Small area prediction and big data visualization:

Analysis of soil losses from sheet and rill erosion on cropland

by

Xiaodan Lyu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Emily J. Berg, Co-major Professor
Heike Hofmann, Co-major Professor
Wayne A. Fuller
Yuyu Zhou
Zhengyuan Zhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

Copyright © Xiaodan Lyu, 2020. All rights reserved.

DEDICATION

To my dear mother, father, and grandmother.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

ACKNOWLEDGMENTS xi

ABSTRACT xiii

CHAPTER 1. GENERAL INTRODUCTION 1
1.1 Soil Erosion Equations 1
1.2 National Resources Inventory 3
1.3 Data Quality and Cognitive Load 4
1.4 Problem Statement and Our Approach 6
1.5 References 8

CHAPTER 2. EMPIRICAL BAYES SMALL AREA PREDICTION UNDER A ZERO-INFLATED LOG-NORMAL MODEL WITH CORRELATED RANDOM AREA EFFECTS 11
2.1 Abstract 11
2.2 Introduction 12
2.2.1 Motivating CEAP RUSLE2 Data 12
2.2.2 Zero-inflated Lognormal Model with Correlated Area Random Effects 13
2.2.3 Related Small Area Procedures for Skewed, Binary, or Zero-inflated data . . 15
2.2.4 Auxiliary Information Acquisition for Small Area Models 17
2.2.5 Outline of Our Approach 18
2.3 Empirical Bayes Prediction for the Zero-inflated Lognormal Model 19
2.3.1 Minimum Mean Square Error Predictor of Small Area Mean 19
2.3.2 Empirical Bayes Predictor of Small Area Mean 21
2.3.3 MSE of the EB Predictor 23
2.4 Simulations 26
2.4.1 Comparison with Alternative Predictors 27
2.4.2 Evaluation of the Parametric Bootstrap MSE Terms 30
2.5 Estimating Sheet and Rill Erosion for the Conservation Effects Assessment Project . 32
2.5.1 Auxiliary Variables and Population Predictions for CEAP 33
2.5.2 CEAP Zero-inflated Lognormal Model Fitting 35
2.5.3 CEAP Empirical Bayesian Predictions 37
2.6 Conclusion 39
2.7 References 40

2.8 Appendix. Supplementary Information . . . 43
2.8.1 Computing Details of MLE Estimation . . . 43
2.8.2 Incorporating Weights . . . 45
2.8.3 Link Function Analysis . . . 48
2.8.4 Monte Carlo Margin of Error . . . 49
2.8.5 Supplementary Figures and Tables . . . 49

CHAPTER 3. AN R SHINY APPLICATION TO DATA QUALITY REVIEW . . . 53
3.1 Abstract . . . 53
3.2 Introduction . . . 54
3.2.1 Motivating NRI Table Review Process . . . 56
3.2.2 Outline of the Paper . . . 60
3.3 User interface . . . 60
3.4 Example of Use . . . 62
3.5 Design and Implementation . . . 64
3.6 Discussion . . . 67
3.7 Conclusion . . . 69
3.8 References . . . 69

CHAPTER 4. VISCOVER: A WEB APPLICATION TO VISUALIZE THE SOIL AND LAND COVER DATA AND THEIR OVERLAY . . . 72
4.1 Abstract . . . 72
4.2 Introduction . . . 73
4.3 Data Description . . . 74
4.4 Motivating Sheet and Rill Erosion Study . . . 77
4.5 User Interface and Example of Use . . . 78
4.6 Design and Implementation . . . 81
4.6.1 Web Scraping . . . 83
4.6.2 Tile Layer . . . 84
4.6.3 Technical Highlights . . . 85
4.7 Discussion . . . 86
4.7.1 Impact on the CEAP Data Analysis . . . 86
4.7.2 Uses of the Soil Map . . . 87
4.7.3 Uses of the Cropland Data Layer by NRI . . . 88
4.8 References . . . 89

CHAPTER 5. INTERACTIVE SHEET AND RILL EROSION MAP OF SOUTH DAKOTA AT A 30-METER SPATIAL RESOLUTION . . . 92
5.1 Abstract . . . 92
5.2 Introduction . . . 93
5.3 Soil Erodibility and Slope Factors . . . 96
5.3.1 Soil Components Inventory . . . 96
5.3.2 Soil Map Units to Points . . . 99
5.3.3 Map of K and S factors . . . 104

5.4 Rainfall and Crop-management Factors . . . 105
5.4.1 Rainfall Factor . . . 105
5.4.2 Crop Management Factor . . . 105
5.5 Estimation and Visualization of Sheet and Rill Erosion . . . 110
5.6 Conclusion and Discussion . . . 114
5.6.1 Impacts on the Data Quality of NRI . . . 114
5.6.2 Future Work . . . 115
5.7 Acknowledgment . . . 116
5.8 References . . . 116
5.9 Appendix. Supplementary Information on SSURGO . . . 117

CHAPTER 6. GENERAL CONCLUSIONS . . . 119

LIST OF TABLES

Page

Table 2.1 Average MSE differences ($\times 10^5$) between the alternative predictors (EB(0), EB predictor assuming $\rho = 0$; PI, plug-in predictor; ZI, zero-ignored MMSE predictor; SI, shifted MMSE predictor) and the EB predictor. The associated Monte Carlo margins of error are presented in parentheses. 28

Table 2.2 Average MSE differences ($\times 10^5$) between the EB(0) predictor and the EB predictor as the true correlation ρ of the random area effects between the positive part and the binary part changes. The associated Monte Carlo margins of error are presented in parentheses. 29

Table 2.3 Average biases ($\times 10^5$) and coverage percentages (CP) of nominal 95% normal theory prediction intervals for the MSE estimators (one-step, bootstrap, semi-bootstrap). The associated Monte Carlo margins of error are presented in parentheses. 31

Table 2.4 The parameter estimates and associated bootstrap standard errors (SE) for the zero-inflated lognormal model fit to the cropland CEAP RUSLE2 data. 35

Table 2.5 Description of the variables in the collected dataset for predicting cropland RUSLE2. 47

Table 2.6 The average ratios (%) of the second term M_2, bias of the leading term M_1, and the cross term M_3 to the fully parametric bootstrap estimate of the MSE of the EB predictor, where the averages are across the areas with the same sample size n_i and $M = 1000$ simulations. The bootstrap size is $B = 100$. 49

Table 2.7 Mapping from CDL categories to NRI Broad Coveruse. 50

Table 5.1 The estimated covariates coefficients, with standard errors in brackets, of the simple linear regression models fitted to the point-level 2006 USLE losses in log scale recorded by NRI using the recorded erosion factors by NRI (left) and the approximated erosion factors by our approach (right). The bottom half of the table shows the R^2, adjusted R^2, the number of observations, and the root mean squared error (RMSE) of each fitted model. 112

Table 5.2 SSURGO table column descriptions. 118

LIST OF FIGURES

Page

Figure 2.1 Left: Histogram of CEAP sampled RUSLE2s in South Dakota. Right: Scatterplot of county proportions of positive RUSLE2s against county mean of log RUSLE2s. 13

Figure 2.2 Standard errors of the EB predictor and the direct estimator (sample mean) of mean cropland RUSLE2 for the South Dakota counties. Standard errors for the EB predictor are square roots of the one-step or semi-boot MSE estimator. The pooled standard error is used for the direct estimator. The comparisons are grouped by sample size labeled on the top. 37

Figure 2.3 Cartogram of the EB predicted county means of cropland RUSLE2 in South Dakota. Darker shade indicates severer soil sheet and rill erosion. Smaller shrinkage indicates smaller coefficient of variance. This figure appears in color in the electronic version of this article. 38

Figure 2.4 Example plot of a SSURGO map unit polygon overlaid with 2006 CDL raster. 46

Figure 2.5 Cartograms of the proportions of zero (top) and sample sizes (bottom) of the cropland CEAP RUSLE2 data in the counties of South Dakota. 51

Figure 2.6 The normal quantile-quantile plot (left) and the standardized residual plot (right) for the marginal (top) and conditional (bottom) residuals from fitting the positive part of the cropland CEAP RUSLE2 data. 51

Figure 2.7 The p-values of the Hosmer-Lemeshow tests with different number of groups for the goodness of fit of the binary part of the cropland CEAP RUSLE2 data. The dotted horizontal line is $p = 0.05$. 52

Figure 2.8 EB predictions plotted against direct estimates of the mean cropland RUSLE2 for the counties in South Dakota. Top-left corner in each panel is the sample Pearson correlation between the two estimators in each sample size group. . 52

Figure 3.1 Flowchart of statistical production involving a data product. The data quality review and improvement process (highlighted in blue) — evaluating (assessing) quality metrics, identifying error sources, and correcting prior interventions, are iterated until the quality of the data product is acceptable. 54

Figure 3.2 Snapshots of estimate tables (national level average annual sheet and rill erosion on non-Federal rural land in tons per acre per year with margins of error presented after a ± sign) from the 2015 (left) and 2012 (right) NRI reports. 59

Figure 3.3 Snapshot of U.S. Difference Table 3a (national level average annual sheet and rill erosion estimates on non-Federal rural land in tons per acre per year) from *iNtr*. The cell values are the absolute relative differences between the 2015 and 2012 NRI estimates. Darker shades of red indicate larger differences warranting closer inspection. 59

Figure 3.4 Overview of the graphical user interface of *iNtr*. 61

Figure 3.5 Hierarchical table review using *iNtr*. Grey arrows show the order of a reviewer's interaction to access more detailed information. Selected cells in each panel are highlighted with a golden border. 63

Figure 3.6 Flowchart of the data processing underlying the review system of *iNtr*. On the top left, the collection of NRI estimate tables is shown. Using key descriptors, such as version identifier, spatial level of aggregation, table identifier, etc., we build a single database (top right). This database forms the foundation of extracting subsets of data feeding into the different visualization panels (bottom). 65

Figure 4.1 A snapshot from *VISCOVER* showing a soil map (lines and alphanumeric symbols in light brown) zoomed near Ada Hayden Lake in the northern Ames. The background layer is the Open Street Map. A blue map marker with an "S" mark appears at each user-clicked location in the map when the toggle at the top left is switched to "S". The tooltip of each map marker contains the identifiers (symbol, key, name, and acres) of the soil map unit containing the pinned location. 75

Figure 4.2 A snapshot from *VISCOVER* showing the 2019 Cropland Data Layer overlaid with the NRCS Soil Map zoomed near Ada Hayden Lake in the northern Ames (right blue area). A color legend appears on the right. A yellow map marker labelled "C" is drawn at each user-clicked location in the map when the toggle at the top left of the map is switched to "C". The tooltip describes the land cover classification at the pinned location. 76

Figure 4.3 Graphical user interface of *VISCOVER*. 78

Figure 4.4 An example overlay in *VISCOVER*: (1) search the geolocation to re-center the map and draw a polygon to define an area of interest; (2) highlight a particular soil map unit by selecting the corresponding row in the Soil Data Layer table; (3) unfold the Cropland Data Layer box to view the overlay results. 80

Figure 4.5 Primary functionalities of *VISCOVER*. It grabs the data from the soil and land cover databases to its user interface on demand by using web services. 82

Figure 5.1 Workflow of integrating the NRI data with other data sources to obtain soil erosion factors at a 30-meter spatial resolution covering South Dakota. . . . 95

Figure 5.2 Partial hierarchy of the NRCS Soil Survey Database. Representative soil erosion K and S factors of a soil component in a soil map unit of a soil survey area are determined by its texture and horizontal depth jointly. . . . 97

Figure 5.3 Left: histogram of the percentages of the major and the minor soil components per soil map unit. Right: the stacked bar plot of the percentages of the three major and the three minor components in the example soil map unit "Janesburg-Regent-Cabba complex". 97

Figure 5.4 The flowchart of the calculations of the NED-derived slopes defined as the average absolute difference among each cell and its eight nearest neighbors: (1) shifting the NED elevation raster toward eight different directions; (2) calculating the absolute differences between the original raster and the eight shifted raster images; (3) stacking and averaging the eight difference raster images. 100

Figure 5.5 The heat map with contours of (1) the NED elevation and (2) the NED-derived slope at 1/3 arc-second resolution covering the soil map unit "Janesburg-Regent-Cabba complex" presented in the bottom. The assignment of the soil components first by (3) slope gradients then by (4) component percentages at a 30-meter resolution. 102

Figure 5.6 The heat maps of the designated K factors (left) and SSURGO-based slope gradients (right) at a 30-meter spatial resolution. The white color indicates undefined values. 104

Figure 5.7 The histogram (left) and Q-Q plot (right) of the standardized residuals after fitting a simple linear model to the recorded R factors in log scale with longitudes and latitudes in the NRI Database. 106

Figure 5.8 Left: the linearly interpolated rainfall (R) factors from the 2006 NRI to the raster image at a 30-meter spatial resolution covering South Dakota. Right: the isoerodent map of South Dakota used in the U.S. Agriculture Book NO.282 for referring average annual values of the R factor. 106

Figure 5.9 The distribution of the recorded C factors by the 2006 NRI for the ten most frequent crop rotation patterns classified by the 2006 and 2007 CDL data in South Dakota. 108

Figure 5.10 The approximated C factors using the best tree model (left) and the best random forest model (right) for the cropland frame at a 30-meter spatial resolution covering South Dakota. The non-cropland cells are in grey. 109

Figure 5.11 The R, K, S and C factors recorded by the 2006 NRI versus the approximated factors by our method. The red dashed are the identity lines, and the top left corners are the Pearson correlation coefficients. Larger dot size indicates the larger number of overlapped points. 110

Figure 5.12 The predicted logarithmic USLE losses on cropland at a 30-meter spatial resolution (left) and aggregated to the maximum at an about 1-kilometer spatial resolution (right). 112

Figure 5.13 The snapshot of the Sheet and Rill Erosion Map (*SREM*) zoomed near Pierre, South Dakota. The location that a user inputs through the search widget is marked by a blue circle in the map. The tooltip at the pinpoint contains the approximated R factor, K factor, soil slope gradient, C factor, and the estimated USLE loss by our approach. 113

ACKNOWLEDGMENTS

I set my heart on thanking my academic advisors, Dr. Heike Hofmann, and Dr. Emily Berg, for their patience, guidance, and confidence in me. Thank them for joining my interest in the application of statistical computing and data visualization to survey data. It is an atypical path with few exemplars. My research projects seem unconventional for a Ph.D. student in Statistics. I used to have doubted myself on my research topics for a thousand times. I genuinely appreciate the encouragement and advocacy of personal style from Dr. Heike and Dr. Emily. Dr. Heike always has the acutest judgment for my position and future direction. Thank Dr. Emily for being available all the time and providing me assistance whenever I need it. She is remarkably knowledgeable of small area estimation literature and able to solve my related puzzles every time.

I would like to thank my committee members: Dr. Wayne Fuller, Dr. Yuyu Zhou, and Dr. Zhengyuan Zhu. It is my honor to have Dr. Wayne Fuller in my Ph.D. committee. He is one of my idols in academia. Thank Dr. Yuyu Zhou for providing helpful suggestions on my project from his professional point of view. I would like to thank Dr. Zhengyuan Zhu for letting me work on the National Resources Inventory data at the Center for Survey Statistics and Methodology (CSSM). It is Dr. Zhu's impetus that I can embark on most of the projects in this dissertation. It is a pleasant experience working with the peer graduate research assistants in Snedecor Hall and the CSSM analysts in the Office and Lab building. To name a few, Tyler Harms and Jie Li have been tremendously helpful and supportive for me during our collaborations.

I would also like to thank Mojtaba Navid, Maryam Rezvani, and Lauren Goodfellow at Autodesk for the career opportunities they have offered me. Their support and endorsement have helped me build massive confidence in my skill sets. I have a much more clear career plan now.

Finally, I would like to thank my friends and family for their care and support. My close friends in Ames endorse me with great mental comfort during my Ph.D. study. My parents have been

providing me with unconditional love since I was born 26 years ago. They have never set a key performance indicator (KPI) for me to strive. They care about my wellness and happiness more than anything. My mother, Liping, is always a kind, lovely, curious, positive, and persistent person even though she has been suffering a lot in her life. She has been spiriting me to overcome many of the difficulties in my life. My grandmother passed away from a heart attack two and a half years ago. I have been missing her much. I had initially planned to share all of my milestones in my life with her. This dissertation is one of them.

I treat every chance in my life as a blessing. I feel lucky for being able to work or live with those lovely people. Thank you all!

ABSTRACT

Assessment of soil erosion benefits both the well-being of people and agricultural production. Sustainable and environmentally friendly agriculture needs to balance short-time production, long-term capabilities, and environmental quality. The overarching applications related to the works in this dissertation are related to the National Resources Inventory (NRI) program. The ongoing NRI surveys collect a wealth of sample data describing natural resources conditions and trends to support national policy-making and enterprise-level landowner decision making on resource conservation practices. Among those natural resources issues, soil erosion assessment is of primary interest to prioritize future soil conservation needs and measure past soil conservation impact. Our effort is aimed at estimation of land use and soil erosion rates, especially sheet and rill erosion, through combined techniques of small area estimation and "big" data visualization.

Small area estimation (SAE) techniques are used to construct model-based estimators when direct survey estimators cannot achieve desired statistical reliability. To account for the zero-contamination and right-skew of the sheet and rill erosion data in our case study, we consider a zero-inflated log-normal model framework and extend the two-part model of Chandra and Chambers (2016) by including an additional parameter to account for significant correlation between the pair of random effects for an area. We develop an empirical Bayes predictor of the area mean that replaces the unknown model parameters in the best predictor, which is guaranteed to be unbiased and have the minimum mean squared error, with consistent parameter estimates. We address the analytic challenges associated with parameter estimation under this model framework by using a maximum likelihood method. Maximum likelihood estimation is challenging because of a need to integrate over a bivariate distribution of the pair of random effects for a county. We transform the bivariate integral to a univariate integral to facilitate numerical integration through a computationally efficient Gauss-Hermite approximation. Computationally efficiency in terms of assessing statistical uncertainty

in the estimates is further enhanced by using the "one-step" MSE estimator, an estimator we propose that does not require resampling. The reliable county-level erosion estimates that are not obtainable from the NRI sample data can be used to prioritize conservation resource allocation at a more granular level. To help practitioners implement our SAE methodology, we develop an R package **saezero**, available at https://github.com/XiaodanLyu/saezero.

Besides the characteristic of *reliability*, there are many other dimensions of data quality, such as *accuracy*, *consistency*, *timeliness*, *usability*, *accessibility*, and *relevance*, which are featured in the quality assurance (QA) process of NRI. The QA process is operationally complex as the involved databases are large in scale. Effective visualization techniques, under the help of well-managed databases, can facilitate the QA process by alleviating the cognitive load and enabling user-data interactions. By using the reactive framework of R **shiny**, we built three web-based graphical tools intended to be used by NRI. The first tool "iNtr", whose public version is available at https://lyux.shinyapps.io/table_review/, is designed to help with the labor-intensive NRI table review process so that data *accuracy* and *consistency* can be checked as much as possible without sacrificing the *timeliness* of the NRI releases. The second tool "VISCOVER", available at https://lyux.shinyapps.io/viscover/, is developed to check the *accuracy* of the auxiliary variables, i.e., public soil and crop-cover data, used in the case study of our SAE methodology. An R package **viscover**, available at https://github.com/XiaodanLyu/viscover, has also been developed by us for practitioners to query the two databases easily. The third tool "SREM", available at https://lyux.shinyapps.io/srem/, presents an interactive sheet and rill erosion map at a 30-meter spatial resolution to enhance the *usability* and *accessibility* of NRI in that the NRI erosion estimates used to be available only at national and state level in the form of printed figures and tables. "SREM" is built upon five databases —- one sheet-and-rill-erosion and four soil-erosion-factor databases we created by assembling the NRI Database and several other public databases by data linkage and statistical modeling.

CHAPTER 1. GENERAL INTRODUCTION

In the 1930s, the Dust Bowl drew people's attention to soil resource conservation and non environmentally friendly land management practices. The 1934 National Erosion Reconnaissance Survey was the first formal study of erosion conducted in the United States. This survey motivated the passage of the Soil Conservation Act of 1935, which promoted the establishment of the Soil Conservation Service, currently known as the Natural Resources Conservation Service (NRCS), an agency of the United States Department of Agriculture (USDA). The Soil and Water Resources Conservation Act of 1977 provided NRCS with an impetus for initiating the National Resources Inventory (NRI) program to monitor the conditions and trends of soil, water, and related resources on non-federal lands. The NRI program has been conducted through a collaboration between NRCS and the Center for Survey Statistics and Methodology (CSSM) at Iowa State University (ISU). The NRI data have been used to assess natural resources health to prioritize conservation treatment needs and measure conservation effort effects. According to Goebel (1998), the 1977 NRI produced the first national quantitative erosion data based on erosion prediction models, i.e., soil erosion equations.

1.1 Soil Erosion Equations

The direct sources of soil erosion include running water, waves, moving ice and wind, etc. The works included in this dissertation focus on analyzing soil loss from water erosion, in particular sheet and rill erosion. The sources of soil loss from water erosion include sheet and rill, ephemeral gully, classical gully, and streambank. Sheet and rill erosion, the first stage of water erosion, is the losses of soil particles within the field due to rainfall events. The combination of sheet and rill erosion contributes most to total water erosion. The universal soil loss equation (USLE) and its variants are widely-used erosion prediction models for approximating long-term average soil

losses from sheet and rill erosion under specific conditions. Since 1930, controlled experiments on field plots and small watersheds have been examining a variety of complex factors affecting soil erosion by rainfall and runoff. The development of soil-loss equations started in about 1940 in the Corn Belt States. The soil slope, soil erodibility, conservation practice, and crop management were found out to be significant factors impacting the soil erosion rate (Zingg et al., 1940; Browning et al., 1947). The Musgrave soil loss equation (Musgrave, 1947) adapted the Corn Belt equation to the Northeastern States by adding a rainfall factor. To further relax the geographic and climatic restrictions in earlier models, the Runoff and Soil-Loss Data Center of the Agricultural Research Service at Purdue University made several major developments, and the improved equation is now usually referred to as USLE. The equation is

$$A = R\,K\,L\,S\,C\,P, \tag{1.1}$$

which says the computed soil loss per unit area (A), can be measured by the rainfall factor (R) multiplied by the soil-erodibility factor (K), the slope-length factor (L), the slope-gradient factor (S), the cropping-management factor(C), and the erosion-control practice factor (P). Graphical methods for determining the values of each of the six factors were created by synthesizing research data. The Agriculture Handbook No. 282 (Wischmeier and Smith, 1965) provided ready-reference tables and graphs to help determine the values of the equation's factors that are appropriate for a particular farm field. The methodologies in Wischmeier and Smith (1965) apply to the states that are to the east of the Rocky Mountains. An improved version of USLE containing revised tables and charts was published in the Agriculture Handbook No. 537 (Wischmeier and Smith, 1978) and expanded the application of USLE to western states and Hawaii. The Revised Universal Soil Loss Equation (RUSLE), published in the Agriculture Handbook No. 703 (Renard et al., 1997) retained the original USLE but provided a computer program to facilitate more involved factor calculations. The cutting-edge erosion-prediction technology is the Revised Universal Soil Loss Equation 2 (RUSLE2), which hybridizes the purely empirically-based USLE for robustness with process-based simulation models for generalization. A more powerful graphical user interface has

also been developed (Renard et al., 2011). The soil erosion equations have been widely used and validated by field experience for more than six decades.

1.2 National Resources Inventory

A permanent sampling frame has been used by NRI to establish a temporal and geospatial database. Since the recent NRI surveys use land use definition and erosion estimation protocols that are closer to the 1982 NRI and different from the 1977 NRI, the 1982 NRI is now considered as the initial time point for the time series data (Goebel, 1998). The 1982 NRI established the NRI Database that has been enriched with the on-going NRI surveys until the present. The NRI sample is a stratified two-stage sample on a county-by-county basis (Nusser and Goebel, 1997). Primary sampling units are, so-called *segments*, typically areas of land of approximately 1/2 by 1/2 square miles except western and northeastern United States. Within each sampled segment three points are selected as secondary sampling units. A "foundation sample" of approximately 800,000 points in 300,000 segments was selected for the 1982 NRI. To support NRI's important goal of estimating the dynamics of change over time, the data gatherers revisit the same points in the original foundation sample in later years. The NRI survey was conducted every five years from 1982-1997 and annually from 2000 through the present. Annual data collection at selected points in the NRI Database, with a different supplementary panel used for each year (Breidt and Fuller, 1999), has been implemented since 2000 to balance statistical efficiency and resource shortage. USLE was deployed by the NRI to estimate sheet and rill erosion before 2008, and RUSLE2 has been used by the NRI since 2008. The NRI erosion estimates are currently available for cropland, Conservation Reserve Program (CRP) land, and pastureland.

Data collection in the NRI is resource-intensive. First, an aerial photograph is taken of each sampled segment. Then, data collectors record the characteristics of the land that they observe in the images. If a point is classified as cropland, CRP, or pastureland, the data collector must further refer to administrative sources to obtain the erosion factors. Resource constraints have limited the statistical and operational aspects of the NRI. Statistically, an initial goal of obtaining reliable

direct estimators at the county level was judged infeasible (U.S. Department of Agriculture, 2018). Instead, the NRI sample size is determined with the objective of obtaining coefficients of variation below 10% for selected variables, including erosion, in Major Land Resource Areas (MLRA). These geographic domains are typically larger than a county but smaller than a state. One of the goals that motivate our work is to use models and auxiliary information to obtain reliable county-level estimates of sheet and rill erosion rates. Operationally, cost-effective tools making use of any technology that can reduce the staffing and training cost to the NRI data quality assurance are highly valued. One of our works is driven by the development of a web-based tool to assist the NRI quality assurance (QA). More details on the NRI QA process are given in section 1.3.

1.3 Data Quality and Cognitive Load

The NRI surveys have been devoting extensive efforts to quality assurance throughout the data collection and post data-collection phases. Those efforts are intended for data quality control and to ensure (1) the errors in data processing, i.e., coding, keying and editing, and file preparation, are reduced to the largest extent (2) the differences in the observed time series data are because of actual changes instead of upgrades in technologies and protocols. In the data collection phase, besides documented instructions and coordinated training exposed to the teams of the NRI data gatherers, many survey software have been developed by NRCS to perform automated edit and error checks on newly collected and all historical data for *consistency* and *accuracy*. Examples include the 1992 NRI Data Entry Software (Natural Resources Conservation Service, 1994), the 1997 NRI Personal Digital Assistants (Nusser et al., 1996), etc. In the post-data-collection phase, taking the undergoing 2017 NRI production as an example, ISU CSSM, in collaboration with USDA NRCS, prepares an initial 2017 NRI database. This database assembles data collected from 1982 to the present. This initial database is then used to produce a collection of preliminary estimate tables of resources, including land cover/use, soil erosion, irrigation, and other natural resources issues for each of the 48 coterminous states, Hawaii and Puerto Rico. A thorough State Review is then conducted to identify questionable preliminary estimates based on the local situations to the

knowledge of the state resources inventory coordinators. The issue tickets submitted through the State Review Tracker System are treated as evidence of potential data quality problems such as an error in data collection, sample weighting, classification interpretation, etc. The analysts at CSSM address those issues by making further investigation, such as inspecting the affected point data and imagery, merging information from auxiliary data sets, etc. NRCS and CSSM conduct a national review of preliminary estimate tables for the U.S. as a whole after the State Review. At the final stage, NRCS publishes official resource estimates obtained from the ultimate NRI Database. The data quality checking in this post-data-collection phase involves a vast amount of workload on the CSSM analysts to correct the data, coding, and algorithm repeatedly, and to check if the data quality improvements succeed. We developed a web-based table review tool for the CSSM analysts to use for alleviating this operational complexity. We give details on this application in chapter 3.

Besides automating the review process, the web-based tool we develop is also a graphical tool that aims at reducing the cognitive load with effective data visualization. Anderson et al. (2011) evaluates the effectiveness of data visualization by task performance and cognitive load. Task performance can be measured by the response time and accuracy rate for completing a task. Cognitive load is related to working memory (Engle, 2002). Working memory is the short-term memory of humans for data retrieval, processing, and integration to make an executive decision (Baddeley, 1992). There are three layers of cognitive load: germane, intrinsic, and extraneous load (Chandler and Sweller, 1991). Germane load involves the learning of new cognitive schema. Intrinsic load depends on the difficulty in the underlying task and cannot be altered by agency. Extraneous load is related to the design of task in terms of how information is presented. Effective visualization techniques can play an essential role in alleviating extraneous cognitive load. For large and complex databases such as the NRI, visualization tools that allow "drill-down" and user-data interaction are particularly useful in reducing the cognitive load.

1.4 Problem Statement and Our Approach

Conservation practices such as terraces, contour farming, and strip cropping were widely adopted during the 1930s and 1940s to reduce erosion. A critical management practice — conservation tillage has been chosen along with other methods since the 1960s and 1970s. To obtain detailed data on cropping patterns, farming activity, and conservation practices that are not available in the NRI data collection process, the Conservation Effects Assessment Project (CEAP) Farmer Survey was initialized. CEAP is a two-phase survey that involves conducting farmer interviews at a subset of the NRI locations classified as cropland or pastureland. A field-scale physical process model, called the Agricultural Policy Environmental Extender (APEX) in Williams and Izaurralde (2006), was used to simulate day-to-day farming operations, weather events, crop cover, soil properties and their interactions to estimate both sheet and rill erosion and sediment loss. A modified version of the universal soil loss equations is contained in the APEX and used to simulate sheet and rill erosion. The business requirements for the CEAP survey are to measure in-depth quantification of environmental benefits (water and soil quality change) from the conservation program funding in the 2002 Farm Bill and to prioritize conservation practice needs at the local level (U.S. Department of Agriculture, 2017).

The USDA-NRCS has published the CEAP regional cropland assessment of sheet and rill erosion for 12 watersheds. Reliable estimates of sheet and rill erosion at more granular geographic levels such as county are valuable for the NRCS to allocate conservation resources. However, the county sample sizes in the CEAP are too small to support reliable direct estimators. Therefore, we resort to small area estimation techniques (Rao and Molina, 2015) to obtain county-level estimates of mean sheet and rill erosion. Statistically, small area estimation addresses estimation problems where the sample size is small in each domain of interest, but the number of domains is massive. For example, chapter 2 describes a case study where the domains of interest are 64 counties of South Dakota, and the county sample sizes vary from 1 to 30. Small area estimation uses models and auxiliary information to obtain estimators that are more precise than what can be obtained from the survey data alone. To obtain small area predictions, we require auxiliary information for the

full population — all cropland points in South Dakota. Environmental characteristics related to soil erosion factors in Equation 1.1 are reasonable auxiliary information to pursue. Two public USDA data sources — the NRCS Soil Survey Database and the 2006 Cropland Data Layer are correlated with the soil-characteristic (K and S) and cropping-management (C) factors separately, and both are available at each point in the population. These data sources are not equipped with standard measures of statistical reliability, such as standard errors and mean square errors. Nonetheless, they offer other benefits, such as gratis, timeliness, and usability. Those benefits are significant, given the data collection is resource-intensive. The NRCS Soil Survey Database has been used by NRI to obtain soil characteristics information since the 1977 NRI. The Cropland Data Layer is produced using satellite imagery, contains crop-specific classifications at a high spatial resolution, and covers the conterminous United States. In chapter 2, we describe how we define a cropland population frame for South Dakota using the Cropland Data Layer according to the NRI definition of cropland. To account for the properties of sheet and rill erosion measurements in South Dakota, we proposed a small area estimation model and developed closed-form predictors of the area means. Details of our methodology and how we contribute to the existing literature in small area estimation are presented in chapter 2.

The synthesis of the NRCS Soil Survey Database and the Cropland Data Layer is computationally intensive. Borrowing the strength of R (R Core Team, 2019) **shiny** (Chang et al., 2018), we develop a web-based user interface called "VISCOVER" (chapter 4) to understand the joint characteristics of the two databases. The graphical and tabular displays presented in "VISCOVER" may be of interest in their own right. "VISCOVER" is publicly available and now featured in the Shiny Gallery[1]. It has demonstrated its value beyond checking the quality of the auxiliary variables used for our small area estimation model. The tool can visualize the intersection of the two databases for the entire United States, which makes it useful for whoever is interested in exploring the two databases.

[1] https://shiny.rstudio.com/gallery/viscover.html

Effective visualization of large databases can significantly reduce the cognitive load on users. Both the NRCS Soil Survey Database and the Cropland Data Layer have a tremendous amount of data points. Our tool "VISCOVER" makes all data points readily accessible with an interactive map. The NRI Database is expanding fast since the frequency of the NRI data collection became once a year. More frequent NRI releases other than every five year turns feasible. For example, there have been two "interim" releases: the 2010 NRI and the 2015 NRI. In need of balance quality assurance and timely reporting, we develop another web-based **shiny** tool called "iNtr" to facilitate the intensive NRI table review process. Details of the development and general applicability of "iNtr" are given in chapter 3. Having seen how effective data visualization promotes the data quality from the user's perspective in terms of *usability* and *accessibility*, we explore the presentation of the NRI data products beyond printed charts and tables. A short-term objective is to produce an interactive map of sheet and rill erosion estimates in 2006 for South Dakota without disclosing the NRI sample identifiers. As such, we synthesize the National Land Cover Database, the NRI Database, the NRCS Soil Survey Database, the National Elevation Database, and the Cropland Data Layer to obtain sheet and rill estimates at each point that is classified as cropland at a 30-meter spatial resolution covering the entire South Dakota. This cropland population frame is of higher spatial resolution than the one we create for the small area predictions in chapter 2. Detailed procedures and results are presented in chapter 5.

1.5 References

Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., and Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. In *Computer graphics forum*, volume 30, pages 791–800. Wiley Online Library.

Baddeley, A. (1992). Working memory: The interface between memory and cognition. *Journal of cognitive neuroscience*, 4(3):281–288.

Breidt, F. J. and Fuller, W. A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 391–403.

Browning, G. M., Parish, C. L., and Glass, J. (1947). Method for determining the use of limitations of rotation and conservation practices in the control of soil erosion in Iowa. *Agronomy Journal*, 39(1):65–73.

Chandler, P. and Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R.* R package version 1.1.0.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23.

Goebel, J. J. (1998). The National Resources Inventory and its role in U.S. agriculture. In *Agricultural Statistics 2000*, pages 181–192. International Statistical Institute, Voorburg, The Netherlands.

Musgrave, G. W. (1947). The quantitative evaluation of factors in water erosion: a first approximation. *Journal of Soil and Water Conservation*, 2:133–138.

Natural Resources Conservation Service (1994). *National Resources Inventory Training Modules.* Natural Resources Conservation Service.

Nusser, S., Thompson, D., and DeLozier, G. (1996). Using personal digital assistants to collect survey data. *JSM Proceedings, Survey Research Methods Section*, pages 780–785.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3):181–204.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. and Molina, I. (2015). *Small Area Estimation.* John Wiley & Sons.

Renard, K. G., Foster, G. R., Weesies, G. A., McCool, D. K., and Yoder, D. C. (1997). A guide to conservation planning with the revised universal soil loss equation (RUSLE). U.S. Department of Agriculture. Agriculture Handbook NO.703.

Renard, K. G., Yoder, D. C., Lightle, D. T., and Dabney, S. M. (2011). Universal soil loss equation and revised universal soil loss equation. *Handbook of erosion modeling*, pages 137–167.

U.S. Department of Agriculture (2017). Effects of conservation practices on water erosion and loss of sediment at the edge of the field: A national assessment based on the 2003-06 CEAP survey and APEX modeling databases. Natural Resources Conservation Service, Washington, DC.

U.S. Department of Agriculture (2018). Summary report: 2015 national resources inventory. Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa.

Williams, J. R. and Izaurralde, R. (2006). The APEX model. In Singh, V. and Frevert, D., editors, *Watershed Models*, chapter 18, pages 437–482. Taylor & Francis Group.

Wischmeier, W. H. and Smith, D. D. (1965). Predicting rainfall erosion losses from cropland east of the rocky mountains: guide for selection for practices for soil and water conservation. U.S. Department of Agriculture. Agriculture Handbook NO.282.

Wischmeier, W. H. and Smith, D. D. (1978). Predicting rainfall erosion losses: a guide to conservation planning. U.S. Department of Agriculture. Agriculture Handbook NO.537.

Zingg, A. W. et al. (1940). Degree and length of land slope as it affects soil loss in run-off. *Agric. Engng.*, 21:59–64.

CHAPTER 2. EMPIRICAL BAYES SMALL AREA PREDICTION UNDER A ZERO-INFLATED LOG-NORMAL MODEL WITH CORRELATED RANDOM AREA EFFECTS

Xiaodan Lyu[1], Emily J. Berg and Heike Hofmann

Department of Statistics, Iowa State University, Ames, IA

Modified from a manuscript published in *Biometrical Journal*[2]

2.1 Abstract

Many variables of interest in agricultural or economical surveys have skewed distributions and can equal zero. Our data are measures of sheet and rill erosion called RUSLE2. Small area estimates of mean RUSLE2 erosion are of interest. We use a zero-inflated lognormal mixed effects model for small area estimation. The model combines a unit-level lognormal model for the positive RUSLE2 responses with a unit-level logistic mixed effects model for the binary indicator that the response is nonzero. In the CEAP data, counties with a higher probability of nonzero responses also tend to have a higher mean among the positive RUSLE2 values. We capture this property of the data through an assumption that the pair of random effects for a county are correlated. We develop empirical Bayes small area predictors and a bootstrap estimator of the MSE. In simulations, the proposed predictor is superior to simpler alternatives. We then apply the method to construct empirical Bayes predictors of mean RUSLE2 erosion for South Dakota counties. To obtain auxiliary variables for the population of cropland in South Dakota, we integrate a satellite derived land cover map with a geographic database of soil properties. We provide an R Shiny application

[1]Corresponding author

[2]This is modified from the accepted version of the following article: Lyu, X., Berg, E. J., and Hofmann, H. (2020). Empirical Bayes small area prediction under a zero-inflated lognormal model with correlated random area effects, *Biometrical Journal*. This article will be published in final form at https://doi.org/10.1002/bimj.202000029. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [https://authorservices.wiley.com/author-resources/Journal-Authors/licensing/self-archiving.html].

called `viscover` (available at https://lyux.shinyapps.io/viscover/) to visualize the overlay operations required to construct the covariates. On the basis of bootstrap estimates of the mean square error, we conclude that the empirical Bayes predictors of mean RUSLE2 erosion are superior to direct estimators.

2.2 Introduction

In small area estimation, direct estimators for domains of interest are considered unreliable due to small sample sizes. Model-based small area estimators (Rao and Molina, 2015) attain efficiency gains relative to design-based estimators through the use of auxiliary information, often assumed known for the full population. Early and influential approaches to small area estimation use linear mixed effects models with symmetric error distributions (Battese et al., 1988; Fay III and Herriot, 1979).

We consider small area estimation for skewed response variables that are contaminated with zeros. Examples of variables that exhibit these distributional properties include household expenditures on durable goods (Tobin, 1958), agricultural production (Dreassi et al., 2014; Karlberg, 2015), and expenditures for medical care. Min and Agresti (2002) provide additional examples in the context of a review of regression models for semi-continuous data. When distributions are skewed and inflated with zeros, the assumptions of the linear mixed effects model may not hold.

2.2.1 Motivating CEAP RUSLE2 Data

The data that motivate our study are from the Conservation Effects Assessment Project (CEAP), a survey conducted from 2003-2006 through a cooperative agreement between the Natural Resources Conservation Service of the United States Department of Agriculture and Iowa State University. The response variables in CEAP are several measures of soil and nutrient loss on crop fields due to different types of water and wind erosion. We focus on a particular measure of sheet and rill erosion obtained by processing survey responses collected in CEAP through a computer model, the Revised Universal Soil Loss Equation - 2 (RUSLE2).

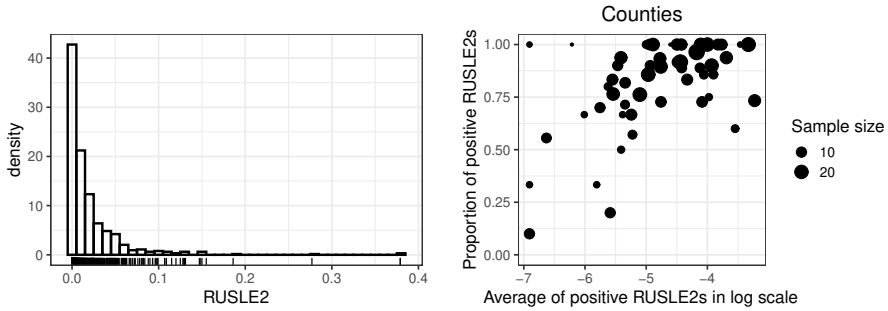

Figure 2.1 Left: Histogram of CEAP sampled RUSLE2s in South Dakota. Right: Scatterplot of county proportions of positive RUSLE2s against county mean of log RUSLE2s.

Figure 2.1 demonstrates the challenges associated with specifying an appropriate small area model for the CEAP RUSLE2 data. The histogram in the left panel of Figure 2.1 shows that the distribution of RUSLE2 is highly skewed right, and approximately 15% of the RUSLE2 measurements are equal to zero. The scatterplot in the right panel of Figure 2.1 shows that counties with higher mean values of positive erosion (in the log scale) are less likely to have observed zeros. An appropriate small area model for the CEAP RUSLE2 data will reflect both of these characteristics: the right skew in the distribution of positive RUSLE2 values and the positive correlation between the county mean of log positive erosion and the county level probability that erosion is nonzero.

2.2.2 Zero-inflated Lognormal Model with Correlated Area Random Effects

We define a zero-inflated lognormal model with correlated random effects to address the challenges in modeling the CEAP RUSLE2 data. Let $i = 1, \ldots, D$ index the small domains (South Dakota counties, for the CEAP application) of interest and $j = 1, \ldots, N_i$ index the units in the population for area i. We assume that a probability sample is selected and let s_i be the set of sampled units for area i with $|s_i| = n_i$. The observed response variable (RUSLE2 in the CEAP application), denoted as y^*_{ij}, satisfies

$$y^*_{ij} = \delta_{ij} y_{ij}, \tag{2.1}$$

where δ_{ij} is a Bernoulli random variable with probability p_{ij} of being 1, and $y_{ij} > 0$. The quantities of interest are the area means defined by $\bar{y}^*_{N_i} = N_i^{-1} \sum_{j=1}^{N_i} y^*_{ij}$ for $i = 1, \ldots, D$. In the model, defined precisely below, we assume y_{ij} satisfies a lognormal mixed effects model and relate p_{ij} to the vector of covariates using a general link function $g(\cdot)$. Assume

$$\log(y_{ij}) = \beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + u_i + e_{ij}, \tag{2.2}$$

and

$$g(p_{ij}) = \alpha_0 + \boldsymbol{z}'_{2,ij}\boldsymbol{\alpha}_1 + b_i, \tag{2.3}$$

where $g(\cdot) : (0,1) \to (-\infty, \infty)$ is a parametric, one-to-one link function, $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, and the pair of random area effects (u_i, b_i) satisfies

$$\begin{pmatrix} u_i \\ b_i \end{pmatrix} \overset{iid}{\sim} \text{BVN}(\mathbf{0}, \boldsymbol{\Sigma}_{ub}), \boldsymbol{\Sigma}_{ub} = \begin{pmatrix} \sigma_u^2 & \sigma_{ub} \\ \sigma_{ub} & \sigma_b^2 \end{pmatrix},$$

where $\sigma_{ub} = \rho\sigma_u\sigma_b$. The pairs of random variables $\{(u_i, b_i) : i = 1, \ldots, D\}$ are mutually independent and are also independent of e_{ij} for $i = 1, \ldots, D$ and $j = 1, \ldots, N_i$. Implications of this dependence structure are that $\delta_{ij} \perp u_i \mid b_i$, and $\delta_{ij} \perp e_{ij} \mid (b_i, u_i)$, where $\perp$ denotes independence. When the link function $g(\mu)$ is chosen to be the logit link $\log(\mu/(1-\mu))$, model (2.3) is called logistic mixed effects model.

The model (2.2-2.3) represents the main features of the data seen in Figure 2.1. The lognormal model (2.2) captures the skewed distribution observed in the left panel of Figure 2.1. The binary part (2.3) accounts for the nontrivial proportion of zeros in the CEAP data set. The important parameter $\rho \in (-1, 1)$, the correlation between b_i and u_i, enables us to represent the correlation between the positive and binary parts observed in the right panel of Figure 2.1 through our model. In a preliminary analysis of the CEAP data, we fit a simplified version of the model (2.2-2.3) with $\rho = 0$. In this "independence model", the lognormal mixed effects model for the positive RUSLE2 values is independent of the logistic mixed effects model for the probability of a nonzero. The sample correlation between random effects from these two independent models for the positive and binary parts suggests that the population correlation is nontrivial. Our model development in this

paper allows us to investigate the significance of the correlation between b_i and u_i formally through a bootstrap interval.

2.2.3 Related Small Area Procedures for Skewed, Binary, or Zero-inflated data

The model (2.2-2.3) advances existing small area literature primarily through the introduction of the correlation parameter ρ in the lognormal framework and secondarily through the development of small area predictors and mean squared error (MSE) estimators in Section 2. We review the literature on which our model and procedure build in this subsection. We first review procedures that address either skewed positive data or binary data individually. We then turn attention to models for zero-inflated data.

Several small area models have been developed for positive, skewed data with support $(0, \infty)$. Most relevant for our work, Berg and Chandra (2014) develop closed form expressions for the empirical Bayes (EB) small area predictor and corresponding mean squared error (MSE) estimator under the assumptions of a lognormal mixed effects model. They apply the mixed effects small area model to construct predictors for a subset of the CEAP data collected in Iowa, a highly cultivated and relatively homogeneous state with a negligible number of observed zeros. For states with greater heterogeneity than Iowa with respect to agriculture, appropriately modeling observed zeros is an important issue for small area estimation. Several studies extend Berg and Chandra (2014). Molina et al. (2018) study MSE estimation for the lognormal small area model of Berg and Chandra (2014) more rigorously, and Zimmermann and Münnich (2018) consider informative sampling. Molina and Rao (2010) and Marhuenda et al. (2017) generalize the lognormal small area model to an arbitrary class of transformations. We considered the Box-Cox family of transformations for the positive component as well as parametrized families of link functions for the binary component. We focus on the log transformation with the logit link function because those transformations fit the CEAP data well, as we demonstrate in Section 2.5. Distributions other than the lognormal distribution for skewed, positive data have also been considered for small area estimation. Berg et al. (2016) compare properties of small area predictors constructed under a mixed effects lognormal model to

those based on a generalized linear mixed model (GLMM) with a gamma response distribution in a simulation study. Jiang (2003) develops EB small area predictors under the assumptions of a general GLMM.

A vast literature also exists on small area estimation for binary data. In the context of a unit-level logistic mixed effects model, González-Manteiga et al. (2007) investigate a bootstrap MSE estimator for a small area predictor of a binary response variable constructed using the penalized quasi-likelihood method of Schall (1991). Hobza and Morales (2016) improve upon González-Manteiga et al. (2007) by using a Laplace approximation for the maximum likelihood estimator and EB prediction. Jiang and Lahiri (2001) develops a rigorous theory for EB prediction under the unit-level logistic mixed effects model, where the parameters are estimated using simulated method of moments. Extensions of the logistic mixed effects model with a single normally distributed random effect have been developed to incorporate a semi-parametric model for the random effects (Marino et al., 2019) and temporal data (Hobza et al., 2018).

Pfeffermann et al. (2008) combine the logistic mixed effects model with a unit-level linear mixed effects model to develop a unit-level small area model for zero-inflated data. In constructing small area estimates of literacy rates, Pfeffermann et al. (2008) use this two-part random effects model to accommodate a large number of observed literacy scores of zero. A logistic mixed effects model is used to describe the probability of a nonzero, and a linear mixed model is used for the nonzero literacy scores. Pfeffermann et al. (2008) use Bayesian methods for inference. Our development addresses many of the challenges associated with a frequentist analysis of a two-part small area model raised in Pfeffermann et al. (2008). The use of a linear model for the positive component, as in Pfeffermann et al. (2008), is undesirable for CEAP because the RUSLE2 data are highly skewed and have nonlinear associations with covariates.

Existing approaches to small area estimation for zero-inflated skewed data include Dreassi et al. (2014) and Chandra and Chambers (2016). Dreassi et al. (2014) develop a zero-inflated gamma model and apply the zero-inflated model to estimate grape wine production in small areas in Italy. Similar to Pfeffermann et al. (2008), Dreassi et al. (2014) use Bayesian methods for inference.

Chandra and Chambers (2016) extend the lognormal model of Berg and Chandra (2014) to a two-part model, where the binary component is independent of the positive, lognormal component. Chandra and Chambers (2016) use penalized quasi-likelihood (PQL) to define a "plug-in" type of predictor for the binary component, similar to González-Manteiga et al. (2007). As discussed in Hobza and Morales (2016), the PQL predictor is not an EB predictor.

When developing a two-part zero-inflated mixed effects model, one issue to consider is the correlation between the random effects in the positive component and the random effects in the binary component. In the model (2.2-2.3), this correlation is represented through the parameter ρ. The Bayesian methods of Dreassi et al. (2014) and Pfeffermann et al. (2008) are flexible enough to allow for a nonzero correlation. Implicit in the "plug-in" approach of Chandra and Chambers (2016) is an assumption that the model for the positive component is independent of the model for the binary component. Frequentist estimation and inference for the two-part model with correlated random effects creates new analytical challenges. Nonetheless, we prefer to adopt a frequentist approach since (1) we lack prior information that would guide an appropriate choice of prior distributions and (2) constructing a posterior distribution for all elements of the large population of cropland in South Dakota is computationally expensive. We address the analytical challenges arising in the frequentist analysis of the zero-inflated lognormal model through our development of a maximum likelihood estimator and an EB predictor in Section 2.3.

2.2.4 Auxiliary Information Acquisition for Small Area Models

To construct EB predictors, we require the covariates to be available for every element of the population. The requirement that the covariates are known for all population elements results from the nonlinear model form. For a linear model (Battese et al., 1988), in contrast, the population means of covariates for each area are sufficient. For nonlinear small area models, the assumption that the covariate is available for all population elements is common. One of the most labor-intensive steps in small area estimation in practice is obtaining the auxiliary variables for the full population. Many applications in small area estimation touch upon that issue lightly (i.e.,

Pfeffermann et al. (2008)) or resort to incomplete auxiliary information, which inflates the MSE of the predictor (Erciulescu and Fuller, 2016).

We provide a thorough treatment of the issue of obtaining population level auxiliary information. We integrate a land cover map derived from satellite imagery with administrative data on soil properties to obtain covariates that are known for the full population and relate to several factors thought to impact erosion. The covariates $\boldsymbol{z}_{1,ij}$ and $\boldsymbol{z}_{2,ij}$ for the positive and binary parts in the model (2.2-2.3) can be the same or different. Besides subject matter and data availability, the choice of covariates for each model component also depends on appropriate variable selection techniques. We use a combination of the science underlying erosion and statistical model comparison techniques to select the acquired covariates for the CEAP data analysis in Section 2.5.

2.2.5 Outline of Our Approach

We develop EB predictors of small area means and corresponding MSE estimators based on the zero-inflated lognormal model defined in (2.2-2.3). The EB predictor has theoretical support because it is an estimator of the minimum MSE unbiased predictor. The EB predictor enables a computationally simple estimator of the leading term in the MSE as well as parametric bootstrap MSE estimators. We present the small area predictors based on the zero-inflated lognormal model with correlated random area effects in Section 2.3. In Section 2.4, we compare our proposed EB predictor to alternatives through simulations. In Section 2.5, we describe how we obtain the auxiliary variables for the full population, apply the method to the CEAP RUSLE2 data collected in South Dakota, and give predictions of county-level average soil erosion with associated standard errors. We develop a visualization tool that demonstrates the overlay operations needed to construct the covariates and also helps verify that the overlay operation works appropriately. The tool is featured in the RStudio Shiny gallery and available online at `https://lyux.shinyapps.io/viscover/`. Concluding remarks are given in Section 2.6. The Appendix provides the computational details for maximum likelihood estimation including the functional form of the score equation. Tables and figures not required to communicate the main ideas are deferred to the Appendix.

We develop a `R` (R Core Team, 2019) package `saezero` (Lyu, 2020) to implement the maximum likelihood estimation and empirical Bayes prediction methods described in this paper.

2.3 Empirical Bayes Prediction for the Zero-inflated Lognormal Model

To define empirical Bayes predictors and MSE estimators under model (2.1), we introduce additional notation. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}', \sigma_u^2, \sigma_b^2, \sigma_e^2, \rho)'$ denote the vector of fixed model parameters to be estimated, where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$ and $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1')'$. As we illustrate through the data analysis, one can add parameters of parametrized families of link functions or transformations.

Assume a sample is selected, and denote the index sets for sampled and nonsampled elements in area i by s_i and $\bar{s}_i$, respectively. Assume that y_{ij}^* is observed for all sampled elements. As discussed in Section 1.3, assume that the covariate $\boldsymbol{z}_{ij} = (\boldsymbol{z}_{1,ij}', \boldsymbol{z}_{2,ij}')'$ is observed for every element in the population, where $\boldsymbol{z}_{1,ij}$ and $\boldsymbol{z}_{2,ij}$, respectively, are the covariates in the models (2.2) for the positive part and in the model (2.3) for the binary part. Denote the observed data as $(\boldsymbol{y}^*, \boldsymbol{z}) = \{y_{ij}^*, j \in s_i, i = 1, \ldots, D\} \cup \{\boldsymbol{z}_{ij} : j = 1, \ldots, N_i; i = 1, \ldots, D\}$. As discussed in Section 1.4, the requirement that the covariate $\boldsymbol{z}_{ij}$ is known for every element of the population results from the nonlinear model form. We explain how we achieve this requirement for the CEAP data analysis in Section 2.5.

2.3.1 Minimum Mean Square Error Predictor of Small Area Mean

A common objective function for defining a predictor in small area estimation is squared error loss. It is straightforward to show that the optimal predictor of $\bar{y}_{N_i}^*$ under squared error loss is $\mathrm{E}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$, the conditional expectation of the population mean given the observed data. We call $\mathrm{E}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$ the minimum mean square error (MMSE) predictor. A synonym for the MMSE predictor used, for example, in Molina and Rao (2010) is the term "best predictor" (BP). The MMSE predictor is also often called the "Bayes" predictor. To stress that the MMSE predictor of $\bar{y}_{N_i}^*$ depends on the unknown $\boldsymbol{\theta}$, we denote the MMSE predictor by $\hat{\bar{y}}_{N_i}^{*\mathrm{MMSE}}(\boldsymbol{\theta})$.

In this section, we derive a form for the conditional expectation defining the MMSE predictor for the specific model (2.2-2.3). The MMSE predictor is difficult to attain because the required

conditional expectation involves a bivariate integral over the distribution $f(u_i, b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$. We present a form for the MMSE predictor as a univariate integral in Theorem 2.1 below. The key idea of the derivation is that the conditional distribution of u_i given b_i and $(\boldsymbol{y}^*, \boldsymbol{z})$ has a standard form. This allows us to transform the bivariate integral defining the MMSE predictor to a univariate integral.

Theorem 2.1. *The minimum MSE predictor of $\bar{y}^*_{N_i}$ under model (2.2-2.3) has the form*

$$\hat{\bar{y}}^{*\text{MMSE}}_{N_i}(\boldsymbol{\theta}) = E\{\bar{y}^*_{N_i} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} = \frac{1}{N_i}\Big[\sum_{j \in s_i} y^*_{ij} + \sum_{j \in \bar{s}_i} E\{y^*_{ij} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}\Big], \tag{2.4}$$

where

$$E\{y^*_{ij} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} = h_{ij}(\boldsymbol{\theta}) \frac{\int \pi_{s_i}(b_i) \exp\left\{-\frac{1}{2v_i}(b_i - m_i)^2\right\} p_{ij}(b_i) \eta(b_i)\, \mathrm{d}b_i}{\int \pi_{s_i}(b_i) \exp\left\{-\frac{1}{2v_i}(b_i - m_i)^2\right\} \mathrm{d}b_i}, \tag{2.5}$$

$$(m_i, v_i) = (\bar{\tilde{r}}_i \rho \sigma_u \sigma_b (\sigma_u^2 + \sigma_e^2/\tilde{n}_i)^{-1}, \sigma_b^2\{(1-\rho^2)\sigma_u^2 + \sigma_e^2/\tilde{n}_i\}(\sigma_u^2 + \sigma_e^2/\tilde{n}_i)^{-1}),$$

$h_{ij}(\boldsymbol{\theta}) = \exp(\beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + \gamma_i \bar{\tilde{r}}_i + \gamma_i \sigma_e^2 \tilde{n}_i^{-1}/2 + \sigma_e^2/2)$, $\pi_{s_i}(b_i) = \prod_{j \in s_i}\left[p_{ij}(b_i)^{\delta_{ij}}\{1 - p_{ij}(b_i)\}^{(1-\delta_{ij})}\right]$, $p_{ij}(b_i) = g^{-1}(\alpha_0 + \boldsymbol{z}'_{2,ij}\boldsymbol{\alpha}_1 + b_i)$, $\eta(b_i) = \exp\{(1-\gamma_i)\rho\sigma_u/\sigma_b b_i\}$, $\gamma_i = (1-\rho^2)\sigma_u^2\{(1-\rho^2)\sigma_u^2 + \sigma_e^2/\tilde{n}_i\}^{-1}$, $\bar{\tilde{r}}_i = \tilde{n}_i^{-1}\sum_{j \in s_i} \tilde{r}_{ij}$, $\tilde{r}_{ij} = \delta_{ij}\{\log(y_{ij}) - \beta_0 - \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1\}$, *and* $\tilde{n}_i = \sum_{j \in s_i} \delta_{ij}$.

Proof. The main steps of the proof are to first find a form for $f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z}))$ and to second find a form for $f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$. By standard properties of bivariate normal distributions, $u_i \mid b_i \sim N(\rho\sigma_u\sigma_b^{-1}b_i, (1-\rho^2)\sigma_u^2)$. Additionally, recall that the assumptions of model (2.2-2.3) imply that $\delta_{ij} \perp e_{ij} \mid b_i$ and $u_i \perp \delta_{ij} \mid b_i$. This allows us to specify the model for $\log(y_{ij})$ as a unit-level model in new parameters. Specifically,

$$\log(y_{ij}) = \beta_0 + \boldsymbol{z}'_{1,ij}\boldsymbol{\beta}_1 + \rho\sigma_u\sigma_b^{-1}b_i + k_i + e_{ij},$$

where $k_i = u_i - \rho\sigma_u\sigma_b^{-1}b_i$, $k_i \sim N(0, (1-\rho^2)\sigma_u^2)$, and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Then standard properties of the linear mixed effects model (i.e., Battese et al. (1988)) imply that $u_i \mid \{b_i, (\boldsymbol{y}^*, \boldsymbol{z})\} \sim N(\tilde{\mu}_{u_i}, \tilde{\sigma}^2_{u_i})$, where $\tilde{\mu}_{u_i} = \gamma_i \bar{\tilde{r}}_i + (1-\gamma_i)\rho\sigma_u\sigma_b^{-1}b_i$ and $\tilde{\sigma}^2_{u_i} = \gamma_i\sigma_e^2/\tilde{n}_i$. When $\tilde{n}_i = 0$ which indicates there is no positive response in area i, it holds that $\gamma_i = 0, \tilde{\mu}_{u_i} = \rho\sigma_u\sigma_b^{-1}b_i$ and $\tilde{\sigma}^2_{u_i} = (1-\rho^2)\sigma_u^2$. To derive the

density of the conditional distribution of b_i given the observed data, note that $f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z}))f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) = f(u_i, b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) \propto f(\boldsymbol{y}_i^* \mid u_i, b_i, \boldsymbol{z})f(u_i \mid b_i, \boldsymbol{z})f(b_i \mid \boldsymbol{z})$ implies

$$f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) \propto \frac{f(\boldsymbol{y}_i^* \mid u_i, b_i, \boldsymbol{z})f(u_i \mid b_i, \boldsymbol{z})f(b_i, \boldsymbol{z})}{f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z}))}, \tag{2.6}$$

where $\boldsymbol{y}_i^* = (y_{i1}^*, \ldots, y_{in_i}^*)'$. Plugging in the known terms on the right side of (2.6) gives $f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) \propto \pi_{s_i}(b_i)\exp\{-(b_i - m_i)^2/(2v_i)\}$. When $\tilde{n}_i = 0$ or $\rho = 0$, it can be shown that $m_i = 0$ and $v_i = \sigma_b^2$. Let $\phi(\cdot)$ denote the probability density function of a standard normal distribution. Then, the conditional probability density function of b_i given the observed data is

$$h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})) = \frac{\frac{1}{\sqrt{v_i}}\pi_{s_i}(b_i)\phi(\frac{b_i - m_i}{\sqrt{v_i}})}{\int \frac{1}{\sqrt{v_i}}\pi_{s_i}(b_i)\phi(\frac{b_i - m_i}{\sqrt{v_i}})\,\mathrm{d}b_i}.$$

We are now able to express the conditional expectation of y_{ij}^* as a univariate integral. By definition,

$$\begin{aligned}
\mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} &= \mathrm{E}\{\delta_{ij}\exp(\beta_0 + \boldsymbol{z}_{1,ij}'\boldsymbol{\beta}_1 + u_i + e_{ij}) \mid (\boldsymbol{y}^*, \boldsymbol{z})\} \\
&= \exp(\beta_0 + \boldsymbol{z}_{1,ij}'\boldsymbol{\beta}_1 + \sigma_e^2/2)\mathrm{E}\{p_{ij}(b_i)\exp(u_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\} \\
&= \exp(\beta_0 + \boldsymbol{z}_{1,ij}'\boldsymbol{\beta}_1 + \sigma_e^2/2)\mathrm{E}[\mathrm{E}\{p_{ij}(b_i)\exp(u_i) \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z})\} \mid (\boldsymbol{y}^*, \boldsymbol{z})] \\
&= h_{ij}(\boldsymbol{\theta})\mathrm{E}\{p_{ij}(b_i)\eta(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\},
\end{aligned}$$

where the second equal sign results from the independence of e_{ij} and e_{ik} for $j \neq k$, and the final equal sign results from the conditional normal distribution of u_i given b_i and the observed data as well as the moment generating function of a normal distribution. The expression for the conditional expectation as a univariate integral in the statement of Theorem 2.1 now follows. □

Observe that we derive a minimum MSE predictor for y_{ij}^* in the original scale of the observed responses, not in the log scale. By double conditioning, $\mathrm{E}[\mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - y_{ij}^*] = \mathrm{E}(\mathrm{E}[\mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})]) = 0$. A correction for the transformation between $\log(y_{ij})$ and y_{ij} is implicit in the development of the minimum MSE predictor.

2.3.2 Empirical Bayes Predictor of Small Area Mean

The minimum MSE predictor is defined through the integral in (2.5) and depends on the unknown $\boldsymbol{\theta}$. An empirical Bayes (EB) predictor is an estimate of the MMSE predictor (Rao and

Molina, 2015, p. 271). Calculating an EB predictor in practice requires estimating the parameter vector $\boldsymbol{\theta}$ and approximating the integral in (2.5). We use maximum likelihood to estimate $\boldsymbol{\theta}$ and approximate the integral in (2.5) using Gauss-Hermite quadrature (Golub and Welsch, 1969).

An advantage of converting bivariate integrals with respect to the conditional distribution of $(u_i, b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})$ to univariate integrals with respect to the distribution of $b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})$ is that one can easily approximate an integral with respect to the distribution $h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$ using a Gauss-Hermite approximation (Smyth, 2014). For an arbitrary function $q(b_i)$, a Gauss-Hermite approximation to $\mathrm{E}\{q(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$ is given by

$$\mathrm{E}_A\{q(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} = \frac{\sum_{k=1}^K \pi_{s_i}(b_{i,k}) q(b_{i,k}) w_k(m_i, v_i)}{\sum_{k=1}^K \pi_{s_i}(b_{i,k}) w_k(m_i, v_i)}, \tag{2.7}$$

where $b_{i,k}$ for $k = 1, \ldots, K$ are specified nodes, and $w_k(m_i, v_i)$ is a weight defined to approximate expectation with respect to a normal distribution with mean m_i and variance v_i. We obtain the nodes and weights using the `R` function `gauss.quad.prob` in the `R` package `statmod` (Giner and Smyth, 2016). We use the approximation (2.7) to define maximum likelihood estimators and to construct EB predictors.

Using the same reasoning used in (2.6) of the derivation of $h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$, we have $f(\boldsymbol{y}_i^* \mid \boldsymbol{\theta}) = [f(u_i \mid b_i, (\boldsymbol{y}^*, \boldsymbol{z})) f(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))]^{-1} f(\boldsymbol{y}_i^* \mid u_i, b_i) f(u_i \mid b_i) f(b_i)$. The terms on the right side of this expression have known forms from the proof of Theorem 2.1. Substitution of these known forms gives the likelihood function $L(\boldsymbol{\theta}) = \prod_{i=1}^D L_i(\boldsymbol{\theta})$ where

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= \frac{\prod_{j \in s_i} [\frac{1}{\sigma_e} \phi(\frac{r_{ij} - u_i}{\sigma_e})]^{\delta_{ij}} \frac{1}{\sqrt{(1-\rho^2)\sigma_u^2}} \phi(\frac{u_i - \rho\sigma_u \sigma_b^{-1} b_i}{\sqrt{(1-\rho^2)\sigma_u^2}}) \pi_{s_i}(b_i) \frac{1}{\sigma_b} \phi(\frac{b_i}{\sigma_b})}{\frac{1}{\tilde{\sigma}_{u_i}} \phi(\frac{u_i - \tilde{\mu}_{u_i}}{\tilde{\sigma}_{u_i}}) \pi_{s_i}(b_i) \frac{1}{\sqrt{v_i}} \phi(\frac{b_i - m_i}{\sqrt{v_i}}) / \int \pi_{s_i}(b_i) \frac{1}{\sqrt{v_i}} \phi(\frac{b_i - m_i}{\sqrt{v_i}}) \mathrm{d}b_i} \\ &= \frac{(1-\gamma_i)^{1/2} (v_i/\sigma_b^2)^{1/2}}{(2\pi\sigma_e^2)^{\tilde{n}_i/2}} \exp\left(\frac{\gamma_i \bar{\tilde{r}}_i^2}{2\sigma_e^2/\tilde{n}_i} + \frac{m_i^2}{2v_i} - \frac{\sum_j \tilde{r}_{ij}^2}{2\sigma_e^2} \right) \int \pi_{s_i}(b_i) \frac{1}{\sqrt{v_i}} \phi(\frac{b_i - m_i}{\sqrt{v_i}}) \mathrm{d}b_i. \end{aligned}$$

A Gauss-Hermite approximation to the log likelihood is then $\ell_A(\boldsymbol{\theta}) = \sum_{i=1}^D \log(L_{i,A}(\boldsymbol{\theta}))$, where $L_{i,A}(\boldsymbol{\theta})$ approximates $L_i(\boldsymbol{\theta})$ by substituting $\int \pi_{s_i}(b_i) {v_i}^{-1/2} \phi((b_i - m_i) {v_i}^{-1/2}) \mathrm{d}b_i$ with $\sum_{k=1}^K \pi_{s_i}(b_{i,k}) w_k(m_i, v_i)$. We define the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ to satisfy $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \ell_A(\boldsymbol{\theta})$. To maximize $\ell_A(\boldsymbol{\theta})$ operationally, we use the `R` function `optim` in the `R` package `stats` (R Core Team, 2019), as explained in more detail in the Appendix. To calculate the empirical Bayes predictor, we replace

the unknown $\boldsymbol{\theta}$ with an estimator and approximate expectations with respect to the distribution $h(b_i \mid (\boldsymbol{y}^*, \boldsymbol{z}))$ using the Gauss-Hermite approximation (2.7). We estimate (2.5), the conditional expectation of a nonsampled y_{ij}^* given the observed data, by

$$\hat{y}_{ij}^{*\text{MMSE}}(\hat{\boldsymbol{\theta}}) = h_{ij}(\hat{\boldsymbol{\theta}})\text{E}_A\{\hat{p}_{ij}(b_i)\hat{\eta}(b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\},$$

where $\hat{p}_{ij}(b_i) = g^{-1}(\hat{\alpha}_0 + \boldsymbol{z}'_{2,ij}\hat{\boldsymbol{\alpha}}_1 + b_i)$ and $\hat{\eta}(b_i) = \exp\{(1-\hat{\gamma}_i)\hat{\rho}\hat{\sigma}_u\hat{\sigma}_b^{-1}b_i\}$. The EB predictor is then defined by

$$\hat{\bar{y}}_{N_i}^{*\text{EB}} = \frac{1}{N_i}\Big\{\sum_{j\in s_i} y_{ij}^* + \sum_{j\in \bar{s}_i} \hat{y}_{ij}^{*\text{MMSE}}(\hat{\boldsymbol{\theta}})\Big\}. \tag{2.8}$$

The term in the EB predictor that provides the predictor of p_{ij} has connections to EB predictors of proportions developed under the unit-level logistic mixed effects model in Hobza and Morales (2016) and Hobza et al. (2018).

2.3.3 MSE of the EB Predictor

Theorem 2.2 gives an expression for the MSE of the EB predictor (2.8) as the sum of the MSE of the MMSE predictor (2.4) and a second term that accounts for the variance of $\hat{\boldsymbol{\theta}}$. We express the MSE of the MMSE predictor as a univariate integral, which we approximate through Gauss-Hermite quadrature after substituting the unknown parameters with estimators. We estimate the second term in the MSE using the parametric bootstrap, a resampling procedure that involves simulating repeatedly from the estimated model (Rao and Molina, 2015, p. 226).

Theorem 2.2. *Under model (2.2-2.3),*

$$\text{MSE}(\hat{\bar{y}}_{N_i}^{*\text{EB}}) = E\{(\hat{\bar{y}}_{N_i}^{*\text{EB}} - \bar{y}_{N_i}^*)^2\} = M_{1i}(\boldsymbol{\theta}) + M_{2i}(\boldsymbol{\theta}),$$

where

$$M_{1i}(\boldsymbol{\theta}) = E\left(\frac{1}{N_i^2}\sum_{j\in\bar{s}_i}\sum_{k\in\bar{s}_i}\Big[E\{y_{ij}^*y_{ik}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\} - E\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}E\{y_{ik}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}\Big]\right),$$

$$E\{y_{ij}^* y_{ik}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\} = \begin{cases} h_{ij} h_{ik} \exp(\tilde{\sigma}_{u_i}^2) E\{p_{ij} p_{ik} \eta(2b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\} & \text{if } j \neq k \\ \\ h_{ij}^2 \exp(\tilde{\sigma}_{u_i}^2 + \sigma_e^2) E\{p_{ij} \eta(2b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z})\} & \text{if } j = k \end{cases},$$

and $M_{2i}(\boldsymbol{\theta}) = E\{(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}} - \hat{\bar{y}}_{N_i}^{*\mathrm{MMSE}})^2\}$.

Proof. Expanding the square defining the MSE, we have $\mathrm{MSE}(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}}) = \mathrm{E}\{(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}} - \bar{y}_{N_i}^*)^2\} = M_{1i}(\boldsymbol{\theta}) + M_{2i}(\boldsymbol{\theta}) + 2M_{3i}(\boldsymbol{\theta})$, where $M_{1i}(\boldsymbol{\theta}) = \mathrm{E}([N_i^{-1} \sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - N_i^{-1} \sum_{j=1}^{N_i} y_{ij}^*]^2)$, and $M_{2i}(\boldsymbol{\theta}) = \mathrm{E}([N_i^{-1} \sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - N_i{}^{-1} \sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}]^2)$. and the cross term $M_{3i}(\boldsymbol{\theta}) = \mathrm{E}([N_i{}^{-1} \sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - N_i{}^{-1} \sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}][N_i{}^{-1} \sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - N_i{}^{-1} \sum_{j=1}^{N_i} y_{ij}^*])$. We consider the cross term $M_{3i}(\boldsymbol{\theta})$. Using double-conditioning,

$$\begin{aligned} M_{3i}(\boldsymbol{\theta}) &= \mathrm{E}\Bigg[\mathrm{E}\Big(\big[\frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - \frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}\big] \\ &\quad \times \big[\frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}^*\big] \mid (\boldsymbol{y}^*, \boldsymbol{z})\Big)\Bigg] \\ &= \mathrm{E}\Bigg[\Big(\frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - \frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}\Big) \\ &\quad \times \mathrm{E}\Big(\big[\frac{1}{N_i}\sum_{j=1}^{N_i} \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\} - \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}^*\big] \mid (\boldsymbol{y}^*, \boldsymbol{z})\Big)\Bigg] = 0. \end{aligned}$$

The leading term $M_{1i}(\boldsymbol{\theta})$ is the MSE of the MMSE predictor (2.4), which has the form $M_{1i}(\boldsymbol{\theta}) = \mathrm{E}[\mathrm{E}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\} - \bar{y}_{N_i}^*]^2 = \mathrm{E}[\mathrm{V}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}]$, where

$$\mathrm{V}\{\bar{y}_{N_i}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\} = \mathrm{E}\left(\frac{1}{N_i^2}\sum_{j\in\bar{s}_i}\sum_{k\in\bar{s}_i}\big[\mathrm{E}\{y_{ij}^* y_{ik}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\} - \mathrm{E}\{y_{ij}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}\mathrm{E}\{y_{ik}^* \mid (\boldsymbol{y}^*, \boldsymbol{z})\}\big]\right).$$

The form for $M_{1i}(\boldsymbol{\theta})$ in Theorem 2.2 now follows from the conditional distributions derived through the proof of Theorem 2.1. □

Before presenting the MSE estimator, we make a couple of remarks on Theorem 2.2. A proof that a cross-term analogous to $M_{3i}(\boldsymbol{\theta})$ for the area-level model is zero is presented on page 141-142 of Rao and Molina (2015). Jiang (2003) uses a decomposition similar to that in Theorem 2.2 for the

MSE of an empirical Bayes small area predictor constructed under a unit-level generalized linear mixed model. Also, see Jiang and Lahiri (2006). The simulations in Section 2.4.2 provide numerical evidence that the MSE is dominated by the leading term $M_{1i}(\boldsymbol{\theta})$, although a formal proof of the orders of the terms in the MSE is beyond the scope of this manuscript.

The estimate of $M_{1i}(\boldsymbol{\theta})$ is an estimate of the observed conditional variance $V\{\bar{y}^*_{N_i} \mid (\boldsymbol{y}^*, \boldsymbol{z})\}$ obtained by substituting the unknown $\boldsymbol{\theta}$ with the MLE $\hat{\boldsymbol{\theta}}$ and using the Gauss-Hermite approximation (2.7) for expectations with respect to the distribution of $b_i \mid (\boldsymbol{y}^*, \boldsymbol{z})$. We call this estimator the "one-step" estimator of the leading term. The "one-step" estimator is defined as

$$\hat{M}_{1i}(\hat{\boldsymbol{\theta}}) = \frac{1}{N_i^2} \sum_{j \in \bar{s}_i} \sum_{k \in \bar{s}_i} \Big[\mathrm{E}_A\{y^*_{ij} y^*_{ik} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} - \mathrm{E}_A\{y^*_{ij} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} \mathrm{E}_A\{y^*_{ik} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \hat{\boldsymbol{\theta}}\} \Big], \tag{2.9}$$

where $\mathrm{E}_A\{y^*_{ij} y^*_{ik} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}$ approximates $\mathrm{E}\{y^*_{ij} y^*_{ik} \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}$ by replacing $\mathrm{E}\{p_{ij}\eta(2b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}$ with $\mathrm{E}_A\{p_{ij}\eta(2b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}$ and $\mathrm{E}\{p_{ij}p_{ik}\eta(2b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}$ with $\mathrm{E}_A\{p_{ij}p_{ik}\eta(2b_i) \mid (\boldsymbol{y}^*, \boldsymbol{z}); \boldsymbol{\theta}\}$ based on the Gauss-Hermite approximation (2.7).

Approximations to $M_{2i}(\boldsymbol{\theta})$ are difficult to derive due to the complexity of the derivatives involved. Therefore, use of the bootstrap, as defined in (2.10) below, to estimate $M_{2i}(\boldsymbol{\theta})$ is more practical. For $b = 1, \ldots, B$, we generate $\{y^{*(b)}_{ij} : i = 1, \ldots, D; j = 1, \ldots, N_i\}$ from the zero-inflated lognormal model (2.1) with the original parameter estimator $\hat{\boldsymbol{\theta}}$. We then obtain a bootstrap sample $\{y^{*(b)}_{ij} : i = 1, \ldots, D; j \in s_i\}$ where s_i is the index set of the originally observed sample units in area i. We let $\hat{\boldsymbol{\theta}}^{(b)}$ denote the vector of parameter estimators obtained from the b-th bootstrap sample. A bootstrap estimator of $M_{2i}(\boldsymbol{\theta})$ is defined by

$$\hat{M}^{\mathrm{Boot}}_{2i} = B^{-1} \sum_{b=1}^{B} (\hat{\bar{y}}^{*(b)\mathrm{EB}}_{N_i} - \hat{\bar{y}}^{*(b)\mathrm{MMSE}}_{N_i})^2, \tag{2.10}$$

where $\hat{\bar{y}}^{*(b)\mathrm{EB}}_{N_i}$ is the EB predictor of $\bar{y}^{*(b)}_{N_i}$ based on $\hat{\boldsymbol{\theta}}^{(b)}$ and $\hat{\bar{y}}^{*(b)\mathrm{MMSE}}_{N_i}$ is the minimum MSE predictor based on $\hat{\boldsymbol{\theta}}$. Both $\hat{\bar{y}}^{*(b)\mathrm{EB}}_{N_i}$ and $\hat{\bar{y}}^{*(b)\mathrm{MMSE}}_{N_i}$ are constructed with the b-th bootstrap sample. We define a semi-bootstrap estimator of the MSE of $\hat{\bar{y}}^{*\mathrm{EB}}_{N_i}$ by

$$\hat{M}^{\mathrm{Semi-Boot}}_{i} = \hat{M}_{1i}(\hat{\boldsymbol{\theta}}) + \hat{M}^{\mathrm{Boot}}_{2i}. \tag{2.11}$$

where $\hat{M}_{1i}(\hat{\boldsymbol{\theta}})$ is the proposed one-step MSE estimator as in (2.9). The MSE estimator (2.11) intentionally ignores the cross term $M_{3i}(\boldsymbol{\theta})$ in the decomposition of the MSE of $\hat{\bar{y}}_{N_i}^{*\mathrm{EB}}$ because this cross term is theoretically 0. Our simulation study confirms that the cross term is indeed negligible. Note that González-Manteiga et al. (2008) consider an estimator with a form analogous to the semi-boot estimator (2.11) in the context of a linear mixed effects model.

The semi-bootstrap estimator of the MSE does not incorporate an estimate of the bias of the one-step MSE estimator for the leading term in the MSE. This bias is defined formally as $\mathrm{E}\{\hat{M}_{1i}(\hat{\boldsymbol{\theta}}) - M_{1i}(\boldsymbol{\theta})\}$. Typically, in small area estimation, the double bootstrap is needed to estimate the bias (Hall and Maiti, 2006). Because we are able to calculate $M_{1i}(\boldsymbol{\theta})$ in one step, we can estimate the bias without requiring the double bootstrap. We assess the importance of this bias and compare the semi-boot estimator of the MSE to a fully parametric bootstrap MSE estimator in simulations.

2.4 Simulations

We evaluate the properties of the proposed EB predictor through Monte Carlo (MC) simulations.We simulate $M = 1000$ populations $\{y_{ij}^* : j = 1, ..., N_i; i = 1, ..., D\}$ from the zero-inflated lognormal model (2.1). We fix the parameters so that the proportion of zeros in the simulation is about 15%, similar to the CEAP data for South Dakota. A single set of covariates $\{z_{ij} : j = 1, ..., N_i; i = 1, ..., D\}$ is generated as mutually independent $N(4.45, 0.055)$ random variables and held fixed across simulations. The covariate z_{ij} is used for both model parts such that $\boldsymbol{z}_{1,ij} = \boldsymbol{z}_{2,ij} = z_{ij}$ The logit link is used for the binary part, and the parameters are $\boldsymbol{\beta} = (-13, 2)', \boldsymbol{\alpha} = (-20, 5)'$, and $(\sigma_u^2, \sigma_e^2, \sigma_b^2) = (0.22, 1.23, 0.52)$. We vary the value of the correlation parameter ρ in the simulations below. There are 64 counties in the CEAP data analysis, so $D = 60$ areas are simulated. In each generated realization, $(N_i, n_i) = (71, 5)$ for 20 of the areas, $(143, 10)$ for another 20 of the areas and $(286, 20)$ for the remaining 20 areas so that $(N, n) = (10000, 700)$. In Section 2.4.1, we evaluate the effect of ignoring the correlation by comparing the EB predictor with simpler alternatives. In Section 2.4.2, we compare the proposed semi-boot

MSE estimator to a more traditional bootstrap MSE estimator and assess the importance of the bias of the estimator of the leading term in the MSE.

2.4.1 Comparison with Alternative Predictors

The four alternative predictors to be compared with the EB predictor (2.8) are computationally simple but lack optimality. The alternative predictors use $\hat{\boldsymbol{\theta}}_0$, the estimator of $\boldsymbol{\theta}$ obtained by fitting the lognormal model (2.2) to the positive responses and independently fitting the model (2.3) to the $\{\delta_{ij} : i = 1, \ldots, D; j = 1, \ldots, n_i\}$. We use the `R` functions `lmer` and `glmer` in the package `lme4` (Bates et al., 2015) for the positive and binary parts, respectively. Restricted maximum likelihood (REML) is used to estimate the parameters of the lognormal component and maximum likelihood is used to estimate the parameters of the binary component. The first alternative predictor is the plug-in predictor (Chandra and Chambers, 2016) defined by replacing $\hat{y}_{ij}^{*\mathrm{MMSE}}(\hat{\boldsymbol{\theta}})$ in (2.8) with $\hat{y}_{ij}^{*\mathrm{PI}} = h_{ij}(\hat{\boldsymbol{\theta}}_0)\hat{p}_{ij}^{\mathrm{PI}}(\hat{\boldsymbol{\theta}}_0)$, where $\hat{p}_{ij}^{\mathrm{PI}}(\hat{\boldsymbol{\theta}}_0) = g^{-1}(\hat{\alpha}_0 + \boldsymbol{z}'_{2,ij}\hat{\boldsymbol{\alpha}}_1 + \hat{b}_i)$ and $\hat{b}_i$ is the predictor of b_i obtained from the penalized quasi-likelihood method of Schall (1991) implemented in the `R` function `glmer`. The second alternative, the "zero-ignored predictor", is the estimated MMSE predictor for the mean of the positive component defined in Berg and Chandra (2014) as $\hat{\bar{y}}_{N_i}^{*\mathrm{ZI}} = \tilde{N}_i^{-1}\{\sum_{j\in s_i} y_{ij}^* + \sum_{j\in \bar{s}_i} h_{ij}(\hat{\boldsymbol{\theta}}_0)\}$, where $\tilde{N}_i = N_i - \sum_{j\in s_i}(1-\delta_{ij}) = N_i - (n_i - \tilde{n}_i)$ is the adjusted population size which excludes the number of zeros in the sample for area i. We refer to the predictor $\hat{\bar{y}}_{N_i}^{*\mathrm{ZI}}$ as the zero-ignored MMSE predictor. Rather than ignoring zeros completely, the third approach begins by adding a small positive constant ϵ, such as the minimum observed positive value, to y_{ij}^*. The third predictor, referred to as the shifted MMSE predictor, is defined by $\hat{\bar{y}}_{N_i}^{*\mathrm{SI}} = \max\{\hat{\bar{y}}_{N_i,\epsilon}^{*\mathrm{MMSE}} - \epsilon, 0\}$, where $\hat{\bar{y}}_{N_i,\epsilon}^{*\mathrm{MMSE}}$ is the MMSE predictor of Berg and Chandra (2014) applied to $y_{ij}^* + \epsilon$. Finally, we define the EB(0) predictor as the EB predictor constructed with $\hat{\boldsymbol{\theta}}_0$ and $\rho = 0$. Thus, EB(0) is an estimated MMSE predictor if the true correlation ρ is 0. The functional forms of the EB(0) and "plug-in" predictors differ only in the predictor of δ_{ij}.

We define the average MC MSE of the EB predictor by

$$\bar{\mathrm{MSE}}_{\mathrm{MC}}(\hat{\bar{y}}_{N_i}^{*\mathrm{EB}}) = \frac{1}{|\{k : N_k = N_i\}|} \sum_{\{k:N_k=N_i\}} M^{-1} \sum_{m=1}^{M} (\hat{\bar{y}}_{N_i}^{*(m)\mathrm{EB}} - \bar{y}_{N_i}^{*(m)})^2, \tag{2.12}$$

where $\hat{\bar{y}}_{N_i}^{*(m)\text{EB}}$ is the EB predictor obtained in the m-th MC simulation, and $\bar{y}_{N_i}^{*(m)}$ is the corresponding population mean simulated in the m-th MC simulation. Note that the average in (2.12) is across MC iterations and areas with the common number of elements. We define the average MSE difference between an alternative predictor (EB(0), PI, ZI and SI) and the EB predictor as

$$\text{MSEDiff}_{\text{MC}}(\hat{\bar{y}}_{N_i}^{*\text{Alt}}) = \frac{1}{|\{k : N_k = N_i\}|} \sum_{\{k:N_k=N_i\}} M^{-1} \sum_{m=1}^{M} (\hat{\bar{y}}_{N_i}^{*(m)\text{Alt}} - \bar{y}_{N_i}^{*(m)})^2 - \bar{\text{MSE}}_{\text{MC}}(\hat{\bar{y}}_{N_i}^{*\text{EB}}), \tag{2.13}$$

where $\hat{\bar{y}}_{N_i}^{*(m)\text{Alt}}$ is an alternative predictor of the i-th area mean in the m-th MC simulation. To account for the variance of $\text{MSEDiff}_{\text{MC}}(\hat{\bar{y}}_{N_i}^{*\text{Alt}})$ as a MC approximation for the true difference, we define a margin of error as 1.96 times the MC standard error. (Detailed definition in Section 2.8.4 in Appendix.) A 95% confidence interval for each average MSE difference $\bar{\zeta}$ is given by $(\bar{\zeta}-\omega, \bar{\zeta}+\omega)$ where ω denotes the corresponding margin of error.

Table 2.1 Average MSE differences ($\times 10^5$) between the alternative predictors (EB(0), EB predictor assuming $\rho = 0$; PI, plug-in predictor; ZI, zero-ignored MMSE predictor; SI, shifted MMSE predictor) and the EB predictor. The associated Monte Carlo margins of error are presented in parentheses.

predictor	EB(0)	PI	ZI	SI
Avg. for $n_i = 5$	1.17 (0.99)	1.20 (0.99)	4.28 (0.96)	71.58 (7.32)
Avg. for $n_i = 10$	0.89 (0.72)	0.90 (0.72)	3.57 (0.70)	78.04 (8.17)
Avg. for $n_i = 20$	0.64 (0.35)	0.63 (0.35)	2.76 (0.36)	71.29 (6.54)

Table 2.1 reports the average MSE difference (multiplied by 10^5) between each alternate predictor and the EB predictor for $\rho = 0.9$. As a reference, the average MC MSEs (multiplied by 10^5) of the EB predictor defined in (2.12) are 24.09, 15.92 and 9.27, respectively for $n_i = 5, 10$ and 20. The relative MSE difference (ratio of the average MSE difference to the average MSE of the EB predictor) tends to increase as the area sample size increases. The zero-ignored MMSE predictor is clearly inefficient with relative MSE differences between 18% and 30%, implying that simply tossing the zeros is not a good approach. Interestingly, the shifted MMSE predictor introduces an even larger error by adding a small positive constant before the log-transformation and corrupting the normality. As a consequence of mis-specification of the correlation structure, the relative MSE

differences for the EB(0) and "plug-in" predictors are about 5% to 7%. The margins of error in Table 2.1 indicate that the increase in MSE of the EB(0) and "plug-in" predictors relative to the EB predictor is significant.

Table 2.2 Average MSE differences ($\times 10^5$) between the EB(0) predictor and the EB predictor as the true correlation ρ of the random area effects between the positive part and the binary part changes. The associated Monte Carlo margins of error are presented in parentheses.

ρ	-0.9	-0.6	-0.3	0	0.3	0.6	0.9
Avg. for $n_i = 5$	2.43 (0.25)	0.97 (0.18)	0.16 (0.12)	-0.14 (0.08)	-0.05 (0.09)	0.45 (0.13)	1.17 (0.19)
Avg. for $n_i = 10$	2.49 (0.23)	0.79 (0.16)	0.01 (0.13)	-0.14 (0.08)	-0.03 (0.07)	0.27 (0.12)	0.89 (0.16)
Avg. for $n_i = 20$	1.54 (0.13)	0.39 (0.10)	-0.02 (0.06)	-0.08 (0.04)	0.02 (0.04)	0.23 (0.06)	0.64 (0.09)

Given that the EB(0) predictor is most competitive with the EB predictor when $\rho = 0.9$, we next assess the effect of the size of ρ on the relative efficiency of the EB predictor to the EB(0) predictor. Table 2.2 contains the average MSE differences (multiplied by 10^5) between the EB(0) and EB predictor for values of ρ ranging from -0.9 to 0.9. When $|\rho| = 0.3$, the MSE of the EB(0) predictor does not differ significantly from that of the EB predictor. As $|\rho|$ increases to 0.6 or 0.9, the EB(0) predictor has larger MSE than the EB predictor. Because the EB predictor is over-parametrized when $\rho = 0$, the EB(0) predictor is superior to the EB predictor when the true correlation is indeed zero.

The comparison to alternative predictors above focuses on predictors constructed under the assumptions of a lognormal model. The remaining predictor for zero-inflated skewed data in the small area literature to which we do not compare is the zero-inflated gamma model presented in Dreassi et al. (2014). We do not include their model in our study because the lognormal and gamma models are two different models for the response distribution. We expect predictors constructed under the assumptions of the lognormal model to outperform predictors constructed under the assumption of a gamma model if the lognormal model is true (and vice versa). We consider an investigation into properties of predictors constructed under a misspecified model assumption to be an area for future work. We focus on predictors constructed under the assumptions of the

lognormal model because the lognormal model appears adequate for the CEAP data, as discussed further in Section 2.5.

2.4.2 Evaluation of the Parametric Bootstrap MSE Terms

We use a traditional parametric bootstrap method to define another MSE estimator besides the one-step and semi-boot MSE estimators. A fully parametric bootstrap MSE estimator is defined by

$$\hat{M}_i^{\text{Boot}} = B^{-1}\sum_{b=1}^{B}(\hat{\bar{y}}_{N_i}^{*(b)\text{EB}} - \bar{y}_{N_i}^{*(b)})^2, \tag{2.14}$$

where $\hat{\bar{y}}_{N_i}^{*(b)\text{EB}}$ is the same as defined in (2.10) and $\bar{y}_{N_i}^{*(b)} = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}^{*(b)}$ is the i-th small area mean based on the b-th bootstrap population. Fully parametric bootstrap estimators such as (2.14) are suggested in Molina and Rao (2010) for transformed parameters, in González-Manteiga et al. (2007) for a logistic mixed effects model, and in González-Manteiga et al. (2008) for a linear mixed effects model.

The bootstrap MSE estimator (2.14) of the EB predictor can be decomposed into three parts as $\hat{M}_i^{\text{Boot}} = \hat{M}_{1i}^{\text{Boot}} + \hat{M}_{2i}^{\text{Boot}} + 2\hat{M}_{3i}^{\text{Boot}}$, where $\hat{M}_{1i}^{\text{Boot}} = B^{-1}\sum_{b=1}^{B}(\hat{\bar{y}}_{N_i}^{*(b)\text{MMSE}} - \bar{y}_{N_i}^{*(b)})^2$, $\hat{M}_{2i}^{\text{Boot}}$ is defined as in (2.10) and a bootstrap estimator of the cross term $\hat{M}_{3i}^{\text{Boot}} = B^{-1}\sum_{b=1}^{B}(\hat{\bar{y}}_{N_i}^{*(b)\text{MMSE}} - \bar{y}_{N_i}^{*(b)})(\hat{\bar{y}}_{N_i}^{*(b)\text{EB}} - \hat{\bar{y}}_{N_i}^{*(b)\text{MMSE}})$. A bootstrap estimator of the bias of the estimator of the leading term in the MSE is defined by $\hat{M}_{1i}^{\text{Bias}} = B^{-1}\sum_{b=1}^{B}\hat{M}_{1i}(\hat{\boldsymbol{\theta}}^{(b)}) - \hat{M}_{1i}(\hat{\boldsymbol{\theta}})$, where $\hat{M}_{1i}(\hat{\boldsymbol{\theta}}^{(b)})$ is the one-step MSE estimator constructed with the original sample and the b-th bootstrap parameter estimate $\hat{\boldsymbol{\theta}}^{(b)}$. The semi-boot MSE estimator (2.11) intentionally ignores the cross-term $M_{3i}(\boldsymbol{\theta})$ because the expected value of the cross-term is zero, as explained in the proof of Theorem 2.2. The one-step MSE estimator (2.9) additionally ignores M_{2i}, the error due to estimating model parameters. Both the fully parametric and semi-boot MSE estimator ignore M_{1i}^{Bias}, the bias of the estimator of the leading term. Therefore, an assessment of the importance of the three MSE components M_{3i}^{Boot}, M_{2i} and M_{1i}^{Bias} is relevant.

The MSE components and MSE estimators are evaluated through simulations, where the simulation configuration is the same as Section 3.1 with $\rho = 0.9$, a MC sample size of 1000, and a

bootstrap sample size of $B = 100$ in each MC sample. We group by sample size and compute the MC means of the ratios of $\hat{M}_{3i}$, $\hat{M}_{2i}$, and $\hat{M}_{1i}^{\text{Bias}}$ to the fully parametric bootstrap MSE estimator (2.14). (We defer full tabular output to Table 2.6 in Appendix.) Accounting for parameter estimation through $\hat{M}_{2i}$ accounts for about 5-6% of the total MSE. The bias of estimating the leading term is less important, as $\hat{M}_{1i}^{\text{Bias}}$ accounts for -2.14%, -2.98%, and -3.06% of $\hat{M}_i^{\text{Boot}}$ on average when the sample size is 5, 10, and 20, respectively. The average ratios for the cross-term $\hat{M}_{3i}$ are approximately -0.2%, supporting the theory that the cross-term is negligible.

Table 2.3 Average biases ($\times 10^5$) and coverage percentages (CP) of nominal 95% normal theory prediction intervals for the MSE estimators (one-step, bootstrap, semi–bootstrap). The associated Monte Carlo margins of error are presented in parentheses.

	One-Step		Bootstrap		Semi-Boot	
	Bias	CP	Bias	CP	Bias	CP
Avg. for $n_i = 5$	-0.66 (0.51)	94.94 (0.33)	0.48 (0.48)	94.54 (0.33)	0.46 (0.53)	95.36 (0.31)
Avg. for $n_i = 10$	-0.96 (0.32)	94.88 (0.32)	-0.00 (0.30)	94.08 (0.33)	-0.14 (0.33)	95.48 (0.30)
Avg. for $n_i = 20$	-0.68 (0.19)	94.47 (0.35)	-0.16 (0.16)	93.78 (0.32)	-0.12 (0.19)	95.37 (0.31)

We next consider the biases for the MC MSE (2.13) and empirical coverages of normal theory prediction intervals for the three MSE estimators (one-step, bootstrap and semi-boot). The analysis of the MC bias and empirical coverage presented in Table 2.3 provides further support for the semi-boot MSE estimator. Accounting for parameter estimation significantly reduces the negative bias of the one-step MSE estimator and keeps the coverage rate close to the nominal 95% level. The bootstrap MSE estimator is unbiased but has significantly lower coverage than the one-step MSE estimator. We believe that the semi-boot MSE estimator is superior to the fully parametric bootstrap MSE estimator because the fully parametric bootstrap MSE estimator unnecessarily estimates the cross-term.

2.5 Estimating Sheet and Rill Erosion for the Conservation Effects Assessment Project

The Conservation Effects Assessment Project (CEAP), detailed in Goebel (2012), is a collection of four types of surveys that assess conditions on cropland, grazing lands, wildlife, and wetlands. The cropland assessment, the focus of our work, monitors soil and chemical loss from crop fields. The sample for the cropland assessment is a subset of locations classified as cropland in a larger survey called the National Resources Inventory (Nusser and Goebel, 1997). Policies that allocate resources at a local level can benefit from small area estimates of variables collected in CEAP. Therefore, we consider county estimation for CEAP in South Dakota.

The response variable of interest for our study (y^*) is a measure of sheet and rill erosion in tons per year obtained from the Revised Universal Soil Loss Equation, version 2 (RUSLE2), one of the outputs of the Agricultural Policy/Environmental eXtender (APEX) model (Williams and Izaurralde, 2006). RUSLE2 is intended to improve upon the Universal Soil Loss Equation (USLE), a traditional measure of sheet and rill erosion, published in Wischmeier and Smith (1965). USLE includes several fundamental factors accounting for soil erosion, some of which we use in our model for the computationally more complex RUSLE2. Specifically, the USLE equation is defined by

$$\mathrm{A} = \mathrm{R}\ \mathrm{K}\ \mathrm{L}\ \mathrm{S}\ \mathrm{C}\ \mathrm{P}, \tag{2.15}$$

where A is USLE soil loss per unit area, R is the rainfall factor, K is the soil erodibility index, L is the slope-length, S is the slope-gradient, C is the cropping-management effect relative to a fallow field, and P is the erosion-control practice factor. The USLE equation (2.15) implies a log-linear relationship between the sheet and rill erosion and the USLE factors. RUSLE2 maintains many of the fundamental concepts of the USLE equation but uses more granular weather information and a more detailed cover-management subfactor.

The proportions of non-zeros and sample sizes of the CEAP RUSLE2 data collected in the 66 counties of South Dakota tend to increase from west to east, as shown in Figure 2.5 of the Appendix. The east of South Dakota has more sampled units than the west, because most of the cropland in

South Dakota is located to the east of the Missouri river. The sample size is less than 5 for most of the counties in the west. Butte county and Lawrence county are considered out of scope because 80% of the lands in those two counties are rangeland or federal land. Among the 64 counties in the cropland population of interest, seven of them have no positive RUSLE2 measurement. The overall proportion of zeros in the sampled cropland RUSLE2 measurements is 15%.

2.5.1 Auxiliary Variables and Population Predictions for CEAP

We require auxiliary variables that are known for the full population of cropland locations in South Dakota. A simple and complete list of crop locations in South Dakota does not exist. Compiling the necessary auxiliary information requires combining three additional sources besides CEAP: National Resources Inventory (NRI), Soil Survey Geographic Database (SSURGO) and Cropland Data Layer (CDL). We use these sources for two purposes. The first is modeling cropland RUSLE2. The second is defining the full population of crop fields for which predictions are required. These three sources provide covariates that are known for the full population of cropland in South Dakota and are related to the USLE factors.

The first covariate is the log of the USLE R-factor, $\log(R)$. We obtain the USLE rainfall factor from NRI. The R-factor is nearly constant within a county due to the uniform climate conditions in the midwestern United States. We define a county level auxiliary variable by the the most frequently recorded NRI R-factor in each county of South Dakota.

We obtain covariates for USLE K and S factors from SSURGO (U.S. Department of Agriculture, 2020b), a detailed map with soil properties for mapunits covering most of the United States. One of the properties is KWFACT, an erosion factor modified by the presence of the rock fragment and ranging from 0.02 to 0.69. We use the values of $\log(K)$, the log of the KWFACT as approximations for the log of the USLE K-factors. Another property is SLOPE_R, the representative slope gradient in percentage. To convert slope gradient to a USLE S-factor, Smith and Wischmeier (1957) proposed $S = (0.43 + 0.30s + 0.043s^2)/6.613$, where s is the slope gradient and S is the USLE S-factor. We use $\log(S)$ as a covariate.

Cropland Data Layer (U.S. Department of Agriculture, 2020a) is a geo-referenced and crop-specific land cover data layer. Given the CEAP sample is collected by the 2003-2006 CEAP survey and South Dakota was not included in the CDL project until 2006, we use the 2006 CDL data for South Dakota. Corn, soybean, spring wheat and winter wheat are the four dominant crops in South Dakota according to the 2006 CDL. Thus we classify each 2006 CDL pixel at a spatial resolution of 56 meters into one of these four categories or the "remainder" category. The crop-based CDL classification is our best available approximation to the USLE C-factor.

These three sources (NRI, SSURGO and CDL) result in seven possible explanatory variables: $\log(R)$, $\log(K)$, $\log(S)$, and four indicator variables for CDL membership in corn, soybeans, spring wheat, or winter wheat. (The baseline category is the remainder of CDL crop categories.) These variables provide $\boldsymbol{z}_{ij}$ in the small area model. Although possible covariates are chosen to be related to USLE factors, RUSLE2 is a complicated function of a lot of factors which may not be available for the full population. The factors we represent through possible covariates are those we are able to populate at each cropland location.

The population of interest is the collection of all crop fields in South Dakota. We define the population of interest to be locations with CDL pixels that are classified as cropland according to the NRI definition of cropland (NRI Broad-use Category 1 or 2, defined in terms of specific crops in Table 2.7 [3] of the Appendix). Our constructed population consists of about 20 million CDL pixels of cropland in 6,615 soil map units. Several soil map units overlap with more than one county, so we overlaid those map unit polygons with the county shapefiles to obtain their intersections. After excluding one sampled point which falls out of the 6,615 soil map units classified as eligible, the sample dataset contains a total of $n = 641$ different geographic locations in the CEAP sample for South Dakota.

Obtaining the soil map unit classification and subsequently defining a prediction for 20 million distinct CDL pixels is computationally prohibitive. To facilitate the computations, we exploit the fact that the predictors are constant within an intersection of a soil map unit, CDL crop type,

[3] The authors thank Gabriel Demuth for his help with the creation of this table.

and county. This simplification allows us to express the predictor as a weighted sum of a smaller number of distinct entities. We refer the reader to Section 2.8.2 of the Appendix for an explanation of how we incorporate weights to reduce the computational complexity for the CEAP analysis.

2.5.2 CEAP Zero-inflated Lognormal Model Fitting

Table 2.4 The parameter estimates and associated bootstrap standard errors (SE) for the zero-inflated lognormal model fit to the cropland CEAP RUSLE2 data.

	Positive Part		Binary Part	
	Estimate	(SE)	Estimate	(SE)
logR	2.19	(0.36)	4.94	(0.72)
logK	0.52	(0.23)		
logS	0.49	(0.08)	0.38	(0.21)
is.soybean			0.71	(0.33)
is.sprwht			0.98	(0.52)
Var: county	0.22		0.47	
Var: residual	1.23			
Correlation	0.77			

We fit the zero-inflated lognormal model (2.1) with the logit link to the observed data $(\boldsymbol{y}^*, \boldsymbol{z})$. To select the covariates, we apply backward variable selection to the model (2.2) for the positive part and to the model (2.3) for the binary part separately. We use $\Delta(\text{AIC}) = 0.5$ as a threshold for variable exclusion. As shown in Table 2.4, the best predictors according to this criterion are *logR*, *logK* together with *logS* for the positive part and *logR*, *logS*, *is.soybean* together with *is.sprwht* for the binary part. When the zero-inflated lognormal model assuming correlation between the random effects with all possible explanatory variables is fit to the CEAP RUSLE2 data, the coefficients associated with the selected covariates remain significant. Table 2.4 contains the maximum likelihood estimates of the coefficients and the corresponding bootstrap standard errors based on the assumed model with estimated ρ. A nominal 95% confidence interval for ρ based on the lower and upper 2.5% percentiles of parametric bootstrap distribution of $\hat{\boldsymbol{\theta}}$ (with bootstrap sample size of 500) is $(0.21, 0.99)$. In consistency with the soil loss equation (2.15), the R-factor, K-factor and S-factor are positively related with soil loss per year. The probability that cropland RUSLE2 is

positive, or there exists soil loss, is estimated to be significantly higher for the crop type of soybean or spring wheat relative to the remainder category.

The simple logit link and log transformations used to fit the model are special cases of parametric families of link functions and transformations. We considered alternate link functions besides the logit for the binary part and transformations in the family of Box-Cox power transformations (Box and Cox, 1964) for the positive part. The analysis detailed in Section 2.8.3 of the Appendix provides no evidence against the simple choices of the logit link and logarithmic transformation. We therefore use the logit link and the logarithmic transformation for the CEAP data analysis.

We assess the goodness of fit of the model using residuals for the positive part and a Hosmer-Lemeshow test for the binary part. We constructed marginal and conditional residuals defined as

$$r_{ij,\text{marg}} = (\log(y_{ij}) - \boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}})/\sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$$

and as

$$r_{ij,\text{cond}} = (\log(y_{ij}) - \boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}} - \hat{E}[u_i \mid (\boldsymbol{y}^*, \boldsymbol{z})])/\sqrt{\hat{\sigma}_e^2}$$

respectively, where $\hat{\boldsymbol{\beta}}$ is the MLE estimate of the regression coefficient for the positive part and $\hat{E}[u_i \mid (\boldsymbol{y}^*, \boldsymbol{z})]$ denotes the estimated conditional expected value of u_i given the data (after integrating over the conditional distribution of b_i). We plotted the residuals against $\boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}}$ and against $\boldsymbol{z}'_{1,ij}\hat{\boldsymbol{\beta}} + \hat{E}[u_i \mid (\boldsymbol{y}^*, \boldsymbol{z})]$. The residual plots (presented in Figure 2.6 of the Appendix) reveal that the RUSLE2 values are rounded so that the same values of $\log(y_{ij})$ are duplicated in the data set. Otherwise, the residual plots show no important trends as a function of the fitted values. The p-values of Shaprio-Wilk tests for normality exceed 0.05, providing essentially no evidence to reject the normality assumption. For the binary part, we construct Hosmer-Lemeshow tests (Hosmer Jr and Lemeshow, 2000) using an estimate of $E[p_{ij} \mid (\boldsymbol{y}^*, \boldsymbol{z})]$ as the predicted probability. The p-values (presented specifically in Figure 2.7 of the Appendix) for group sizes ranging from 5 to 15 are all above 0.05. These analyses provide support for the zero-inflated lognormal model.

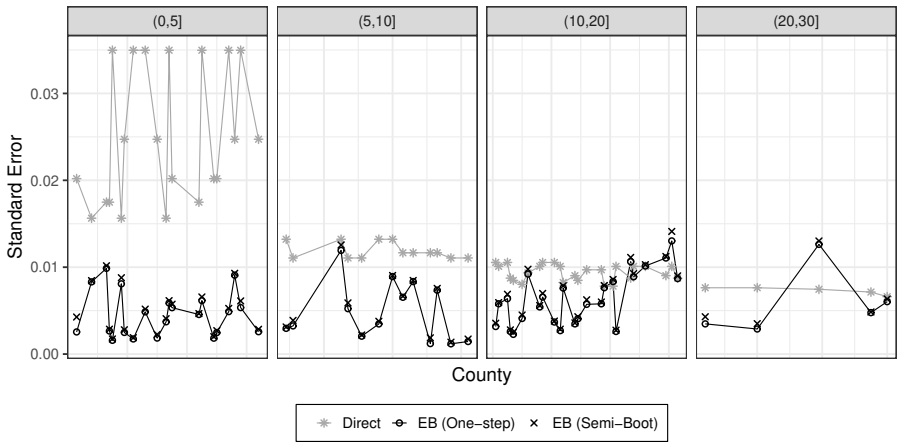

Figure 2.2 Standard errors of the EB predictor and the direct estimator (sample mean) of mean cropland RUSLE2 for the South Dakota counties. Standard errors for the EB predictor are square roots of the one-step or semi-boot MSE estimator. The pooled standard error is used for the direct estimator. The comparisons are grouped by sample size labeled on the top.

2.5.3 CEAP Empirical Bayesian Predictions

We obtain the EB predicted mean cropland RUSLE2 in the 64 sampled counties of South Dakota with the fitted zero-inflated lognormal model with correlated random effects. We compare the EB predictors and corresponding MSE estimators to direct estimators and associated standard errors. The direct estimator is the sample mean. Scatterplots of the EB predictors (presented in Figure 2.5 of the Appendix) against the direct estimators show a strong, linear association. The strength of the association increases as the area sample size increases, an expected trend because increasing the sample size should reduce the effect of modeling on the predictors. We define a standard error of the direct estimator by $\text{SE}_{\text{pool}}(\hat{\bar{y}}^*_{N_i}) = \sqrt{S^2/n_i}$, where S^2 is the pooled within-county variance given by $S^2 = (n-D)^{-1}\sum_{i=1}^{D}\sum_{j=1}^{n_i}(y^*_{ij} - \bar{y}^*_{i.})^2$ and $\bar{y}^*_{i.} = n_i^{-1}\sum_{j=1}^{n_i} y^*_{ij}$. We use the pooled variance S^2 because the sample variance for the i-th county is undefined if the sample size n_i is 1. Figure 2.2 shows the standard errors of the direct estimators along with the square root of the one-step MSE estimator and the square root of the semi-boot MSE estimator constructed with bootstrap size $B = 100$. The standard errors computed by the one-step method are close to the semi-boot

method. With either the one-step or the semi-boot standard error, as depicted in Figure 2.2, the proposed EB predictor is more precise than the direct estimator across different sample sizes. The reduction in standard error tends to be greater for the group of counties with smaller sample sizes.

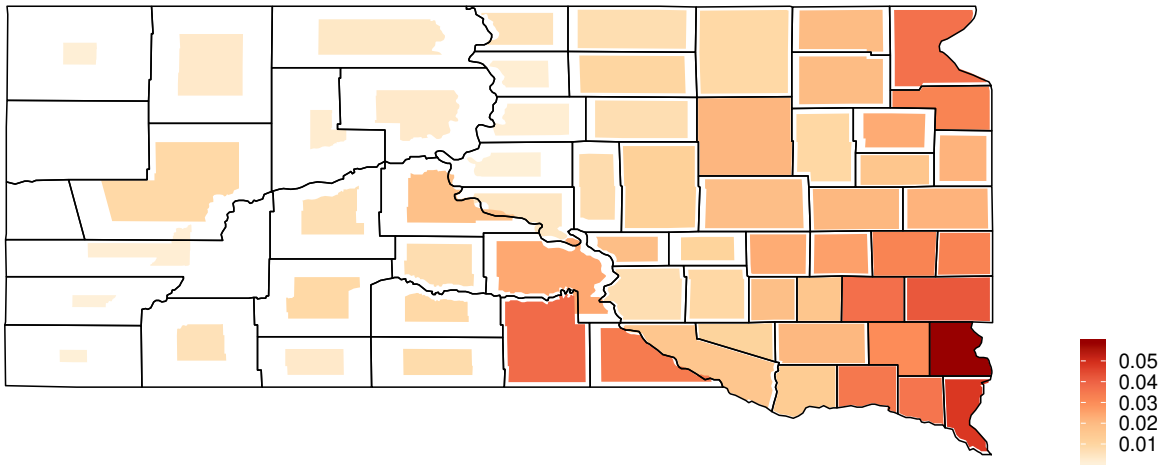

Figure 2.3 Cartogram of the EB predicted county means of cropland RUSLE2 in South Dakota. Darker shade indicates severer soil sheet and rill erosion. Smaller shrinkage indicates smaller coefficient of variance. This figure appears in color in the electronic version of this article.

Figure 2.3 gives a cartogram of the EB predicted mean cropland RUSLE2 in the 64 sampled counties of South Dakota. The darkness of the shade corresponds to the magnitude of the prediction. The relative size of the shaded region to the total area of the county is inversely related to the estimated coefficient of variation (CV), a graphical technique intended to draw the reader's attention toward more reliable estimates. The predicted erosion tends to be highest in magnitude and most reliable (smallest estimated CV) in the eastern, particularly the southeastern, portion of South Dakota. The increase in predicted erosion from west to east occurs partly because rainfall and the prevalence of soybeans are greater in the east than in the west of South Dakota. The decrease in

estimated CV from west to east likely occurs because the CEAP sample is more dense in more highly cultivated areas.

2.6 Conclusion

A zero-inflated lognormal model with correlated random area effects has been proposed for modeling soil loss data, and empirical Bayesian small area estimators have been derived from the model. By deriving the conditional distribution of the random components in the model, we are able to use the maximum likelihood method to estimate the whole parameter vector even with the correlation coefficient ρ. The model assumptions appear reasonable for the data analysis of the cropland CEAP RUSLE2 measurements collected in South Dakota. The correlation of the random area effects between the two parts is estimated to be positive and moderately large. The model based EB predictors of area means have values near the direct estimator (sample mean) when the sample size is relatively large. The standard errors of the EB predictor are estimated to be much smaller than the direct estimator, especially for counties with only a few samples.

In the simulation study, we compare our proposed EB predictor of area means with the EB(0), plug-in, zero-ignored MMSE and shifted MMSE predictor. Simply tossing all the zeros in the sample would result in poor estimation results, and shifting the responses by the amount of the smallest positive quantity leads to even larger mean square errors. The plug-in estimator has comparable performance with the EB(0) predictor, both of which assume independence between the two parts and have moderately larger MSE than the proposed EB predictor when the independence assumption is violated. For the proposed EB predictor, we give three ways to estimate the mean square error: an analytic estimator and two bootstrap estimators. The analytic approximation is based on the observed conditional variance. The two bootstrap MSE estimators take into account the variance due to parameter estimation. The semi-bootstrap MSE estimator that adds the variance due to parameter estimation to the estimator of the conditional variance outperforms the other MSE estimators. This MSE estimator differs from the full parametric bootstrap MSE estimator because it uses the fact that the prediction error is uncorrelated with the difference between the

EB predictor and the MMSE predictor. For both the simulation and the data analysis, the sample size is large enough that the observed conditional variance and the semi-boot MSE estimator are close.

In the Appendix, we also show that our proposed estimators can accommodate any parametrized link functions for the binary part of the model as well as parametrized transformations for the positive part. Although a generalized link function for the positive part may prevent one from using the analytic form of the proposed EB predictor in the paper, the conditional distributions of the random components are implicitly using the link functions. Therefore, one can easily adapt the simulation-based method of Molina and Rao (2010) to obtain the predictions. In the data analysis, it is shown that the proposed predictor can be easily adapted to incorporate weights for averaging across the individual level predicted values.

The data analysis also suggests areas for future work. The observed cropland RUSLE2 measurements are rounded to the three decimal places. Modifying our EB approach to account for the discrete nature of the data is an area for future study. The lognormal model for the positive component, though reasonable for South Dakota, may not hold for all states and response variables of interest. Investigation of compromises between the fully parametric approach considered here and the semiparametric quantile regression model of Berg and Lee (2019) is a possible avenue for future research.

2.7 References

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, 68(2):357–363.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Berg, E. and Chandra, H. (2014). Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis*, 78:159–175.

Berg, E., Chandra, H., and Chambers, R. (2016). *Small Area Estimation for Lognormal Data*, chapter 15, pages 279–298. John Wiley & Sons, Ltd.

Berg, E. and Lee, D. (2019). Small area prediction of quantiles for zero-inflated data and an informative sample design. *Statistical Theory and Related Fields*, 3(2):114–128.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Chandra, H. and Chambers, R. (2016). Small area estimation for semicontinuous data. *Biometrical Journal*, 58(2):303–319.

Dreassi, E., Petrucci, A., and Rocco, E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in tuscany. *Biometrical Journal*, 56(1):141–156.

Erciulescu, A. L. and Fuller, W. A. (2016). Small area prediction under alternative model specifications. *Statistics in Transition new series*, 17(1):9–24.

Fay III, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.

Giner, G. and Smyth, G. K. (2016). statmod: probability calculations for the inverse gaussian distribution. *R Journal*, 8(1):339–351.

Goebel, J. (2012). Statistical methodology for the NRI-CEAP cropland survey.

Golub, G. H. and Welsch, J. H. (1969). Calculation of gauss quadrature rules. *Mathematics of computation*, 23(106):221–230.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics & data analysis*, 51(5):2720–2733.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78(5):443–462.

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):221–238.

Hobza, T. and Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of official statistics*, 32(3):661–692.

Hobza, T., Morales, D., and Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *Test*, 27(2):270–294.

Hosmer Jr, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. John Wiley & Sons.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111(1-2):117–127.

Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53(2):217–243.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15(1):1.

Karlberg, F. (2015). Small area estimation for skewed data in the presence of zeroes. *Statistics in Transition new series*, 4(16):541–562.

Lyu, X. (2019). *viscover: Visualize SSURGO and CDL and their Overlay*. R package version 0.1.0.

Lyu, X. (2020). *saezero: Small Area Estimation under a Zero Inflated Lognormal Model with Correlated Random Area Effects*. R package version 0.1.0.

Marhuenda, Y., Molina, I., Morales, D., and Rao, J. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):1111–1136.

Marino, M. F., Ranalli, M. G., Salvati, N., Alfò, M., et al. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, 13(2):1166–1197.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, 1(1):7–33.

Molina, I., Martin, N., et al. (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics*, 46(5):1961–1993.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3):181–204.

Pfeffermann, D., Terryn, B., and Moura, F. A. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology*, 34(2):235–249.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Smith, D. D. and Wischmeier, W. H. (1957). Factors affecting sheet and rill erosion. *Eos, Transactions American Geophysical Union*, 38(6):889–896.

Smyth, G. K. (2014). Polynomial approximation. *Wiley StatsRef: Statistics Reference Online*.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 26(1):24–36.

U.S. Department of Agriculture (2020a). National Agricultural Statistics Service Cropland Data Layer. Published crop-specific data layer [Online]. USDA-NASS, Washington, DC.

U.S. Department of Agriculture (2020b). Soil Survey Geographic (SSURGO) Database. Soil Survey Staff, Natural Resources Conservation Service.

Williams, J. R. and Izaurralde, R. (2006). The APEX model. In Singh, V. and Frevert, D., editors, *Watershed Models*, chapter 18, pages 437–482. Taylor & Francis Group.

Wischmeier, W. H. and Smith, D. D. (1965). Predicting rainfall erosion losses from cropland east of the rocky mountains: guide for selection for practices for soil and water conservation. U.S. Department of Agriculture. Agriculture Handbook NO.282.

Zimmermann, T. and Münnich, R. T. (2018). Small area estimation with a lognormal mixed model under informative sampling. *Journal of Official Statistics*, 34(2):523–542.

2.8 Appendix. Supplementary Information

2.8.1 Computing Details of MLE Estimation

To maximize the likelihood, we use the method `BFGS` in the `R` function `optim`, which is a quasi-Newton algorithm that uses information about the objective function and its gradient. To improve the speed of the optimization, we supply `optim` with a function to calculate the gradient vector to find the fastest ascending direction. The functional form for the gradient vector is given in subsubsection 2.8.1.1 below. In the `R` code, we express the functions defining the log likelihood

and its gradient in terms of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_e, \sigma_u, \sigma_b, t)'$ where $t = \tan(\rho\pi/2)$. This transformation for ρ avoids a need to specify bounds for the parameter space. To obtain starting values for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ_e, σ_u, and σ_b, we begin with the elements of $\hat{\boldsymbol{\theta}}_0$, the parameter estimator from fitting the two model parts independently as described in Section 2.4.1. We then bound the starting value of σ_b below by $\sqrt{10^{-5}}$ because the gradient vector is undefined if the starting value for σ_b is zero. The `R` functions `lmer` and `glmer` furnish predictors of u_i and of b_i, where the predictor of u_i is the best linear unbiased predictor, and the predictor of b_i is calculated using PQL. We use as the starting value for ρ the sample correlation between the initial predictors of b_i and u_i. If the starting value for σ_b is set to the bound of $\sqrt{10^{-5}}$, the starting value for ρ is set to zero.

2.8.1.1 Gradient vector

We give the functional form for the gradient of the log likelihood $l_i(\boldsymbol{\theta}) = \log(L_i(\boldsymbol{\theta}))$. Denote $\boldsymbol{Z}_{1i}$ as the model matrix and $\tilde{\boldsymbol{r}}_i = (\tilde{r}_{i1}, \ldots, \tilde{r}_{in_i})'$ as marginal residual vector of model (2.2) for area i. It can be shown that $\partial\bar{\tilde{r}}_i/\partial\boldsymbol{\beta} = -\bar{\tilde{\boldsymbol{z}}}_{1i}$, $\partial m_i/\partial\boldsymbol{\beta} = m_i/\bar{\tilde{r}}_i\bar{\tilde{\boldsymbol{z}}}_{1i}$ where $\bar{\tilde{\boldsymbol{z}}}_{1i} = \tilde{n}_i^{-1}\boldsymbol{Z}_{1i}'\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{in_i})'$. Denote $\boldsymbol{Z}_{2i}$ as the model matrix of model (2.3) for area i. It can be shown that $\partial\pi_{s_i}/\partial\boldsymbol{\alpha} = \pi_{s_i}(b_i)\boldsymbol{Z}_{2i}'\boldsymbol{u}_i$, where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{in_i})'$, $u_{ij} = (g^{-1})'(\alpha_0 + \boldsymbol{z}_{2,ij}'\boldsymbol{\alpha}_1 + b_i)p_{ij}^{-\delta_{ij}}(1 - p_{ij})^{\delta_{ij}-1}(2\delta_{ij} - 1)$ and $(g^{-1})'(\cdot)$ is the derivative function of $g^{-1}(\cdot)$. When $g(\cdot)$ is the logit function, u_{ij} simplifies to $p_{ij}^{1-\delta_{ij}}(1 - p_{ij})^{\delta_{ij}}(2\delta_{ij} - 1)$. The derivatives of the log likelihood function with respect to the fixed effect coffcient vectors are

$$\begin{aligned}
\frac{\partial l_i}{\partial\boldsymbol{\beta}} &= -\frac{\gamma_i\bar{\tilde{r}}_i}{\sigma_e^2/\tilde{n}_i}\bar{\tilde{\boldsymbol{z}}}_{1i} - \frac{m_i}{v_i}\frac{\partial m_i}{\partial\boldsymbol{\beta}} + \frac{1}{\sigma_e^2}\boldsymbol{Z}_{1i}'\tilde{\boldsymbol{r}}_i + \frac{\int \pi_{s_i}(b_i)(-\frac{b_i-m_i}{\sqrt{v_i}})\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}\frac{1}{\sqrt{v_i}}\frac{\partial m_i}{\partial\boldsymbol{\beta}} \\
\frac{\partial l_i}{\partial\boldsymbol{\alpha}} &= \frac{\int(\frac{\partial\pi_{s_i}}{\partial\boldsymbol{\alpha}})\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}.
\end{aligned}$$

For derivatives with respect to the variance components, it can be derived that $\partial\gamma_i/\partial\sigma_u = 2\gamma_i(1 - \gamma_i)/\sigma_u$, $\partial\gamma_i/\partial\sigma_e = -2\gamma_i(1 - \gamma_i)/\sigma_e$, $\partial m_i/\partial\sigma_u = (m_i/\sigma_u)(\sigma_e^2/\tilde{n}_i - \sigma_u^2)/(\sigma_e^2/\tilde{n}_i + \sigma_u^2)$, $\partial m_i/\partial\sigma_b = m_i/\sigma_b$, $\partial m_i/\partial\sigma_e = -2m_i(\sigma_e/\tilde{n}_i)/(\sigma_e^2/\tilde{n}_i + \sigma_u^2)$, $\partial v_i/\partial\gamma_i = -v_i\rho^2/\{1-(1-\gamma_i)\rho^2\}$, $\partial v_i/\partial\sigma_b = 2v_i/\sigma_b$. The derivatives of v_i use the fact that $v_i = \sigma_b^2(1 - \rho^2)/\{1 - (1 - \gamma_i)\rho^2\}$ is an equivalent expression

of v_i as in Theorem 2.1. Then

$$\begin{aligned}
\frac{\partial l_i}{\partial \sigma_u} &= \frac{1}{2}\Big[\frac{-1}{1-\gamma_i}\frac{\partial \gamma_i}{\partial \sigma_u} + \frac{\bar{\tilde{r}}_i^2}{\sigma_e^2/\tilde{n}_i}\frac{\partial \gamma_i}{\partial \sigma_u} + \frac{2m_i}{v_i}\frac{\partial m_i}{\partial \sigma_u} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial \sigma_u}\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial \sigma_u}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i} \\
\frac{\partial l_i}{\partial \sigma_e} &= \frac{1}{2}\Big[\frac{-1}{1-\gamma_i}\frac{\partial \gamma_i}{\partial \sigma_e} - \frac{2}{\sigma_e/\tilde{n}_i} + \frac{\bar{\tilde{r}}_i^2}{\sigma_e^2/\tilde{n}_i}\frac{\partial \gamma_i}{\partial \sigma_e} - \frac{2\gamma_i\bar{\tilde{r}}_i^2}{\sigma_e^3/\tilde{n}_i} \\
&\quad + \frac{2m_i}{v_i}\frac{\partial m_i}{\partial \sigma_e} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial \sigma_e} + 2\sigma_e^{-3}\sum_j \tilde{r}_{ij}^2\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial \sigma_e}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i} \\
\frac{\partial l_i}{\partial \sigma_b} &= \frac{1}{2}\Big[\frac{2m_i}{v_i}\frac{\partial m_i}{\partial \sigma_b} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial \sigma_b} - \frac{2}{\sigma_b}\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial \sigma_b}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}
\end{aligned}$$

where $\partial\{(b_i-m_i)/\sqrt{v_i}\}/\partial\sigma = -\{\partial m_i/\partial\sigma + (b_i-m_i)/(2v_i)\partial v_i/\partial\sigma\}/\sqrt{v_i}$ for $\sigma = \sigma_u, \sigma_e$ or σ_b. For derivative with respect to ρ, we have $\partial\gamma_i/\partial\rho = -2\gamma_i(1-\gamma_i)\rho/(1-\rho^2)$, $\partial v_i/\partial\rho = -2\rho\gamma_i\sigma_b^2/\{1-(1-\gamma_i)\rho^2\}$, $\partial m_i/\partial\rho = \bar{\tilde{r}}_i\sigma_u\sigma_b/(\sigma_u^2+\sigma_e^2/\tilde{n}_i)$. Then

$$\begin{aligned}
\frac{\partial l_i}{\partial \rho} &= \frac{1}{2}\Big[\frac{-1}{1-\gamma_i}\frac{\partial \gamma_i}{\partial \rho} + \frac{\bar{\tilde{r}}_i^2}{\sigma_e^2/\tilde{n}_i}\frac{\partial \gamma_i}{\partial \rho} + \frac{2m_i}{v_i}\frac{\partial m_i}{\partial \rho} - \frac{m_i^2}{v_i^2}\frac{\partial v_i}{\partial \rho}\Big] \\
&\quad + \frac{\int \pi_{s_i}(b_i)[-(\frac{b_i-m_i}{\sqrt{v_i}})\frac{\partial}{\partial \rho}(\frac{b_i-m_i}{\sqrt{v_i}})]\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}{\int \pi_{s_i}(b_i)\frac{1}{\sqrt{v_i}}\phi(\frac{b_i-m_i}{\sqrt{v_i}})\mathrm{d}b_i}
\end{aligned}$$

where $\partial\{(b_i-m_i)/\sqrt{v_i}\}/\partial\rho = -\{\partial m_i/\partial\rho + (b_i-m_i)/(2v_i)\partial v_i/\partial\rho\}/\sqrt{v_i}$. The implementation of this MLE method inputs $t = \tan(\rho\pi/2)$ instead of ρ to the `optim` function. Therefore, the last element of the gradient vector is $\partial l_i/\partial t = (\partial l_i/\partial\rho)(\partial\rho/\partial t)$ where $\partial\rho/\partial t = 2/\{\pi(1+t^2)\}$.

2.8.2 Incorporating Weights

2.8.2.1 Population covariates for CEAP

While the sampling unit is a single location, a unit in the dataset defining the population is a single soil map unit. We overlaid the SSURGO soil polygons with the 2006 CDL raster to obtain

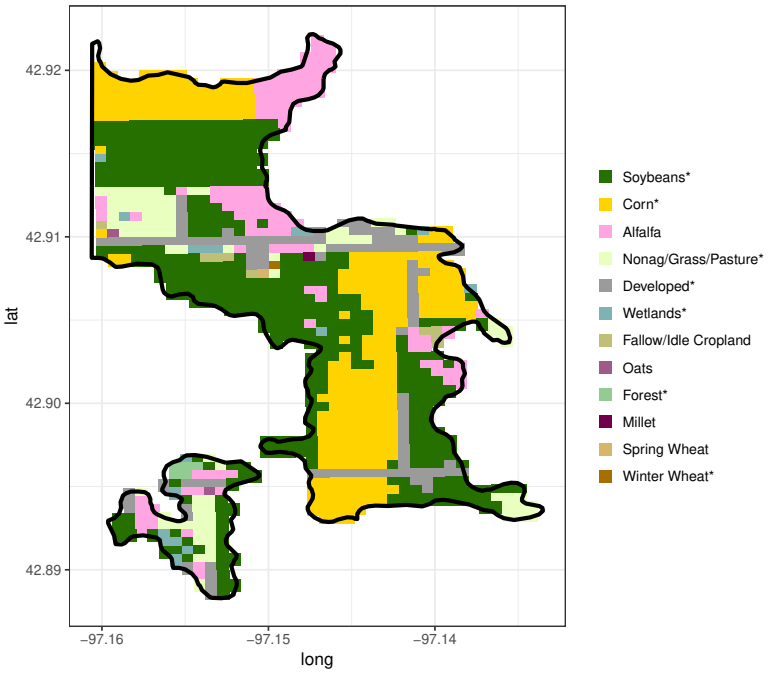

Figure 2.4 Example plot of a SSURGO map unit polygon overlaid with 2006 CDL raster.

their "joint" distribution: the area of selected CDL crops in each soil map unit. We developed an `R` package called `viscover` (Lyu, 2019) to query SSURGO and CDL data. Figure 2.4 shows the raw CDL pixels falling within an example soil map unit polygon. This overlay operation results in a population-level dataset that we can use for prediction. Table 2.5 summarizes the structure of the dataset that we use for prediction. This dataset is available in our `R` package `saezero` (Lyu, 2020).

2.8.2.2 Adapted empirical Bayes predictor

Let $i = 1, \ldots, 64$ index the county, where area i contains N_i crop pixels. Let y_{ij}^* be connected with the cropland RUSLE2 measurement for the j-th crop pixel in county i, which could be zero or positive. Let $\boldsymbol{z}_{ij}$ denote the covariate vector associated with crop pixel j in county i. For area i, the available data is $\{y_{ij}^* : j \in s_i\} \cup \{\boldsymbol{z}_{ij} : j = 1, \ldots, N_i\}$. For a single soil map unit, we further split the map unit into 5 segments. Each segment is composed of all pixels of one type: corn, soybean, winter wheat, spring wheat, or the remainder. Let $g = 1, \ldots, G_i$ index the soil map unit segments in area i, and Q_{ig} be the CDL pixel counts of the g-th soil map unit segment in county i. Then for

Table 2.5 Description of the variables in the collected dataset for predicting cropland RUSLE2.

Variable	Definition
mukey	the identity number of the soil map unit
county	the county in which the map unit locates
logR	log-scale county-level R-factor
logK	log-scale K-factor of the soil map unit
logS	log-scale S-factor of the soil map unit
counts.crop	CDL pixel counts of all crop categories within the soil map unit
counts.corn	CDL pixel counts of corn within the soil map unit
counts.soybean	CDL pixel counts of soybean within the soil map unit
counts.sprwht	CDL pixel counts of spring wheat within the soil map unit
counts.wtrwht	CDL pixel counts of winter wheat within the soil map unit

county i the total crop pixel count is $N_i = \sum_{g=1}^{G} Q_{ig}$ and $w_{ig} = Q_{ig}/N_i$ is the crop-acre percentage of the g-th soil map unit segment in county i. Because the unit in our data set representing the population is a soil map unit segment instead of a field, we express the predictors as a function of map unit segment acres. When the population parameter of interest is the average cropland RUSLE2 in tons per crop-pixel (56m $\times$ 56m) for each county, which is $\bar{y}^*_{N_i} = {N_i}^{-1} \sum_{j=1}^{N_i} y^*_{ij}$, the EB predictor of $\bar{y}^*_{N_i}$ is

$$\hat{\bar{y}}^{*\mathrm{EB}}_{N_i} = \frac{1}{N_i}\Big(\sum_{j\in s_i} y^*_{ij} + \sum_{j\in \bar{s}_i} \hat{y}^{*\mathrm{EB}}_{ij}\Big) \approx \frac{1}{N_i}\sum_{j=1}^{N_i} \hat{y}^{*\mathrm{EB}}_{ij}, \tag{2.16}$$

where the approximation holds because $n_i << N_i$. An approximation analogous to (2.16) is used in the context of the logistic mixed effects model in Hobza and Morales (2016). Since the covariates vector $\boldsymbol{z}_{ij}$ is now constant within a soil map unit segment, the EB prediction $\hat{y}^{*\mathrm{EB}}_{ij}$ is also constant. Therefore, we can write

$$\hat{\bar{y}}^{*\mathrm{EB}}_{N_i} \approx \frac{1}{N_i}\sum_{g=1}^{G_i} Q_{ig}\hat{y}^{*\mathrm{EB}}_{ig} = \sum_{g=1}^{G_i} w_{ig}\hat{y}^{*\mathrm{EB}}_{ig}.$$

Hobza and Morales (2016) and Hobza et al. (2018) use a similar simplification for the EB predictor as a weighted sum across entities with the same covariate value in their development of predictors for the logistic mixed effects model. Similarly, the variance of the EB predictor can be estimated

by

$$\hat{M}_{1i} \approx \frac{1}{N_i^2}\sum_{j=1}^{N_i}\sum_{k=1}^{N_i}\Big[E\{y_{ij}^* y_{ik}^*|(\boldsymbol{y}^*,\boldsymbol{z})\} - \hat{y}_{ij}^{*\mathrm{EB}}\hat{y}_{ik}^{*\mathrm{EB}}\Big]$$
$$= \frac{1}{N_i^2}\sum_{j=1}^{N_i}\sum_{k=1}^{N_i}E\{y_{ij}^* y_{ik}^*|(\boldsymbol{y}^*,\boldsymbol{z})\} - \Big(\sum_{g=1}^{G_i} w_{ig}\hat{y}_{ig}^{*\mathrm{EB}}\Big)^2.$$

where

$$E\{y_{ij}^* y_{ik}^*|(\boldsymbol{y}^*,\boldsymbol{z})\} = \begin{cases} h_{ig}h_{ig'}\exp(\tilde{\sigma}_{u_i}^2)E\{p_{ig}p_{ig'}\eta(2b_i)|(\boldsymbol{y}^*,\boldsymbol{z})\} & g\neq g', j\neq k \\ h_{ig}^2\exp(\tilde{\sigma}_{u_i}^2)E\{p_{ig}^2\eta(2b_i)|(\boldsymbol{y}^*,\boldsymbol{z})\} & g=g', j\neq k \\ h_{ig}^2\exp(\tilde{\sigma}_{u_i}^2+\sigma_e^2)E\{p_{ig}\eta(2b_i)|(\boldsymbol{y}^*,\boldsymbol{z})\} & g=g', j=k \end{cases}$$

Therefore,

$$\frac{1}{N_i^2}\sum_{j=1}^{N_i}\sum_{k=1}^{N_i}E\{y_{ij}^* y_{ik}^*|(\boldsymbol{y}^*,\boldsymbol{z})\}$$
$$=\frac{1}{N_i^2}\Big[\sum_{g=1}^{G_i}\sum_{g\neq g'}Q_{ig}Q_{ig'}h_{ig}h_{ig'}\exp(\tilde{\sigma}_{u_i}^2)E\{p_{ig}p_{ig'}\eta(2b_i)|(\boldsymbol{y}^*,\boldsymbol{z})\}$$
$$+\sum_{g=1}^{G_i}Q_{ig}(Q_{ig}-1)h_{ig}^2\exp(\tilde{\sigma}_{u_i}^2)E\{p_{ig}^2\eta(2b_i)|(\boldsymbol{y}^*,\boldsymbol{z})\}$$
$$+\sum_{g=1}^{G_i}Q_{ig}h_{ig}^2\exp(\tilde{\sigma}_{u_i}^2+\sigma_e^2)E\{p_{ig}\eta(2b_i)|(\boldsymbol{y}^*,\boldsymbol{z})\}\Big]$$

The approach above generalizes beyond this CEAP application to any situation where the covariates are known at an aggregate level where the EB predictions are constant.

2.8.3 Link Function Analysis

We consider a parametrized family of transformations for the positive part in the analysis of the CEAP RUSLE2 data to test the hypothesis of a logarithmic transformation. Specifically, the commonly used Box-Cox transformation function is considered, which is denoted as

$$\ell(y|\lambda) = \begin{cases} \frac{y^\lambda-1}{\lambda}, & \lambda\neq 0; \\ \log(y), & \lambda = 0. \end{cases}$$

Then the likelihood function is $L_i(\boldsymbol{\theta},\lambda) = \prod_{j\in s_i} y_{ij}^{\lambda-1}L_i(\boldsymbol{\theta})$ where $L_i(\boldsymbol{\theta})$ is the same as described in Section 2.2 except $\tilde{r}_{ij} = \ell(y_{ij}|\lambda) - \beta_0 - \boldsymbol{z}_{1,ij}'\boldsymbol{\beta}_1$. Applying the profile likelihood method gives a 95%

confidence interval of λ as $(-0.039, 0.080)$ which indicates there is no significant evidence to reject the hypothesis that $\lambda = 0$. Thus, the logarithmic transformation for the positive part is reasonable for the CEAP data.

The same technique applies to estimating the link function for the binary part. The logit link is contained in the Aranda-Ordaz (1981) transformation defined by

$$g(\mu|\phi) = \log\{\frac{(1-\mu)^{-\phi} - 1}{\phi}\}.$$

It simplifies to the logit in the case of $\phi = 1$. Our analysis of the profile log likelihood for the link function parameter gave a 95% confidence interval of ϕ as $[0, 1.62)$, suggesting no reason to reject the logit link.

2.8.4 Monte Carlo Margin of Error

Each MC mean shown in Tables 2.1-2.3 is an average across 20 areas with the same sample size and $M = 1000$ Monte Carlo simulations. Denote by $\zeta^{(m)}$ one average across 20 areas with the sample sample size based on the m-th Monte Carlo simulation. Each MC margin of error presented in parentheses is calculated as $1.96\sqrt{M^{-1}\sum_{m=1}^{M}(\zeta^{(m)} - \bar{\zeta})^2}/\sqrt{M}$, where $M = 1000$ is the number of Monte Carlo simulations and $\bar{\zeta} = M^{-1}\sum_{m=1}^{M}\zeta^{(m)}$.

2.8.5 Supplementary Figures and Tables

Table 2.6 The average ratios (%) of the second term M_2, bias of the leading term M_1, and the cross term M_3 to the fully parametric bootstrap estimate of the MSE of the EB predictor, where the averages are across the areas with the same sample size n_i and $M = 1000$ simulations. The bootstrap size is $B = 100$.

	$\hat{M}_{2i}^{\text{Boot}}/\hat{M}_i^{\text{Boot}}$	$\hat{M}_{1i}^{\text{Bias}}/\hat{M}_i^{\text{Boot}}$	$\hat{M}_{3i}^{\text{Boot}}/\hat{M}_i^{\text{Boot}}$
Avg. for $n_i = 5$	4.93	-2.14	-0.14
Avg. for $n_i = 10$	5.65	-2.98	-0.21
Avg. for $n_i = 20$	6.38	-3.06	-0.26

Table 2.7 Mapping from CDL categories to NRI Broad Coveruse.

NRI Code	Source	Category
1	NRI	Cropland, Cultivated
	CDL	Corn (1), Cotton (2), Rice (3), Sorghum (4), Soybeans (5), Sunflower (6), Peanuts (10), Tobacco (11), Sweet Corn (12), Pop or Orn Corn (13), Mint (14), Barley (21), Durum Wheat (22), Spring Wheat (23), Winter Wheat (24), Other Small Grains (25), Dbl Crop WinWht/Soybeans (26), Rye (27), Oats (28), Millet (29), Speltz (30), Canola (31), Flaxseed (32), Safflower (33), Rape Seed (34), Mustard (35), Alfalfa (36), Camelina (38), Buckwheat (39), Sugarbeets (41), Dry Beans (42), Potatoes (43), Sugarcane (45), Sweet Potatoes (46), Misc Vegs & Fruits (47), Watermelons (48), Onions (49), Cucumbers (50), Chick Peas (51), Lentils (52), Peas (53), Tomatoes (54), Herbs (57), Fallow/Idle Cropland (61), Aquaculture (92), Triticale (205), Carrots (206), Asparagus (207), Garlic (208), Cantaloupes (209), Honeydew Melons (213), Broccoli (214), Peppers (216), Greens (219), Squash (222), Dbl Crop WinWht/Corn (225), Dbl Crop Oats/Corn (226), Lettuce (227), Pumpkins (229), Dbl Crop Lettuce/Durum Wht (230), Dbl Crop Lettuce/Cotton (232), Dbl Crop Lettuce/Barley (233), Dbl Crop Durum Wht/Sorghum (234), Dbl Crop Barley/Sorghum (235), Dbl Crop WinWht/Sorghum (236), Dbl Crop Barley/Corn (237), Dbl Crop WinWht/Cotton (238), Dbl Crop Soybeans/Cotton (239), Dbl Crop Soybeans/Oats (240), Dbl Crop Corn/Soybeans (241), Cabbage (243), Cauliflower (244), Celery (245), Radishes (246), Turnips (247), Eggplants (248), Gourds (249), Dbl Crop Barley/Soybeans (254)
2	NRI	Cropland, Non-cultivated
	CDL	Other Hay/Non Alfalfa (37), Other Crops (44), Caneberries (55), Hops (56), Clover/Wildflowers (58), Cherries (66), Peaches (67), Apples (68), Grapes (69), Christmas Trees (70), Other Tree Crops (71), Citrus (72), Pecans (74), Almonds (75), Walnuts (76), Pears (77), Pistachios (204), Prunes (210), Olives (211), Oranges (212), Pomegranates (217), Nectarines (218), Plums (220), Strawberries (221), Apricots (223), Dbl Crop Lettuce/Cantaloupe (231), Blueberries (242), Cranberries (250)

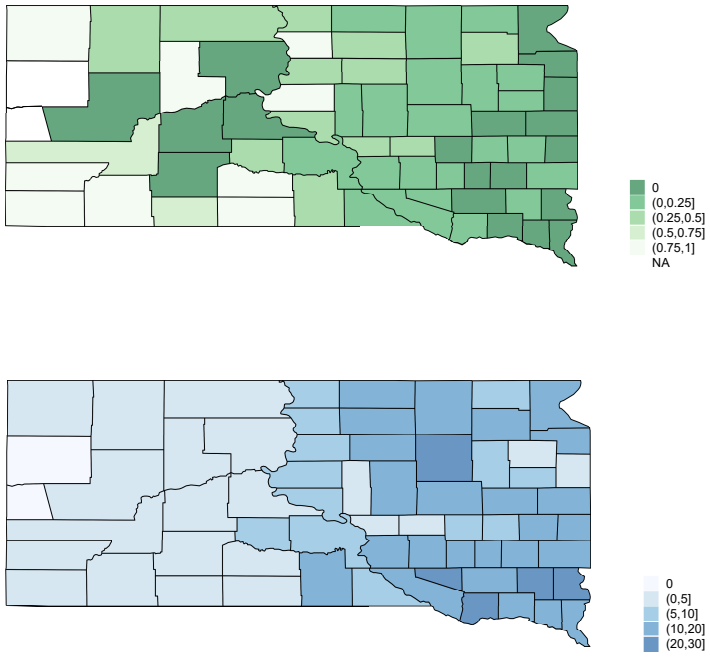

Figure 2.5 Cartograms of the proportions of zero (top) and sample sizes (bottom) of the cropland CEAP RUSLE2 data in the counties of South Dakota.

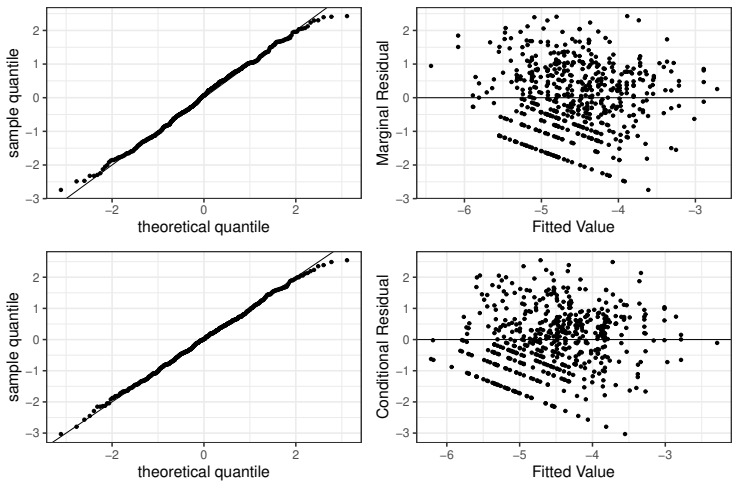

Figure 2.6 The normal quantile-quantile plot (left) and the standardized residual plot (right) for the marginal (top) and conditional (bottom) residuals from fitting the positive part of the cropland CEAP RUSLE2 data.

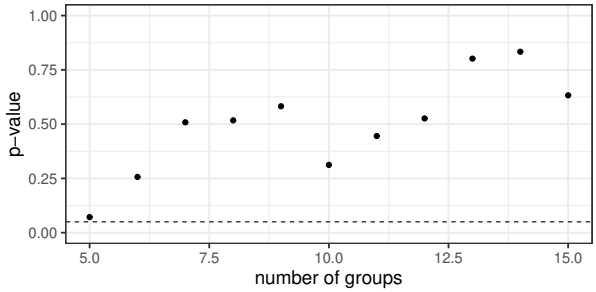

Figure 2.7 The p-values of the Hosmer-Lemeshow tests with different number of groups for the goodness of fit of the binary part of the cropland CEAP RUSLE2 data. The dotted horizontal line is $p = 0.05$.

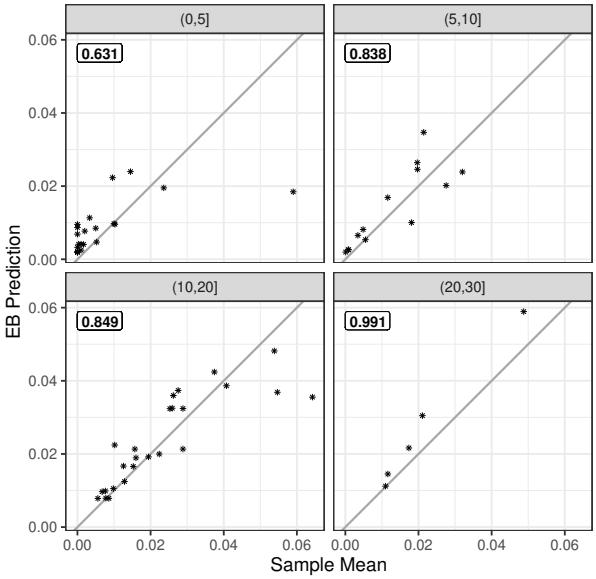

Figure 2.8 EB predictions plotted against direct estimates of the mean cropland RUSLE2 for the counties in South Dakota. Top-left corner in each panel is the sample Pearson correlation between the two estimators in each sample size group.

CHAPTER 3. AN R SHINY APPLICATION TO DATA QUALITY REVIEW

Xiaodan Lyu[1], Heike Hofmann[1], Emily J. Berg[1] and Jie Li[2]

[1]Department of Statistics, Iowa State University, Ames, IA

[2]Center for Survey Statistics and Methodology, Iowa State University, Ames, IA

Modified from a manuscript to be submitted to *R Journal*

3.1 Abstract

Data quality assurance benefits the ultimate decision-making. The process of data quality review and improvement, for statistical production involving a data product such as the National Resource Inventory (NRI), is usually iterative. It requires evaluating quality metrics and correcting identified errors frequently until a program-specific quality standard is met. R Shiny tools can address the operational challenges in solving real data quality problems with functionalities of data manipulation and visualization. Data manipulation aids automate the computation of quality metrics once they are formulated by subject matter experts and ready to be calculated from the database. Effective graphical presentation of the data reduces the cognitive load on reviewers and improves their task performance in terms of accuracy and speed. This paper introduces a Shiny-based web app called *iNtr* to demonstrate the practical use of R Shiny in the ongoing NRI data quality review process. This paper describes the use case, graphical user interface, and data-processing workflow of *iNtr*. For the protection of security and confidentiality of the unreleased NRI data, we present a modified version of the web app that allows comparing the released 2015 and 2012 NRI data products. The public version of the tool is accessible at https://lyux.shinyapps.io/table_review/.

3.2 Introduction

Conceptually, data quality is the capability of data in terms of effectiveness, economics, and timeliness to support decision-making (Karr et al., 2006). Industries and governmental agencies have been devoting plenty of efforts to improve data quality. Telecommunications companies lean on high-quality streaming data to deliver high-quality service to customers. Huh et al. (1990) describes the implementation of quality control, data engineering, and process management techniques in AT&T's data quality improvement program. High-quality health facility data support health system planning. The World Health Organization (2017) developed the Data Quality Review Toolkit and provided guidelines and Excel-based tools to aid the institutionalization of routine data quality assessments. High-quality population count data help plan bill funding. Bushery et al. (2003) proposed a specific contention process for quality assurance of the 2010 United States Census. High-quality employment data help monitor labor market. The quality report of the European Union Labour Force Survey 2017 (European Statistics, 2019) followed the guidelines in the European Statistics handbook (European Statistics, 2015) and assessed data quality in several dimensions, including accuracy, timeliness, etc. High-quality natural resources and environmental data support informative agricultural and environmental policymaking. Natural Resources Conservation Service (1994) described the use of data entry software to automate checking the edit and compatibility during the data collection in the National Resources Inventory surveys.

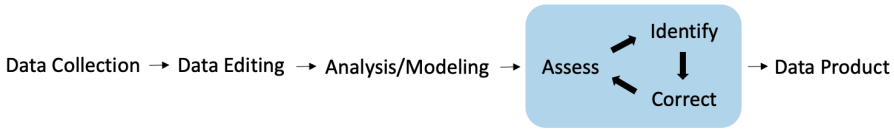

Figure 3.1 Flowchart of statistical production involving a data product. The data quality review and improvement process (highlighted in blue) — evaluating (assessing) quality metrics, identifying error sources, and correcting prior interventions, are iterated until the quality of the data product is acceptable.

We depict a general data production pipeline in Figure 3.1. Any large-scale and complex operation that involves a data product typically requires assessing and improving data quality at

each phase of the complex production process. Prior to a data product, there are usually a series of interventions: (1) collecting data either by human, such as experiment and survey, or by machine, such as automate sensors; (2) editing the collected data, either non-statistically, such as synthesizing and consolidating data from various sources, or statistically, such as imputing missing values and inspecting outliers; (3) analyzing the data with statistical modeling. The third step includes data quality strategies such as adjusting uncertainties in estimates or predictions to reflect the deficiency in the collected data. Since statistical production can be viewed conceptually the same as physical industrial production (Wang, 1998), data quality review and improvement (highlighted in blue color in Figure 3.1) is important for building confidence and credibility in the data product to be delivered. It is the last chance to carry out the inspection before the time point, which Deming (2018) called "too late".

The evaluation studies of data quality are usually complex to implement (Biemer, 2010). The actual data quality review and improvement process are iterative. After identifying the root cause of "bad-quality" data, corrections are made to prior interventions, which elicits another round of data quality review and improvement until certain quality criteria are met. The National Institute of Statistical Sciences Data Quality Workshop in 2000 (Karr et al., 2001) called for "software tools that solve real data quality problems". An effective **shiny** (Chang et al., 2018) tool can benefit data quality review in various aspects. The R package **shiny** is designed to implement a reactive framework that allows users to directly and immediately interact in their web-browser with data and charts provided by the underlying R program. The **shiny** tool can automate computing the quality metrics once there is any update in previous steps, and reduce cognitive load on reviewers to boost their efficiency through effective graphical displays.

Cognition is the process of knowledge acquisition and reasoning. Anderson et al. (2011) uses cognitive load to evaluate visualization effectiveness. Effective information representation in a way that is easy to read and interpret can reduce the amount of mental effort to complete a task and thus improve performance (Engle, 2002). Visual techniques have been used in other research studies to aid data quality improvement. Karr et al. (2006) suggest the visual techniques contained

in the Exploratory Data Analysis (Tukey, 1977) approach are helpful for identifying errors in administrative data. Williams and Berg (2013) propose using heat maps to assess alternative weighting procedures in calibration, and calibration addresses statistical data quality concerns for survey data.

3.2.1 Motivating NRI Table Review Process

The National Resources Inventory (NRI) program is a longitudinal survey that monitors status and trend of characteristics related to natural resources and agriculture on non-federal U.S. land. NRI is supported by the Natural Resources Conservation Service (NRCS) of the United States Department of Agriculture (USDA) and conducted through a collaboration between NRCS and the Center for Survey Statistics and Methodology (CSSM) at Iowa State University. The 1982 NRI starts to establish and maintain a database containing around 800,000 points across the United States. Developed upon the NRI Database, an NRI data product includes aggregate estimates of the area of land in each of several broad use categories, change in land use, average annual water and wind erosion on cropland and pastureland, among others, at national and state levels.

The NRI data are collected from a complex survey process (Nusser and Goebel, 1997). The primary sampling units in the NRI are *segments* of the land of approximately 1/2 by 1/2 square miles. The NRI data collection begins with aerial photographs of sampled segments. Data collectors visually interpret the aerial photographs and then delineate areas by features such as urban areas, roads, and water bodies on the image of each sampled segment. The secondary sampling units are points selected in the sampled segments. For points classified in certain categories, such as cropland and pastureland, additional information is obtained from administrative sources. To aid the estimation of change over time, the NRI data collectors observes the same sampling units over time since 1982. A foundation sample of 300,000 segments is observed every five years from 1982 through 1997. To support annual estimates, NRI transitioned to a supplemented panel design (Breidt and Fuller, 1999) in 2000. In the supplemented panel design, a core panel of approximately 40,000

segments is observed every year, and rotation panels, each with approximately 30,000 segments are observed less frequently.

The supplemental panel design introduces missing data, and the NRI data collection procedures involve multiple modes of collection. The collected data are synthesized through a complex estimation procedure. Imputation is used to create a complete time-series for every point. Calibration procedures are used to maintain pre-specified state-level estimates. Imputation and calibration herein are part of the efforts to improve data quality from a statistical perspective. Such estimates are produced for each of the 48 coterminous states, Hawaii, Puerto Rico, and the U.S. as a whole. An NRI data product presents survey estimates in the form of tabular displays by topic and by area together in an NRI report (U.S. Department of Agriculture, 2018, 2015). Before released, an NRI data product is vetted in a formal table review. Taking the 2015 NRI as an example, the table review process involves evaluating quality metrics of pre-released data products in a way as described in the following two paragraphs.

Type I evaluation An NRI data product includes estimates in 5-year intervals. NRCS releases a new set of NRI estimates every 3-5 years. This means that, for example, the released 2015 NRI product (U.S. Department of Agriculture, 2018) contains a new set of estimates for 2015 beyond all estimates available in the released 2012 NRI product (U.S. Department of Agriculture, 2015). Even though these products share estimates for a lot of the same years, some of these estimates vary between NRI products. The variation is mainly due to data edits, algorithmic improvements, or methodological adjustments. It is also important to realize that large revisions of previously published estimates can potentially corrode the public's trust and should, therefore, be done very carefully. The difference between a 2015 version and the 2012 version helps quantify whether the variation is reasonable and acceptable.

Type II evaluation Although the official release of the NRI data product at a specific year contains only one set of estimates, the production of the estimates is an iterative process because of continuous quality review and improvement. As a result, many versions of the data products will

be generated before the final release. The 2015 NRI data product, published in September 2018, evolved from a total of eleven versions. Each evolution is necessary after errors are identified from a careful table review, and corrections are made to the NRI Database. Differences between a newer 2015 version and an older 2015 version help verify whether the data quality improvements succeed.

Review procedure Type I and II evaluations are used to reflect the quality of an NRI data product. While some of these differences are plausible and necessary, it is important to check whether the entries in the output product (estimate tables) reflect the intended updates. Any unexpected changes revealed by the evaluations indicate potential data quality problems. And differences larger than their benchmarks may illuminate errors in the collected data. Each NRI data product has a root in the NRI Database containing about 800,000 data records. For large differences, we need to see the contributions from smaller areas to find the root cause in the NRI Database and remedy the errors accordingly. Thus an inspection consists of a hierarchical investigation from a national overview to the state level and then, if necessary, to the county level. Once an anomaly has been identified at the county level, it would be feasible to look into the affected points related to that county in the NRI Database and make corresponding corrections after verification of actual errors. Next, we give an example to illustrate how to review an estimate table in a pre-released 2015 NRI data product. On the left of Figure 3.2 is the 2015 NRI estimate table of national-level average annual sheet and rill erosion on non-Federal rural land by year in tons per acre per year. Besides the individual estimate presented at the top of each cell, an NRI estimate table also provides measures of statistical uncertainty in the form of margin of error (standard error times 1.96) at the bottom of each cell. The 2012 NRI also contains this set of estimates except for the bottom row, as shown in the right of Figure 3.2. The difference table in Figure 3.3 furnishes a Type I evaluation of this NRI estimate table and shows the absolute relative differences of the two versions of estimates in the 2015 and 2012 NRI data products. (The bottom row containing the estimated national level annual sheet and rill erosion for 2015 does not exist in this difference table for Type I evaluation but does exist for Type II evaluation.) A similar difference table of standard errors also needs to be reviewed but is not shown in Figure 3.3.

	Cropland			CRP land	Pastureland
Year	Cultivated	Non-Cultivated	Total		
1982	4.18 ±0.04	0.72 ±0.05	3.82 ±0.03	–	1.00 ±0.02
1987	3.82 ±0.04	0.77 ±0.06	3.50 ±0.03	2.05 ±0.19	0.93 ±0.02
1992	3.29 ±0.03	0.67 ±0.05	2.97 ±0.03	0.55 ±0.04	0.89 ±0.02
1997	2.96 ±0.03	0.73 ±0.03	2.67 ±0.02	0.35 ±0.02	0.79 ±0.02
2002	3.03 ±0.04	0.76 ±0.05	2.70 ±0.03	0.36 ±0.03	0.73 ±0.02
2007	2.90 ±0.05	0.77 ±0.04	2.59 ±0.04	0.37 ±0.02	0.68 ±0.02
2012	2.95 ±0.04	0.74 ±0.05	2.64 ±0.04	0.37 ±0.02	0.64 ±0.02
2015	3.03 ±0.05	0.72 ±0.05	2.71 ±0.05	0.39 ±0.03	0.62 ±0.02

	Cropland			CRP land	Pastureland
Year	Cultivated	Non-Cultivated	Total		
1982	4.15 ±0.04	0.71 ±0.05	3.79 ±0.03	--	1.00 ±0.02
1987	3.80 ±0.04	0.75 ±0.05	3.47 ±0.03	2.05 ±0.15	0.93 ±0.02
1992	3.26 ±0.03	0.64 ±0.03	2.94 ±0.03	0.56 ±0.04	0.89 ±0.02
1997	2.95 ±0.02	0.66 ±0.03	2.65 ±0.02	0.37 ±0.02	0.80 ±0.02
2002	3.02 ±0.03	0.70 ±0.04	2.69 ±0.03	0.37 ±0.03	0.75 ±0.02
2007	2.91 ±0.04	0.68 ±0.04	2.58 ±0.04	0.37 ±0.02	0.69 ±0.02
2012	2.99 ±0.05	0.67 ±0.04	2.66 ±0.04	0.40 ±0.02	0.69 ±0.03

Figure 3.2 Snapshots of estimate tables (national level average annual sheet and rill erosion on non-Federal rural land in tons per acre per year with margins of error presented after a ± sign) from the 2015 (left) and 2012 (right) NRI reports.

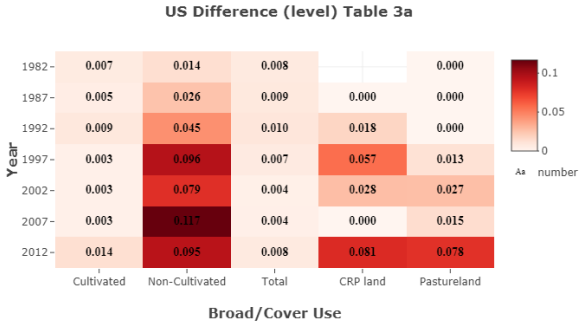

Figure 3.3 Snapshot of U.S. Difference Table 3a (national level average annual sheet and rill erosion estimates on non-Federal rural land in tons per acre per year) from *iNtr*. The cell values are the absolute relative differences between the 2015 and 2012 NRI estimates. Darker shades of red indicate larger differences warranting closer inspection.

The volume of information to be cognized so as to complete the NRI table review task is overwhelming. For instance, every state has the sheet and rill erosion estimate table, as shown in the left of Figure 3.2, and there are 84 such tables (of different topics) in total for each state in the 2015 NRI product alone. It's known that 12 versions (including the final version) have been produced for the 2015 NRI. Reviewers need to go through 12 Type I evaluations and 11 Type II evaluations for each of the 84 estimate tables at both the national level and state level (50 NRI "states": 48 contiguous states, Hawaii, and Puerto Rico). In total, around 197,000 difference tables (estimate and standard error are counted separately) with about the same dimension as the one in Figure 3.3 need to be reviewed. Before *iNtr* is developed, the reviewer used to inspect the vast number of difference tables using an Excel-based tool that displays difference tables by version and by area (or topic) in piles of spreadsheets. The objective of our work is to optimize the table review process by developing a graphical tool to reduce the cognitive load on reviewers.

3.2.2 Outline of the Paper

The NRI table review task is an integral component of the NRI quality assurance program. We developed an R **shiny** tool called *iNtr*, named after **i**nteractive **N**RI **t**able **r**eview, to aid the review of the pre-released NRI data products. We give the introduction to the user interface of *iNtr* in section 3.3, and an example of use in section 3.4. We give the technical details of the tool development in section 3.5. The source code is available on GitHub[1]. In section 3.6, we discuss more general implications of our studies. Even though *iNtr* has specific applicability to NRI, we explain how our experience furnishes general guidelines to consider when developing a similar **shiny** tool. Concluding remarks are given in section 3.7.

3.3 User interface

Figure 3.4 depicts an overview of the graphical user interface of *iNtr*. This user interface is developed with the R package **shinydashboard** (Chang and Borges Ribeiro, 2018). The R function

[1] https://github.com/XiaodanLyu/shinyreview

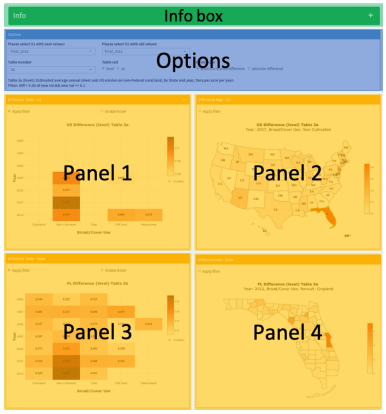

Figure 3.4 Overview of the graphical user interface of *iNtr*.

`box` helps arrange the visual components in a neat order respecting a reading habit from top to bottom and left to right. The arrangement of the boxes is important. Both space proximity and color similarity can enforce a visual correlation (Few, 2006). It is sensible to put top-left the most important component, such as the instruction texts or input widgets. At the top of the application is a green collapsible Info box containing a brief tutorial on how to use the tool. In the middle is a blue Options box containing input widgets, i.e., dropdown menus and radio buttons to choose parameter settings (as shown at the top of Figure 3.5). Given the NRI table review is iterated over versions, tables, and variables (estimate or standard error) as described in subsection 3.2.1, the parameter settings include two table versions (old and new), one table identifier, and one variable to choose. Additionally, the Options box asks the reviewer to choose which specific quality metric to compute, absolute relative difference or absolute difference. The default choice is the absolute relative difference to old estimation. The absolute difference is there in case of sanity check for any edge case that is not captured by the relative absolute difference, for example, the relative difference is not defined when the old estimation is 0.

The **shiny** tool *iNtr* can function as a data dashboard as described in Biemer (2010) to compare the same data source from different views and "drill down" into the data so as to conduct root cause analysis of error sources. The "drill-down" operations are presented in a **shiny** tool through linkage among visual components, i.e., tables and charts. At the bottom of the user interface are four

yellow boxes representing the main output: Panels 1 to 4 show the main charts and tables of the review tool. After a reviewer decides the parameters in the Options box, a hierarchical inspection from the national level to the state level and then county level is realized through the reviewer's interactions from Panels 1 to 4 as indicated by the grey arrows in Figure 3.5 and illustrated later in the following section. We are using the same color for these four panels on purpose to visually highlight the connection. Panel 1 shows a national level difference table where the review process begins. Panel 3 shows a state-level difference table (one of the 50 NRI "states") to be reviewed. Reviewer's click selection in Panel 1 triggers a corresponding breakdown of a national level difference into state-level differences depicted in a U.S. map. A similar interaction rule applies between Panel 3 and Panel 4, where a state-level difference is broken down into county-level differences, whereas the choice of one out of the 50 NRI "states" in Panel 3 is dependent on reviewer's click selection on the U.S. map in Panel 2.

3.4 Example of Use

The tool *iNtr* reduces the reviewer's cognitive load in the following aspects. Firstly, there are table-specific screening criteria to flag important differences of either estimates or standard errors between two versions. For example, a cell of estimate in the difference Table 3a (Figure 3.3) is indicated to have potential data-quality problem if the resulting absolute relative difference is greater than 5% and the new estimate is no less than 0.1. Those criteria are determined by domain knowledge experts. Differences below the table-specific criterion are considered as indicators of probable "good-quality" data and hidden to reduce the cognitive burden on reviewers. Panel 1 in Figure 3.5 shows the same difference table as given in Figure 3.3 except that it has been screened to show only those cells with (1) a minimum absolute relative difference greater than 0.05 and (2) a new estimate of no less than 0.1. Secondly, with larger differences highlighted in bolder colors, now it is easier to locate the cell with the most striking difference indicated by the darkest shade of red than inspecting the numbers in a spreadsheet. Thirdly, the "drill-down" feature in *iNtr* makes root cause analysis (hierarchical inspection) much more efficient than going over multiple spreadsheets

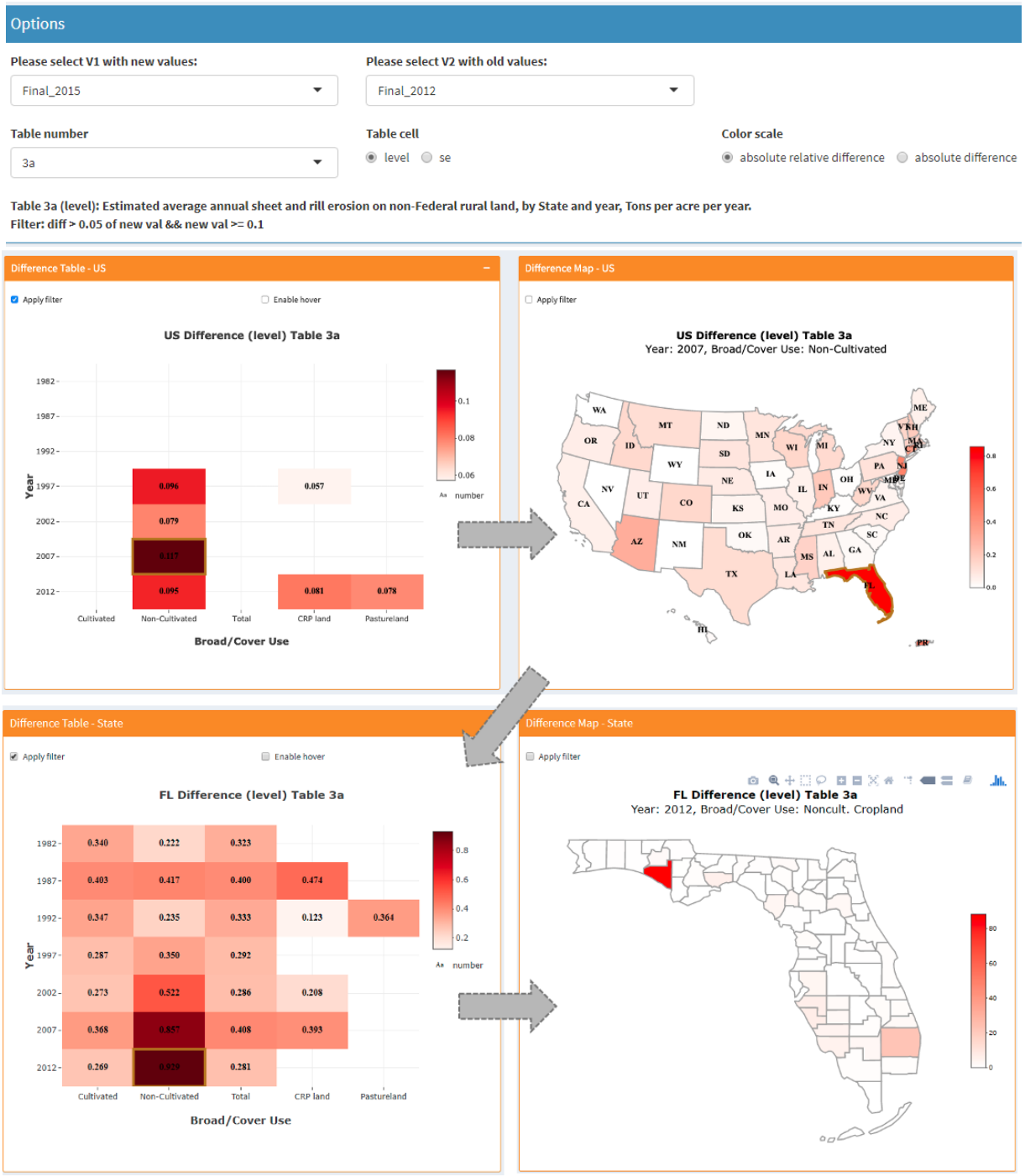

Figure 3.5 Hierarchical table review using *iNtr*. Grey arrows show the order of a reviewer's interaction to access more detailed information. Selected cells in each panel are highlighted with a golden border.

saved in different digital folders. Once the reviewer selects this cell (2007, Non-Cultivated) by clicking on it in Panel 1 as Figure 3.5 shows, Panel 2 presents a choropleth map with the 48 contiguous states, Hawaii, and Puerto Rico shaded with differences associated with the clicked cell. In this map, Florida immediately draws one's attention. The shading indicates that Florida shows the biggest departure from the previous estimate warranting a closer look at the county level. By click-selecting Florida in the national map, a state-level difference table of Florida appears in Panel 3, and a choropleth map of Florida county-level differences appear in Panel 4[2] once a table-entry has been click-selected in Panel 3. From the state map in Panel 4, one can see which counties have contributed most to the overall large difference in the state. (Since the county-level estimates are not published, we are unable to make the comparison between the released 2015 and 2012 NRI products at the county level. Thus, Panel 4 does not exist in the public version of *iNtr* but exists in the version used internally at CSSM.) On the contrast, county-level inspections using the legacy Excel-based table review tool are not implemented due to operational complexity. In contrast, *iNtr* improves the reviewer's performance in terms of time needed to complete the table review task through effective information presentation. After a county is flagged, the action item for the reviewer is to dig into the NRI Database outside *iNtr*, checking metadata, and correcting any error existing in the prior interventions. Once the NRI Database has been rectified, a newer version of the 2015 NRI product is made out of the improved database and added to *iNtr* for another set of Type I and Type II evaluations.

3.5 Design and Implementation

Any web-application based on the **shiny** framework (Chang et al., 2018) has to explicitly define the *user interface*, i.e., the graphic appearance of the application, and the *server* function, which connects user-data interactions, such as clicks or hovers, to the displayed charts and tables. Data processing is a fundamental third piece in the setup of the **shiny** application *iNtr*. This data processing "wrangles" the relatively unstructured NRI estimate tables into one underlying *database*

[2]Panel 4 in Figure 3.5 is for illustrative purpose only and should not be interpreted as the true absolute relative differences between the 2015 and 2012 NRI estimates for the counties in Florida.

that provides the content for all of the charts and tables displayed in the user interface. The user interface has been delineated in section 3.3. We elaborate on the other two parts in the following paragraphs.

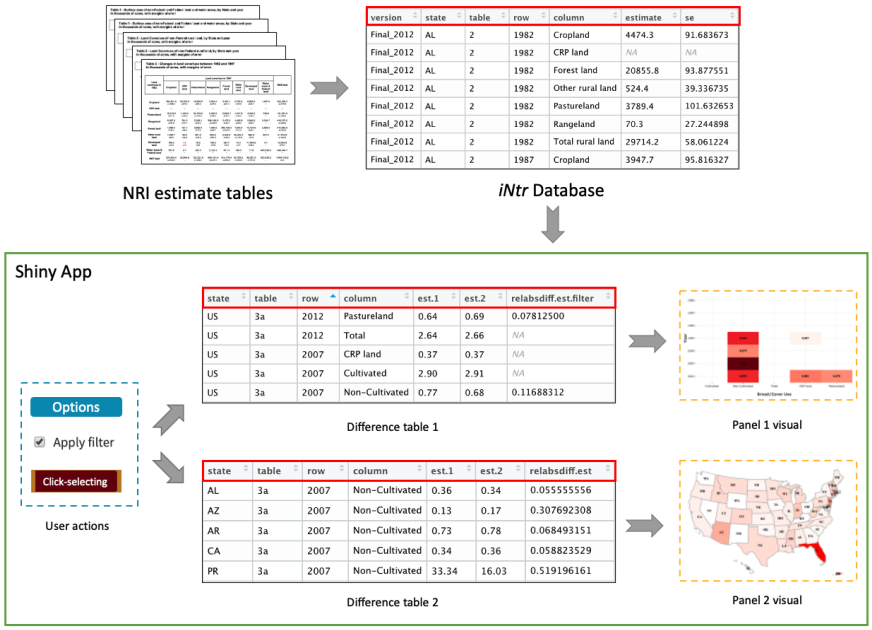

Figure 3.6 Flowchart of the data processing underlying the review system of *iNtr*. On the top left, the collection of NRI estimate tables is shown. Using key descriptors, such as version identifier, spatial level of aggregation, table identifier, etc., we build a single database (top right). This database forms the foundation of extracting subsets of data feeding into the different visualization panels (bottom).

One underlying database Figure 3.6 shows the work flow of *iNtr* data processing. Data in any NRI product are published in a plethora of estimate tables that vary in every aspect, such as topic, dimensions, spatial level of aggregation (U.S. or one of the 50 NRI "states"), etc. What the tables have in common, though, is that they show estimates with margins of error in a tabular format with the same (internal) table identifier throughout the NRI products. For instance, the table, as shown in the left of Figure 3.2 is showing annual estimates of sheet and rill erosion on non-Federal rural land in every version of NRI product. This consistency allows

us to create a single database containing all the information across all versions of released and unreleased NRI products to be compared. When designing a schema for the *iNtr* Database, we prefer a tidy long-form (Wickham, 2014) to assist the data wrangling steps (more details on this application in the next paragraph). To reshape the data into a(n almost) tidy long-form as shown in Figure 3.6, we use version label, spatial level of aggregation, internal table identifier, row index, and column index as five key descriptors to uniquely describe an estimate with its standard error in any NRI estimate table. Another advantage of this schema design is that the *iNtr* Database is so scalable that any newly created version of NRI estimates can be easily appended to this database to support forthcoming review. The *iNtr* Database only contains the original NRI estimates. The differences (evaluation metrics) are calculated on demand because a concise database is preferred for the purpose of reducing the memory load on the R session. This single database containing all NRI estimates is essential for connecting Panels 1 to 4 by employing user-data interactions with "interesting" subsets of the data. These interactions are implemented in the form of selection filters to zoom into the relevant data, as described below.

Selection filters For *iNtr*, there are two main types of user-data interactions — manipulating input widgets in the Options box and click-selecting visuals in Panels 1 to 4. In order for *iNtr* to compute the differences on demand, two versions of NRI estimate tables need to be paired first. User selection of parameter settings in the Options box screens two different versions, one table and one variable out of the *iNtr* Database. A diagram for explaining how to render the visuals in Panels 1 and 2 is presented in the green box of Figure 3.6. Our goal is to first present a national difference table in Panel 1 (as shown at the top right). To achieve that, the *iNtr* server, i.e., the underlying R program, renders an intermediate difference table for Panel 1, as shown in Figure 3.6 when the input widgets in the Options box are set to default. This difference table only contains the calculated differences between the two versions of Table 3a at the national level. Each cell in the national difference table is then uniquely located by `row`, and `column` indexes and shaded based on their values. Once the reviewer click-selects a table cell such as "(2007, Non-cultivated)" in Panel 1, a filter of the row and column identifiers is passed to Panel 2. The second step is to present

a U.S. map of state-level differences in Panel 2 to trace the contribution to that national-level difference from a smaller geographical region (a state in this case). The intermediate difference table rendered by the *iNtr* server for Panel 2, as shown in the bottom middle of Figure 3.6), is obtained by pulling together the 50 computed differences for the 50 NRI "states" with the chosen table, row, and column identifiers. The `state` information in this difference table helps put in place state-level differences in the form of color shade in the U.S. map. Next, in order to present a state-level difference table to be reviewed in Panel 3, each polygon in the U.S. map serves as a filter option. We let the reviewer's click-selecting any state in the U.S. map of Panel 2 decides which state to show in Panel 3. The user-data interaction and selection filter from Panels 3 to 4 work in the same fashion as from Panels 1 to 2. To summarize, the intermediate difference tables of Panels 1 and 3 anchor `state` but vary `row` and `column` while the opposite holds for Panels 2 and 4. The anchor point of Panel 1 is invariant (`state == "US"`), while the anchor points for Panels 2 to 4 are determined by reviewer selection in the preceding panel. The varying identifier(s) serves as "coordinate" for assigning data values to their "spots" in the output visual. A database of county-level NRI estimates is used internally for supporting Panel 4 and has a similar schema to the database presented in Figure 3.6.

3.6 Discussion

In this section, we discuss the benefits of data quality and the broader impacts of the **shiny** application *iNtr*. We first discuss how *iNtr* improves the quality of the NRI data products. We then discuss how our experience provides guidelines to develop similar **shiny** applications.

Maximizing data quality within cost and time constraints is the objective of optimizing data quality management (Biemer, 2010). R **shiny** tools are inherently cost effective. R **shiny** is free and open-source. Development of R **shiny** tools requires little experience with web development. Most of our time and effort is spent on the design and implementation of the user interface and data processing workflow. Our work is dependent on the effort from what Karr et al. (2006) called multiple disciplines of data quality: (1) well designed and managed NRI database by the

computer scientists; (2) sophisticated survey design and estimation by statisticians; (3) data quality metrics determined by the domain knowledge experts. With those prerequisites, *iNtr* was able to be designed, developed, validated, and productionized within the 2015 NRI table review cycle. The CSSM IT team use an internal server to host *iNtr* and limit the tool access for security purpose, which greatly helps with the adoption of the tool. By enabling an efficacious root cause analysis, *iNtr* provides insight into where to correct the errors and improves the *accuracy* of the NRI data. Rather than reducing information, *iNtr* is programmed to effectively filter and display relevant information intended to reduce the cognitive load on reviewers. The tool *iNtr* was used vigorously before the release of the 2015 NRI product. Based on the user feedback we received, *iNtr* has helped to identify issues that the most experienced staff would have been unable to find previously. The cognitive load reduction introduces a more efficient process, so *iNtr* indirectly helps ensure that the NRI product is released on time, thus improving the data quality in the dimensions of *timeliness* and *punctuality*. The tool *iNtr* is currently being used for the 2017 NRI. The NRI staff plan to completely migrate from the legacy table review tool to *iNtr* in the near future.

There are some general steps that we follow when developing an R **shiny** tool from scratch. The first step is to design a user interface. The R package **shinydashboard** is helpful for creating a clean layout, and good visual practices such as those described in Few (2006) help avoid caveats in dashboard design. The second step is to figure out the crucial visuals and make them interactive. The difference tables and maps are foremost for *iNtr*. Before putting all visuals together in a tool, it is recommended to design each visual component under one set of input values using some static plotting tools such as **ggplot2** (Wickham, 2016). Once a working plot has been designed, it can be made interactive by connecting the inputs or mouse events with the plot object. To reduce the cognitive load, we hide auxiliary information from the output plot and have it accessed through tooltips once the reviewer hovers or clicks a data point. Tooltips are supported by a couple of R packages among which **plotly** (Sievert et al., 2018) is quite popular and the R function `ggplotly` can transform a `ggplot` object into interactive `plotly` object directly. The third step is to link certain visual components. The linkage allows drilling down data from an aggregate level to granular

levels. The linkage can be achieved by making the input data of one visual component reactive to the mouse event, i.e., clicking or hovering, of the other visual component(s).

3.7 Conclusion

Evaluating and improving data quality in statistical production is complex and difficult. The tool *iNtr* is a program-specific data quality tool supporting the NRI table review. The tool *iNtr* helps (1) automate measuring data quality metrics; (2) identifying root causes of anomaly; (3) guide the data quality improvement; (4) ensure efforts to improve data quality come into effect. While the impacts of the tool *iNtr* are specific to NRI, many of the general principles underlying the application apply more broadly.

3.8 References

Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., and Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. In *Computer graphics forum*, volume 30, pages 791–800. Wiley Online Library.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848.

Breidt, F. J. and Fuller, W. A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 391–403.

Bushery, J., Reichert, J. W., and Blass, R. F. (2003). U.S. Census 2010 quality assurance strategy. In *JSM Proceedings, Survey Research Methods Section*, pages 749–754. American Statistical Association.

Chang, W. and Borges Ribeiro, B. (2018). *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.1.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

Deming, W. E. (2018). *Out of the Crisis*. MIT press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23.

European Statistics (2015). ESS handbook for quality reports. Luxembourg: Publications Office of the European Union.

European Statistics (2019). Quality report of the European Union Labour Force Survey 2017. Luxembourg: Publications Office of the European Union.

Few, S. (2006). *Information dashboard design*. O'reilly Sebastopol, CA.

Huh, Y., Keller, F., Redman, T. C., and Watkins, A. (1990). Data quality. *Information and software technology*, 32(8):559–565.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173.

Karr, A. F., Sanil, A. P., Sacks, J., and Elmagarmid, A. (2001). Workshop report: Affiliates workshop on data quality. Technical Report, National Institute of Statistical Sciences.

Natural Resources Conservation Service (1994). *National Resources Inventory Training Modules*. Natural Resources Conservation Service.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3):181–204.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2018). *plotly: Create Interactive Web Graphics via 'plotly.js'*. R package version 4.8.0.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley, Reading, MA.

U.S. Department of Agriculture (2015). Summary report: 2012 national resources inventory. Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa.

U.S. Department of Agriculture (2018). Summary report: 2015 national resources inventory. Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Williams, M. and Berg, E. (2013). Incorporating user input into optimal constraining procedures for survey estimates. *Journal of Official Statistics*, 29(3):375–396.

World Health Organization (2017). Data quality review: a toolkit for facility data quality assessment. Licence: CC BY-NC-SA 3.0 IGO.

CHAPTER 4. VISCOVER: A WEB APPLICATION TO VISUALIZE THE SOIL AND LAND COVER DATA AND THEIR OVERLAY

Xiaodan Lyu, Heike Hofmann and Emily J. Berg

Department of Statistics, Iowa State University, Ames, IA

Modified from a manuscript to be submitted to *PLOS One*

4.1 Abstract

The interactions between the pedosphere and land covers have important implications for many environmental processes, such as soil genesis, crop production, soil erosion, etc. The Soil Survey Geographic Database (SSURGO) and the Cropland Data Layers (CDL) contains the soil profiles and the crop-specific land cover classifications data separately at a high spatial resolution covering the United States. To our knowledge, no user interface exists that displays the vast spatial-temporal information contained in the two data sources simultaneously. Based on the **R** **shiny** reactive framework, we develop a web-based application *VISCOVER* to present the two data layers jointly through interactive graphical displays. We enable a tabular presentation summarizing the distribution of crop species for any user click-selected soil segments on the map. *VISCOVER* is originated from a small area estimation project. The model-based small area estimation procedures integrate auxiliary information obtained from the SSURGO and CDL databases. We develop and use *VISCOVER* as a proof of concept to explain what the auxiliary data look like, how we obtain the model covariates, and how uncertain the information might be. The tool is built upon an **R** package we develop to query the two databases by taking advantage of the web services. *VISCOVER* is an open-source project. The source code of the **R** package and the web app is accessible at `https://github.com/XiaodanLyu/viscover`. The user interface of *VISCOVER* is available at

https://lyux.shinyapps.io/viscover. The **shiny** developer community recognizes the technical innovations of *VISCOVER* and have it featured in the Shiny Gallery[1].

4.2 Introduction

Many environmental and agricultural issues depend on an understanding of the interactions between land covers and soil characteristics. The pedosphere is of primary importance to agricultural production. The productivity or fertility information based on soil taxonomy aids precision farming. The joint spatial distribution of soil taxonomy and land covers aids in discovering the soil classes that are better suited for individual plants, especially the cultivated crops that can feed the expanding world population. Sustainable agricultural production requires timely conservation and scientific management of soil resources. According to the soil loss prediction models (Renard et al., 2011), the site-specific conservation practices aimed at reducing soil loss are dependent on the regional-level information about the soil properties, such as slope, erodibility, erosion tolerance, etc. Besides soil texture, vegetation is another crucial factor in the construction or genesis and destruction or erosion processes of soil. Therefore, an overlay of the soil map and the crop-specific land cover map facilitates the exploration into which crop patterns cause more soil loss under the circumstance of water and wind erosion.

The two agencies, Natural Resources Conservation Service (NRCS) and National Agricultural Statistics Service (NASS), of the United States Department of Agriculture (USDA), have been developing and maintaining relevant databases for the public. The NRCS Soil Survey program publishes detailed soil map data. The NASS Cropland Data Layer (CDL) program provides crop-species classifications at a high spatial resolution. Details about the two data sources are given in section 4.3.

NRCS has been conducting the National Resources Inventory (NRI) program in collaboration with the Center for Survey Statistics and Methodology (CSSM) at Iowa State University since 1977. The NRI program augments the NRCS Soil Survey program and produces resource estimates, such

[1] https://shiny.rstudio.com/gallery/viscover.html

as land covers, sheet and rill erosion, wind erosion, etc., through ongoing longitudinal survey studies. The Conservation Effects Assessment Project (CEAP) samples a subset of the NRI points that are classified as cropland or pastureland. It is an on-site study intended to quantify different types of soil and nutrient loss from cropland and pastureland. A data collector of CEAP visits a sampled point and collects detailed information about crop management and conservation practices through a farmer interview. The farmer interview data are then combined with the administrative data, historical weather data, and soil characteristics data. The full set of data is used to produce a regional-level assessment of soil erosion and chemical runoff. The CEAP soil erosion study that motivates this project is described in section 4.4.

The web-based application *VISCOVER* takes its name from "**vis**ualizing **s**oil and **c**ropland data and their **over**lay". An example use of the tool is described in section 4.5. The technical details of the tool development are given in section 4.6. We discuss the general applicability of *VISCOVER* in section 4.7.

4.3 Data Description

Soil Survey Geographical Database The USDA NRCS has been conducting the National Cooperative Soil Survey (Soil Science Division Staff, 2017) since 1896. The observation unit of the soil survey is a pedon, a three-dimensional soil sample that can show the characteristics of all its horizons. The pedons are analyzed in the soil laboratories using standardized methods to obtain chemical, physical, mineralogical, and morphological data. The data are then evaluated to determine the taxonomy, a category in a hierarchical classification, of a particular soil series. Besides classification, another essential goal of the soil survey is the mapping of soils. The Soil Survey Geographical Database (SSURGO), one of the digital soil interpretation databases maintained by NRCS, contains a tremendous amount of tabular and spatial data of soil profiles. The models of soil patterns are essential for map-making (Arnold, 1999). A soil map unit is a geographical entity, a collection of area segments, decided based on the similarity in the use, management, and location of soils (Arnold, 1999). A soil map unit is uniquely identified by its survey area, usually a county,

and an alphanumeric symbol. The soil map unit named "Water" in Figure 4.1 consists of all the water area in the Story county of Iowa, such as the Ada Hayden Lake in northern Ames and the Lake LaVerne at the Iowa State University campus. According to the information displayed in the tooltip, the NRCS Soil Survey estimates about 1,803 acres of the landscape in the Story County is water.

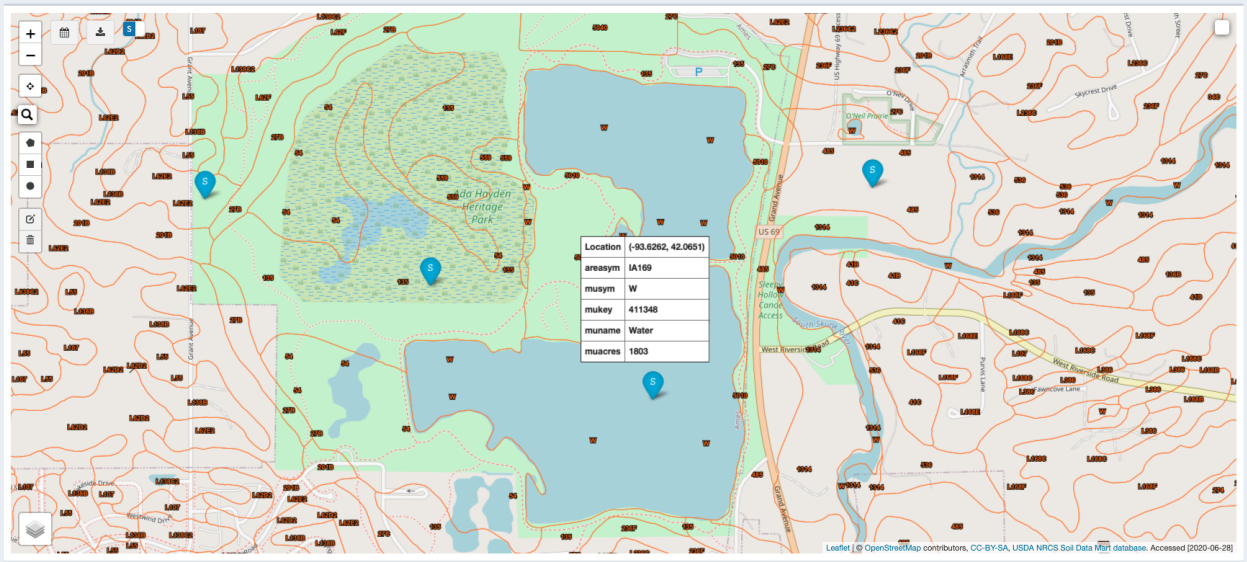

Figure 4.1 A snapshot from *VISCOVER* showing a soil map (lines and alphanumeric symbols in light brown) zoomed near Ada Hayden Lake in the northern Ames. The background layer is the Open Street Map. A blue map marker with an "S" mark appears at each user-clicked location in the map when the toggle at the top left is switched to "S". The tooltip of each map marker contains the identifiers (symbol, key, name, and acres) of the soil map unit containing the pinned location.

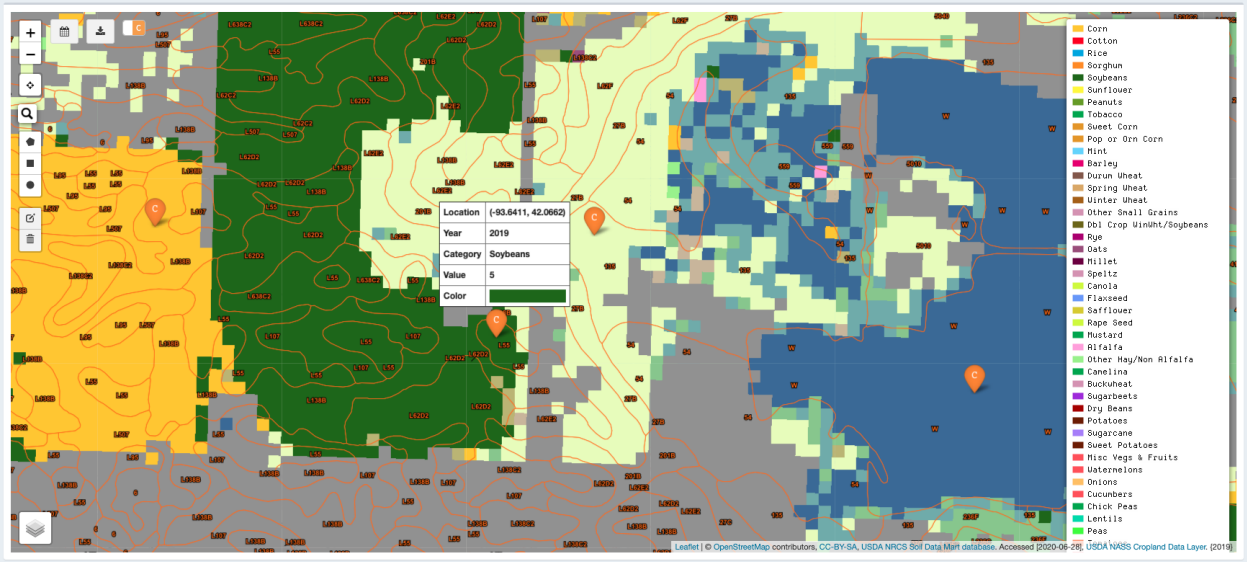

Figure 4.2 A snapshot from *VISCOVER* showing the 2019 Cropland Data Layer overlaid with the NRCS Soil Map zoomed near Ada Hayden Lake in the northern Ames (right blue area). A color legend appears on the right. A yellow map marker labelled "C" is drawn at each user-clicked location in the map when the toggle at the top left of the map is switched to "C". The tooltip describes the land cover classification at the pinned location.

Cropland Data Layer The CDL is a geo-referenced crop-specific land cover data layer produced using moderate-resolution satellite imagery and extensive validation (Boryan et al., 2011). The CDL data were first available in 1997, only for Arkansas. Over time, as the CDL became operational, more states were included. The CDL has been published annually for the contiguous United States since 2009. The CDL data are available in the format of a raster-based image. For the year 2009 and earlier, the ground resolution of the CDL raster is 56 meters. For the year 2010 and newer, the ground resolution is 30 meters. Each pixel in the CDL raster is associated with an integer-valued category code ranging from 0 to 254. Each category code represents a land cover class with a unique Red-Green-Blue color code. The category codes, class names, and color legend

(in the right of Figure 4.2) are consistent for all states and all CDL years. Figure 4.2 is a snapshot of *VISCOVER* presenting the 2019 CDL zoomed near the area as depicted in Figure 4.1. The map indicates there is a soybean field (in green) and a cornfield (in yellow) near the Ada Hayden Lake (blue area in the right). The loamy-textured soils, medium-textured soils with functionally-equal contributions of sand, silt, and clay, are often considered ideal for agriculture. It takes less effort for farmers to cultivate them, and they can be highly productive for crop growth (Parikh and James, 2012). We can tell from Figure 4.2 that most of the corn and soybean pixels of the CDL fall within the soil map units with a symbol starting with "L" which indicates loamy soil texture.

4.4 Motivating Sheet and Rill Erosion Study

One of the estimation objectives in CEAP is the small area estimation (SAE) to obtain estimates of population parameters at each domain of interest when the sample size in each domain is too small to allow reliable design-based estimates. SAE uses model-based estimators to incorporate information from the auxiliary data that are known for the full population (Rao and Molina, 2015). The auxiliary variables play an essential role in the functional form of the predictor. The accuracy and reliability of the auxiliary variables have implications for the quality of the ultimate predictions. Our goal, as described in chapter 2, is to assess county-level average sheet and rill erosion rates on the cropland of South Dakota. The significance of soil properties and crop-specific land cover to soil erosion, and the accessibility of the SSURGO and CDL data as shown in Figure 4.1 and Figure 4.2, motivate us to obtain covariates by linking the two databases with the survey data.

According to the description in subsubsection 2.8.2.2, an overlay operation applied to the SSURGO and CDL data is necessary to obtain the model covariates, i.e., the CDL pixel counts of soybean, sprint wheat, and other crops, per soil map unit. The ultimate auxiliary data set we obtain by overlaying the two data sets is contained in our **R** (R Core Team, 2019) package **saezero** (Lyu, 2020) and named as `Xaux`. As responsible scientists, we would like to have our methods and data verified and explained to the public and the scientific community. The SAE methodology we propose can be validated and peer-reviewed by looking into our manuscript and the source code in

our **R** package. The auxiliary data sources — the SSURGO and CDL data, as well as the overlay operation required for integration, are complex and prone to errors and inconsistencies. We pursue capable graphical display to (1) explain the two data sources and the overlay operation; (2) reveal the quality of the auxiliary variables and the integrity of the overlay operation. This work benefits us in verifying the feasibility of linking the two databases and the accuracy of the overlay operation. For the user interested in working with the auxiliary data set `Xaux`, *VISCOVER* helps with their understanding of the data generation and characteristics. For the users involved in making an executive decision based on our small area predictions, the tool informs them of the reliability concern about the covariates used for producing the outcomes to help them make a fully informed decision.

4.5 User Interface and Example of Use

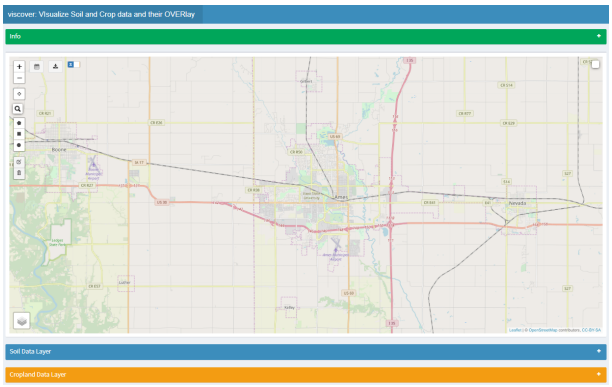

Figure 4.3 Graphical user interface of *VISCOVER*.

Figure 4.3 shows the layout of *VISCOVER*. The base layer of the map in the middle is the open Street Map (Haklay and Weber, 2008). In the map box, the first column of top left are several widgets, i.e., zoom in/out, GPS location, search, draw tools, and edit tools, provided by the **R** packages **leaflet** (Cheng et al., 2018) and **leaflet.extras** (Karambelkar and Schloerke, 2018). Upon start, the tool is centered at the Snedecor Hall, the Department of Statistics at Iowa State University. By zooming out, users can, for example, view the United States as a whole on the

map. The GPS location and the search button are particularly useful in customizing the map and enhance the relevance of the tool by letting user center the map at their current location or any particular location they key in with the search widget. The draw and edit tools are used to allow the user to draw an arbitrary polygon in the map to indicate their area of interest. At the first row of the top left are a calendar button, a download button, and a toggle switch we design for exploring the SSURGO and CDL data at any particular locations. An example will be shown later to illustrate the functionality of those buttons. The layer widget at the bottom left enables users to choose one of the available CDL tiles as a base layer and any soil data layers to be overlaid with the selected base layer. The image at the top of Figure 4.4 depicts an example use of the search and layer widgets.

Figure 4.2 shows a snapshot of the map when the user selects the 2019 Cropland Data Layer. The user can obtain land cover interpretations by referring to the legend on the right of the map. However, the cognitive load (Chandler and Sweller, 1991) is too heavy on the user to associate all land cover classes with their colors simultaneously because of a large number of categories and the limited working memory (Engle, 2002) of the human brain. The cognitive burden is further severed as some of the colors tend to be close to one another. Therefore, the tool *VISCOVER* is designed to allow putting place markers at specific points of interest by clicking the corresponding locations in the map. A single click will produce a yellow marker labeled with a "C" as shown in Figure 4.2. A tooltip is displayed on mouse-hover. The tooltip of the click-selected CDL pixel helps users get intelligible land cover classification immediately. To view the changes in land cover classifications, the user can re-select another CDL year using the layer widget at the bottom left the map. The user has to hit the calendar button to let *VISCOVER* know a different CDL year is now interested. Switching from "C" to "S" on the toggle at the top left of the map, as Figure 4.1 shows, allows users to change the referred database from CDL to SSURGO. At this time, a single click will draw a blue marker labeled "S," and the tooltip includes the location given as geographic longitude and latitude as well as several specific descriptors for each soil map unit. The user can export the CDL

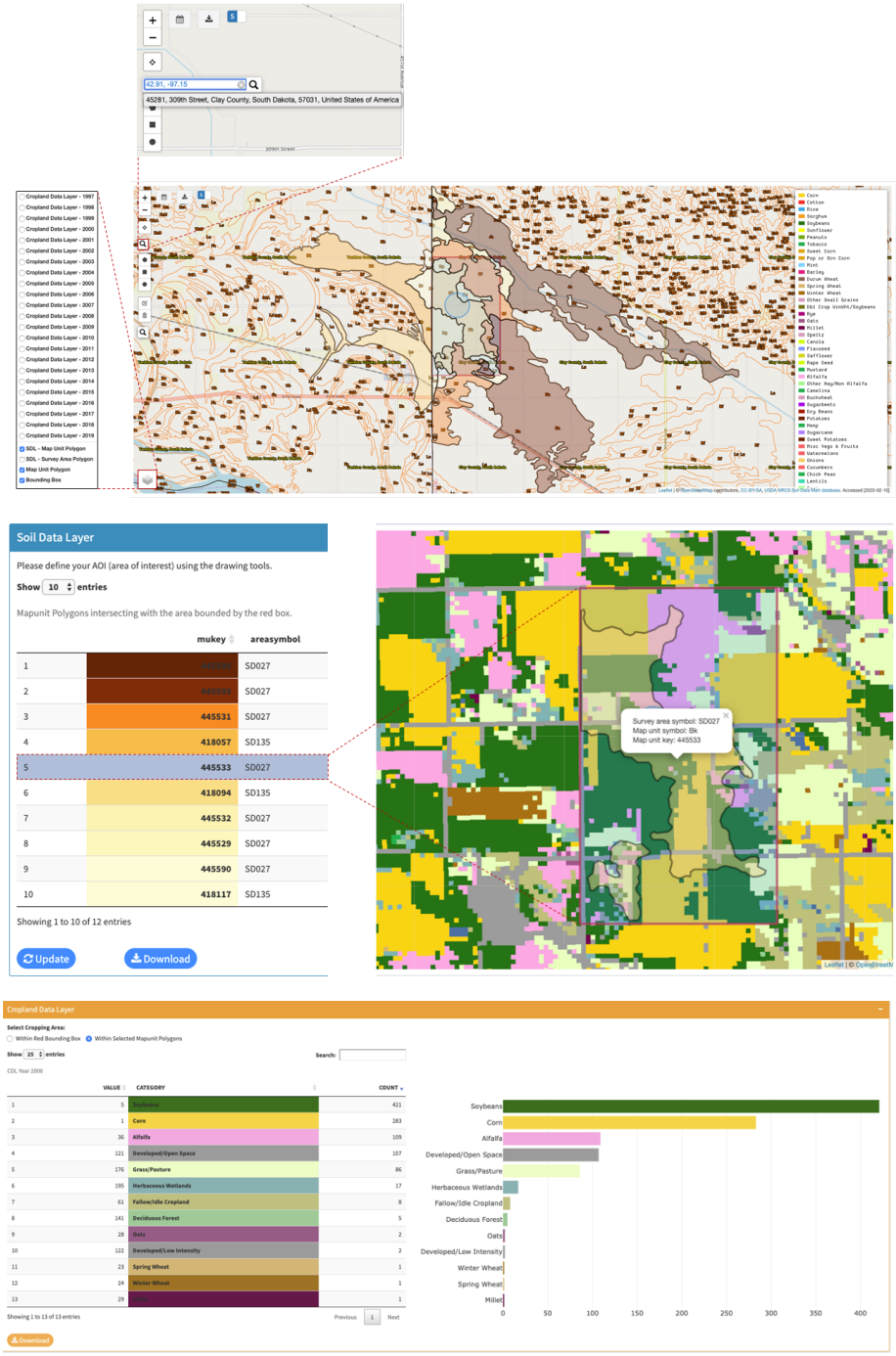

Figure 4.4 An example overlay in *VISCOVER*: (1) search the geolocation to re-center the map and draw a polygon to define an area of interest; (2) highlight a particular soil map unit by selecting the corresponding row in the Soil Data Layer table; (3) unfold the Cropland Data Layer box to view the overlay results.

(or SSURGO) features at the marked locations by hitting the download button. To exclude any unwanted points, the user can remove a map marker by clicking it again.

Next, we go through a specific overlay example. For the map unit, with an identifier key of 445544, in soil survey area SD027 — Clay County of South Dakota, we are interested in obtaining the frequency of each crop cover class. We first recenter the map by inputting a location in this map unit, identified by longitude and latitude, using the search widget. A light blue circle centered at the selected area will be drawn in the map, as shown at the top of Figure 4.4. Then we draw a rectangle in the map using the draw tools and hitting the "Update" button in the Soil Data Layer box to ask *VISCOVER* to scrape data from the SSURGO database on the fly. Once the download is complete, all the soil map unit polygons overlapping with the red bounding box will be presented on the map.

Meanwhile, a data table containing the features of the scraped map units is shown in the Soil Data Layer box. Since we are only interested in the soil map unit 445544, highlighting the corresponding row in the data table will hide other deselected soil map units, as shown in the middle of Figure 4.4. In this example, we are interested in the 2006 CDL for the sheet and rill erosion study described in section 4.4. A data table and a bar plot of the tabulated CDL pixels ordered by descending frequency within the soil map unit 445544 can be found in the Cropland Data Layer box, as shown at the bottom of Figure 4.4. We can learn from the bar plot that the 2006 CDL classified most of the points within this soil map unit as soybean, followed by corn and alfalfa. By referring to the map, we can verify that the results from the overlay operation are correct.

4.6 Design and Implementation

The Web Soil Survey (Soil Survey Staff, 2005) is a user interface maintained by NRCS to interact with the NRCS Soil Survey Database. The CropScape (Han et al., 2012) is a web-based tool by NASS for users to interact with the Cropland Data Layers. Neither of them is letting users look at the two databases jointly. Both of them have a complicated graphical user interface.

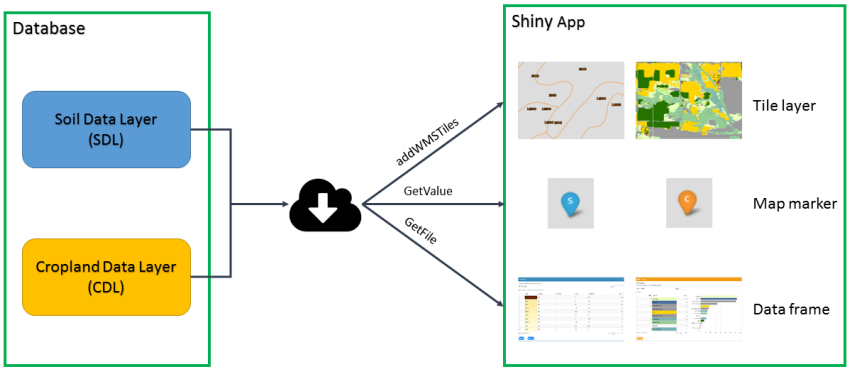

Figure 4.5 Primary functionalities of *VISCOVER*. It grabs the data from the soil and land cover databases to its user interface on demand by using web services.

The tool *VISCOVER* has three primary functionalities as shown in Figure 4.5. From the SSURGO, also denoted as Soil Data Layer (SDL) in Figure 4.5, and the CDL databases, we let *VISCOVER* request real-time data and display maps with *tile layers*, pinpoints with *map markers*, and overlay results with *data frames*.

We write an **R** function `TileinPoly` for overlaying a raster object and a spatial polygon object. It utilizes the method "over" in the **R** package **sp** (Pebesma and Bivand, 2018) to retrieve attributes from the CDL raster tile at those locations falling within a soil map unit. To our knowledge, there is no available tool to overlay raster and spatial polygons directly in **R**. One has to transform raster to spatial points first. Besides the `over` method, the `st_join` function in the package **sf** (Pebesma, 2018) can also be used to obtain the intersection. The on-demand pinpoints and overlay are effectuated under the condition that we can crawl data from the two databases upon the user's request. We describe how we query the two databases in subsection 4.6.1 and how we load the map tile layers into the map in subsection 4.6.2.

4.6.1 Web Scraping

In this section, we denote "GetValue" as to query a data vector for a location and "GetFile" as to query both spatial and tabular data for a small area. For soil data, we exploit the **R** package **soilDB** (Beaudette et al., 2018), which provides ready-to-use **R** functions for extracting soil information from Soil Data Access (SDA). The SDA website is a suite of web services and applications which facilitate querying tabular and spatial soil data from the SSURGO database on demand for an area of interest of any size with a limit of up to 250,000 features. Specifically, we utilize two functions called `mapunit_geom_by_ll_bbox` and `SDA_query`. The **R** function `mapunit_geom_by_ll_bbox` fetches map unit geometry with a bounding box identified by longitudes and latitudes. `SDA_query` is another **R** function for submitting a Structured Query Language (Hursch et al., 1988) query to SDA and returning a data frame. Given a point with the World Geodetic System 1984 (WGS84) coordinate, we use a circle centered at that point with a radius as short as 0.1 meters in the query to return associated features at the soil map unit level. The returned features are approximately soil interpretations at that point.

The CDL web geoprocessing service includes `GetCDLFile` and `GetCDLValue`. The web geoprocessing service deals with web requests and is supported by the underlying ArcGIS server — a mapping and analytics platform. `GetCDLFile` returns a CDL raster file given a spatial extent while `GetCDLValue` returns corresponding CDL category code, color code and class name given a geographic coordinate. Those requests require all coordinates to be in the coordinate reference system (CRS) projection of USA Contiguous Albers Equal Area Conic (USGS version), and the query to the CropScape is limited to areas less than 2 million square kilometers. To fulfill the functionalities described in 4.5, we are interested in obtaining the land cover interpretations in a small area of any size and at any point location. To our knowledge, the existing **R** packages such as **cdlTools** (Chen, 2018) are not designed for downloading the CDL data at such a small scale. Therefore, we write our own **R** functions `GetCDLFile` and `GetCDLValue` to web scrape the CDL data. `GetCDLValue` returns point-level CDL features given a WGS84 coordinate. Since the web request returns an Extensible Markup Language (Bray et al., 2000) file, we use the **R** package **XML** (Lang and the

CRAN Team, 2018) to clean the results and return a list containing the CDL code, category and color. `GetCDLFile` is used to download CDL raster given a small spatial extent in the same way as `mapunit_geom_by_ll_bbox`. A temporary raster image is downloaded locally and then read into **R** using the `raster` function in the **R** package **raster** (Hijmans, 2017).

4.6.2 Tile Layer

We use the **leaflet** framework to set up the map and to support those functionalities listed in Figure 4.5. Leaflet (Agafonkin, 2019) is one of the most widely-used open-source JavaScript libraries for interactive maps. The **R** package **leaflet** is a tool to integrate and control Leaflet maps in **R**. Both the SSURGO and CDL support the Web Map Service (WMS) such that they can be loaded into the **leaflet** map as tile layers using the **R** function `addWMSTiles`. The WMS (de La Beaujardiere, 2006) provides a simple web interface for requesting map images that can be displayed in a browser application from one or more distributed geospatial databases. It is critical to specify the correct CRS in tile options to submit a legitimate request to the WMS. The following code example shows how to accommodate the CRSs.

```
leaflet() %>%
      addWMSTiles(
        baseUrl = "https://sdmdataaccess.sc.egov.usda.gov/Spatial/SDM.wms",
        layers = "MapunitPoly", group = "SDL - Map Unit Polygon",
        options = WMSTileOptions(format = "image/png", transparent = TRUE,
                                 crs = leafletCRS(crsClass = "L.CRS.EPSG4326")),
        attribution = sprintf('<a href="https://sdmdataaccess.sc.egov.usda.gov">
                               USDA NRCS Soil Data Mart database</a>.
                               Accessed [%s]', Sys.Date()))
```

The successful implementation of the WMS in **leaflet** needs a thorough study of the database documentation. The instruction page of the SDA contains information about the base URL of SSURGO and the name of the available data layers. *VISCOVER* displays two of the soil data layers, i.e., survey area polygon and map unit polygon. The CropScape developer guide contains

instructions for the available web data services and web geoprocessing services of the CDL data. The base URL of CDL and the names of the raster layers can be found there.

4.6.3 Technical Highlights

VISCOVER is now featured in the Shiny Gallery. The Shiny Gallery highlights example **shiny** applications from the community's top developers for **shiny** users to learn from. The technical merits of *VISCOVER* are the followings. It combines the visualization of spatial data and web scraping coherently. The web scraping or crawling is quite popular nowadays as there are more and more good-quality data sources freely available online. *VISCOVER* demonstrates that the web scraping of large scale spatial data can be well utilized through a user interface containing an interactive map. *VISCOVER* is one of the few **shiny** applications that enable a plethora of user-data interactions with the build-in map, such as locating, searching, selecting, clicking, hovering, drawing, and downloading. **R** (R Core Team, 2019) has been overshadowed by other computer languages for its slow computation and low-memory-capacity, which makes it a less appealing tool for working with the data of large scale such as SSURGO and CDL. On the other hand, **R** contains a powerful library of visualization tools, which plays a vital role in enabling the communication of the vetted data and analysis to the end-users. *VISCOVER* breaks the "big" data manipulation into "small" pieces through real-time data request upon the user's click-selection of a small number of points or a small area of interest. At the same time, the ability to allow the user to preview the data through a chunk of WMS images catches the user's attention and inspires their interest to make in-depth exploration.

Besides **R shiny**, one can also use ArcGIS to develop web applications that enable user interactions with large-scale spatial datasets. One incomparable advantage of **shiny** over ArcGIS is the convenience of endorsing the user interface with state-of-art statistical data analysis. For example, one potential and interesting use case of *VISCOVER* to NRI is to provide small area predictions of natural resources such as soil loss from sheet and rill erosion for any click-selected area by land managers. It is promising for us to solve this problem by implementing existing and

under-development statistical methodologies in small area estimation, such as the one described in chapter 2. Most of the statistical methodologies are accessible in related R packages or programs. We can use those R tools straightforwardly for adapting *VISCOVER* to this specific use case under the **shiny** framework Chang et al. (2018).

4.7 Discussion

VISCOVER explores the visualization of large-scale spatial data, i.e., SSURGO and CDL, using **R shiny**. Its original purpose is to serve as a diagnostic tool to assess the eligibility and accuracy of an overlay operation needed to define the covariates for a unit-level small area estimation model applied to the soil erosion assessment of the counties in South Dakota. Its impact on the CEAP data analysis is discussed in subsection 4.7.1. Its extended purpose is to facilitate the public's understanding of the interactions between land covers and soil properties covering the United States. *VISCOVER* can be used for various purposes as it is. For example, students and instructors can use *VISCOVER* as an instrumental tool to learn the landscape around them during a field trip. A variety of users, such as data analysts, land managers, and land operators, can use *VISCOVER* to obtain information from the SSURGO and CDL databases on-site. *VISCOVER* can also be adapted or further developed by modifying the source code to suit program-specific requirements or needs. subsection 4.7.2 describes other use cases of the soil map, and subsection 4.7.2 presents the use case of the land cover data layers by the NRI State Review.

4.7.1 Impact on the CEAP Data Analysis

We develop *VISCOVER* to visually assess the quality, in terms of *accuracy*, *coherence*, and *reliability*, of the SSURGO and CDL data, as they pertain to the auxiliary data sources of our small area estimation models. The presentation of the auxiliary information in a refreshing way has an indirect positive impact on the quality of our predictions. It helps build the *credibility* of our predictions in that it allows users to see the data and follow our line of reasoning. The *accuracy* and *reliability* of our predictions can be further enhanced as *VISCOVER* provides an accessible

user interface for subject matter experts to validate our choices of covariates against their domain knowledge. *VISCOVER* reveals some *coherence* problems by helping us identify the discrepancies and inconsistencies between the SSURGO and CDL data. For instance, Figure 4.2 shows the blue CDL pixels of class "water" do not always line up with the soil map units labeled "W" (water). The soil map units with a dominant soil component among "water", "badlands", "rock outcrop", "rubble land", and "pits" are known to be not suitable for crop production. Because SSURGO delineates the boundary of the actual fields more accurately than CDL, and the conditions of such kind of soil map units are less subject to change, we exclude all the points within those soil map units from our CEAP-Cropland population frame regardless of their CDL land cover classes.

4.7.2 Uses of the Soil Map

In addition to crop production and soil conservation, the soil map can also be used for urban planning and wetland conservation. The knowledge of the soil types on-site helps a construction project decide whether the soils can hold up buildings. The identification of hydric soils can determine the location of wetlands. The "Locate Me" button in *VISCOVER* would be of potential use to those situations as there is increasing use of digital devices in the scene. The tool *VISCOVER* extends the *accessibility* and *usability* of the digital NRCS Soil Map. It is web-based, thus compatible with many portable devices, such as laptops, tablets, and smartphones, and any operating systems, as long as they are equipped with a working web browser.

At a global scale, soil maps and databases are considered as useful information for assessing land productivity and soil erosion. The Food and Agriculture Organization (FAO) and the Educational, Scientific, and Cultural Organization (UNESCO) of the United Nations maintains a 40-years-old digital soil map of the world at a scale of 1:5 million. The Harmonized World Soil Database (Nachtergaele et al., 2010) at a one-kilometer spatial resolution synthesizes worldwide updates of soil information, such as the Soil Map of China at a scale of 1:1 million (Shi et al., 2004), with the FAO/UNESCO Soil Map of the World. A future global digital soil mapping project (Sanchez et al., 2009) is aiming at a revised global product from the FAO/UNESCO Soil Map. The FAO

Soils Portal[2] contains a list of existing regional and national soil maps downloadable in a raster format. Feasible and transferable agro-technology is essential to provide technical assistance to less developed countries or areas. The NRCS Web Soil Survey (Soil Survey Staff, 2005) is a powerful user interface to present soil data. On the other hand, the way *VISCOVER* presents the NRCS Soil Map can be a simple alternate solution to develop a functioning user interface of soil maps. The development of *VISCOVER* is based on the **shiny** framework and the supporting web services. Those soil maps of other countries or regions, once equipped with the web services, can be made readily accessible through a web browser by adapting our source code, available at https://github.com/XiaodanLyu/viscover.

4.7.3 Uses of the Cropland Data Layer by NRI

VISCOVER can be used in the NRI State Review process, as described in chapter 1. A specific example is to use *VISCOVER* to query the affected NRI points questioned by the State Resources Inventory Coordinators. For example, a state says they do not have a vineyard to their local knowledge, but NRI has a non-zero estimate for it. In that case, the NRI analysts at CSSM can query out the relevant points. In doing that, they use the latitudes and longitudes to search for those points in *VISCOVER* and quickly get the corresponding CDL land cover uses for the years they are interested in. Currently, an effort has been devoted to customizing *VISCOVER* to fulfill more specific needs of the NRI State Review. Some of the functionalities under development include allowing a batch query to the CDL for a list of NRI points and multiple years, verifying whether there is any CDL pixel of a particular land cover class such as wheat in a county for several years, etc. The self-service functionality of *VISCOVER* reduces the workload of the ArcGIS analysts at CSSM and also boost the task efficiency of the NRI analysts since they can query the results instantly using the web-based user interface of *VISCOVER*. As such, *VISCOVER* helps reduce the staffing and training cost of NRI.

[2] http://www.fao.org/soils-portal/en/

4.8 References

Agafonkin, V. (2019). *leaflet: an open-source JavaScript library for mobile-friendly interactive maps*. JavaScript library version 1.6.0.

Arnold, R. W. (1999). The soil survey: past, present and future. U.S. Department of Agriculture, Natural Resources Conservation Service.

Beaudette, D., Skovlin, J., and Roecker, S. (2018). *soilDB: Soil Database Interface*. R package version 2.0-1.

Boryan, C., Yang, Z., Mueller, R., and Craig, M. (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer program. *Geocarto International*, 26(5):341–358.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., et al. (2000). Extensible markup language (XML) 1.0.

Chandler, P. and Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

Chen, L. (2018). *cdlTools: Tools to Download and Work with USDA Cropscape Data*. R package version 0.13.

Cheng, J., Karambelkar, B., and Xie, Y. (2018). *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.0.0.

de La Beaujardiere, J. (2006). OpenGIS® Web Map Server Implementation Specification. Version 1.3.0.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23.

Haklay, M. and Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

Han, W., Yang, Z., Di, L., and Mueller, R. (2012). Cropscape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture*, 84:111–123.

Hijmans, R. J. (2017). *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7.

Hursch, C. J., Hursch, J. L., and Hursch, C. J. (1988). *SQL, the Structured Query Language*. Tab Books.

Karambelkar, B. and Schloerke, B. (2018). *leaflet.extras: Extra Functionality for 'leaflet' Package*. R package version 1.0.0.

Lang, D. T. and the CRAN Team (2018). *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.98-1.12.

Lyu, X. (2020). *saezero: Small Area Estimation under a Zero Inflated Lognormal Model with Correlated Random Area Effects*. R package version 0.1.0.

Nachtergaele, F., van Velthuizen, H., Verelst, L., Batjes, N., Dijkshoorn, K., van Engelen, V., Fischer, G., Jones, A., and Montanarela, L. (2010). The harmonized world soil database. In *Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010*, pages 34–37.

Parikh, S. J. and James, B. R. (2012). Soil: the foundation of agriculture. *Nature Education Knowledge*, 3(10):2.

Pebesma, E. (2018). *sf: Simple Features for R*. R package version 0.6-0.

Pebesma, E. and Bivand, R. (2018). *sp: Classes and Methods for Spatial Data*. R package version 1.2-7.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Renard, K. G., Yoder, D. C., Lightle, D. T., and Dabney, S. M. (2011). Universal soil loss equation and revised universal soil loss equation. *Handbook of erosion modeling*, pages 137–167.

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A. E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A. B., McKenzie, N. J., de Lourdes Mendonça-Santos, M., et al. (2009). Digital soil map of the world. *Science*, 325(5941):680–681.

Shi, X., Yu, D., Warner, E., Pan, X., Petersen, G., Gong, Z., and Weindorf, D. (2004). Soil database of 1: 1,000,000 digital soil survey and reference system of the Chinese genetic soil classification system. *Soil Survey Horizons*, 45(4):129–136.

Soil Science Division Staff (2017). *Soil survey manual*. USDA Handbook 18. Government Printing Office, Washington, D.C.

Soil Survey Staff (2005). Web soil survey. Natural Resources Conservation Service, U.S. Department of Agriculture.

CHAPTER 5. INTERACTIVE SHEET AND RILL EROSION MAP OF SOUTH DAKOTA AT A 30-METER SPATIAL RESOLUTION

Xiaodan Lyu, Emily J. Berg, Heike Hofmann and Zhengyuan Zhu

Department of Statistics, Iowa State University, Ames, IA

Modified from a manuscript to be submitted for publication

5.1 Abstract

In this chapter, we use a statistical model to construct interpretable predictions of sheet and rill erosion in tons per acre per year at a 30-meter spatial resolution covering the cropland of South Dakota. The underlying model mimics the structure of the Universal Soil Loss Equation (USLE), a widely accepted approximation for sheet and rill erosion. USLE models sheet and rill erosion as a product of several factors related to climate, soil properties, cropping management, and conservation practices. To obtain the auxiliary variables that approximate the USLE factors, we integrate several data sources that cover expansive geographic domains without sacrificing spatial granularity. The National Land Cover Dataset (NLCD) provides a 30-meter grid on which we construct erosion predictions. We obtain information on soil erodibility and slope gradients determined for the dominant soil components in each soil map unit from the Soil Survey Geographic Dataset (SSURGO). We derive a measure of slope steepness from the National Elevation Dataset (NED). The NED-derived slopes are then aligned with the SSURGO soil slopes to transform soil data from map-unit level to grid-cell level. We use interpolation to down-scale the USLE rainfall factor from NRI sampled locations to NLCD grid cells. We use a random forest model to predict the USLE crop management factor using auxiliary data from a satellite-derived land-cover map called the Cropland Data Layer (CDL). We obtain approximated USLE soil erodibility, soil slope, and rainfall factors at every NLCD grid cell, and crop management factors only at the grid cells

classified as cropland by NLCD. We use a combination of geographic overlays and statistical models to transform the approximations for the USLE factors into predictions of sheet and rill erosion at a 30-meter spatial resolution. We develop a web-based interactive map tool to present the approximated USLE factors and the estimated sheet and rill erosion rates. The tool named Sheet and Rill Erosion Map (SREM) is available at https://lyux.shinyapps.io/srem/.

5.2 Introduction

The National Resources Inventory (NRI) is a longitudinal survey that monitors natural resources on the non-federal land of the United States. A primary product of the NRI is a set of estimates of average sheet and rill erosion rates on cropland, Conservation Reserve Program (CRP) land, and pastureland at the national and state level, such as those depicted in Figure 3.2. The NRI visualization products have been focusing on only cropland, because the spatial pattern of soil erosion on cropland is more complex and interesting than that on CRP land and pastureland. The NRI made a preliminary attempt to estimate and visualize the sheet and rill erosion rates on cropland at a one-mile spatial resolution[1]. This visualization product is obtained by interpolating the recorded sheet and rill erosion rates at NRI cropland sample sites within a 50-mile radius. The interpolation operation makes strong implicit assumptions about the spatial distribution of soil erosion. Furthermore, this product does not reflect, through explicit assumptions, the physical processes that govern erosion.

A traditional measure of sheet and rill erosion is the Universal Soil Loss Equation (USLE). The USLE equation is defined as

$$A = R\,K\,L\,S\,C\,P, \tag{5.1}$$

which measures the soil loss from sheet and rill erosion per unit area A as the multiplication of rainfall factor R, soil-erodibility factor K, slope-length factor L, slope-gradient factor S, cropping-management factor C, and erosion-control practice factor P (Wischmeier and Smith, 1965). NRI uses USLE to collect sheet and rill erosion data before 2008. Then the survey transited to a

[1] https://www.nrcs.usda.gov/Internet/NRCS_RCA/maps/erosion_animation15.gif

more advanced computer model called RUSLE2, short for Revised Universal Soil Erosion Equation, Version 2 (Renard et al., 2011). RUSLE2 keeps the fundamental empirical equation structure of the USLE approach but incorporates process-based equations to allow for enhanced functionalities such as a more granular time step. The NRI Database contains about 800,000 sample point locations that describe the field conditions across the United States from 1982 to the present (U.S. Department of Agriculture, 2018). Each point in the NRI Database has an estimation weight and those points classified as cropland, CRP, or pastureland have a number of soil loss data elements.

The goal of our project is to create a visualization product, which improves the current NRI visualizations in spatial resolution, erosion estimation, usability, and accessibility. Starting with the state South Dakota and the year 2006, we collect covariates related to the USLE factors and use the recorded USLE losses in the NRI Database as responses to predict the sheet and rill erosion rate at each cropland location. Cropland, according to the definition of NRI, is a land-use category that includes areas used for the production of adapted crops for harvest. The 2006 NRI recorded the sheet and rill erosion rates at 2,572 cropland sample points in South Dakota. Figure 5.1 gives an overview of how we use various data sources to approximate USLE factors at all 30-meter grid cells classified as cropland by the National Land Cover Data (NLCD). We explain our approach to data integration in the following paragraphs.

The National Land Cover Data (NLCD) is created by the Multi-Resolution Land Characteristics Consortium of the United States Geological Survey (USGS). The NLCD products include the raster images which provide a spatial reference to the land cover classifications such as cultivated crops, water, urban, forest, wetland, and so on. The 2006 NLCD has a 30-meter spatial resolution and covers the conterminous United States. We use the 2006 NLCD as a base layer to define a population frame $\mathcal{D} = \{(i, j) : i = 1, \ldots, N_i, j = 1, \ldots, N_j\}$, where i and j denote the cell row and column indexes respectively on a raster image. The spatial resolution of $\mathcal{D}$ is 30 meters covering South Dakota. There are about 300 million cells in $\mathcal{D}$. For the cell $(i, j) \in \mathcal{D}$, we denote $\gamma_{ij} = 1$ if the cell is classified as "cultivated crops" by the 2006 NLCD and 0 otherwise. To predict the USLE loss on

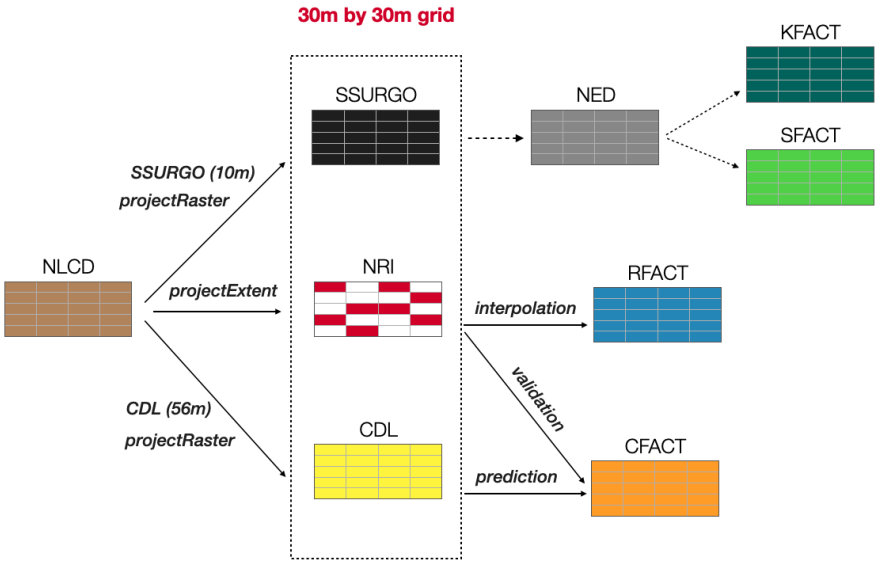

Figure 5.1 Workflow of integrating the NRI data with other data sources to obtain soil erosion factors at a 30-meter spatial resolution covering South Dakota.

cropland, we need to find the soil erosion factors at the grid cells where $\gamma_{ij} = 1$. About 1/3 of the landscape is classified as cropland by NLCD in South Dakota.

In a preliminary data analysis, we explore which erosion factors explain the most variance in the recorded USLE losses by the 2006 NRI. After removing two points measured to have zero USLE loss, we find 40% of the variance in the logarithmic USLE losses can be explained by a linear model regressed with the NRI recorded R, K, and S factors in log scale denoted as `logR`, `logK`, and `logS`. The log transformations are implemented because the USLE equation indicates the soil loss rate has a multiplicative relationship with the soil erosion factors. The proportion of explainable variance in the responses increases to 98% if another covariate — the recorded C factor in log scale denoted as `logC` is added to the model. Under this circumstance, we decide that the R, K, S, and C factors are sufficient to predict the USLE losses at the grid-cell level. The P factors are unobtainable. To our knowledge, there is no practical way to collect the data related to conservation practices, such as tillage and terrace, at all cropland locations. The L factors are not collected either because the related feature in the SSURGO database has a large number of missing values.

The K and S factors are obtained by linking the Soil Survey Geographic Database (SSURGO) with the National Elevation Data (NED). The section 5.3 gives the details of the data linkage. The subsection 5.4.1 describes how we obtain the approximated R factors by interpolating the records in the NRI Database. The subsection 5.4.2 presents the way we use the Cropland Data Layers (CDL) to approximate the C factors. The approximations are validated using the recorded C factors in the NRI Database. The section 5.5 presents the fitting results of our erosion prediction method and the final visualization product. We discuss the impact of our visualization product on the data quality of NRI and future work in section 5.6.

5.3 Soil Erodibility and Slope Factors

The SSURGO database maintained by the Natural Resources Conservation Service (NRCS) contains detailed information on the soil resources in the United States. The most granular geo-reference information in SSURGO is available at the soil map unit level. A soil map unit outlines homogeneous soil segments within a soil survey area, usually a county. The auxiliary information related to the erosion factors contained in the SSURGO database, i.e., soil erodibility, slope length, and slope gradient, are the characteristics of the soil components residing in a soil map unit. The subsection 5.3.1 describes how we obtain a complete list of representative soil components for South Dakota from SSURGO. The exact geo-locations of the interesting soil components are not available. To transfer the soil data from a map unit level to grid-cell level, we align the SSURGO slope per soil component with the slope derived from the NED data per cell in the population frame $\mathcal{D}$. The subsection 5.3.2 presents the alignment procedure.

5.3.1 Soil Components Inventory

Figure 5.2 shows part of the hierarchical data structure in the SSURGO database. A detailed description of the table schema can be found in Table 5.2 in the Appendix. A soil survey area indexed by legend key (`lkey`) and area symbol is usually a single county. Each soil survey area consists of a list of soil map units. A soil map unit indexed by a key identifier (`mukey`) delineates

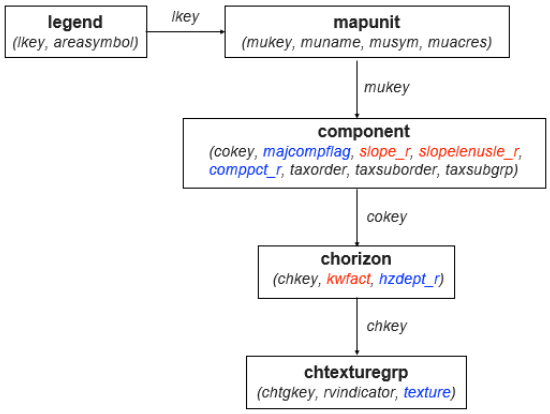

Figure 5.2 Partial hierarchy of the NRCS Soil Survey Database. Representative soil erosion K and S factors of a soil component in a soil map unit of a soil survey area are determined by its texture and horizontal depth jointly.

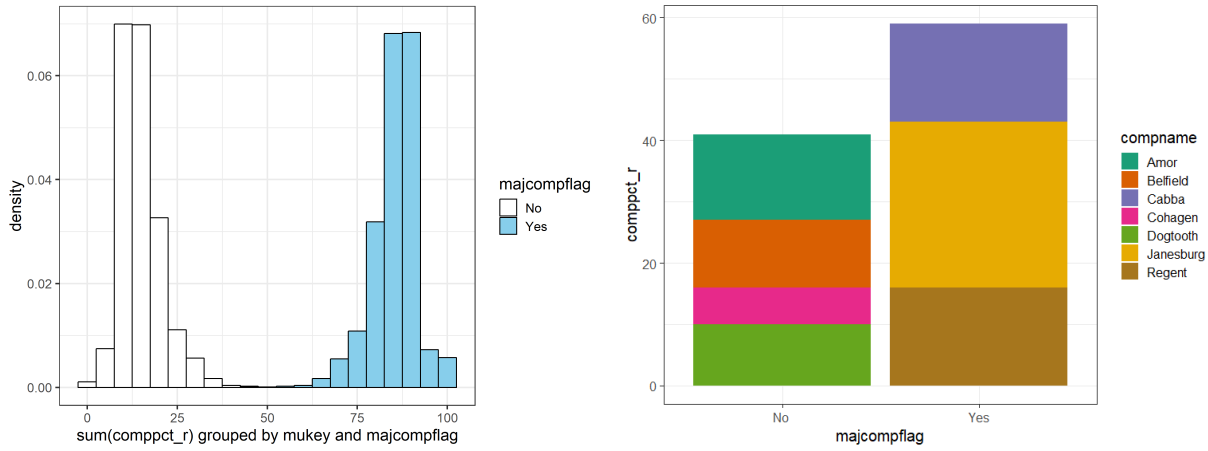

Figure 5.3 Left: histogram of the percentages of the major and the minor soil components per soil map unit. Right: the stacked bar plot of the percentages of the three major and the three minor components in the example soil map unit "Janesburg-Regent-Cabba complex".

areas with homogeneous soil components. The soil components in a soil map unit are identified as "major" or "minor" based on their percentages. The left panel of Figure 5.3 shows that the sum of the percentages (`comppct_r`) of the major soil components (`majcompflag` = "Yes") in a soil map unit is typically no less than 50%. In Figure 5.3, the example soil map unit, called "Janesburg-Regent-Cabba complex", is named after its three major components. The representative slope (`slope_r`) in the soil component table gives the characteristic slope gradient s of a soil component. The slope gradient can then be used to derive the USLE slope gradient factor S. For instance, Wischmeier and Smith (1965) used the following relation

$$S = \frac{0.43 + 0.30s + 0.043s^2}{6.613}, \tag{5.2}$$

proposed by Smith and Wischmeier (1957). The USLE slope-length factors (`slopelenusle_r`) recorded for the soil components in the SSURGO is impracticable for us to use because it is subject to a missing rate of around 65% for the major soil components in South Dakota. The feature, named as `kwfact` in SSURGO, quantifies the USLE soil-erodibility factor K. We can typically determine the K factor of a soil component by its kind, name, and texture. The texture of a soil component varies by the horizon. Typically, the top horizon decides characteristic soil texture per soil component. For non-histosol soil components, however, the top mineral horizon is chosen to determine representative soil texture. Therefore, the K factor of a soil component is located by its component-horizon identifier in the SSURGO database. The algorithm 1 in the Appendix presents our queries to the SSURGO database for a complete list of major soil components and their appropriate K and S factors. We inventory 7,378 soil map units, correspondingly 11,729 major soil components, in a data table. For simplicity, we record at most four major soil components per soil map unit. We can attain K and S factors for all soil components with a few exceptions. Slope gradients are undefined for any water component. The USLE K factors are not appropriate and thus not populated for any histosol or most "miscellaneous" soil components such as water, rock outcrop, badland, and so on.

5.3.2 Soil Map Units to Points

As soil erosion K and S factors have been determined for almost every major soil component, next, we assign those soil components to the cells in the population frame $\mathcal{D}$. The K and S factors for each cell are then decided according to the soil component. One way to distribute the soil components within a soil map unit is to follow the distribution of the slope gradients. As such, we align the soil slope gradients with the slopes we derive from the National Elevation Data (NED). The following paragraphs describe how we designate the NED-derived slopes and then align them with the SSURGO slope gradients.

5.3.2.1 NED-derived slope

The NED data maintained by USGS include a set of standard digital elevation models (DEMs) at various horizontal resolutions. The 1/3 arc-second DEM is the spatial data set that has the highest resolution and covers the 48 conterminous states. 1/3 arc-second is approximately 10 meters in the north-south direction. We denote the NED raster with a 1/3 arc-second resolution as $\mathcal{D}'$. We denote the NED elevation at the cell of the row r and the column c as λ_{rc}. We define a NED-derived slope β_{rc} as

$$\beta_{rc} = \frac{1}{8} \sum_{|r-r'|\leq 1, |c-c'|\leq 1} |\lambda_{rc} - \lambda_{r'c'}|, \tag{5.3}$$

which is the average absolute difference on elevation between the cell (r, c) and its eight nearest neighbors in the NED raster. Figure 5.4 shows how to conduct this computation in practice. We first shift the original NED raster toward eight different directions by 10 meters. Then we calculate the absolute differences between the shifted raster images and the original raster. Lastly, we average the eight resulted difference raster images. The obtained raster image excludes the most edged rows and columns. Figure 5.5 depicts the NED elevation heat map and the corresponding derived-slope heat map zoomed in to a rectangular area that covers the example map unit "Janesburg-Regent-Cabba complex".

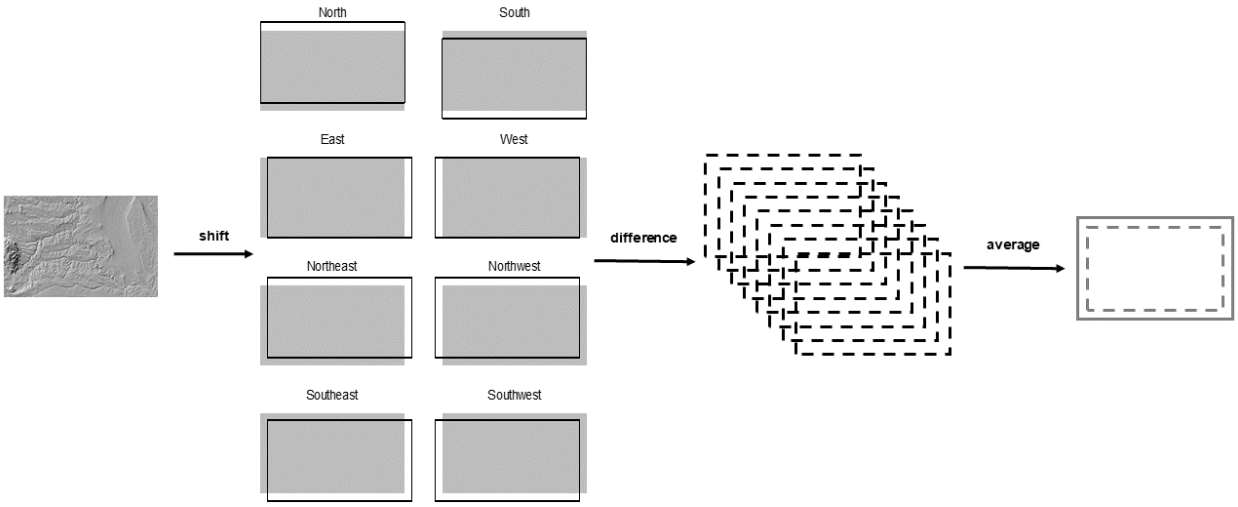

Figure 5.4 The flowchart of the calculations of the NED-derived slopes defined as the average absolute difference among each cell and its eight nearest neighbors: (1) shifting the NED elevation raster toward eight different directions; (2) calculating the absolute differences between the original raster and the eight shifted raster images; (3) stacking and averaging the eight difference raster images.

In geography, a more standard approach to derive slope from elevation is to use the maximum of the elevation differences, defined as

$$\beta_{rc}^{\text{Alt}} = \max_{|r-r'|\leq 1, |c-c'|\leq 1} |\lambda_{rc} - \lambda_{r'c'}|. \tag{5.4}$$

We consider the choice of the exact slope-derivation method to be inconsequential because the derived slopes are used as auxiliary information to attain appropriate transformation of the soil data including soil components, SSURGO soil slopes, and erodibility factors from map-unit level to grid-cell level as described in subsubsection 5.3.2.2. We expect the NED-derived slopes under Equation 5.3 and Equation 5.4 result in similar assignments of soil components, thus similar final maps of K factors and SSURGO-based slope gradients as presented later in Figure 5.6.

5.3.2.2 Slope alignment

The gridded SSURGO (gSSURGO) data are similar to SSURGO but available in a raster image format. It provides the spatial reference to the soil map units at a 10-meter resolution. The gSSURGO raster can be projected onto our population frame $\mathcal{D}$ using the `R` (R Core Team, 2019) function `projectExtent` in the package `raster` (Hijmans, 2017) so that we can link each grid cell with a soil map unit. We denote the matching soil map unit key for the cell (i, j) as τ_{ij}. We use the following procedure to line up the soil slopes between SSURGO and NED.

For $m = 1, \ldots, M$, suppose there are O_m major components in the soil map unit with a key denoted as $\tau^{(m)}$. For the m-th map unit, we denote $k^{(m,p)}$, $s^{(m,p)}$, and $\rho^{(m,p)}$ as the SSURGO-based K factor, slope, and percentage associated with the p-th soil component in out inventory. We sort the components by their percentages such that $s^{(m,1)} \leq s^{(m,2)} \leq \ldots \leq s^{(m,O_m)}$. If two components have the same slope, we sort them by ascending percentages. To define the USLE K and S factors at any cell $(i, j) \in \mathcal{D}$, we repeat the following steps for $m = 1, \ldots, M$,

1. Find the set of cells falling within the m-th map unit, denoted as $\mathcal{D}^{(m)} = \{(i, j) : \tau_{ij} = \tau^{(m)}\}$.

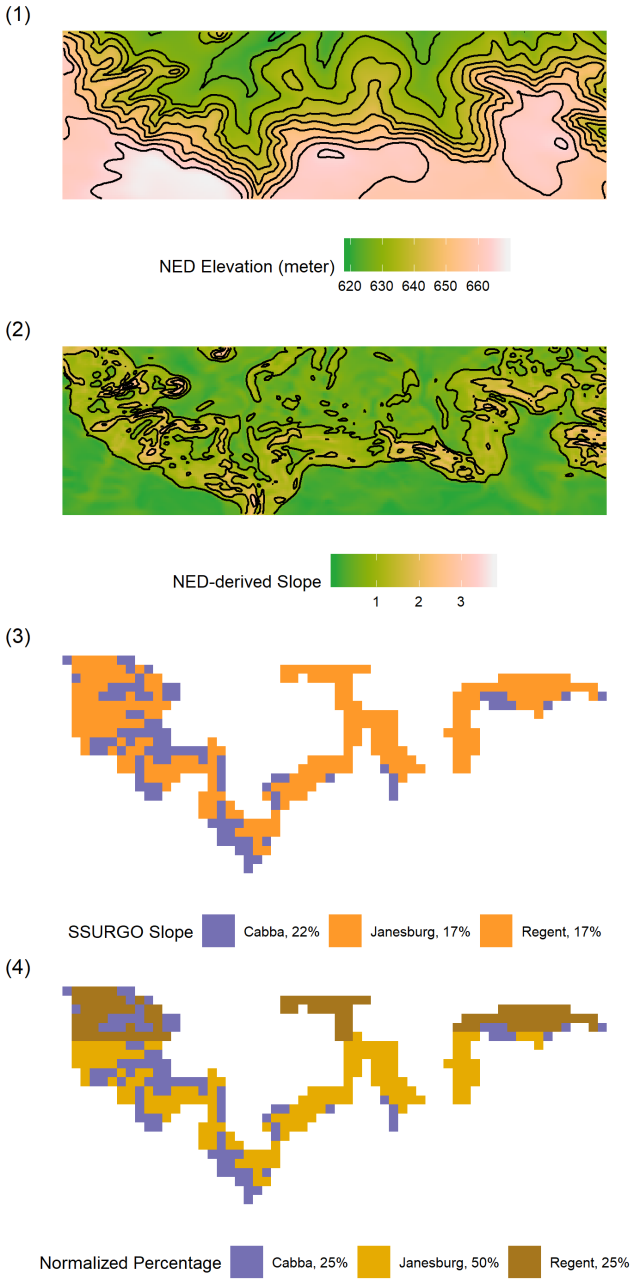

Figure 5.5 The heat map with contours of (1) the NED elevation and (2) the NED-derived slope at 1/3 arc-second resolution covering the soil map unit "Janesburg-Regent-Cabba complex" presented in the bottom. The assignment of the soil components first by (3) slope gradients then by (4) component percentages at a 30-meter resolution.

2. Link every cell in $\mathcal{D}^{(m)}$ with an NED-derived slope by referring to the raster $\mathcal{D}'$. Let N_m denote the number of cells in $\mathcal{D}^{(m)}$, and $\beta^{m,(l)}, l = 1, \ldots, N_m$ denote the order statistics of the NED-derived slopes associated with $\mathcal{D}^{(m)}$.

3. Let $c(l) = (i_l, j_l)$ be the cell associated with $\beta^{m,(l)}$ for $l = 1, \ldots, N_m$. For the component $p = 1, \ldots, O_m$ in the m-th map unit, denote the cumulative component percentage as $\omega^{(m,p)} = \sum_{r=1}^{p} \rho^{(m,r)} / \sum_{r=1}^{O_m} \rho^{(m,r)}$ and define $\omega^{(m,0)} = 0$; define $b_{m,p} = [N_m \omega^{(m,p)}]$ where the notation $[a]$ denotes the operation of rounding a to the nearest integer; denote the set of cells associated with the p-th component as $\mathcal{D}^{(m,p)}$.

   (a) If there is no tie among $\{s^{(m,p)} : p = 1, \ldots, O_m\}$, $\mathcal{D}^{(m,p)} = \{c(b_{m,p-1} + 1), \ldots, c(b_{m,p})\}$.

   (b) If there is any tie among $\{s^{(m,p)} : p = 1, \ldots, O_m\}$, for example, $s^{(m,q)} = s^{(m,q+1)}$, order the grid cells $c(b_{m,q-1} + 1), \ldots, c(b_{m,q+1})$ first from top to bottom by row then from left to right by column. Finally, assign the first $b_{m,q} - b_{m,q-1}$ cells to $\mathcal{D}^{(m,q)}$ and the last $b_{m,q+1} - b_{m,q}$ cells to $\mathcal{D}^{(m,q+1)}$.

4. For $p = 1, \ldots, O_m$, assign the S and K factors as $S_{ij} = [0.43 + 0.30 s^{(m,p)} + 0.043 \{s^{(m,p)}\}^2] / 6.613$ and $K_{ij} = k^{(m,p)}, \forall (i, j) \in \mathcal{D}^{(m,p)}$.

Step 3 in the above procedure is used to preserve the correlation between the SSURGO slopes and the NED-derived slopes as well as the spatial proximity of soil components. Around 49% of the 7,378 map units in South Dakota have only one major component. If the m-th map unit has only one major component with K factor $k^{(m)}$ and S factor $s^{(m)}$, the preceding procedure is equivalent to assigning the K and S factors to all the cells linked with this soil map unit, denoted as $k_{ij} = k^{(m)}, s_{ij} = s^{(m)}$ for all $(i, j) \in \mathcal{D}^{(m)}$. Around 26% have multiple major soil components but only one distinct SSURGO-based slope gradient, in which case the assignment can be determined solely based on the row and column order of the cells in $\mathcal{D}^{(m)}$ without referring to the NED-derived-slope raster $\mathcal{D}'$. Around 22% have multiple major components but no tie among the SSURGO-based slopes, in which case the assignment can be determined solely based on the order of the NED-derived slopes associated with $\mathcal{D}^{(m)}$. The example map unit in Figure 5.3 has three major components,

i.e., “Janesburg”, “Regent”, and “Cabba”, and the SSURGO-based slopes are 17%, 17%, and 22% correspondingly. The cumulative percentages of the three components are about 50%, 75%, and 100%. We assign the component “Cabba” to the cells with the top 25% largest NED-derived slopes as depicted in purple in plot (3) of Figure 5.5. Then we assign the component “Regent” in brown to the first consecutive 1/3 of the remaining cells and the component “Janesburg” in yellow to the residual 2/3.

5.3.3 Map of K and S factors

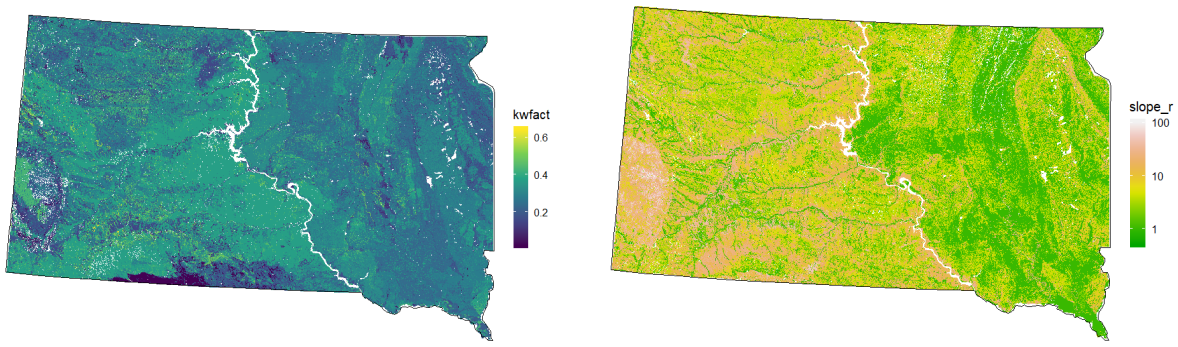

Figure 5.6 The heat maps of the designated K factors (left) and SSURGO-based slope gradients (right) at a 30-meter spatial resolution. The white color indicates undefined values.

Figure 5.6 shows the maps of the designated USLE K factors and SSURGO-based slope gradients under the population frame $\mathcal{D}$. Because the SSURGO-based K factor and slope gradient are not defined for water components, white color is painted over the water areas such as the Missouri River. As both SSURGO-based K factor and slope gradient are positively correlated with the soil erosion rate, it can be inferred from the two heat maps that the mountainous western side to the Missouri river is generally more fragile to soil erosion.

5.4 Rainfall and Crop-management Factors

For the rainfall and the crop-management factors, to our knowledge, there is no such census dataset as SSURGO to populate the erosion factors for each population unit. Therefore, we resort to fitting a model to the recorded R and C factors separately in the NRI Database and incorporating relevant auxiliary information from administrative data sources so as to approximate the grid-cell level rainfall and crop-management factors.

5.4.1 Rainfall Factor

The Midwestern States are known to have a much more uniform distribution of long-term rainfall conditions than the other regions of the United States. We linearly interpolate the recorded R factors by NRI to the population frame $\mathcal{D}$. The 2006 NRI did data collection at 5,205 sample points, including both cropland and non-cropland, in South Dakota. We consider the recorded R factors of zeros or over 550 as outliers and remove them. We fit a simple linear regression model to the logarithmic R factors with the longitudes and latitudes as covariates. The log transformation of the R factors helps eliminate heteroskedasticity. The model fitting result shows around 97% of the variance in the logarithmic R factors can be explained by this linear model. Figure 5.7 shows the distribution of the standardized residuals is bell-shaped and centered around zero. The points in the quantile plot almost fall on a straight line with a few exceptions at the tails, which suggests the normal assumption is validated. Therefore, we consider the approximation of R factors in this way is reasonable. The final interpolation result is shown in Figure 5.8. The interpolated R factors show a decreasing trend from southeast to northwest. This trend agrees with the direction of change in the isoerodent map presented in Wischmeier and Smith (1965). The isoerodent map shows the average annual values of the R factors recorded in the 1960s.

5.4.2 Crop Management Factor

To our knowledge, there is no available database to link C factors with every cell in $\mathcal{D}$. Since the C factors are related to land covers, we explore the correlation between the C factor and

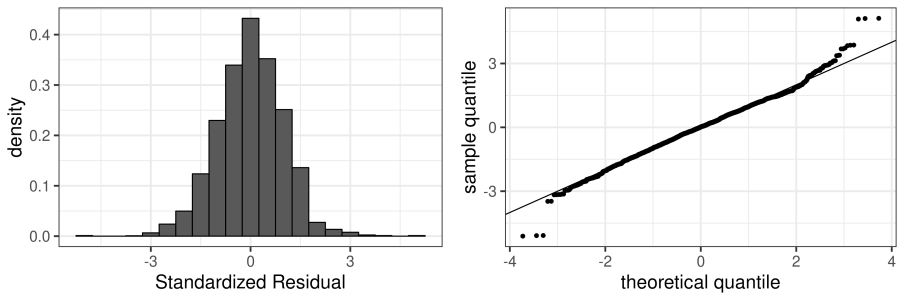

Figure 5.7 The histogram (left) and Q-Q plot (right) of the standardized residuals after fitting a simple linear model to the recorded R factors in log scale with longitudes and latitudes in the NRI Database.

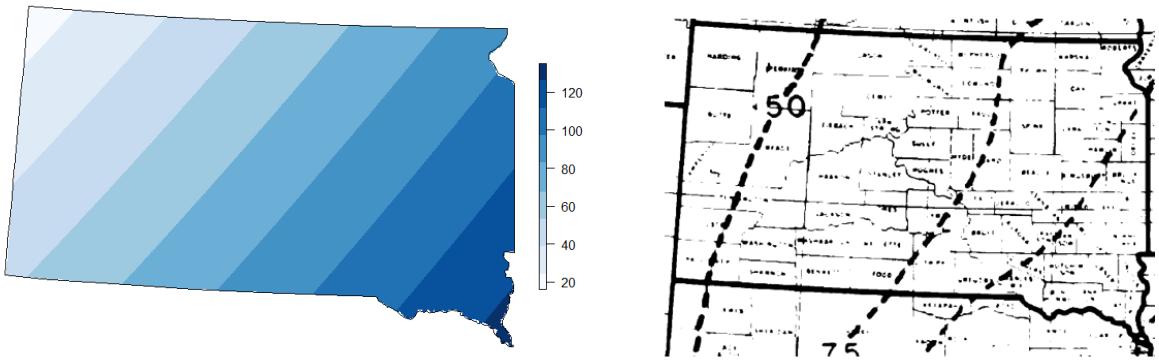

Figure 5.8 Left: the linearly interpolated rainfall (R) factors from the 2006 NRI to the raster image at a 30-meter spatial resolution covering South Dakota. Right: the isoerodent map of South Dakota used in the U.S. Agriculture Book NO.282 for referring average annual values of the R factor.

the auxiliary data such as crop type and crop rotation pattern using the 2006 NRI and the 2006 CDL data. The 2006 NRI records C factors at each of the 2,572 cropland sample locations. The Cropland Data Layer (CDL) is a crop-specific land cover data layer published annually by the National Agricultural Statistical Service (NASS) of USDA. To approximate the C factors for the year 2006, ideally, we would like to use the CDL data in and before the year 2006. However, South Dakota is not included in the CDL program until 2006. Therefore, we have to use the CDL data in and after 2006, assuming that the crop rotation patterns before and after 2006 keep the same. The CDL data have a 56-meter spatial resolution in and before the year 2009 and 30-meter after the year 2009. To line up the spatial resolutions, we project the 2006-2009 South Dakota CDL raster images onto the 30-meter raster $\mathcal{D}$ using a nearest neighbor method.

Even though both NLCD and CDL are related to land covers, the CDL data contain crop-species classifications. Therefore, we can obtain useful information related to cropping management from CDL. Both the NRI Database and the CDL data indicate corn and soybean were the dominant crop species in South Dakota around the year 2006. Based on the empirical distribution of the crop types and some previous studies on C factor (Wischmeier and Smith, 1965), we re-categorize the CDL crop types into Corn (C), Soybeans (S), Grass/Pasture (G), Small Grains (SG), and Others (O). The small grains collection includes spring wheat, winter wheat, durum wheat, barley, oats, rye, and the other wild relatives. We compare the empirical distribution of the recorded C factors by the 2006 NRI among the ten most frequent crop rotation patterns classified by the 2006 and 2007 CDL data. Figure 5.9 shows the locations classified as continuous small grains ("SG-SG") tend to be associated with lower C factors than the sites of at-least-one-year soybeans or corn. The locations classified as permanent grass/pasture ("G-G") seem to have the lowest C factors, among others.

Denote the observed C factor as ζ_s and the predicted C factor as $\hat{\zeta}_s$ at location s, $s = 1, \ldots, n$, where n is the number of cropland sample points in the 2006 NRI. To help select the important features and measure the performance of different models to approximate C factors, we used the

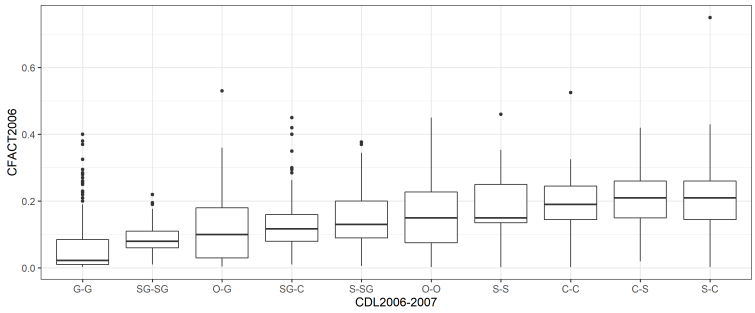

Figure 5.9 The distribution of the recorded C factors by the 2006 NRI for the ten most frequent crop rotation patterns classified by the 2006 and 2007 CDL data in South Dakota.

R^2 metric defined as

$$R^2 = 1 - \sum_{s=1}^{n} (\hat{\zeta}_s - \zeta_s)^2 / \sum_{s=1}^{n} (\zeta_s - \bar{\zeta})^2, \tag{5.5}$$

where $\bar{\zeta} = s^{-1} \sum_{s=1}^{n} \zeta_s$. Here R^2 is interpreted as the proportion of variance in the observed C factors that can be explained by the evaluated model. Considering we need to predict C factor at almost 100 million cells in $\mathcal{D}$, we prefer a computationally simple model such as the ANOVA regression tree model (Breiman et al., 1984). The random forest model (Breiman, 2001) is a bagged tree regression model to reduce variance while maintaining low bias of the tree model using bootstrapping technique. For covariates, we use re-categorized CDL 2006-2009 classifications to account for crop rotation patterns and geolocation, i.e., longitude and latitude, to account for spatial correlation. Using the C factors recorded by the 2006 NRI as the response variable, we conduct repeated 10-fold cross-validation and tune the complexity parameter in the tree model with a tune length of 50. The model with the largest R^2 has 46 leaf nodes and an R^2 of 0.367. On the other hand, with a bootstrap sample size of 1,000, the random forest model with the largest R^2 randomly selects five of the six input variables at each node for splitting and has an out-of-bag R^2 of 0.494. Therefore, we prefer the random forest model to the regression tree model in terms of prediction accuracy.

Another reason we prefer the random forest model instead of the regression tree model is that the random forest predictions are more continuous in space and appear more reasonable on the

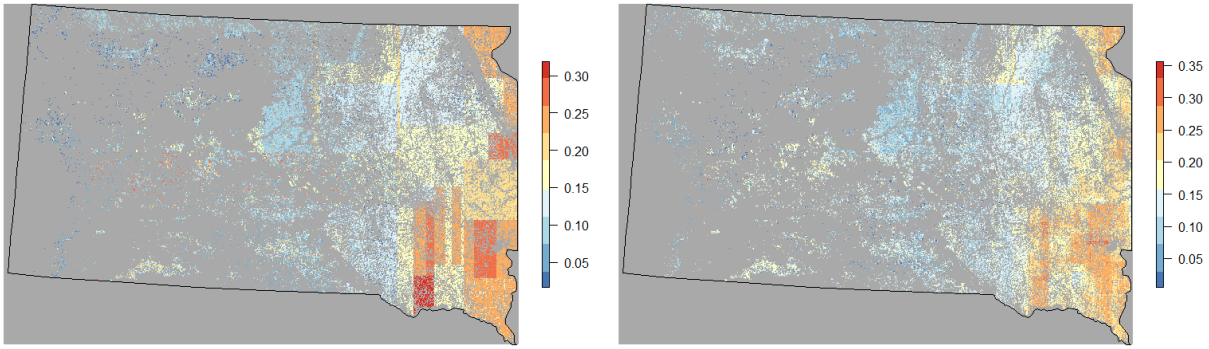

Figure 5.10 The approximated C factors using the best tree model (left) and the best random forest model (right) for the cropland frame at a 30-meter spatial resolution covering South Dakota. The non-cropland cells are in grey.

map. The left plot in Figure 5.10 shows the predicted C factors based on the final tree model. The rectangular shape in the southeastern areas is due to those tree predictions with splits on solely longitude or latitude. It seems misleading in terms of the spatial pattern since the actual C factors change much more smoothly in space. The random forest model averages across the bootstrapped small tree predictions. It produces more visually reasonable results, as shown in the right plot in Figure 5.10. Since the model is trained using the NRI cropland samples, the model predictions are calculated at the set of cropland cells $\{(i,j) \in \mathcal{D} : \gamma_{ij} = 1\}$ indicated by the non-grey areas in Figure 5.10.

By exploring the 2006 CDL data using the tool *VISCOVER* that is described in chapter 4, we find that the predicted large C factors in the east of South Dakota correspond to the vast areas of soybeans and corn fields. Generally, the acreages of soybeans and corn decrease from southeast to northwest, while the acreages of grass increases according to CDL. This spatial pattern agrees with the overall trend of the predicted C factors. On the west side of South Dakota, there are some scattered fields of small grains (mostly wheat) or other crops, which might explain the predicted large C factors in the western South Dakota, depicted by the sprinkled red dots in Figure 5.10. Even though the final random forest model we adopt can only explain 49% of the variance in the

recorded C factors by the 2006 NRI, the approximated C factors are the best we could obtain given the data availability and computation difficulty.

5.5 Estimation and Visualization of Sheet and Rill Erosion

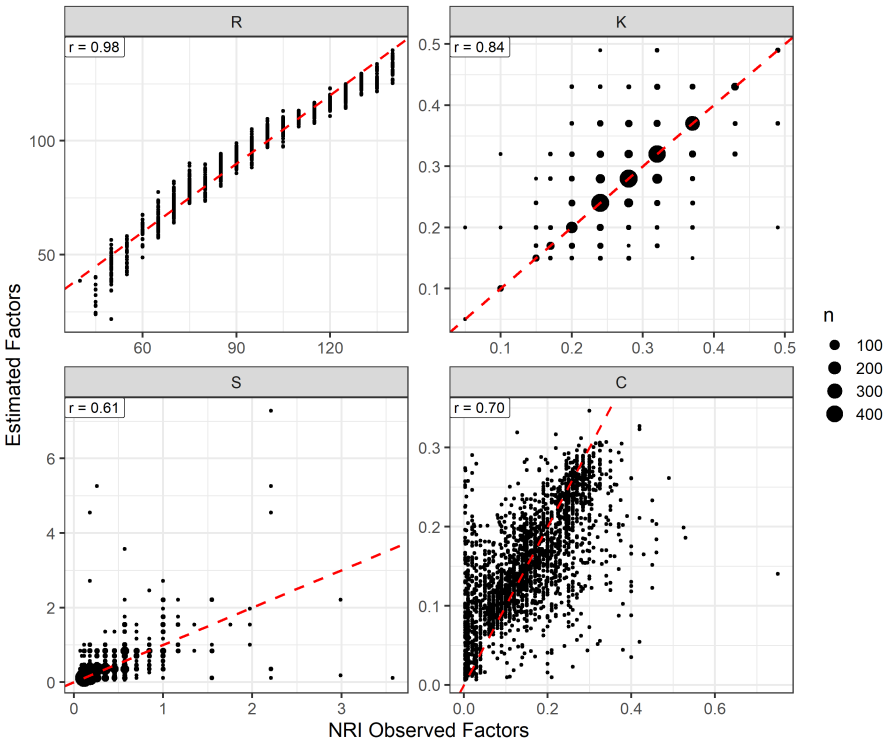

Figure 5.11 The R, K, S and C factors recorded by the 2006 NRI versus the approximated factors by our method. The red dashed are the identity lines, and the top left corners are the Pearson correlation coefficients. Larger dot size indicates the larger number of overlapped points.

Figure 5.11 presents the comparison between the NRI records and our approximations of the four erosion factors. The Pearson correlation coefficient between the NRI data and our approximations is higher for the R and K factors than the S and C factors. The correlation for the R factors is almost 1. The correlation between the assigned K factors and the NRI recorded K factors is as high as 0.84, and the correlation is 0.61 for the S factors between the two sources. The correlation between our approximated C factors and the NRI recorded C factors is 0.70.

Next, we regress the recorded sheet and rill erosion rates by NRI with the USLE factors. We assume the USLE loss y_i at location i satisfies

$$\log(y_i) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}_1 + e_i \tag{5.6}$$

where $x_i = (r_i, k_i, s_i, c_i)'$ are the corresponding logarithmic R, K, S, and C factors, denoted as `logR`, `logK`, `logS` and `logC`, at location i, and $e_i, i = 1, \ldots, n$ follows an independent and identical normal distribution with mean 0 and variance σ_e^2. We fit model (5.6) to the USLE losses recorded by the 2006 NRI with two sets of covariates separately — the soil factors recorded by NRI and the soil factors approximated by our approach. For fitting the model with approximated USLE factors, we remove one point which is linked with undefined K factors and slope gradient in SSURGO. The fitting results are presented in Table 5.1. Unsurprisingly, all the soil erosion factors are positively associated with the USLE losses. The effect sizes are approximately one for all the erosion factors in log scale which agrees with the USLE equation (5.1). The intercept term accounts for the average L and P factors. Using the approximated erosion factors leads to more uncertainty in the parameter estimation than using the NRI erosion factors. However, the four erosion factors obtained by our approach are still highly significant in predicting the USLE loss. We test the significance of the sampling weights under model (5.6) with the presence of the other four predictors to check whether there is any issue of informative sampling (Verret et al., 2015). It turns out the existing covariates have captured the characteristics of the sampling weights, if any, in explaining the variance in the responses.

Under model (5.6), the empirical Bayes predictor of the USLE loss at location $(i, j) \in \mathcal{D}$ where $\gamma_{ij} = 1$ is given by

$$\hat{y}_{ij} = \exp(\hat{\beta}_0 + \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}_1 + \hat{\sigma}_e^2/2), \tag{5.7}$$

where $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}_e)$ denotes a consistent estimator of the model parameters, such as the maximum likelihood estimator. Using the formula (5.7) and the estimated model parameters presented in Table 5.1, we calculate the predicted USLE losses at each cell on the cropland frame. We present a heat map of the predicted USLE losses on cropland at 30-meter resolution, as depicted in the left of Figure 5.12. We present the aggregated predictions at a lower resolution of about one kilometer in

Table 5.1 The estimated covariates coefficients, with standard errors in brackets, of the simple linear regression models fitted to the point-level 2006 USLE losses in log scale recorded by NRI using the recorded erosion factors by NRI (left) and the approximated erosion factors by our approach (right). The bottom half of the table shows the R^2, adjusted R^2, the number of observations, and the root mean squared error (RMSE) of each fitted model.

	NRI Factors		Approximated Factors	
logR	0.97	$(0.02)^{***}$	0.80	$(0.09)^{***}$
logK	0.98	$(0.02)^{***}$	1.02	$(0.09)^{***}$
logS	1.12	$(0.01)^{***}$	0.75	$(0.03)^{***}$
logC	1.00	$(0.00)^{***}$	1.16	$(0.04)^{***}$
R^2	0.98		0.45	
Adj. R^2	0.98		0.45	
Num. obs.	2570		2569	
RMSE	0.21		1.04	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

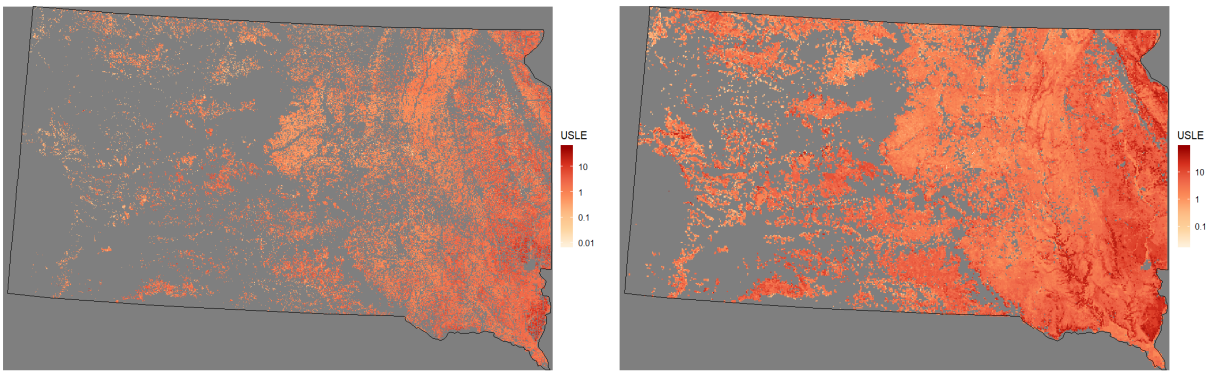

Figure 5.12 The predicted logarithmic USLE losses on cropland at a 30-meter spatial resolution (left) and aggregated to the maximum at an about 1-kilometer spatial resolution (right).

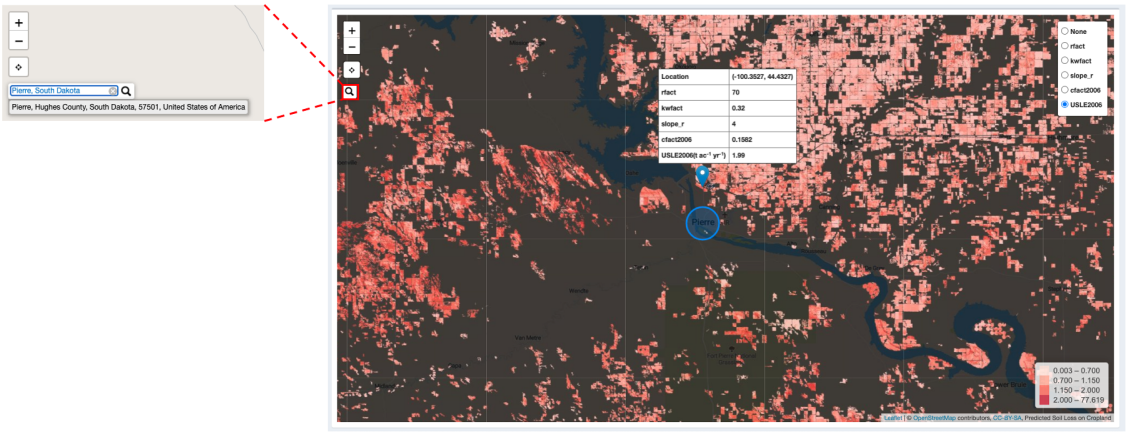

Figure 5.13 The snapshot of the Sheet and Rill Erosion Map (*SREM*) zoomed near Pierre, South Dakota. The location that a user inputs through the search widget is marked by a blue circle in the map. The tooltip at the pinpoint contains the approximated R factor, K factor, soil slope gradient, C factor, and the estimated USLE loss by our approach.

the right of Figure 5.12. A heat map with a lower resolution, depicted in darker red in Figure 5.12, helps highlight the areas of more conservation needs. The heat map in the right of Figure 5.12 is obtained by aggregating the erosion estimates to the maximum at a lower spatial resolution of 990 meters. In practice, this can be achieved by using the `aggregate` function in the R package `raster` (Hijmans, 2017) and set the argument `fact` to be 33 and the argument `fun` to be `max`.

Using the same techniques as used in *VISCOVER*, we develop a web-based interactive tool — Sheet and Rill Erosion Map (*SREM*) as depicted in Figure 5.13 based on the R **shiny** (Chang et al., 2018) framework. The Web Map Service (de La Beaujardiere, 2006) is implemented with the Mapbox Studio[2]. Similar to *VISCOVER*, *SREM* allows user to search for any particular location. Currently, the grid-cell level soil erosion estimates are only available for South Dakota. The example introduced in Figure 5.13 searches for Pierre, the capital city of South Dakota. The map is then automatically zoomed in to that location and marks the target location with a blue circle. Besides the estimated 2006 USLE losses in tons per acre per year, other available layers in *SREM* include the approximated R, K, C factors, as well as the approximated soil slope gradients.

[2] https://studio.mapbox.com/

A map marker is placed at a location near the capital city, and the tooltip shows the associated soil erosion characteristics we obtain at that grid-cell.

5.6 Conclusion and Discussion

Starting from the well-known and widely-used USLE equation, we integrate a collection of geostatistics data including NLCD, SSURGO, NED, CDL with NRI to gather population-level soil erosion factors, specifically R, K, S, and C factors. The NLCD data are used to create a cropland population frame at a 30-meter spatial resolution by distinguishing the cultivated cropland cells. Due to the uniformity of the climate conditions in the Midwest States, the linear interpolations can explain more than 90% of the variance of the NRI recorded R factors in log scale. The SSURGO data contain high-resolution information about soil properties. By designating grid-cell level slopes from the NED data and aligning them with the SSURGO slope gradients, we can allocate the soil components at a 30-meter spatial resolution covering South Dakota, then assign the component-specific K and S factors to each grid cell. Using the 2006-2009 CDL data, we can associate the recorded C factors by the 2006 NRI with crop type, crop rotation, longitude, and latitude. Finally, a preliminary interactive map of the predicted USLE losses in tons per acre per year on the cropland frame is developed and made available at https://lyux.shinyapps.io/srem/.

5.6.1 Impacts on the Data Quality of NRI

Discussions of data quality often focus on collection and estimation. In this thesis, Chapter 2 exemplifies how models and auxiliary information can improve estimation accuracy. Chapters 3 and 4 relate more directly to the data collection stage. Equally important is the quality of disseminated information. In this chapter, we invoke the powers of big data and interactive graphics to produce sheet and rill erosion maps with the important qualities of *interpretability*, *usability*, and *accessibility*.

The National Resources Inventory (NRI) currently uses a crude interpolation procedure to estimate sheet and rill erosion on cropland at a one-mile spatial resolution. The NRI publishes

these interpolated erosion values through static maps that display the distribution of the predicted erosion on cropland across the coterminous United States. The interpolation procedure rests on strong implicit statistical assumptions and utilizes no auxiliary information about the processes governing erosion. Further, the static character of the maps and the relatively high level of spatial aggregation limit the extent of information that a user can glean.

We improve upon NRI's current erosion maps in three important quality dimensions: *interpretability*, *usability*, and *accessibility*. Our predictions of sheet and rill erosion are interpretable because they derive from the widely accepted USLE equation. To enhance the *usability* and *accessibility* of our product, we present the erosion predictions through a publicly available interactive map. Specifically, the R **shiny** application Sheet and Rill Erosion Map (SREM) provides an interactive geographic display of the erosion predictions. The fine level of spatial granularity (30-meter resolution) coupled with the capability of user interaction equip SREM with a high level of *usability*. Additionally, SREM is publicly available, ensuring the *accessibility* of our erosion predictions to any interested analyst. This chapter defines the process of building the SREM tool for delivering accessible, useful, and interpretable erosion predictions to the public.

5.6.2 Future Work

Our approach can be applied to other states and other years given the auxiliary data are nationwide available. There are a few exceptions. For example, the CDL data are not available for Hawaii or Puerto Rico. Many challenges exist in this topic. For instance, we could not obtain year-specific SSURGO data because the NRCS Soil Survey Database has been updated from time to time, and the old archived versions are not accessible to the public. Most of the computational difficulty of this application lies in making predictions at hundreds of millions of grid cells. For future studies, the balance between model efficiency and computational efficiency could be investigated. To reduce the computational burden, we can compare lower spatial resolutions such as 56 meters as in the 2006 CDL with the 30 meters resolution. To boost model efficiency, we can investigate other models to produce more accurate C factor approximation. Raw satellite imagery data and deep machine

learning algorithms are potential resources for improving the approximation of the C factors and investigating the feasibility of approximating the P factors by identifying conservation practices, such as terraces, contour, strip, etc., from the satellite imagery. The approximated USLE factors by our method deviate from the true values. More sophisticated models that take into account the approximation errors could help measure the statistical reliability of the ultimate erosion estimates.

5.7 Acknowledgment

The authors thank Harvey Terpstra for his consultancy in determining the K factors of soil components from the SSURGO database.

5.8 References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. (1984). Classification and regression trees (belmont, ca: Wadsworth international group). *Biometrics*, 40(3):17–23.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

de La Beaujardiere, J. (2006). OpenGIS® Web Map Server Implementation Specification. Version 1.3.0.

Hijmans, R. J. (2017). *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renard, K. G., Yoder, D. C., Lightle, D. T., and Dabney, S. M. (2011). Universal soil loss equation and revised universal soil loss equation. *Handbook of erosion modeling*, pages 137–167.

Smith, D. D. and Wischmeier, W. H. (1957). Factors affecting sheet and rill erosion. *Eos, Transactions American Geophysical Union*, 38(6):889–896.

U.S. Department of Agriculture (2018). Summary report: 2015 national resources inventory. Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa.

Verret, F., Rao, J., and Hidiroglou, M. A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41(2):333–347.

Wischmeier, W. H. and Smith, D. D. (1965). Predicting rainfall erosion losses from cropland east of the rocky mountains: guide for selection for practices for soil and water conservation. U.S. Department of Agriculture. Agriculture Handbook NO.282.

5.9 Appendix. Supplementary Information on SSURGO

Algorithm 1: Query K and S Factors From the Soil Survey Geographical Database

Result: A list of USLE K and S factors of major soil components in South Dakota.

SELECT `mukey` FROM `mapunit` LEFT JOIN `areasymbol` ON `lkey` WHERE `areasymbol` start with "SD" AS `sdmu`;

SELECT `cokey`, `slope_r`, `taxorder`, `taxsubgrp`, `taxsuborder` FROM `component` LEFT JOIN `sdmu` ON `mukey` WHERE `component.majcompflag` = "Yes" AS `sdco`;

Calculate the USLE S factor with Equation 5.2;

while *Calculating USLE K factor for* `sdco.cokey` **do**

  if *(*`taxorder` *= "Histosols"AND* `taxsubgrp` *contains "Histic") OR* `taxsuborder` *= "Histels"* **then**

    SELECT `chkey` FROM `chorizon` WHERE `hzdept_r` = 0 AS `sdch` ;

  else

    SELECT `chkey` FROM `chorizon` LEFT JOIN `chtexturegrp` ON `chkey` WHERE `chtexturegrp.texture` not end with "MUCK", "PEAT", "MPT", "HPM", "MPM" or "SPM" AND `chorizon.hzdept_r` is the minimum AS `sdch`;

  end

  SELECT `kwfact` FROM `chorizon` WHERE `chkey` = `sdch.chkey`.

end

Table Name	Column Name	Column Label	Data Type
legend	lkey	legend key	integer
legend	areasymbol	area symbol	string
mapunit	mukey	mapunit key	integer
component	cokey	component key	integer
component	majcompflag	major component	boolean
component	compname	component name	string
component	compkind	component kind	choice
component	comppct_r	representative component %	integer
component	slope_r	representative slope gradient	float
component	slopelenusle_r	slope length USLE	integer
component	taxorder	taxonomic order	choice
component	taxsuborder	taxonomic suborder	choice
component	taxsubgrp	taxonomic subgroup	choice
chorizon	chkey	coponent horizon key	integer
chorizon	kwfact	soil erodibility factor	choice
chorizon	hzdept_r	representative top depth	integer
chtexturegrp	chtgkey	chorizon texture group key	integer
chtexturegrp	texture	texture modifier and class	string

Table 5.2 SSURGO table column descriptions.

CHAPTER 6. GENERAL CONCLUSIONS

In this dissertation, the data we analyze are pertinent to the sheet-and-rill erosion on cropland; the outcomes we produce are applicable to the National Resources Inventory (NRI); the problems we study are around the theme of data quality. We study the data quality of the NRI in the dimensions of *accuracy*, *reliability*, *comparability*, *usability*, and *accessibility*.

The model-based small area predictor we develop in chapter 2 pertains to the *reliability* of county-level erosion estimates. The *accuracy* of statistical estimates is implied in the quality of the associated covariates and recorded responses. Therefore, we develop a web-based map tool *VISCOVER* (chapter 4) to inspect the soil and land cover data used in our application and use it as a proof of concept to demonstrate the feasibility of linking the two databases. The NRI data are collected from survey samples intended to construct estimates of parameters for the full population. The NRI data are of higher quality inherently than the two auxiliary data sets we use for the small area prediction in chapter 2. The NRI data are scrutinized at every phase of the survey process for quality assurance (QA). The Table Review and State Review are two integral parts of the NRI QA process. We develop *iNtr* (chapter 3), a cost-effective table review tool equipped with effective data visualizations for the use of NRI. It enhances the efficacy of the NRI Table Review process. *VISCOVER* has also shown its *usability* in aiding the NRI State Review by providing a web user interface to link the land cover data with the NRI points.

VISCOVER strengthens the *comparability*, *usability* and *accessibility* of the soil and land cover databases, which raises the data quality from user's perspective. The public version of *iNtr*, available at `https://lyux.shinyapps.io/table_review/`, augmented the *comparability* of the NRI data in that it exhibits the coherence between the 2015 NRI and the 2012 NRI. In an effort to increase the *usability* and *accessibility* of the NRI products, we develop a web-based interactive map tool *SREM* (chapter 5) to present a high-resolution erosion map and to facilitate the com-

munication of the NRI data to the public. The underlying point-level sheet and rill erosion factors and rates at a 30-meter spatial resolution are obtained using the NRI data available at the sampled locations and other relevant auxiliary data available for the full population. User-data interactions are enabled in the three tools, *iNtr*, *VISCOVER*, and *SREM*, and built upon the R **shiny** framework. The user-data interactions are effectuated by filtering the underlying data set(s).

A set of multi-disciplinary skills are positioned in the data analysis cataloged in this dissertation. The skill of database query and the cognition of domain knowledge mainly play a role in the phase of the pre-statistical-modeling for data acquisition, data linkage, and sanity check. The knowledge of probability and statistics theories is of use to address the mathematical challenges in the phase of statistical estimation. At the time of the post-statistical-model, the experience of database design and user interface design, wrapped in tool development, is of value to converse the veiled auxiliary data and statistical analysis to the community. We consider tool development as important as methodological development. The R package **saezero** we develop assists technical users in implementing the complicated statistical methodology conveniently. A well-designed user interface aids a non-technical user in exploring large data sets and obtaining insight against their expertise. This skill set supports our involvement in the entire sheet and rill erosion data analysis pipeline. The logistics of one phase might enhance the understanding and development of another, which ultimately benefits the data product as a whole.