IOWA STATE UNIVERSITY
Department of Statistics

JSM 2018
July 31, 2018

Xiaodan Lyu, Emily Berg, Heike Hofmann

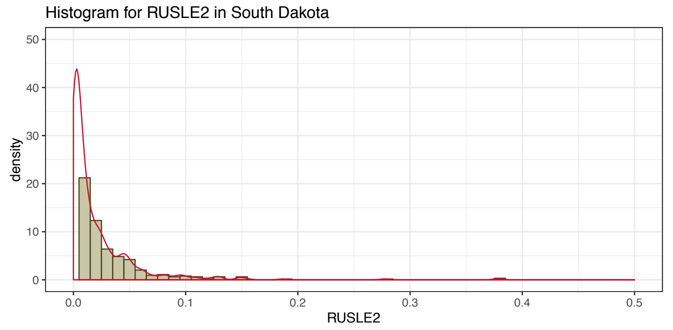
Empirical Small Area Prediction of Sheet and Rill Erosion Using a Zero-inflated Lognormal Model

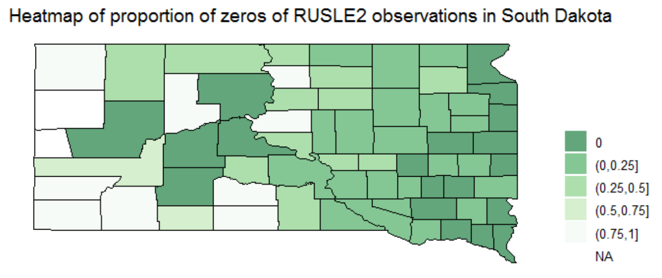
Introduction

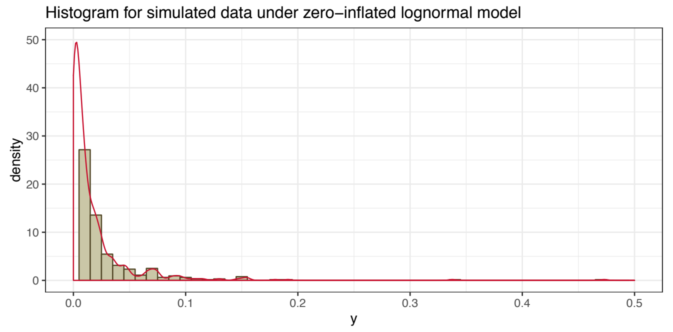
- Small area estimation widely used when sample sizes too small for direct estimation.
- Skewed data w/ zeros: Conservation Effects Assessment Project Sheet and rill erosion (RUSLE2) data in South Dakota contains about 15% zeros.
- Small area predictors and MSE estimators for a lognormal model have closed-form expressions (Berg and Chandra, 2014). Can we extend this to a zero-inflated model?
- How does empirical Bayes compare to the plug-in predictor (Chandra and Chambers, 2016)?

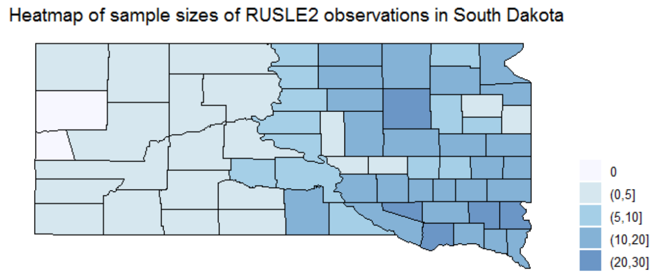
Zero-inflated Lognormal Model

- Let $i = 1,...,D$ index areas, $j = 1,...,N_i$ index units in each area.
- Variable of interest: $y^*_{ij} = y_{ij}\delta_{ij} \geq 0$
  - $\delta_{ij} = 0$ if observed value is positive, 0 otherwise.
  - population mean: $\bar{y}^*_{N_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} y^*_{ij}$.
- Observed data: $\{y^*_{ij}, i = 1,\ldots,D, j \in s_i\} \cup \{z_{ij} : i = 1,\ldots,D, j = 1,\ldots,N_i\}$
- Positive part: $\log(y_{ij}) = \beta_0 + z'_{1ij}\boldsymbol{\beta}_1 + u_i + e_{ij}$
- Binary part: $\delta_{ij} \sim \textsf{Bernoulli}(p_{ij}), g(p_{ij}) = \alpha_0 + z'_{2ij}\boldsymbol{\alpha}_1 + b_i, g(\cdot)$ is a parametric link function.
- $(u_i, b_i, e_{ij}) \sim N(\mathbf{0}, \textsf{diag}(\sigma^2_u, \sigma^2_b, \sigma^2_e))$









Small Area Prediction

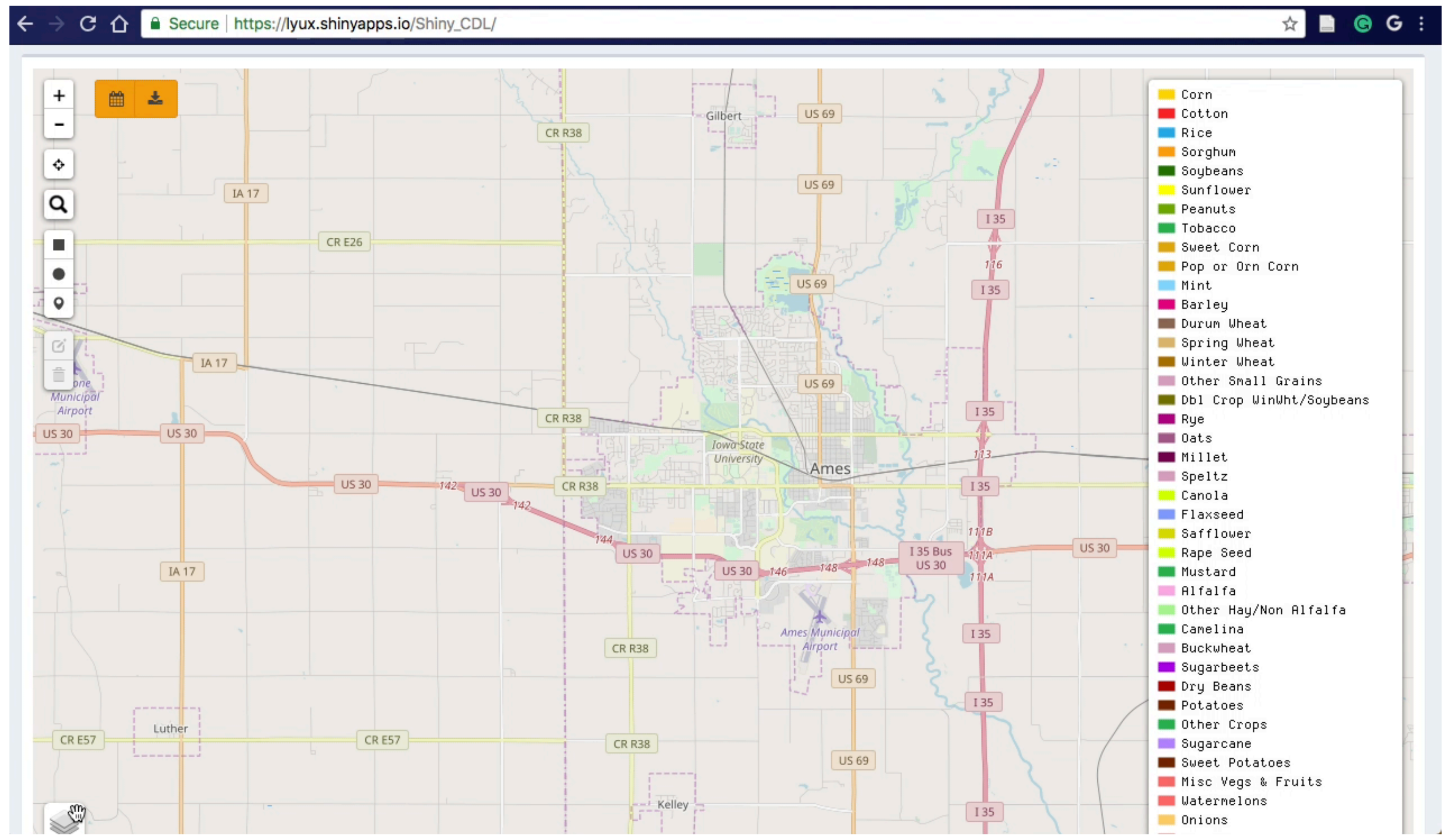
- Use empirical Bayes method to predict population means at small area level.
- For MSE estimator, we propose:
  - a close-formed analytic "one-step" estimator ignoring variance due to parameter estimation.
  - parametric bootstrap estimator incorporating variance due to parameter estimation and bias of the "one-step" estimator of leading term.

Conservation Effects Assessment Project (CEAP)

- Response variable y*: sheet and rill erosion, as measured by the Revised Universal Soil Loss Equation (RUSLE2), an update of a model for sheet and rill erosion called USLE.
- Possible explanatory variables related to the USLE:

Variable	Source	Denifition
logR	NRI	log-scale county-level R-factor
logK	Soil Survey	log-scale K-factor of the soil map unit containing the location
logS	Soil Survey	log-scale S-factor of the soil map unit containing the location
is.corn	2006 CDL	1 if it's corn
is.soybean	2006 CDL	1 if it's soybean
is.sprwht	2006 CDL	1 if it's spring wheat
is.wtrwht	2006 CDL	1 if it's winter wheat

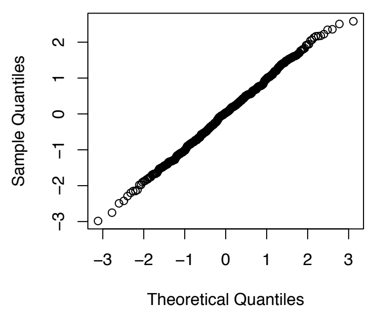
- Visualize an overlay operation required to collect auxiliary information:

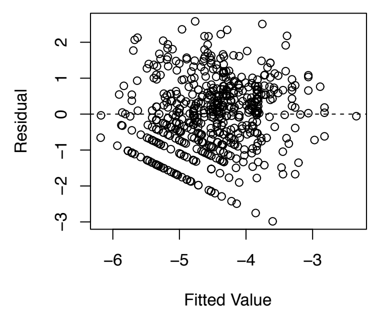


- Model assessment:
  - Consider county random effect for both positive and binary part.
  - Backward variable selection applied to the fixed effects with a threshold of $\Delta(\textsf{AIC}) = 0.5$.
  - For the binary part, the Hosmer-Lemeshow test shows no significant lack of fit.
  - Lognormal-logistic model fitting result and standardized residual plot for the positive part:

	Positive Part	Binary Part
logR	2.08 (0.36)***	5.04 (0.73)***
logK	0.48 (0.23)*	
logS	0.48 (0.07)***	0.38 (0.21)·
is.soybean		0.70 (0.33)*
is.sprwht		0.95 (0.50)·

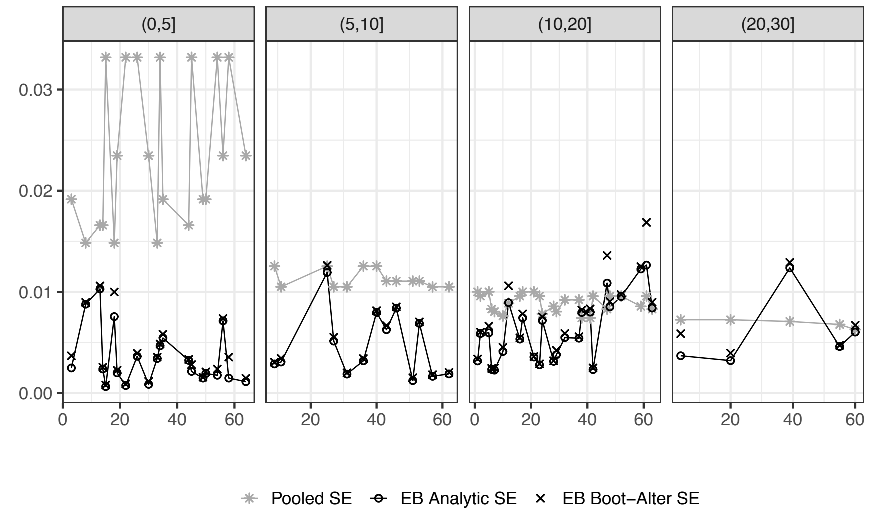
****p < 0.001, **p < 0.01, *p < 0.05, ·p < 0.1*

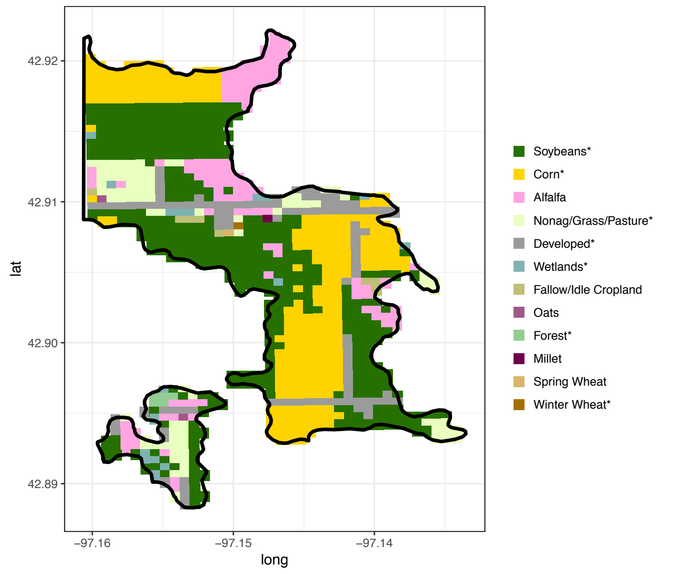




CEAP Empirical Bayesian Predictions

- Population element: a CDL pixel classified as cropland within a county in a CEAP state.
- Incorporating weights: predicted population mean is an average across soil mapunit segments weighted by crop acreage.
- Comparison of standard errors and example of a soil mapunit overlaid with 2006 CDL:





Simulation Results

Relative MSE of PI/ZI predictor to EB predictor:

	size = 5		size = 10		size = 20	
Link	PI	ZI	PI	ZI	PI	ZI
logit	1.003	1.445	1.002	1.789	1.000	2.781
probit	1.007	1.362	1.002	1.702	1.000	2.348
cauchit	1.001	1.478	0.999	1.904	1.001	2.796

PI: plug-in predictor, ZI: zero ignored MMSE predictor

Simulation study on the proposed one-step MSE estimator

	size = 5		size = 10		size = 20	
Link	RB	CP	RB	CP	RB	CP
logit	0.0	94.5	-3.6	94.9	-0.1	94.9
probit	4.1	94.6	-3.8	94.8	-3.5	94.7
cauchit	-1.2	94.3	-3.4	94.9	-1.9	94.7

RB: relative bias, CP: coverage probability

Summary

- We developed EB predictors based on a zero-inflated lognormal for SAE:
  - EB and plug-in have similar efficiency, unless data extremely sparse.
  - For D = 60, the "one-step" MSE estimator is a reasonable approximation.
  - For D = 30, the bootstrap MSE estimator: RB 2%~3%, CP 94%~96%.
  - EB predictor is typically more efficient than direct estimators in terms of MSE in CEAP application.
- For data analysis purposes, we combined three additional sources besides CEAP: National Resources Inventory (NRI), National Cooperative Soil Survey and USDA National Agricultural Statistics Service Cropland Data Layer (CDL).
- Future work:
  - Modifying our EB approach to account for the discrete nature of the data.
  - Investigate extensions to more flexible distributional forms, such as a GB2 distribution or quantile regression model.

References: [1] Berg, E. and Chandra, H., 2014. Small area prediction for a unit-level lognormal model. Computational Statistics & Data Analysis, 78, pp.159-175.
[2] Chandra, H. and Chambers, R., 2016. Small area estimation for semicontinuous data. Biometrical Journal, 58(2), pp.303-319.