Predicting RUSLE2 and Visualizing Covariates

Xiaodan Lyu

Joint work with Heike Hofmann, Emily Berg

26 October 2017
Navigation

1. Motivation
2. Objective
3. Soil Data
4. Cropland Data Layer
5. Procedure
6. Results
7. Visualization with Shiny
8. Discussion
9. Reference
Motivation
Conservation Effects Assessment Project (CEAP)

Assessments in CEAP are carried out at national, regional and watershed scales.

The four national assessments: CEAP-Cropland, CEAP-Grazing Lands, CEAP-Wildlife and CEAP-Wetlands.
• In each CEAP-Croplands survey, the farm fields sampled are derived from a subset of NRI cropland sample points.

• At each point, the NRI maintains data on site-specific soils and topography.

• The response variables are several measures of soil and nutrient loss on crop fields.

• We focus on a particular measure of sheet and rill erosion obtained from the Revised Universal Soil Loss Equation - 2 (RUSLE2).
Universal Soil Loss Equation

\[ A = R K L S C P \]

- **A**, computed soil loss per unit area
- **R**, rainfall factor, number of *erosion-index* units in a normal year's rain
- **K**, soil erodibility factor, erosion rate per unit of *erosion-index*
- **L**, slope-length factor, soil loss ratio to a 72.6-foot length field
- **S**, slope-gradient factor, soil loss ratio to a 9-percent slope field
- **C**, cropping-management factor, soil loss ratio to a fallow field
- **P**, erosion-control practice factor

*The erosion index is a measure of the erosive force of specific rainfall.*
RUSLE2 Technology

- USLE was limited in its application to experimental data.
- RUSLE2 keeps the USLE basic formulation of the unit plot.
- RUSLE2 uses improved cover-management subfactor relationship, expanding the application of USLE.
- RUSLE2 measurement in CEAP is highly skewed and contaminated with zeros, which motivated our proposed zero-inflated log-normal model.
Objective
Objectives

- NRI data is not our target population, but subset of the population.

- To predict the population mean of rainfall-erosion losses from cropland at county level.
  - target population: all cropland soils in South Dakota.
  - to find an appropriate frame for prediction.
  - to distinguish cropland soil from non-cropland soil.
  - collect auxiliary information (RKLSCP) for the whole population.
  - implement EB zero-inflated log-normal model to make small area prediction.
Soil Data
SSURGO database

- U.S. Land Resource Hierarchy Diagrams

- The information contained in the SSURGO database is
  - collected by the National Cooperative Soil Survey;
  - available for most areas in the United States;
  - gathered by walking over the land and observing the soil.
Soil Survey Area

- The SSURGO database can be viewed in the Web Soil Survey or downloaded in ESRI shapefile format.
  - shapefiles are readable by R function `rgdal::readOGR`.
- The extent of a SSURGO dataset is a soil survey area, which may consist of a single county, multiple counties or parts of multiple counties.
- There are 67 soil survey areas in South Dakota.
  - 57 survey areas (SD003-SD137) in South Dakota cover a single county;
  - 10 survey areas (SD600-SD603, SD606-SD613) do not.
### Table 1: 10 survey areas and their associated counties.

<table>
<thead>
<tr>
<th>AREASYM</th>
<th>CTYFIP</th>
<th>CTYNAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD600</td>
<td>93</td>
<td>Meade</td>
</tr>
<tr>
<td>SD601</td>
<td>93</td>
<td>Meade</td>
</tr>
<tr>
<td>SD602</td>
<td>61, 67</td>
<td>Hanson, Hutchinson</td>
</tr>
<tr>
<td>SD603</td>
<td>15, 17</td>
<td>Brule, Buffalo</td>
</tr>
<tr>
<td>SD606</td>
<td>33, 103</td>
<td>Custer, Pennington</td>
</tr>
<tr>
<td>SD607</td>
<td>33, 103</td>
<td>Custer, Pennington</td>
</tr>
<tr>
<td>SD610</td>
<td>71</td>
<td>Jackson</td>
</tr>
<tr>
<td>SD611</td>
<td>71</td>
<td>Jackson</td>
</tr>
<tr>
<td>SD612</td>
<td>71, 103, 113</td>
<td>Jackson, Pennington, Shannon</td>
</tr>
<tr>
<td>SD613</td>
<td>113</td>
<td>Shannon</td>
</tr>
</tbody>
</table>
10 survey areas Map
Map Units

- The soil map outlines areas called map units.
- A soil survey area is a group of map units.
- Each map unit may contain one to three major components and some minor components.
- The map units are typically named for the major components.
Mapunit Example Plotted Using R

MUKEY = 417991; MUSYM = "Pr";

MUNAME = "Prosper-Stickney complex, 0 to 2 percent slopes".
Two Useful Component Properties

- The **component** soils have unique properties for each map units.

- **KWFAC**
  - "Erosion factor Kw (whole soil)"
  - modified by the presence of rock fragment;
  - range from 0.02 to 0.69.

- **SLOPE_R**
  - slope gradient between two points = \( \frac{\text{difference in elevation}}{\text{distance}} \)
  - three separate values in the database: SLOPE_L, SLOPE_H, SLOPE_R;
  - SLOPE_R indicates the expected value of this attribute for the component.
  - range from 0 to 100.
  - Slope Factor \( S' = \frac{0.43 + 0.30s + 0.043s^2}{6.613} \) where \( s \) is the slope gradient.
The higher the value, the more susceptible the soil is to sheet and rill erosion by water.
SLOPE_R, SD612, Badlands National Park
Findings

- prediction frame: a list of **mapunits** in SSURGO database.
  - To get mapunits grouped by county, the soil map needs to be intersected with county shapefile.

- auxiliary information available: K factor and SLOPE_R (related to S factor).
  - for each mapunit, properties of the **dominating component** on the selected **horizon** would be used.
  - For Histosol soils (wet organic soils), the top (first) horizon is selected. For all other soils, the top mineral horizon is selected.
  - picking the actual horizon for the soil would need to manipulate the **mapunit component** table, **component horizon** table and **component taxonomic classification** table.
Cropland Data Layer
• United States Department of Agriculture (USDA), National Agricultural Statistics Service (NASS).

• provide timely, accurate, and useful statistics in service to U.S. agriculture.

• CDL data available free at the CropScape portal.
CDL data

- A raster, geo-referenced, crop-specific land cover data layer created annually for the continental United States.

- Produced using moderate-resolution satellite imagery and extensive validation.

- CDL data for crop year 2009 and earlier has a ground resolution of 56 meters. For CDL year 2010 and newer, the ground resolution is 30 meters.

- Available in GeoTIFF(.TIF) file format
  - .tif file: readable by R function `raster::raster`; each pixel has an integer value ranging from 0 to 254.
  - .tif.vat.dbf file: readable by R function `foreign::read.dbf`; contains information about each value, including CLASS_NAME and COLOR (RED, GREEN, BLUE, OPACITY).
2006 South Dakota CDL color legend

2006 South Dakota
Land Cover Categories (by decreasing acreage)

AGRICULTURE
- Pasture/Grass
- Corn
- Soybeans
- Fallow/Idle Cropland
- Spring Wheat
- Winter Wheat
- Alfalfa
- Sunflower
- Millet
- Sorghum
- CATS
- Safflower
- Barley
- Peas
- Rye

NON-AGRICULTURE
- Developed
- Forest
- Wetlands
- Shrubland
- Water
- Barren

Dry Beans
Durum Wheat
Other Crops
Flaxseed
Other Small Grains
Misc Vegetables & Fruits
Double Crop Winter/Winter/Soybeans
Lentils
Clover/Wildflowers

24 / 52
CDL Example Plotted Using R
Findings

- CDL raster could be used for classification to identify cropland by land broad use.
  - Developed, Forest, Wetlands, Shrubland, Water, Barren, Grass/Pasture and Perennial Ice/Snow has land broad use of non-cropland.

- Statistics generated by CDL data could contribute to C factor in USLE.
  - Principle Components of the crop category pixel counts at county level are used.

- Cropland acreage data could be derived from the CDL and depend upon pixel counting.
Procedure
Implementation

1. Export SSURGO soils map in South Dakota into a shapefile.

2. Download CDL raster file from year 2006 to year 2016 for South Dakota from USDA-NASS.

3. Iterate first by county then by mapunits
   - Crop the CDL raster to the extent of the bounding box of the target mapunit;[1]

   - `raster::rasterToPoints` transforms rater file into spatial point object;

   - Overlay the mapunit polygon with the cropped CDL points;[2]

   - Table pixel count by CDL category grouped by each mapunit.


Problems

1. Large size of statewide soils map shapefile.
   - South Dakota soils map shapefile has size of 1.64 GB.

2. Large number of pixels associated with CDL raster.
   - Cropping the raster to the extent of a mapunits with 440 small polygons gives 1,295,385 pixels.
   - The most complex mapunit shape in South Dakota is composed of 10,087 small polygons.

3. 10 irregular surveys areas.
   - We are interested in small area predictors at county level.

4. Different coordinate reference system (CRS) attached to spatial objects.
   - Both crop and over requires identical CRS.

5. Expecting long computing time.
   - 7501 unique mapunits in total for South Dakota.
Solutions

1. Split the shapefile into separate shapefiles grouped by soil survey areas.

2. Omit background pixels and pixels with missing values in CDL data.[1]

3. Download the latest county boundary shapefiles from the website of US census Bureau, then overlay them with the shapefiles of the 10 survey areas.

4. R function `sp::spTransform` allows map projection and datum transformation.[2]

5. Using 4 cores and parallelized computing, it only takes about 3 hours to overlay all the mapunits with CDL for South Dakota.[3]

[1] `raster::rasterToPoints` allows subsetting the values at the same time.


[3] Package `multidplyr` is used.
Example overlaid image

SD602 MUKEY 417991 overlaid with 2006 CDL raster.
Example overlaid image

Hanson MUKEY 417991 overlaid with 2006 CDL raster.
Example CDL pixel counts

<table>
<thead>
<tr>
<th>CODE</th>
<th>CLASS_NAME</th>
<th>PIXEL_COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Soybeans</td>
<td>12874</td>
</tr>
<tr>
<td>1</td>
<td>Corn</td>
<td>12214</td>
</tr>
<tr>
<td>176</td>
<td>Grass/Pasture</td>
<td>5836</td>
</tr>
<tr>
<td>121</td>
<td>Developed/Open Space</td>
<td>1749</td>
</tr>
<tr>
<td>24</td>
<td>Winter Wheat</td>
<td>1472</td>
</tr>
<tr>
<td>61</td>
<td>Fallow/Idle Cropland</td>
<td>813</td>
</tr>
<tr>
<td>36</td>
<td>Alfalfa</td>
<td>699</td>
</tr>
<tr>
<td>195</td>
<td>Herbaceous Wetlands</td>
<td>550</td>
</tr>
<tr>
<td>23</td>
<td>Spring Wheat</td>
<td>543</td>
</tr>
<tr>
<td>141</td>
<td>Deciduous Forest</td>
<td>196</td>
</tr>
</tbody>
</table>

Table 2: 2006 CDL pixel count for the mapunit 417991 in Hanson.
Results
Cropland Population

- Non-water map units with positive cropland acreage and with name not starting with Badland, Rock outcrop, Rubbleland or Pits make up the population frame.

- Covariates at mapunit level:
  - SFACT and KFACT: corresponding with SLOPE_R and KWFAC in the soils data.

- Covariates at county level:
  - R factor: mode of the R factors obtained from NRI data in each county given its homogeneity.
  - C factor: principle components of the CDL data grouped by county.

- EB Prediction weighted by cropland acreages for each map units.
  \[ \hat{y}_{U_i}^{*}_{EB} = \bar{y}_{U_i}^{*}_{MMSE}(\hat{\theta}) \approx \sum_{j \in N_i} \hat{y}_{ij}^{*EB} \ast w_{ij} \text{ where } w_{ij} \text{ is the cropacre percentage of the } j_{th} \text{ mapunit in county } i.\]
### 10 map units with the largest cropland areas

<table>
<thead>
<tr>
<th>county</th>
<th>MUKEY</th>
<th>cropacre</th>
<th>SLOPE_R</th>
<th>KWFACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lyman</td>
<td>2615602</td>
<td>164625</td>
<td>4</td>
<td>0.37</td>
</tr>
<tr>
<td>Hutchinson</td>
<td>417943</td>
<td>151589</td>
<td>1</td>
<td>0.24</td>
</tr>
<tr>
<td>Edmunds</td>
<td>2798620</td>
<td>133443</td>
<td>3</td>
<td>0.28</td>
</tr>
<tr>
<td>Lincoln</td>
<td>416664</td>
<td>118890</td>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>Beadle</td>
<td>354587</td>
<td>99732</td>
<td>4</td>
<td>0.28</td>
</tr>
<tr>
<td>Day</td>
<td>417076</td>
<td>96946</td>
<td>4</td>
<td>0.24</td>
</tr>
<tr>
<td>Sully</td>
<td>353058</td>
<td>94878</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>McCook</td>
<td>418018</td>
<td>91733</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Edmunds</td>
<td>2798562</td>
<td>90905</td>
<td>4</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Note: attribute *cropacre* is dependent upon pixel counts.
SAE Zero-Inflated Lognormal Model

For the sample point $j$ of the area (county) $i$ in South Dakota, denote

- $y_{ij}^*$: observed RUSLE2 measurement, no less than zero;
- $z_{ij} = (z_{1ij}, \ldots, z_{4ij})'$: covariate vector (logarithmic RKSC);
- $p_{ij}$: probability that $y_{ij}^*$ is greater than zero.
SAE Zero-Inflated Lognormal Model (Cont’d)

We assume for \( i = 1, \ldots, D; j = 1, \ldots N_i \)

\[
y_{ij}^* = y_{ij} \delta_{ij}
\]

where

\[
\log(y_{ij}) = \beta_0 + z'_{ij} \beta_1 + u_i + e_{ij}
\]

and \((u_i, e_{ij})' \overset{i.i.d.}{\sim} N\left(0, \text{diag}(\sigma_u^2, \sigma_e^2)\right)\).

In addition, \( \delta_{ij} \sim Bernoulli(p_{ij}) \), where

\[
\text{logit}(p_{ij}) = \alpha_0 + z'_{ij} \alpha_1 + b_i
\]

and \( b_i \overset{i.i.d.}{\sim} N(0, \sigma_b^2) \), independent of \((u_i, e_{ij})'\).
EB Predicted Soil Loss Map

County Level for South Dakota
Comparison with predictions under NRI frame
## Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y_R_FACT</td>
<td>2.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Y_K_FACT</td>
<td>1.75</td>
<td>0.32</td>
</tr>
<tr>
<td>Y_S_FACT</td>
<td>0.84</td>
<td>0.07</td>
</tr>
<tr>
<td>P_R_FACT</td>
<td>5.31</td>
<td>0.76</td>
</tr>
<tr>
<td>P_S_FACT</td>
<td>0.57</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Point Estimates and Standard deviation based on the assumed model.
Spatial Effect Test

exponential spatial correlation structure

Observations with different counties are assumed to be uncorrelated. Within the same county, assume the correlation between two measurement error observations a distance $r$ apart is

$$corr(e_{ij}, e_{ik}) = exp(-r/d)$$

where $r$ is chosen to be the euclidean distance applied to the longitude and latitude of the sample points.

A likelihood ratio test for hypothesis $H_0 : d = 0$ vs $H_1 : d > 0$ gives a p-value of 0.1411. So we conclude there is no significant spatial effect among the individual samples within the same county.
Spatial Effect Test (Cont'd)

Moran’s I Test

- Moran’s I statistic is given by

\[ I = \frac{n}{s_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

where \( s_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \), \( w_{ij} = 1 \) if \( i, j \) are neighbors and \( w_{ij} = 0 \) if \( i, j \) are not neighbors or \( i = j \).

- For random data, \( E(I) = -\frac{1}{n-1} \) and the variance has a complicated form.

- Null hypothesis: No spatial association, i.e. \( X_i \) iid

- For positively correlated data with similar neighbors, \( I \) is close to 1; for negatively correlated data with dissimilar neighbors, \( I \) is close to −1.
To choose the neighbor criterion to be used, we use R function `spdep::poly2nb` to create neighbors for polygon contiguities, using heuristics identifies polygons sharing boundary points as neighbors.

Rook-style Polygon Contiguity Neighbors.
Spatial Effect Test (Cont'd)

Apply Moran's I test to BLUP of random effects $\hat{u}$.

R output for global Moran's I under randomization

```r
## Moran I test under randomisation
## data:  est$ui
## weights: nb2listw(est.cov_nb)
## Moran I statistic standard deviate = 0.2156, p-value = 0.4147
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.001387866      -0.015873016       0.006409809
```

The null hypothesis of no spatial correlation is accepted.

**Remark:** $\hat{u}_i$ is only available for counties with at least one positive observation.
Visualization with Shiny
Beta Version

available at shinyapps.io
Discussion
Discussion

- Other resources for Covariates:
  - R factor - annual precipitation;
  - C factor - rotation type and land cover use derived from CDL.

- Application of overlaying:
  - Automated crop specific area frame stratifications based on CDL.
  - Assessment in CEAP at watershade scale.

- Existing visualization tools based on SSURGO:
  - Interfaces to SoilWeb - California Soil Resource Lab :: SoilWeb Apps
Reference
References


- Cooper, K., 2011. Evaluation of the Relationship between the RUSLE R-Factor and Mean Annual Precipitation.