Applications of High Resolution Soil Data for Crop Insurance Rating and Yield Distribution Estimation

Soil Renaissance, Economics of Soil Health Workshop
September 21-22, 2015

Joshua D. Woodard, PhD
Assistant Professor
Zaitz Faculty Fellow in Agribusiness and Finance
Dyson School of Applied Economics and Management
Cornell University

Soil Data and Insurance

- The advent of "Big Data" in agriculture has led to increased interest in large-scale empirical applications which previously were impractical or impossible
- Scalability component has brought to light new policy options
- Currently, soil data are not explicitly taken into account when estimating insurance premium rates for Federal Crop Insurance Program (FCIP), but have recently begun collection.
- In the U.S., the Federal Crop Insurance Program now serves as the cornerstone agricultural policy and mode of subsidization, to the tune of around $10 billion in expected costs annually.
- High resolution soil and other data are available from different agencies, just currently not used for rating/pricing
- RMA and began collecting data matched to individual contiguous farming parcels (FSA Common Land Units) since around 2009 (100% target reporting by 2016)

Motivation, Purpose, Objectives

- Can available soil data be used to refine crop yield distribution estimates, and rates? If so, why do that?
- Investigate the feasibility of using high resolution soil data in the modeling of field and farm level yields using some major high availability datasets, namely, the gSURRGO soil dataset from NRCS.
- Employ Common Land Unit field maps in the public domain (last released in 2008) to investigate degree of rate differentials implied and compare to what is currently captured in rates
- Proof of concept and motivation for RMA Rating modification
- Scalability and operational considerations
- *Expecting perfect rules and rates from RMA always through time is probably not a realistic (or even fair) expectation. Nonetheless, it is important to keep working toward that target.*

Contributions and Implications

- Several studies highlight importance of intra-regional variability (e.g., Woodard, 2014; Claassen and Just, 2011; Lobell, Ortiz-Monasterio, and Falcon, 2007; Popp, Rudstrom, and Manning, 2005; Rudstrom et al., 2002, among others)
- Very little in terms of actually linking soil data back to field level yield data explicitly on large scale, nor evaluation of insurance implications
- Implications and Applications
  - Conservation, adverse selection, adverse incentive, and insurance
  - TransitionalYield determination FCIP
  - APH determination under FCIP
  - Rating in areas with less experience data or for new products

Why focus on operationalizing large scale, high resolution yield distribution estimation?

- Reliable and scalable frameworks for estimating yield distributions critical prerequisite for being able to design and implement actual policies to reflect technology advancements in conservation
- Yield distribution encompasses or formalizes characteristics like:
  - Expected yield, and yield potential
  - Variability in yields, or yield risk more generally
  - How probabilities of different yield outcomes change under different conditions, practices etc.
  - Insurance rates and expected losses

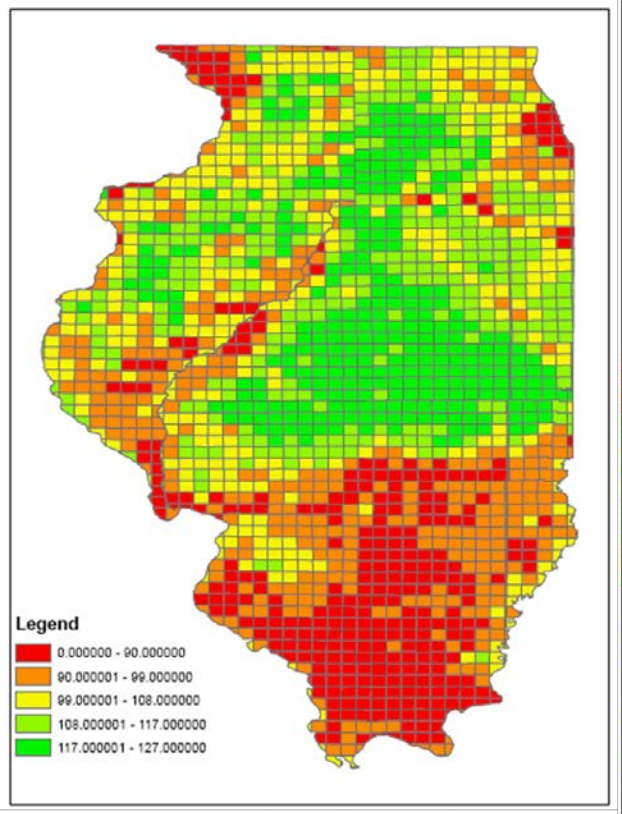
Why focus on operationalizing large scale, high resolution yield distribution estimation?

- Must have foundation upon which to build before scaling
  - What does yield distribution look like already?
  - If I adopt a new practice, how does that change soil over time, and by implication the conditional yield distribution?
  - How does that change expectation of insurance losses, distribution of profits, etc?
  - If this is linkable to scalable data and models, now can operationalize things like changes to insurance rates changes or rules changes

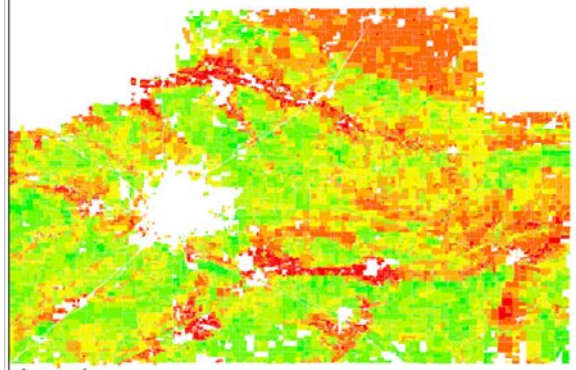
Data and Methods

- 1) IL FBFM Farm Level Data (Standard FE OLS and Cond. Weibull FE)
  - Yield and Acreage (farms with 20+ years of data)
  - Attached Soil Productivity Ratings using IL Soil Bulletin 810/Circular 1156 soil type designations
  - The CLU shapefiles were used in conjunction with SURRGO soil to construct IL Soil Bulletin 810/Circular 1156 ratings as with FBFM data
- 2) National County Level Analysis with Gridded SURRGO data; explore using richer set of soil factors (57 published attributes in Valu1 table from NRCS)
- 3) Crop Insurance Rating Analysis and RMA System Comparison
- Weather data from PRISM database and PDSI data from NCDC
- Cropland data layers
- *Missing link to facilitate national scalability: RMA Yield Data in APH Databases linked to CLU (more on this later)*

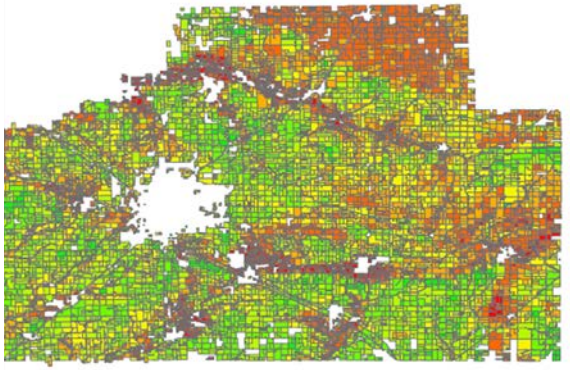
IL Soil Productivity Ratings (810 Average)



McLean County, IL SPR 810 Avg by Common Land Unit (CLU)



McLean County, IL Corn Expected Yieldby Common Land Unit (CLU), 2009 Base Year



FBFM IL Farm Yield Regression Results Models (County Fixed Effects and Trends)

Variables	Model 1	Model 2	Model 3	Model 4
SOIL	1.166***	1.082***	1.068***	3.944***
	(0.013)	*(0.211)*	*(0.204)*	*(0.275)*
SOIL²		0.000	0.000	-0.001
		(0.001)	*(0.001)*	*(0.001)*
TEMP	-11.328***	-11.328***	66.431***	81.941***
	(0.077)	*(0.077)*	*(1.592)*	*(1.950)*
TEMP²			-1.674***	-1.766***
			(0.035)	*(0.036)*
PREC	0.056***	0.056***	0.326***	0.440***
	(0.001)	*(0.001)*	*(0.004)*	*(0.010)*
PREC²			0.000	0.000
			(0.000)	*(0.000)*
SOIL x TEMP				-0.103***
				(0.007)
SOIL x PREC				-0.001
				(0.000)
N	121416	121416	121416	121416
Adj. R^2	0.536	0.536	0.564	0.565
σ^2	578.453	578.457	544.053	542.752

Note: TREND and INTERCEPT terms are shown at their means.

FBFM IL Standard Deviation Models

Variables	Model 1	Model 2
SOIL	-0.117***	0.809***
	(0.013)	*(0.216)*
SOIL^2		-0.004***
		(0.001)
INTERCEPT	37.802***	-11.834
	(1.827)	*(11.714)*
N	3861	3861
Adj. R^2	0.172	0.175
σ^2	19.222	19.135

Note: County Level Fixed Effects, TREND and INTERCEPT terms are shown at their means.

Additional Approaches: Conditional Weibull Distribution

- PDF:

$$f(y \mid a,b) = ba^{-b} y^{b-1} e^{-(y/a)^b}$$

- Conditional parameter models:

$$a(\mathbf{x}_a, \boldsymbol{\beta}_a) = g_a(\mathbf{x}_a, \boldsymbol{\beta}_a)$$

$$b(\mathbf{x}_b, \boldsymbol{\beta}_b) = g_b(\mathbf{x}_b, \boldsymbol{\beta}_b)$$

Why Weibull?

- Good evidence that corn yields exhibit left skewness
- Normal dist. tends to underestimate rates in this case
- Lower bound at zero (prevents negative yields)
- Conditional setup allows us to work with the underlying weather distribution
- Allows for modeling of differences in the response rates of yields to weather stressors over time
- Allows assessment of losses conditional on a *weather event*, or unconditional on weather (i.e., over all possible weather events) and under different distributions for weather

Conditional Weibull Distribution

- Conditional Mean:

$$\mu(y \mid a, b, \mathbf{x}, \boldsymbol{\beta}_a, \boldsymbol{\beta}_b) = \frac{a(\mathbf{x}, \boldsymbol{\beta}_a)}{b(\mathbf{x}, \boldsymbol{\beta}_b)} \Gamma\left(\frac{1}{b(\mathbf{x}, \boldsymbol{\beta}_b)}\right)$$

- Conditional Variance:

$$\sigma^2(y \mid a, b, \mathbf{x}, \boldsymbol{\beta}_a, \boldsymbol{\beta}_b) = a(\bullet)^2 \left(\Gamma\left(\frac{b(\bullet)+2}{b(\bullet)}\right) - \Gamma\left(\frac{b(\bullet)+1}{b(\bullet)}\right)^2 \right)$$

Conditional Weibull Distribution

- Conditional Mean Elasticity:

$$\varepsilon_k^{\mu} = \frac{\partial \ln(\mu(y \mid a, b, \mathbf{x}, \boldsymbol{\beta}_a, \boldsymbol{\beta}_b))}{\partial \ln(x_k)} = \frac{\partial \mu(\bullet)}{\partial x_k} \cdot \frac{x_k}{\mu(\bullet)}$$

$$= -\frac{\left(x_k\, a(\bullet)\Psi_0\left(\frac{1}{b(\bullet)}\right) + x_k\, a(\bullet) b(\bullet)\right)\left(\frac{\partial b(\bullet)}{\partial x_k}\right) - x_k\, b(\bullet)^2\left(\frac{\partial a(\bullet)}{\partial x_k}\right)}{a(\bullet) b(\bullet)^2}$$

Conditional Weibull Distribution

- Conditional Variance Elasticity:

$$\varepsilon_k^{\sigma^2} = \frac{\partial \ln(\sigma^2(\bullet))}{\partial \ln(x_k)} =$$

$$-\left\{\left(2x_k a(\bullet)\left(\frac{\partial b(\bullet)}{\partial x_k}\right)\Psi_0\left(\frac{b(\bullet)+2}{b(\bullet)}\right)-2x_k b(\bullet)^2\left(\frac{\partial a(\bullet)}{\partial x_k}\right)\right)\Gamma\left(\frac{b(\bullet)+2}{b(\bullet)}\right)\right.$$
$$\left.+\left(2x_k b(\bullet)^2\left(\frac{\partial}{\partial x_k}a(\bullet)\right)-2x_k a(\bullet)\left(\frac{\partial b(\bullet)}{\partial x_k}\right)\Psi_0\left(\frac{b(\bullet)+1}{b(\bullet)}\right)\right)\Gamma\left(\frac{b(\bullet)+1}{b(\bullet)}\right)^2\right\}$$

$$\times\frac{1}{a(\bullet)b(\bullet)^2\Gamma\left(\frac{b(\bullet)+2}{b(\bullet)}\right)-a(\bullet)b(\bullet)^2\Gamma\left(\frac{b(\bullet)+1}{b(\bullet)}\right)^2}$$

Conditional Weibull Distribution: Recovery of Unconditional

$$f(y \mid \hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b, g(\bullet))_{\breve{\mathbf{x}}} = \int \left[f(y \mid \breve{\mathbf{x}}, \hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b, g(\bullet)) \cdot g(\breve{\mathbf{x}}) \right] \mathbf{d}\breve{\mathbf{x}}$$

where, $\hat{\boldsymbol{\beta}}_a$ and $\hat{\boldsymbol{\beta}}_b$ estimated parameters, and $g(\breve{\mathbf{x}})$ is the joint distribution of $\breve{\mathbf{x}}$, which may or may not be the same as the distribution implied by the data $\mathbf{x}$ used to fit the model.

Cond. Weibull Analysis-Data and Methods

- Producer-level Corn yield data from the Illinois Farm Business Farm Management database (FBFM)
- Period:1972-2008.
- 30,467 yield observations from 5 contiguous counties
- Variables:
  - *TREND:* Time trend to proxy for technology gains
  - *ACRE*: Acreage
  - *SOIL*: Soil productivity
  - *WEATHER:* Palmer Drought Severity Index (PDSI)

Data and Methods (cont.)

- Quadratic specification for conditional parameter models

$$\mathbf{x} = [TREND, LN(ACRE), LN(SOIL), WEATHER]$$

- Recovery of Unconditional:
  - Evaluate *SOIL* and *ACRE* at medians
  - *WEATHER* distribution is estimated for two periods: 1895-2009 and 1980-2009
- Parameters estimated via Maximum Likelihood
- Bootstrap employed to generate standard errors for elasticities

Parameter Estimates for Conditional Weibull Model

	Weibull Model Parameters	
Conditioning Variable	β_a	β_b
INTERCEPT	61.311***	191.683***
TREND	-3.011***	-1.354***
LN(ACRE)	3.578	-3.987***
LN(SOIL)	-60.798***	-82.997***
WEATHER	36.648***	-2.742***
$TREND^2$	0.064***	0.011***
*TREND * LN(ACRE)*	-0.124***	0.006
*TREND * LN(SOIL)*	0.645***	0.230***
TREND * WEATHER	**0.084*** **	**0.029*** **
$LN(ACRE)^2$	0.290**	-0.080***
*LN(ACRE) * LN(SOIL)*	0.065	1.196***
*LN(ACRE) * WEATHER*	-0.036	0.095***
$LN(SOIL)^2$	16.011***	9.097***
*LN(SOIL) * WEATHER*	-7.351***	0.493***
$WEATHER^2$	-1.633***	-0.060***

Weather Unconditional Production Elasticities, 2008 Technology, (1895-2009 PDSI Distribution)

Risk Measure	*TREND*	*ACRE*	*SOIL*
E(Y)	0.8293***	0.0829***	2.5924***
	(0.0168)	(0.0145)	(0.0684)
σ(Y)	**-0.3615*** **	-0.1375	-2.7435***
	(0.0986)	(0.1041)	(0.2697)
σ(Y) / E(Y)	**-1.1810*** **	-0.2202**	-5.2010***
	(0.1037)	(0.1104)	(0.2798)

Rating and Distribution Analysis

- Using estimated conditional distribution models, downscale to estimate CLU specific distributions and generate implied insurance rates.
- Generate RMA rates by APH (approx. as expected yield) for each CLU using published rating methodology (publicly available Web API available online at agfinance.dyson.cornell.edu)
- Compare rate differentiation across rating method which takes into account soil explicitly, with RMA system
- Filter CLU's by recent Cropland Data Layer to focus only on Corn
- Models are estimated using entire state data, but for exposition here, will focus on a major county, McLean, IL

Actuarial Rate Impact Analysis

- Expected Loss Cost Ratio (Rate):

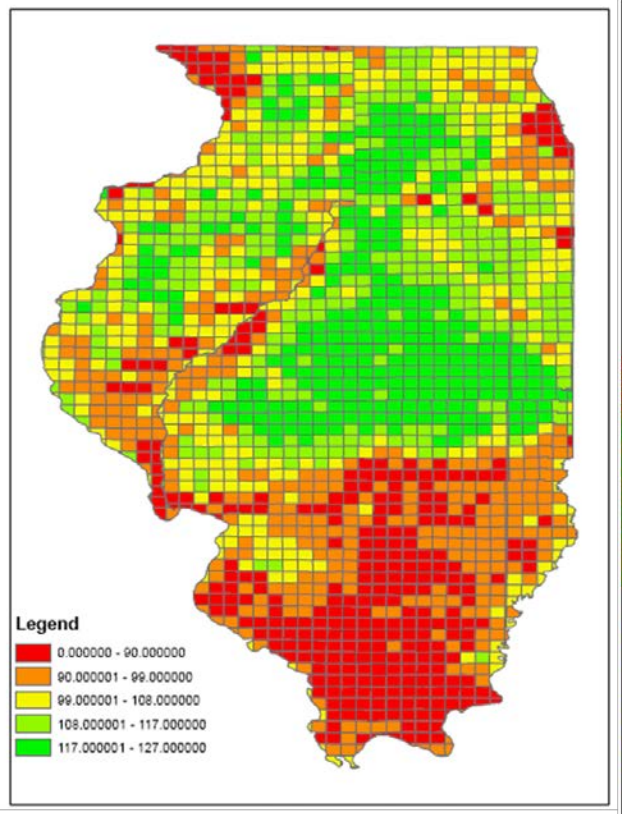
$$E(LCR) = \int_0^{E(Y)\cdot Cov} Max(0, E(Y)\cdot Cov - y)\cdot f(y)dy \Big/ \left(E(Y)\cdot Cov\right)$$

$E(Y)$: Expected yield

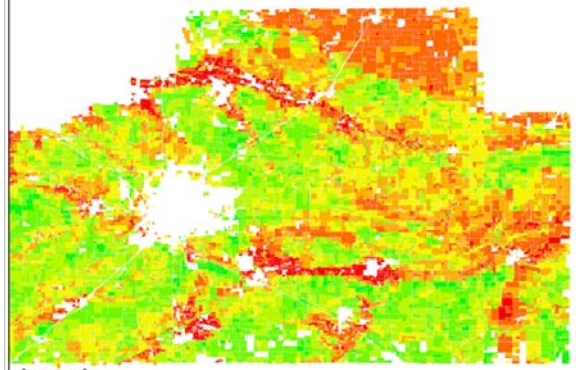
Cov : Coverage level

$f(y)$: Yield distribution

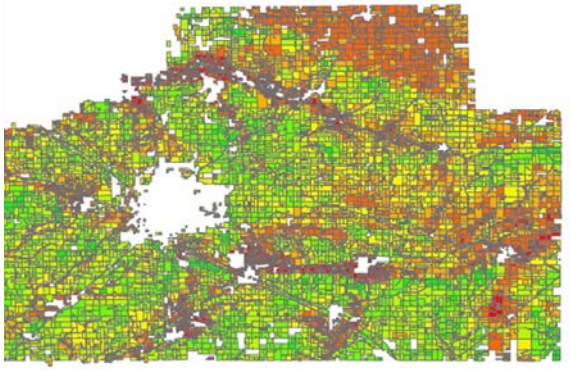
IL Soil Productivity Ratings (810 Average)

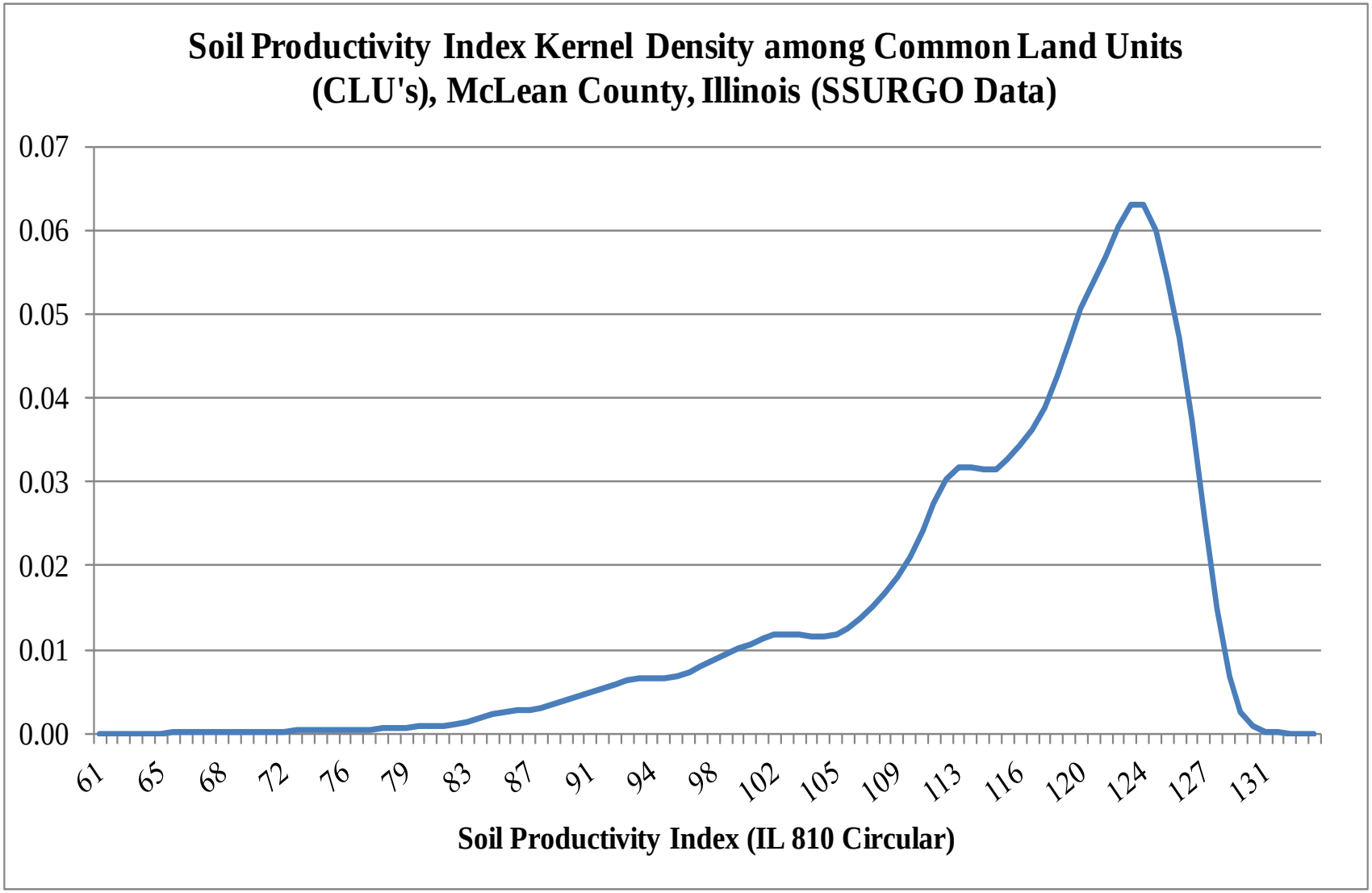


McLean County, IL SPR 810 Avg by Common Land Unit (CLU)

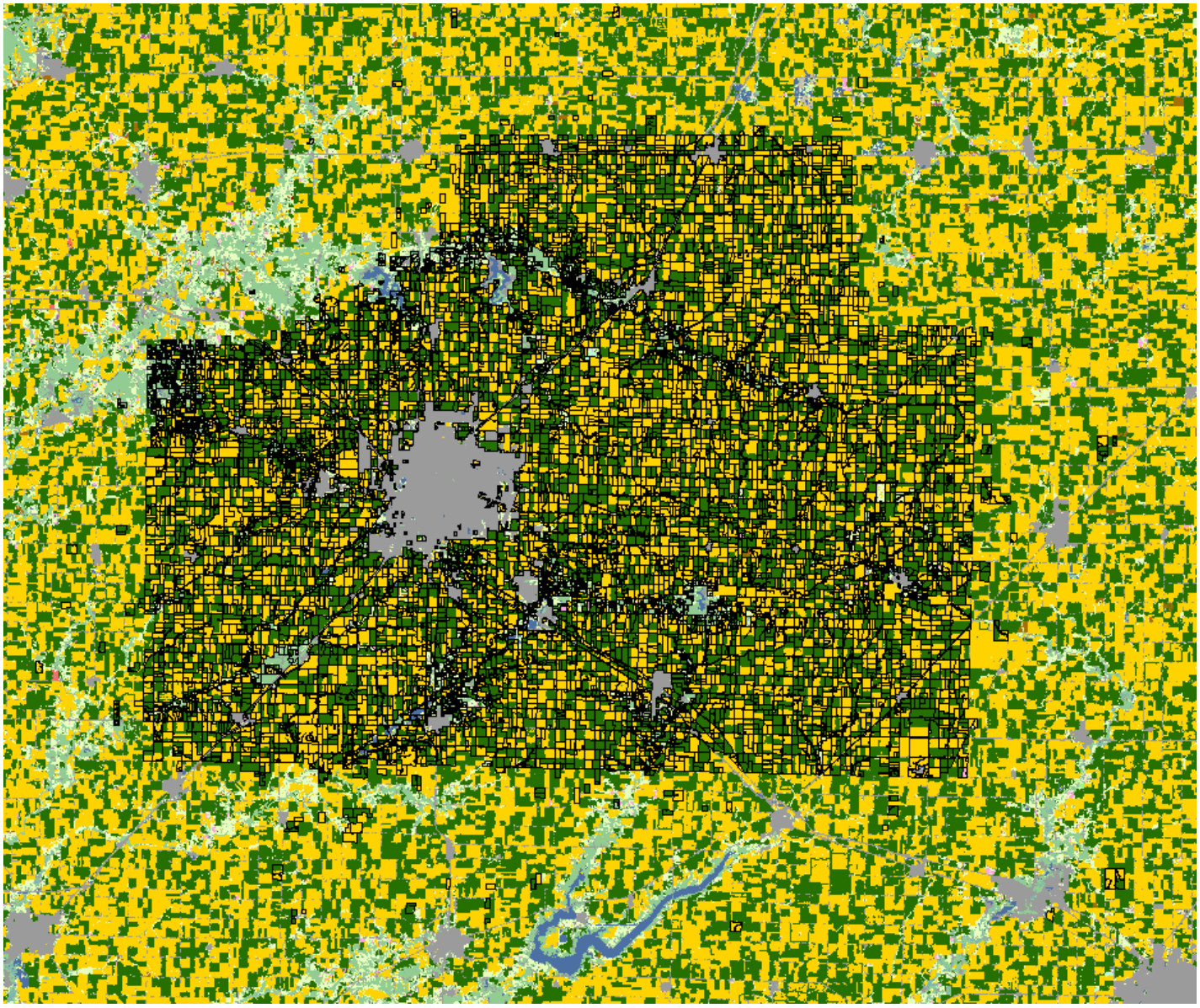


McLean County, IL Corn Expected Yieldby Common Land Unit (CLU), 2009 Base Year

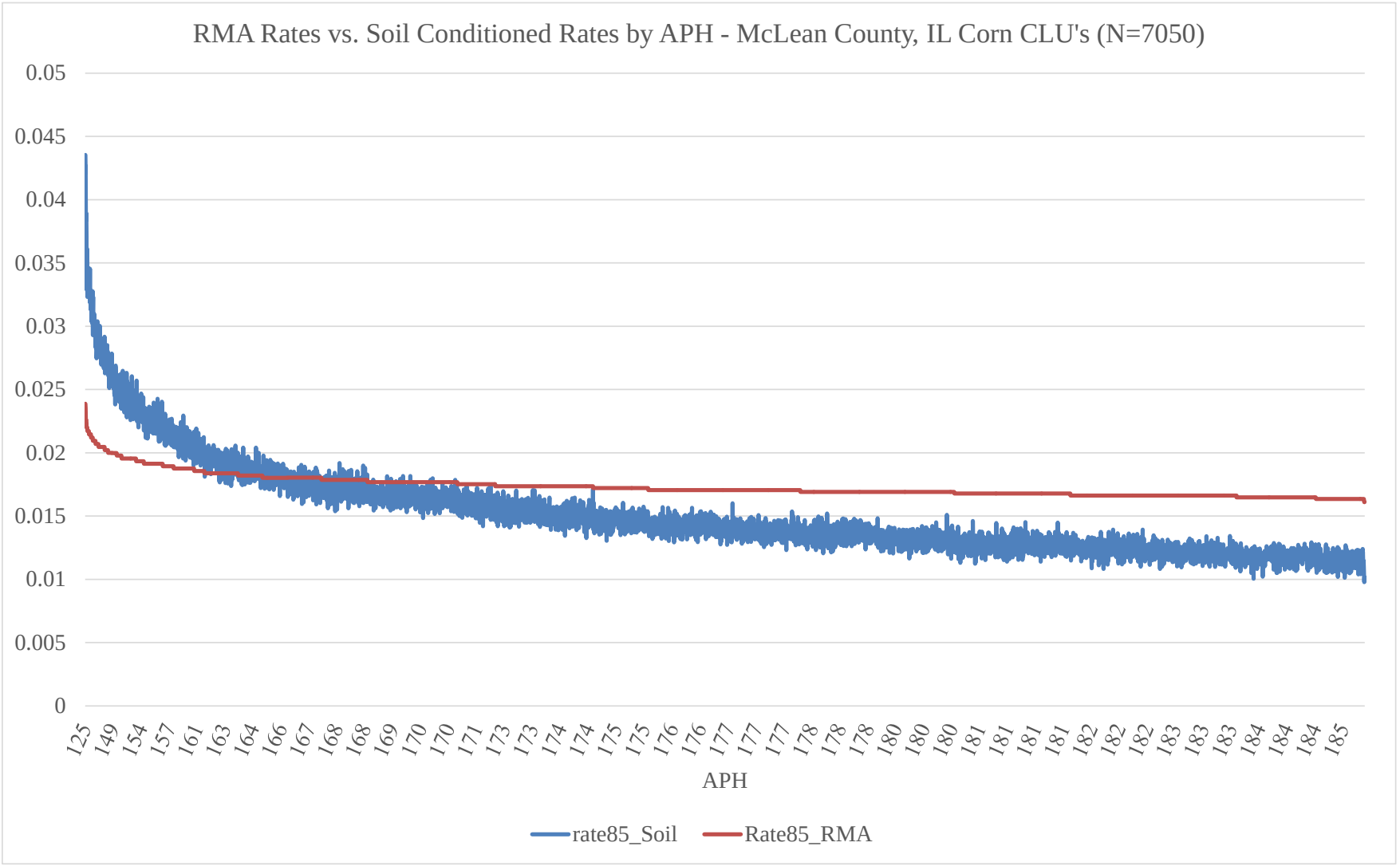



Soil Productivity Index Kernel Density among Common Land Units (CLU's), McLean County, Illinois (SSURGO Data)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
61
65
68
72
76
79
83
87
91
94
98
102
105
109
113
116
120
124
127
131
Soil Productivity Index (IL 810 Circular)

Filter CLUs by Cropland Data Layer



Could Incorporating Soil Information Improve RMA Rating Accuracy?



National Yield Analysis with gSURRGO

- ***Soil Variables***
- Soil data are grouped in 8 main categories (NRCS): AWS, TK, SOC, TKs, NCCPI, ROOTZ, DROUGHTY, PWSL. The variable names are defined as follows.
- AWS (Available Water Storage)
- TK (Thickness used in the Available Water Storage calculation)
- SOC (Soil Organic Carbon)
- TKs (Thickness used in the Soil Organic Carbon calculation)
- NCCPI (National Commodity Crop Productivity Index)
- ROOTZ (Root Zone Depth)
- DROUGHTY (Drought vulnerable soil landscapes)
- PWSL (Potential Wetland Soil Landscapes)

National Level Models-Available Water Storage

Variable	*Coefficient*	*t-statistic*	*Average Effect*
AWS0_5	2.37	21.78	3.03
AWS5_20	1.13	26.71	4.55
AWS20_50	0.75	34.51	7.78
AWS50_100	0.49	42.19	9.17
AWS100_150	0.33	35.20	10.07
AWS150_999	-0.01	-2.39	-0.53
AWS0_20	0.83	26.32	4.45
AWS0_30	0.63	29.59	5.47
AWS0_100	0.27	40.35	8.55
AWS0_150	0.17	41.30	10.65
AWS0_999	0.09	28.33	7.66

National Level Models- **tk_a:** Thickness (cm) used in the AWS calculation

Variable	*Coefficient*	*t-statistic*	*Average Effect*
TK0_5a	8.44	11.89	3.40
TK5_20a	2.77	11.66	3.33
TK20_50a	1.22	10.46	2.94
TK50_100a	0.73	12.21	3.04
TK100_150a	0.60	15.02	2.68
TK150_999a	-0.07	-7.92	-1.82
TK0_20a	2.09	11.72	3.35
TK0_30a	1.39	11.69	3.33
TK0_100a	0.40	11.98	3.25
TK0_150a	0.28	14.30	3.48
TK0_999a	-0.01	-1.49	-0.29

National Level Models-Soil Organic Carbon (g C per square meter)

Variable	*Coefficient*	*t-statistic*	*Average Effect*
SOC0_5	0.012	32.08	6.30
SOC5_20	0.005	38.63	8.20
SOC20_50	0.004	49.70	8.95
SOC50_100	0.003	33.42	5.68
SOC100_150	0.003	20.11	3.58
SOC150-999	0.000	0.88	0.15
SOC0_20	0.004	37.85	7.97
SOC0_30	0.003	42.78	8.88
SOC0_100	0.001	43.98	8.18
SOC0_150	0.001	41.97	7.88
SOC0_999	0.001	39.83	7.32

National Level Models

Variable	*Coefficient*	*t-statistic*	*Average Effect*
TK0_5s	9.98	13.87	3.99
TK5_20s	3.29	13.86	3.95
TK20_50s	0.56	7.99	1.44
TK50_100s	0.27	7.83	1.19
TK100_150s	0.39	12.93	1.77
TK150_999s	-0.07	-7.85	-1.81
TK0_20s	2.48	13.89	3.97
TK0_30s	1.48	13.22	3.64
TK0_100s	0.20	9.08	1.73
TK0_150s	0.15	11.19	1.96
TK0_999s	0.00	-0.60	-0.11
Musumcpcts	0.32	16.10	2.57

- **nccpi2cs:**National Commodity Crop Productivity Index -CORN and SOYBEANS.
- **nccpi2sg:**National Commodity Crop Productivity Index -SMALL GRAINS.
- **nccpi2co:**National Commodity Crop Productivity Index -COTTON.
- **nccpi2all:**National Commodity Crop Productivity Index -OVERALL.
- **pctearthmc:**National Commodity Crop Productivity Index -map unit percent earthy major components.
- **rootznemc:**Root Zone Depth (cm) -earthy major components.
- **rootznaws:**Root Zone Available Water Storage (mm) -earthy major components.
- **droughty:**Droughty Soil Landscapes -earthy major components.
- **pwsl1pomu:**Potential Wetland Soil Landscapes.
- **musumcpct:**Map Unit summed component percentage (representative value).

National Level Models

Variable	*Coefficient*	*t-statistic*	*Average Effect*
NCCPI2cs	72.86	61.17	13.80
NCCPI2sg	78.70	43.63	9.58
NCCPI2co	-14.94	-15.86	-0.20
NCCPI2all	78.47	65.86	14.08
Pctearthmc	0.40	20.16	2.29
Rootznemc	0.18	29.65	4.90
Rootznaws	0.14	45.03	9.89
Droughty	-23.12	-40.99	-7.24
Pwsl1pomu	0.01	4.49	0.60
Musumcpct	0.11	1.99	0.57

Summary of Findings

- Soil strongly predictive of expected yield and risk (more generally, distribution shape), and not trivial differences at high resolution
- Even in areas with relatively high quality soil and homogeneity such as Central IL, **strong evidence that current rating methods which use only APH as basis for intra-county rate do not fully capture observable heterogeneity in rates across soil quality**
- One would expect this might lead to adverse selection/incentive

Conclusion and Future Research

- Demonstrate proof of concept that use of high resolution soil data may be useful for improved estimation of rates (results here are probably the lower bound of benefits)
- Several implications for policy design and performance
  - Reduced adverse selection and taxpayer costs
  - Environmental benefits
- Analysis of RMA insurance loss data at CLU level, and incorporation within existing RMA rating system is critical, as is researcher data access
  - Help RMA improve rating accuracy and actuarial soundness, reduce adverse selection and adverse incentive
  - Not just a rating problem
  - Not just rates, but also underwriting rules
- False precision concerns?
- What about other indexes or soil measures (note need to be general)?

Moving Forward

- Data need to be made available for research through university and other systems (with confidentiality and privacy protections)
- Potentially one of the largest untapped (or under-tapped) data resources for scaling soil economics research and bridging to policy solutions
- Secure data warehouse for integrating administrative data for **analytical purposes**
  - Would require strong support and directives from leadership, forethought on security and privacy issues, etc., and many precedents exist

Moving Forward

- Hybridization of biophysical or mechanistic models with large scale statistical models
- With good estimates of soil conditional yield distributions, may be possible does adoption of new practices affect distribution
- Dynamic effects over time (slowly changing conditional distribution)
- Combining large scale distributional models and data with trial data in order to scale (see e.g., case on Skip Row insurance, Woodard et al., 2012 AJAE)