables; $\{v_i\}$ and $\{e_{ij}\}$ are all independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$

An Example

- Estimation of the number of hectares of corn for 12 Iowa counties based on the 1978 June Enumerative Survey and satellite data.
- y_{ij}: the number of hectares of corn in the jth segment of the ith county as reported in the June Enumerative Survey.
- $x_{ij}' = (1, x_{1ij}, x_{2ij})$, where x_{1ij} (x_{2ij}) is the number of *pixels* classified as corn (soybean) in the jth segment of the ith county.
- $\bar{X}' = (1, \bar{X}_{1i}, \bar{X}_{2i})$, where $\bar{X}_{1i}$ $(\bar{X}_{2i})$ is the mean number of pixels per segment classified as corn (soybean) for county i.

Unit Level Model with Big Data

Gershunskaya and Lahiri 2011

Model:

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij},$$

where

- $v_i \overset{iid}{\sim} N(0, \tau^2)$
- $e_{ij} \overset{iid}{\sim} (1 - z_{ij})N(0, \sigma_1^2) + z_{ij}N(0, \sigma_2^2)$
- z_{ij} is the mixture part indicator random variable with

$$z_{ij}|\pi \overset{iid}{\sim} \text{Bin}(1, \pi)$$

"...D.J. Finney once wrote about the statistician whose client comes in and says, "Here is my mountain of trash. Find the gems that lie therein." Finney's advice was to not throw him out of the office but to attempt to find out what he considers "gems". After all, if the trained statistician does not help, he will find some one who will...." David Salsburg, ASA Connect Discussion

THANK YOU!