

# Statistical Methodology for the Conservation Effects Assessment Project Cropland Farmer National Surveys

Patrick E. Flanagan, Ph.D.<sup>1</sup>  
National Statistician  
Natural Resources Conservation Service  
U. S. Department of Agriculture

February 2021

---

<sup>1</sup> Most of the CEAP Sampling Overview and the CEAP1 Survey Design Sections written by J. Jeffrey Goebel (2009)

## I. Introduction

### A. Purpose

The purpose of the National Assessment for Cropland (CEAP-Cropland) is to estimate the environmental benefits and effects of conservation practices applied to cultivated cropland and cropland enrolled in long-term conserving cover (e.g., the Conservation Reserve Program).

The CEAP-Cropland Component of the National Assessment has three specific goals:

- Estimate the effects of conservation practices currently present on the landscape.
- Estimate the need for conservation practices and the potential benefits of additional conservation treatment.
- Simulate alternative options for implementing conservation programs on cropland in the future.

The ultimate goal of CEAP-Cropland is to report conservation effects in terms that represent recognizable outcomes, such as cleaner water and soil quality enhancements that will result in more sustainable and profitable production over time.

### B. Background

The Conservation Effects Assessment Project (CEAP) was initiated in 2002 as a means by which to analyze societal and environmental benefits gained from the 2002 Farm Bill's substantial increase in conservation program funding. Though the National Resources Inventory (NRI) collected data on agricultural land, including erosion and conservation practices, links between those data were not directly discernable. In addition, many details about management practices, use of chemicals, and other aspects that affect the environment were not collected. The CEAP Cropland Farmer Surveys were designed to collect that data and link them via physical process models to environmental effects necessary to achieve the CEAP goals.

## II. Survey Sample Designs

### A. Overview

The objective of the NRI-CEAP Cropland Survey was to obtain additional site specific data needed to utilize the field-level process model APEX to estimate field-level effects of conservation practices. The process model was run for a sub-sample of NRI sample points; inputs for a sample point included historical NRI site specific data, data obtained from the NRI-CEAP Cropland Survey for the agricultural field where the sample point is located, additional information on conservation practices from Field Office records, soil properties and characteristics associated with the particular soil at the sample point location, and climate data

associated with the sample point location. The input data associated with a particular point describe a “representative field;” outputs from the process model runs include losses of materials (such as sediment and chemicals) from this field and changes in condition (such as accumulation of carbon). These outputs are used to estimate both on-site and off-site effects.

The APEX model outputs can be treated like other NRI variables; the site specific results for each sample point can be aggregated or averaged for some meaningful portion of the landscape using statistical weights. The statistical (survey) weight for an NRI sample point is the acreage value assigned to that sampling unit based upon the sampling design and certain control figures [derivation of weights for the NRI-CEAP Cropland Survey is discussed in Section VI, Estimation Procedure]. The APEX model outputs also serve as inputs into hydrologic models that simulate transport of water, sediment, and chemicals from the land into and through stream networks and eventually into estuaries and oceans. The NRI-CEAP data and the models can then be used to estimate changes in in-stream concentration of sediment and chemicals that result from changes in land management.

The sampling strategy utilized for the NRI-CEAP Cropland Survey was to select a sub-sample of NRI sampling units from the NRI Foundation Sample; in particular, a subset of sample points was selected from those sampling units used for the 2002 and 2003 Annual NRI surveys. Sampling strategies for the NRI Foundation Sample, Annual NRI surveys, and the NRI-CEAP survey are discussed below. The NRI sampling structure provided a natural framework for the data collection and modeling activities needed to support the CEAP national cropland assessment; it also provided efficiency to the process because sample locations were already identified and significant data already existed for these sites. The full collection of NRI sample sites provides a statistically credible representation of the diversity of soils, climate, cropping systems, and natural resource issues for the Nation’s agricultural lands. Data collection activities were spread over a four-year period because of financial constraints and operational considerations. A different set of sample points was selected for each year. The goal was to develop a data base that supported statistical analysis of the benefits of conservation practices at the national and regional levels.

## B. 2003 – 2006 Survey (CEAP1)

The target population for the NRI-CEAP Cropland Survey was all land in the 48 contiguous states that is classified by NRI as having a land cover/use of “cultivated cropland” or “land in CRP.” Cultivated cropland is defined by NRI as “land in row or close-grown crops, including hayland and pastureland in rotation with row or close-grown crops;” land in CRP is “land that was under a Conservation Reserve Program (CRP) contract.”

The sampling approach utilized for the NRI-CEAP Cropland Survey was to select a sub-sample of Annual NRI sample points. In particular, the sample comes from sampling units selected initially for the 2002 and 2003 Annual NRI surveys. The sampling strategy developed for the farmer surveys included:

- Collect data for 20,000 sample sites over a four year period, in order to obtain a full representation of the diversity of cropping systems, resource concerns, farming activities, conservation practices, soils, climate, and other natural resource conditions on cultivated cropland; and to obtain insight into implementation of conservation systems associated with the 2002 Farm Bill. [sample sites are cropland fields associated with NRI sample points; the Foundation NRI sample contains about 200,000 cropland points].
- Sampling and data collection for 2003 and 2004 were to focus on developing a good base-line for the most predominant cropping and conservation systems, to make sure that credible statistical analyses could be made on a national basis for all U. S. cultivated cropland.
- Sampling and data collection for 2005 and 2006 were to have a complementary focus: (a) to obtain data for areas and systems that are less extensive but usually more environmentally sensitive (vulnerable); and (b) to obtain data on actual changes in conservation systems and practices that occurred due to implementation of 2002 Farm Bill provisions – data collection in 2005 and 2006 provided a fuller and broader perspective, since some practices were not installed until after 2003.

An NRI sample point is used to identify a field in order to determine land cover/use and management systems; similar protocols are used to determine the natural or inherent features, such as soil type or erosion equation factors. The NRI utilizes points as the sampling units rather than farms or fields; land use and land unit boundaries change frequently in some parts of the country, and factors such as soil type do not follow human-induced boundaries such as land unit boundaries. Sample point coordinates are known based upon Digital Ortho-Photo Quadrangle (DOQ) base maps and standards. The temporal nature of desired results was handled in several ways: (i) the NRI-CEAP farmer survey collected site specific data for several years, and historical NRI data are available for each sample point; (ii) conservation practices, other agricultural management systems, and acts of nature have long-term effects upon the environment – the process models used to quantify effects produce results by year and season; (iii) the Annual NRI utilizes a supplemented panel survey design, wherein each year's sample includes a Core Panel (sampling units observed each year) and a Supplemental (or rotating) Panel – this provides the flexibility to revisit sample units over the course of time.

### Sample for 2003 Survey

The sample for the 2003 NRI-CEAP Farmer Survey was selected from the 2002 Annual NRI sample points classified as having a land cover/use of either cultivated cropland or land in CRP for the 2002 growing season. In particular, the samples were selected from the supplemental panel P02, as follows:

- (a) Any sample point in P02 classified as "land in CRP" for 2002 was included.
- (b) Sample points classified as "cultivated cropland" were selected as follows:
  - it was determined which segments in P02 contained at least one point classified as "cultivated cropland" for 2002
  - within each of those segments, one point classified as "cultivated cropland" in 2002 was selected randomly.
- (c) For South Dakota and North Dakota, one-half of these points were not sampled; systematic sampling was used to select half of the points. The sampling rate was reduced due to lack of available interviewers within these two states.

- (d) An additional 333 points were removed from the sample because they represented farm operators that had also been selected for the ARMS-II survey. These samples were removed from the survey so that respondent burden for ARMS-II would not be affected. An initial examination of these overlap samples indicated that no bias should be expected; the samples were distributed across the country in proportion to cropland occurrence. This will be verified as part of a post-survey statistical evaluation of non-response, which will utilize historical NRI information and operator information collected from NRCS field offices.

Sample sizes by state are presented in Table 6. The sample included 2,236 CRP sample points and 9,580 cultivated cropland points.

#### Sample for 2004 Survey

The sample for the 2004 NRI-CEAP Cropland Survey was selected from the 2003 Annual NRI sample points classified as having a land cover/use of either cultivated cropland or land in CRP for the 2003 growing season. In particular, the samples were selected from the supplemental panel P03, as follows:

- (a) Any sample point in P03 classified as “land in CRP” for 2003 was included.
- (b) Sample points classified as “cultivated cropland” were selected as follows:
  - it was determined which segments in P03 contained at least one point classified as “cultivated cropland” for 2003
  - within each of those segments, one point classified as “cultivated cropland” in 2003 was selected randomly.

The sample included 2,268 CRP sample points and 10,148 cultivated cropland points.

#### Sample for 2005 Survey

The sample for the 2005 NRI-CEAP Cropland Survey was selected from the 2003 Annual NRI sample points classified as having a land cover/use of either cultivated cropland or land in CRP for the 2003 growing season. In particular, the samples were selected from the Core Panel P00, as follows:

- (a) Any sample point in P00 classified as “land in CRP” for 2003 was included.
- (b) Sample points classified as “cultivated cropland” were selected as follows:
  - it was determined which segments in P00 contained at least one point classified as “cultivated cropland” for 2003
  - within each of those segments, one point classified as “cultivated cropland” in 2003 was selected randomly.
- (c) The following randomization process was used to eliminate all cropland sample points in 10 states:
  - Minnesota and Wisconsin were paired [placed in Stratum A]; each was given an equal chance of selection. Minnesota was kept in the sample and Wisconsin was selected for elimination.
  - North Dakota and South Dakota were paired [placed in Stratum B]; each was given an equal chance of selection. South Dakota was kept in the sample and North Dakota was selected for elimination.
  - The states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut were combined into a New England Grouping. New York and the New

England Grouping were paired [placed in Stratum C]; each was given equal chance of selection. New York was kept in the sample and the New England Grouping was selected for elimination.

- The states of Montana, Colorado, Wyoming, Utah, and New Mexico were grouped [placed in Stratum D]; each was given an equal chance of selection. Colorado, Montana, and Utah were kept in the sample; Wyoming and New Mexico were selected for elimination.
- (d) Sample sizes for cultivated cropland were reduced in 11 states, as follows:
- randomization techniques were utilized that reduced the sample by one-third in four states: Kansas; Minnesota; North Carolina; Ohio
  - randomization techniques were utilized that reduced the sample by one-half in two states: South Dakota; Texas
  - randomization techniques were utilized that reduced the sample by two-thirds in five states: Illinois; Indiana; Iowa; Missouri; Nebraska
- (e) No cropland points in Florida, Nevada, and West Virginia were included for the 2005 survey; problems had been encountered in the 2003 and 2004 surveys. These three states were included for the 2006 survey.

Sample sizes by state are presented in Table 6. The sample included 3,893 CRP sample points and 7,489 cultivated cropland points. The sample size for cultivated cropland was about 25% less than for each of the earlier years; less funding was available for conducting farmer interviews.

### Sample for 2006 Survey

The primary objective for sampling in 2006 was to provide a greater ability to make regional-level assessments (rather than just national), particularly by Major River Basin. Stratified sampling techniques were used to concentrate on fields in the most environmentally sensitive (or vulnerable) areas in order to provide more precise estimates of the effects of conservation in areas where the impacts of conservation are the greatest; sampling in 2003, 2004, and 2005 provided appropriate representation for predominant situations that covered 90% of the cropland base. Funding existed to conduct approximately 6,000 farmer interviews for cultivated cropland fields; no additional tracts of CRP land were selected.

Each county was ranked relative to its potential for soil and nutrient loss from cropland, by using the National Nutrient Loss and Soil Carbon (NNLSC) database which contains estimates based upon EPIC model runs for 1997 NRI cropland sample points [see Potter et al (2006)]. The NNLSC database used general information on farming practices that was imputed onto the NRI cropland sample points. County level estimates were derived for: wind erosion, waterborne sediment, nitrogen loss in sediment, phosphorus loss dissolved in runoff, nitrogen loss dissolved in runoff, and nitrogen loss dissolved in leachate. County vulnerability rankings were derived using these seven factors as follows:

- A county was classified with vulnerability rank 1 if it had an estimated value for at least one factor in the top 10%; for wind erosion, the factor needed to be in the top 3% of all counties because 85% of all counties do not have significant cropland wind erosion. This category contained 658 counties.

- A county was classified with vulnerability rank 2 if it was not classified as vulnerability rank 1 but had an estimated value for at least one factor in the top 20% [top 5% for wind erosion]. This category contained 385 counties.
- A county was given a vulnerability rank 3 if its vulnerability could not be estimated from the NNLSC database and it contained at least 20,000 acres of cultivated cropland. This category included 70 counties.
- Counties with low and very low vulnerability according to these seven factors were given vulnerability ranks 4 and 5 respectively. There were 736 counties with rank 4 and 1,255 counties with rank 5.

The sample for the 2006 NRI-CEAP Farmer Survey came from 2003 Annual NRI sample points that had not been selected for previous farmer surveys. Each state and county had a different assortment of available cultivated cropland sample points relative to the county vulnerability rankings described above. The 2006 sample is not a stand-alone sample as are the samples for the three previous years. Some areas had no probability of selection for the 2006 survey; the 2006 results can only be used in conjunction with data collected for previous survey years.

For the 2003, 2004, and 2005 NRI-CEAP Farmer Surveys, sample points were spread out across states and counties as much as possible given the nature of the 2002 and 2003 Annual NRI samples. For example, only one cultivated cropland point per sample segment was selected for the farmer surveys; this spread out the sample and also greatly reduced the chance that the same farmer or operator was included in the sample more than one time in a given year. This was a restriction put in place following discussions with USDA-NASS and the Office of Management and Budget (OMB) in an effort to reduce respondent burden. For the 2006 sample, it was necessary to select some sample points in sample segments that had been used for the 2004 or 2005 sample.

One of the basic methods of sample selection for 2006 was as follows:

- determine which segments in P00 and P03 had at least two points classified as cultivated cropland in 2003
- if the segment had two points classified as cultivated cropland in 2003 and the county had vulnerability rank less than 4, select the sample point not used for either the 2004 or 2005 survey
- if the segment had three points classified as cultivated cropland in 2003 and the county had vulnerability rank less than 4, randomly select one of the two sample points not used for either the 2004 or 2005 survey
- no sample points were selected in counties with vulnerability rank 4 or 5.

This procedure was used for Alabama, Arizona, California, Colorado, Kentucky, Michigan, Mississippi, New Jersey, North Carolina, Oklahoma, Oregon, Tennessee, Utah, Virginia, and Washington. The modified procedure used for Arkansas, Georgia, Idaho, Louisiana, Maryland, Pennsylvania, and South Carolina was that only sample points from P03 were used.

Florida, Maine, Massachusetts, Nevada, New Mexico, Vermont, and West Virginia used sample points in all P00 segments not used for the 2005 survey. For Indiana, Iowa, and Nebraska, sample points were selected from all P00 segments not used for 2005 for counties with rank 1, and half in counties with rank 2; for Delaware, Missouri, North Dakota, Wisconsin, and

Wyoming, all eligible P00 points were selected except only half in counties with rank >3. For Kansas and Texas, sample points were selected from all P00 segments not used for 2005 for counties with rank < 4; and sample points were selected from all eligible P03 segments in counties with rank 1, and half were selected for counties with rank 2 or 3. For Minnesota and South Dakota, all eligible sample points in counties with rank 1 or 2 were selected, and half of the P00 rank 4 or 5 sample points. For Connecticut, half of the P00 points were selected. For Illinois, sample points in P00 segments not used for 2005 were used in counties with rank 1; sample points were selected for half of the segments in counties with rank > 1. For Montana, all eligible sample points in counties with rank 3 were selected; sample points were selected from segments in half of the eligible P03 counties with rank 4 or 5. For Ohio, all eligible points in segments in counties with rank 1 and 2 were selected, except for half of the P03 segments with rank 2. For New York, all eligible points in segments in counties with rank 1 and 2 were selected, except for P00 segments with rank 2. No sample points were selected in New Hampshire and Rhode Island.

### C. 2015 – 2016 Survey (CEAP2)

#### Frame

A point is included in the frame if the most recent collected land cover/use (LCU) satisfies one of the following conditions:

- LCU > 0 and LCU < 200 and LCU ≠ 7
- LCU in {211, 212, 213}
- LCU = 200 and not range
- LCU = 410

The states included in the frame are the coterminous 48 states (not including DC).

Aquaculture (171) is treated just like other cropland LCUs.

Points that are classified as urban or roads in 1997 and as 200-213 in the most recent collected year are not eligible. The last set of points removed contains 376 points, one of which is 213 in the most recent year and the rest of which are 200. These 376 points are removed because NRI editing procedures change the a collected LCU of 200-213 to urban.

Each point is classified into one of five mutually exclusive and exhaustive LCU groups. With  $t$  representing the most recent year, the LCU categories obtained from this program in item 2 are defined as follows:

- 1 = LCU( $t$ ) in 1 – 20 (high value specialty crops)
- 2 = LCU( $t$ ) in 141-144 and LCU( $t-1$ ), LCU( $t-2$ ), LCU( $t-3$ ) not in the set 11 – 116
- 3 = LCU( $t$ ) in 200-213 and LCU( $t-1$ ), LCU( $t-2$ ), LCU( $t-3$ ) not in set 11-116
- 4 = LCU( $t$ ) in 21-116, 170, 171, 180, or LCU( $t$ ) in 141-144 and at least one of LCU( $t-1$ ), LCU( $t-2$ ), LCU( $t-3$ ) in the set 11-116, or LCU( $t$ ) in 200-213 and at least one of LCU( $t-1$ ), LCU( $t-2$ ), LCU( $t-3$ ) in the set 11-116
- 5 = LCU 410 (CRP)



The combination of groups 1, 2, and 4 below approximates the NRI definition of cropland (cultivated and non-cultivated combined). Category 3 approximates the NRI definition of pasture. (A perfect classification of NRI points into broaduses is not possible with the data available because of the 2004 protocol change.) The five LCU categories are aggregated to the following three groups.

- Cropland (1): LCU category of 1, 2, 4
- Pasture (3): LCU category of 3
- CRP (4): LCU category of 5

The CRP category is labeled 4 for consistency with the original code of 410

Ten CEAP production regions are defined for the 2015-2016 National survey. Figure 1 shows a map of the ten production regions. The specifications for the regions are outlined in Table 1.

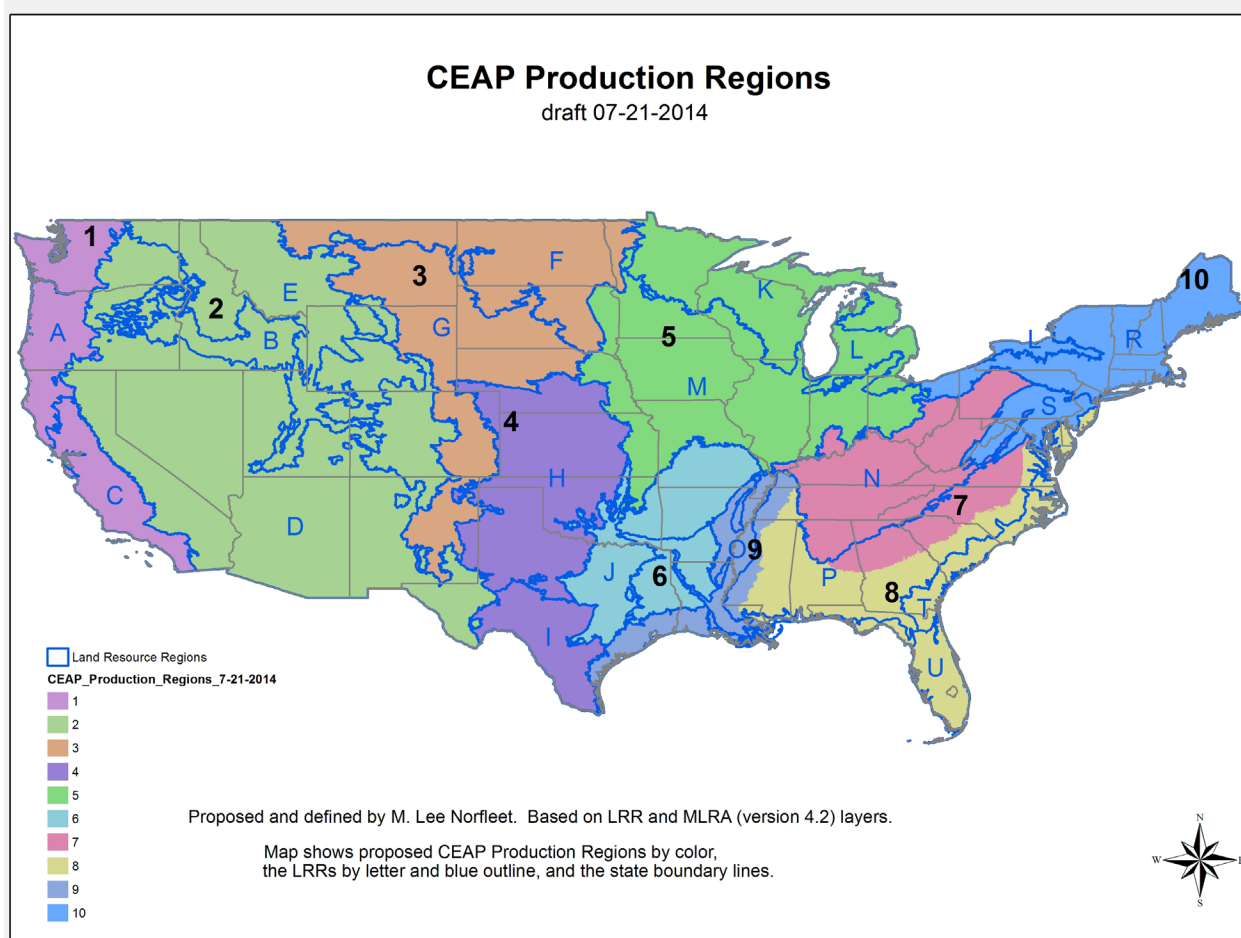


Figure 1: CEAP Production Regions

	Working Name	LRR	Refinements	Concept or Rationale
1	Pacific Coast	A, C	None	High value irrigated crops. Water quantity issues

2	Irrigated West	B, D, E	None	Water quantity resource concern note the non Irr portions
3	Northern Plains Wheat Belt	F, G	Dryland wheat predominant Wind ero.	Dominated by wheat and other small grains
4	Southern Plains Wheat/Cotton	H, I	Aquifer driven	Wheat cotton region, wind erosion and aquifer depletion
5	Corn Belt	K, L, M	Less 109 from LRR L	Corn soy dominated region
6	South Central Pasture and Crop	J, P. and western N	P west of MS River	Animal ag and manure on crop/pasture hay common
7	East Central Pasture and Crop	Eastern portion N	Plus MLRA 136	Animal ag and manure on crop/pasture hay common
8	Southeast Mid Atlantic Coastal Plain	T, P, U	East of MLRA 134	Split from other for less rain intensity and courser soil in general
9	Lower MS River and Texas Gulf	O, T, and P	West with MLRA 134	Silty soils with cotton rice and cane. Intense rain
10	Northeast	R, S, + 101 from L		

Table 1: Description of 10 CEAP Production Regions for Sampling

### Sample Size by CEAP Region

Let  $\hat{A}_h$  be the estimate of the area in CEAP-eligible categories in CEAP region  $h$  in the year 2010. The estimated areas are obtained from the 2010 pointgen. Due to topological errors in the CEAP region shapefile, not all NRI points are in a CEAP region. The total area corresponding to points classified in eligible LCUs that are not in a CEAP region is 83,800 acres. The total estimated area in CEAP-eligible categories is 480,436,800 acres. Because the area associated with missing CEAP regions is only 0.017% of the estimated area based on points located in a CEAP region polygon, the area with missing CEAP regions is ignored for the sample size calculation. Define a target sample size for CEAP region  $h$  by  $n_h$ , the result of accumulating and rounding the  $\tilde{n}_h$ , where

$$\tilde{n}_h = \frac{N \hat{A}_h^{0.5}}{\sum_{h=1}^H \hat{A}_h^{0.5}}$$

and  $N = 45,000$ : This yields the sample sizes in Table 3 below. The square root allocation is often used as a compromise between equal allocation, optimal for individual area estimates, and proportional allocation, optimal for the total of the regions combined (Bankier, 1988)

CEAP.Reg	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
$n_h$	2420	3881	5733	5651	8701	4070	4120	3611	3700	3113

Table 3: Target sample sizes by CEAP region

A more detailed discussion of the sampling, adjustment to the totals, and analysis or results are provided in Appendix B.

## Change to the 2016 CEAP Sample

A point is included in the initial frame if the point's latest observed LCU satisfies at least one of the following:

- $> 0, < 200$ , not 7
- 211 – 212
- 200 and not range
- 410

We refer to the above set of LCUs as the eligible set. We add several flags to use at various stages of the processing. We call the following LCU-ineligible:

- BDB-ineligible:  $!(\text{collect14}=0 \text{ or } (\text{collect14}=200 \text{ and } \text{range14} = 0))$  Points ineligible according to the information available in collect for 2014.
- REMOVED
- FEDERAL
- Urban Change (latest LCU = 200 and 1997 = urban)
- Latest LCU not eligible set (This can occur if new data were collected or if data were edited.)

We call the following NRI-ineligible:

- LA no 1997 data
- In 2001 not 2014
- Killed
- Salvaged

We replace points that are in the current 2016 sample and are ineligible. We define a point to be ineligible if it is either LCU-ineligible or NRI-ineligible. Points in the 2013 or 2014 sample are not replaced and are not used as replacements in this step. The removed points are replaced using the following algorithm:

- For psus containing removed points that also contain least one other eligible point, randomly select one of the eligible points from the psu.
- Remaining psus do not contain at least one other eligible point. Define a stratum to be an intersection of a CEAP region, segment LCU group (defined for the original 2015-2016 sample), sample class, and state. Tabulate the number of psus to replace in the stratum. Select one psu at random from the psus containing at least one eligible point in the stratum. Randomly select one point from the eligible points in the selected psu. Prior CEAP points are selected with priority when selecting points from segments.

We maintain the one point per segment rule. Segments containing points in the 2013, 2014, or 2015 samples are not eligible for selection as replacements. Likewise, segments containing at least one usable point for 2016 are not eligible as replacements.

A total of 789 points are replaced in this fashion. This procedure replaces all but 44 of the points that need to be replaced. We select 44 additional crop points in the next step to compensate for these 44 points.

#### Selection of Additional Crop Points

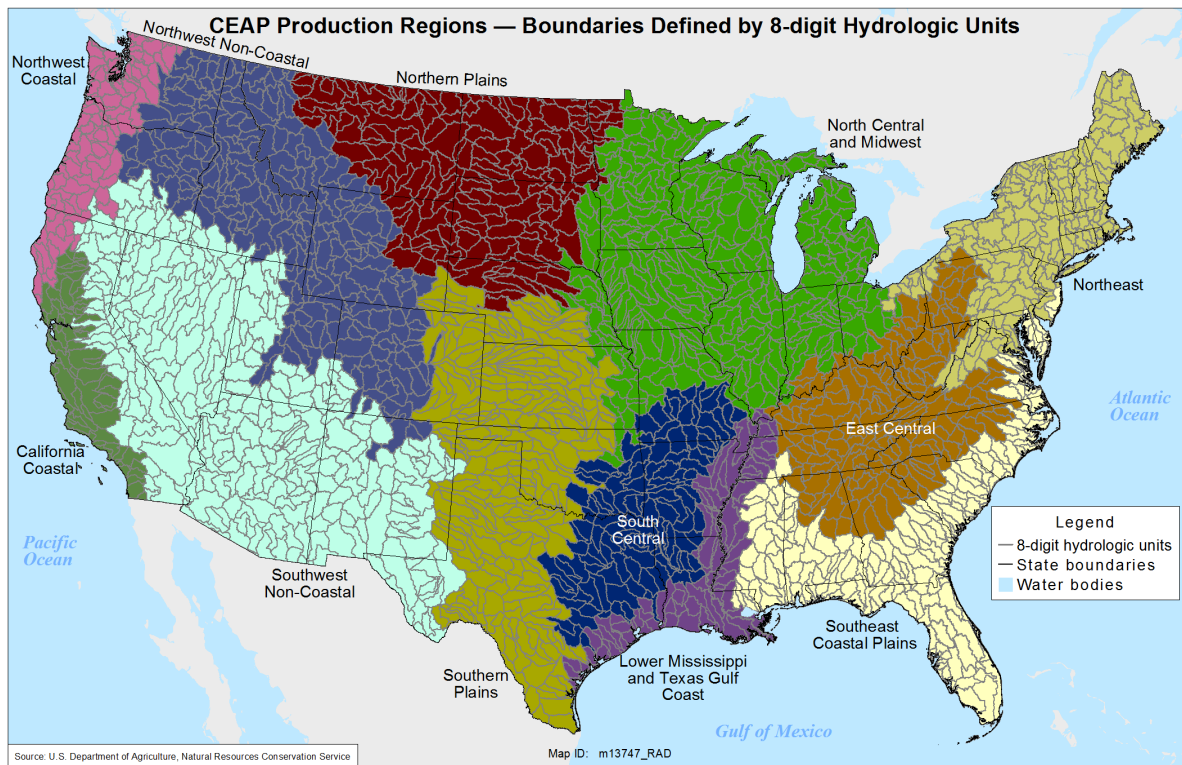
To reach 25,000 total points, we select 1376 additional crop points. Two segments containing eligible crop points are selected in region 6, and randomly selected points in these two segments are added. The crop sample sizes for the remaining CEAP regions are approximately in proportion to the estimated crop area in the CEAP region. Within a CEAP region, we define strata by intersections of sample classes and states for segment group 1 (crop). The number of segments to select from segment group 1 is proportional to 2 the number of segments in the stratum. Segments such that all points are LCU-ineligible are removed from the counts for this allocation. A simple random sample of segments of the specified size is selected from each stratum. The CEAP region sizes are (37, 94, 238, 227, 513, 2, 59, 65, 89, 53) for regions (1, 2, 3, 4, 5, 6, 7, 8, 9, 10). For the original sample, we used a compromise of one-per-stratum and two-per-stratum sampling within the strata. We use simple random sampling here for simplicity. We randomly select one point from the eligible points in the selected psu, with prior CEAP points selected with priority.

### III. Estimation

#### A. Overview

In both CEAP1 and CEAP2, points were selected from the NRI using stratification, which gives each point a different probability of selection. In addition, the NRI itself used a similar process. To produce unbiased estimates from such a multistage sample, each point is given a weight. Its base weight is computed from the inverse of its probability of selection taking all stages into account. Then, weights must be adjusted for in-scope nonresponse in a classification manner that mitigates the impact on nonresponse bias. Finally, weights are adjusted to known (or nearly known) totals such as total cultivated cropland by region.

After the sample selection for CEAP2, the CEAP regions were redefined as shown below. Since estimates were to be created by these regions, both CEAP1 and CEAP2 weights were adjusted based on these regions.



## B. CEAP1

### Introduction

The Annual NRI estimation procedure combines information from several sources to produce a final data set composed of records containing information for the years 1982, 1987, 1992, 1997, 2000, and annually thereafter. Each record represents data elements for a sample point; an estimation weight is attached to each record. For each NRI survey year, data are collected at both the segment level and at the point level. The areas measured for small water features, roads and railroads, and urban and built-up lands are converted to point data during the estimation process. Each of these created points is given an initial weight based on the area in the segment and the probability that the segment is included in the sample; imputation is used for unobserved data elements in order to complete the data record for these created points. Initial weights for created points and for observed points are adjusted during the estimation process using ratio adjustments and small area estimation. Control totals for surface area, federal land, and large water areas, derived from GIS databases, are maintained throughout the process. Finally, the weights are adjusted using iterative proportional scaling (raking) so that the new data base produces acreage estimates for broad cover/use categories for historical years that closely match previously published estimates [see Fuller (1999)].

## Development of Estimation Weights for NRI-CEAP

Estimation weights for the NRI-CEAP1 cultivated cropland sample points were developed in a manner consistent with development of weights for the Annual NRI. Weights for other river basins will be developed in a similar fashion although some additional ratio adjustment procedures may be utilized, for example, for irrigated conditions. Estimation weights for points identified as “land in CRP” were basically those derived for the Annual NRI data base.

The procedure for points identified as cultivated cropland follows:

- Calculate initial weights, where  $W_{\text{Init},q,k,j}$  is the initial weight for point j, where point j falls within 6-digit hydrologic unit q and has cropping system k

$$W_{\text{Init},q,k,j} = A_{q,k,j} / (p_{q,k,j} * m_{q,k,j}), \text{ where:}$$

$$A_{q,k,j} = \text{size of segment } (q,k,j) \text{ in acres,}$$

$$p_{q,k,j} = \text{probability that segment } (q,k,j) \text{ is in the sample,}$$

$$m_{q,k,j} = \text{number of sample points in segment } (q,k,j)$$

- Make the first adjustment to the initial weights

$$W_{\text{Adj1},q,k,j} = (W_{\text{Init},q,k,j}) * (Y_k / X_k), \text{ where:}$$

$$Y_k = \text{estimated acres of cultivated cropland in cropping system } k, \\ \text{based upon 2003 Annual NRI}$$

$$X_k = \sum_{q,j} W_{\text{Init},q,k,j}$$

- Make the second adjustment to the initial weights

$$W_{\text{Adj2},q,k,j} = (W_{\text{Adj1},q,k,j}) * (T_q / Z_{1,q}), \text{ where:}$$

$$T_q = \text{estimated acres of cultivated cropland in 6-digit} \\ \text{hydrologic unit } q, \text{ based upon 2003 Annual NRI}$$

$$Z_{1,q} = \sum_{k,j} W_{\text{Adj1},q,k,j}$$

- Make the third adjustment to the initial weights

$$W_{\text{Adj3},q,k,j} = (W_{\text{Adj2},q,k,j}) * (Y_k / X_{2,k}), \text{ where:}$$

$$X_{2,k} = \sum_{q,j} W_{\text{Adj2},q,k,j}$$

- Make the fourth adjustment to the initial weights

$$W_{\text{Adj4},q,k,j} = (W_{\text{Adj3},q,k,j}) * (T_q / Z_{3,q}), \text{ where:}$$

$$Z_{3,q} = \sum_{k,j} W_{\text{Adj3},q,k,j}$$

- Designate the final adjusted weight for point (q,k,j) to be the estimation weight,  $W_{0,q,k,j}$

#### Development of Replicate Weights for Estimating Variances

A form of jackknife variance estimation is utilized for the Annual NRI because of the rather complex nature of the estimation procedure. The Annual NRI survey process is a type of two phase sampling, since the samples represent a subsample of segments selected from the 1997 NRI sample. The replication method used for the NRI is a form of the “delete-a-group jackknife” [see Kott (2001)]. The goal of the variance estimation procedure for an Annual NRI data set is to construct a set of H modified weights for each observation, which allows computation of H replicate estimates for a variable y. A variance estimate can then be calculated for an NRI estimate, say  $\hat{Y}$ , as follows:

$$\text{var}(\hat{Y}) = \sum_h c_h * (\hat{Y}_h - \ddot{Y})^2, \text{ where}$$

$c_h$  is a constant determined by the replication procedure

$\hat{Y}_h$  is the  $h^{\text{th}}$  replicate estimate for Y, and

$$\ddot{Y} = H^{-1} \sum_h \hat{Y}_h$$

For the 2003 Annual NRI and the NRI-CEAP cropland survey,  $H = 29$  is used. To define the replicates, a form of systematic sampling was used with the 1997 NRI sample units to create 29 groups of samples of approximately equal size. The same set of replicates is used for both the 2003 Annual NRI and the NRI-CEAP cropland database. This means that an estimation process can be established so that variance estimates based upon the larger sample can be retained within the smaller data base, if certain regression and/or ratio techniques are utilized.

The first set of replicate weights for the NRI-CEAP data set is derived as follows:

- Calculate initial weights for the point (q,k,j) by modifying the estimation weight,  $W_{0,q,k,j}$ , as follows:

$$\begin{aligned} W_{\text{Init},1,q,k,j} &= 0, & \text{if point (q,k,j) is in replicate \#1} \\ &= (29/28) * W_{0,q,k,j}, & \text{otherwise} \end{aligned}$$

- Make the first Adjustment to the Initial Weights

$$W_{\text{Adj}1,1,q,k,j} = (W_{\text{Init},1,q,k,j}) * (Y_k / X_{1,k}), \text{ where:}$$

$Y_k$  = estimated acres of cultivated cropland in cropping system k,  
based upon 2003 Annual NRI

$$X_{1,k} = \sum_{q,j} W_{\text{Init},1,q,k,j}$$

- Make the second adjustment to the initial weights

$$W_{\text{Adj}2,1,q,k,j} = (W_{\text{Adj}1,1,q,k,j}) * (T_q / Z_{1,q}), \text{ where:}$$

$T_q$  = estimated acres of cultivated cropland in 6-digit

hydrologic unit  $q$ , based upon 2003 Annual NRI

$$Z_{1,q} = \sum_{k,j} W_{Adj1, 1,q,k,j}$$

- Make the third adjustment to the initial weights

$$W_{Adj3,1, q,k,j} = (W_{Adj2,1, q,k,j}) * (Y_k / X_{1,2,k}), \text{ where:}$$

$$X_{1,2,k} = \sum_{q,j} W_{Adj2,1, q,k,j}$$

- Make the fourth adjustment to the initial weights

$$W_{Adj4, 1,q,k,j} = (W_{Adj3, 1,q,k,j}) * (T_q / Z_{1,3,q}), \text{ where:}$$

$$Z_{1,3,q} = \sum_{k,j} W_{Adj1,1, q,k,j}$$

- Designate the final adjusted value for point  $(q,k,j)$  to be the first replicate weight,  $W_{1, q,k,j}$

A similar process is used for each of the remaining 28 replicates. Each point  $(q,k,j)$  then has an estimation weight,  $W_{0, q,k,j}$ , and a set of 29 replicate weights,  $\{ W_{h, q,k,j} : h=1,2, \dots, 29 \}$ , that are used for variance estimation.

## 2020 Weight Adjustments

For comparisons between CEAP1 and CEAP2, the CEAP1 weights were given a final ratio adjustment to the 2003 totals for cultivated cropland in each of the 12 current CEAP regions taken from the 2017 NRI database.

## C. CEAP2

Preserves state controls exactly for states with ratio adjustment factors in a specified interval.

1. Compute base weight as defined as follows:

For point  $j$  in segment  $l$ , define the base weight by

$$w_{0,l,j} = p_{j,l}^{-1} A_l \frac{N_{f,g(i)} m_{cc,i}}{n_{s,g(i)} m_i \tilde{m}_{cc,i}},$$

where

- $P_{f,i}$  is the foundation probability or segment  $i$ ;
- $A_i$  is the area of segment  $i$ ;
- $N_{f,g(i)}$  and  $n_{s,g(i)}$ , respectively, are the number of segments in the frame and sample in the group  $g$  containing segment  $l$ ;
- $m_{cc,i}$  is the number of CEAP2 eligible (real) points in segment  $i$ ;
- $m_i$  is the number of real points in segment  $i$ ; and
- $\tilde{m}_{cc,i}$  is the number of points in the CEAP2 sample in segment  $i$ .



The number of CEAP2 points in a segment ( $\tilde{m}_{cc,i}$ ) is 1 except for segments in the 2013 or 2014 regional CEAP surveys, where the number can exceed 1.

The groups are defined as intersections of the 2015 CEAP2 regions (see sample design discussion about regions), the 7 aggregated sample classes, and states.

The frame is the union of the frame used for the 2015 sample and the revised frame used for the revised 2016 sample.

2. Nonresponse adjustment. Bound ratios by 4. Cells defined as intersections of...
  - Two broaduse groups (based on pgen)
    - Cropland = Cultivated, noncultivated
    - Pasture
  - HUC 4 ( $\approx 200$  HUC 4's)
    - PGEN HUC 4 definition
  - 2015 CEAP Regions
    - Five erosion categories "quintiles"
    - National quintiles instead (unweighted, year 2012 from 2012 pgen):
3. Combine Rhode Island cultivated with Connecticut cultivated, and combine Nevada cultivated with Nevada non-cultivated
  - No CEAP point is classified as Rhode Island cultivated
  - One CEAP point is classified as Nevada cultivated, and this point has broaduse non-cultivated cropland through 2008. (The point is a new rotation point that changes to cultivated cropland when it is sampled in 2009.)
4. Ratio adjustment at state level x 3 NRI broaduses (cultivated, non-cultivated, pasture). NO BOUNDS on ratios. With no bounds on ratios and combining states as in step 3, national level estimates are preserved.
5. Truncate weights from step 4 to remain in [Median/4, Median\*4] by CEAP region
6. Repeat ratio adjustment at state level x 3 NRI broaduses (cultivated, non-cultivated, pasture). No bounds on ratios. Bound ratios to remain in [0.75, 1.25]. Call the weights that result from this step 6 "W2."
7. Use raking (successive ratio adjustments) to control to 2015 broaduse estimates (cultivated, non-cultivated, pasture separately) by 20CEAP region and HUC-2. We control to CEAP region and HUC-2 margins, not intersections. The total number of controls is  $3(C + H)$ , where C is the number of CEAP regions and H is the number of HUC-2's.
  - Use the broaduse designation for the year 2015 from the 2017 pointgen
  - Hold weights for points not classified as crop or pasture fixed at W2 from step 6
  - Only ratio adjust weights for points classified as cultivated crop, non-cultivated crop, or pasture in 2015 based on the 2017 pointgen.
8. Final adjustment was made after correcting some HUC8 designations to the 12 CEAP Regions.

#### Replicate variance estimation:

The replicate weight procedure starts with the weights from step 6. Sort all points (note CEAP sample has 1 point per segment) by state and by geoorder within stage. The point in position  $j$  is assigned the replicate number  $r = (j - 1) \bmod 29 + 1$ . We set the  $r^{th}$  replicate weight for a

point equal to 0 if the point is assigned replicate number  $r$ . Otherwise, we set the weight to  $W_2$  from step 6. We repeat step 7 with the replicates.

#### D. Change Between Surveys

An important analysis of CEAP2 vs. CEAP1 is whether a characteristic measured in a specified geographic area shows a statistically significant change. Since most of these measures are sums, or averages constructed from a nontrivial number of sample points, the Central Limit Theorem applies. If the absolute difference of the two,  $|\Delta_{12}|$ , is greater than  $z_\alpha \hat{\sigma}_{12}$  ( $|\Delta_{12}| > z_\alpha \hat{\sigma}_{12}$ ), where

- $z_\alpha$  is the  $\alpha$  level standard normal z score for the chosen type I error (often  $z_\alpha = 1.96$ , with  $\alpha = 0.25$ ), and
- $\hat{\sigma}_{12}$  is the standard error of the difference, further discussed below, then the difference is regarded as statistically significant.

An estimate of the standard error of the difference discussed above is computed using the formula

$\hat{\sigma}_{12} = \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2}$ , where

- $\hat{\sigma}_1^2$  is an estimate of the variance of the CEAP1 estimate, using replicated weights (see below) and  $\hat{\sigma}_1$  is the square root of that variance;
- $\hat{\sigma}_2^2$  is an estimate of the variance of the CEAP2 estimate, using replicated weights (see below) and  $\hat{\sigma}_2$  is the square root of that variance; and
- $\hat{\rho}_{12}$  is an estimate of the covariance of the two estimates (see below).

Given a characteristic estimate or observation  $y_i$  at each of the points of interest in a geographic area ( $y_i$  could be a dichotomous [0,1] variable indicating in or not in a category), the estimates above are calculated as follows using the  $R = 29$  replicate weights:

- $\hat{\sigma}_1^2 = \sum_{r=1}^R (t_{1(r)} - \hat{t}_1)^2$ , where  $t_{1(r)} = \sum_S w_{1i(r)} y_{1i}$ , for  $r = 1..R$ , and  $\hat{t}_1 = \frac{1}{R} \sum_{r=1}^R t_{1(r)}$
- $\hat{\sigma}_2^2 = \sum_{r=1}^R (t_{2(r)} - \hat{t}_2)^2$ , where  $t_{2(r)} = \sum_S w_{2i(r)} y_{2i}$ , for  $r = 1..R$ , and  $\hat{t}_2 = \frac{1}{R} \sum_{r=1}^R t_{2(r)}$
- $\hat{\rho}_{12} = \sum_{r=1}^R (t_{1(r)} - \hat{t}_1)(t_{2(r)} - \hat{t}_2)$ , where
  - $t_{1(r)} = \sum_{S^*} w_{1i(r)} y_{1i}$ , for  $r = 1..R$ , and  $\hat{t}_1 = \frac{1}{R} \sum_{r=1}^R t_{1(r)}$ ,
  - $t_{2(r)} = \sum_{S^*} w_{2i(r)} y_{2i}$ , for  $r = 1..R$ , and  $\hat{t}_2 = \frac{1}{R} \sum_{r=1}^R t_{2(r)}$ , and
  - $S^*$  is the subset of the sample points in both CEAP1 and CEAP2.

## References

- Bankier M. D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42(3), 174-177.
- Breidt, F.J. & W.A. Fuller (1999) Design of supplemented panel surveys with application to the National Resources Inventory, *Journal of Agricultural, Biological, and Environmental Statistics*, 4(4): 391 – 403.
- Fuller, W.A. (1999) Estimation procedures for the United States National Resources Inventory, *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, 39 – 44.
- Goebel, J.J. (2009). *Statistical Methodology for the NRI-CEAP Cropland Survey*, Natural Resource Conservation Service, Washington, D.C.
- Kott, P.S. (2001) The delete-a-group jackknife, *Journal of Official Statistics*, 17: 521 – 526.
- Nusser, S.M. & J.J. Goebel (1997) The National Resources Inventory: a long-term multi-resource monitoring programme, *Environmental and Ecological Statistics*, 4(3):181- 204.
- Potter, S.R., S. Andrews, J.D. Atwood, R.L. Kellogg, J. Lemunyon, L. Norfleet, D. Oman (2006) *Model Simulation of Soil Loss, Nutrient Loss, and Change in Soil Organic Carbon Associated with Crop Production*, Natural Resources Conservation Service, USDA, Washington, D.C.
- Schnepf, M. (2016). *A History of Natural Resource Inventories Conducted by the USDA's Soil Conservation Service and Natural Resources Conservation Service*, Natural Resources Conservation Service, Washington, D.C.

## Appendix A. The National Resources Inventory (NRI)

### Introduction to the NRI

The current National Resources Inventory evolved from a need for information to guide decisions about resources conservation after the “Dust Bowl” of the 1930s. After evolving through several iterations, the National Resources Inventory was formally mandated in the 1972 Rural Development Act and its current design began in 1982. In 2000, it converted from a 5-year collection to an annual design. Throughout that time period its scope expanded from a heavy focus on cropland erosion to a much wider assessment of resources described herein. A detailed account of the history of the NRI can be found on the NRI Website, <http://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/nri/>, at the “History of the NRI” link. That contains the report, “A History of Natural Resource Inventories Conducted by the USDA’s Soil Conservation Service and Natural Resources Conservation Service” compiled by Max Schnepf for the Soil and Water Conservation Society in 2008 and updated by Dr. Patrick Flanagan in 2016.

As of FY 2021, the National Resources Inventory (NRI) program houses a database of surface-level information about the non-Federal natural resources of the United States of America and provides the infrastructure and overall process to collect updated information about those resources. The information consists of characteristics of land, that which covers it, including water, and how it is used. The database is a longitudinal data set containing variables from 1982, 1987, 1992, 1997, and annually from 2000 through 2017. The variables consist of raw collected data, data derived from the raw data, estimates, and administrative data for a two-stage sample of geographic areas, called segments, and sample points on the ground within those segments. At this point, the NRI covers the 48 conterminous States, Hawaii, Puerto Rico, and the Virgin Islands for all of the aforementioned years and Alaska for 2007.

### NRI Goals and Objectives

The primary goal of the NRI will be to comply with the initial mandate from the Rural Development Act of 1972 that directed the Secretary of Agriculture “to carry out a land inventory and monitoring program to include, but not be limited to, studies and surveys of erosion and sediment damages, flood plain identification and utilization, land use changes and trends, and degradation of the environment resulting from improper use of soil, water, and related resource conditions.”

The primary objective of the NRI is to provide natural resource managers, policy makers, and the public with scientifically valid, timely, and relevant information on natural resources and the environment. The NRI is unique because of its established linkages to NRCS soil survey data. Information about specific properties and characteristics of the soil and surrounding landscapes is utilized to develop NRI data elements and interpretations.

NRCS operates the NRI program on the basis of rigorous, scientifically developed sample survey (statistical) principles and protocols. To that end, the NRI –

- utilizes the independent, objective expertise of internationally recognized experts in survey statistics via a cooperative agreement with the Center for Survey Statistics and Methodology (Iowa State University)
- utilizes probability sampling techniques to ensure that results are scientifically credible
- follows strict quality assurance protocols
- protects the integrity and confidentiality of the data collection
- provides databases and statistical summaries that allow data users to make statistically valid analyses and inferences

## The NRI Sample Design and Selection

### Target Universe

The NRI target universe is the land area of the United States of America and its territories, where land area includes land covered by anything including water. The exception is coastal territorial water. Portions of water along the coast are included in the target universe, but only to the extent that they have the potential to change to land area or become part of the estuarine system. Many large bays are included that are primarily interior to the coastline, e.g., Chesapeake, Delaware, San Francisco, and Mobile bays. Most gulfs are not included, e.g., Gulf of Maine. Islands off of the coast are included, but the water areas surrounding them are not. The Great Lakes and Saint Lawrence Seaway are treated the same way as the oceans. Since the NRI is a longitudinal data set, the Universe is the above over time from 1982 to the present at specific time intervals: 1982, 1987, 1992, 1997, and yearly 2000 – 2017.

### NRI Foundation Sample

The Foundation NRI sample is a two-stage stratified area sample of all States, Puerto Rico, and the Virgin Islands. The primary sampling units (PSUs) are areas of land called “segments.” The segments in the sample were selected from a collection of grids covering all land and water area in the target universe. Within the sample segments, points were selected in the geographically balanced random process described below. For most segments, three points are selected, but that varies to some degree dependent on the segment size. The foundation sample for 1997 contained 300,000 segments and about 800,000 points. See Nusser and Goebel (1997) for a more complete description of the survey. The samples each year from 2000 to 2017 are core and rotation subsamples of about 72,000 segments selected from the 1997 “foundation” sample. The annual sampling process is further described below.

### Selection of Sample Primary Sampling Units (PSUs)

The NRI evolved into a longitudinal data collection going back to the same sources of data over and over to get both cross-sectional data for each release and have the ability to compare the data over time to assess change at local levels. The sources of data for the 2017 NRI were almost entirely selected for the 1982 NRI, so the sampling details below reflect sample selection in 1982.

### The Sampling Frame

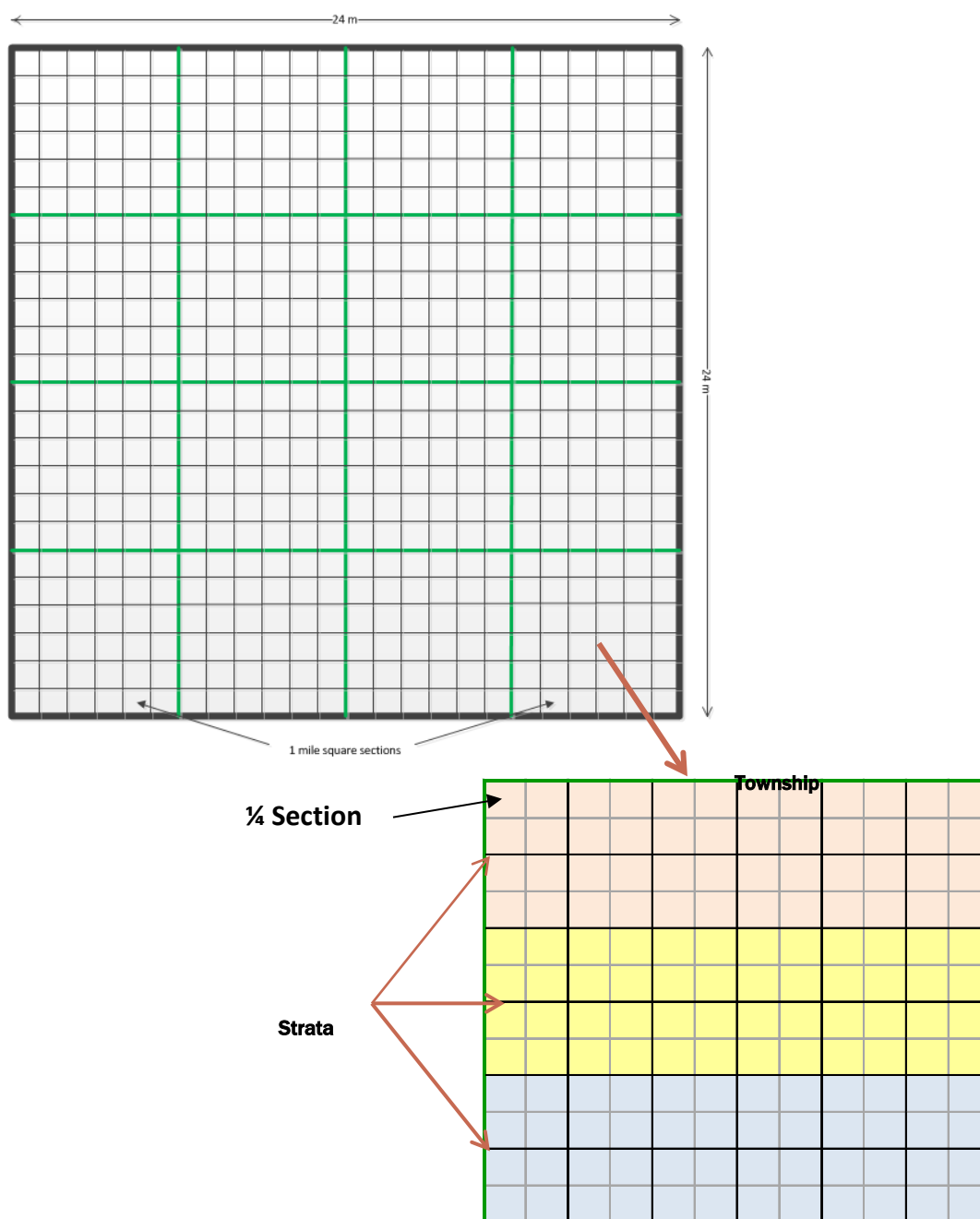
The surveys from which the NRI data is collected are entirely area frames with a two-stage selection, intending that data collection take place at each stage level. The first stage is a selection of primary sampling units (also called segments), which are subsets of each of the 3,100+ counties in the 48 contiguous states, Hawaii, Puerto Rico, and the Virgin Islands. To construct the first stage frame, each county is first divided up into non-overlapping portions ranging in size from 40 to 640 acres.

## Defining the PSUs and PSU Strata

Defining the PSUs, was done as follows:

Standard County. For those parts of the country defined by the Public Land Survey System (PLSS) and for a standard county that is square and 24 miles on each side, the county would be divided into 16 square townships, each 6 miles on a side. Each township is then divided into 36 sections, each one mile on a side. The sections are numbered from 1 to 36 starting in the Northeast corner and proceeding back and forth horizontally in a serpentine manner. For sampling, 3 strata of 12 sections are then formed in each township, with the two top rows being one stratum, the second two rows as the second stratum and the last two rows as the third stratum. Each of the sections is then divided into four PSUs, each  $\frac{1}{2}$  mile on a side. See diagrams below.

PLSS Standard County



PLSS Non-Standard Counties. In irregularly shaped (non-square) PLSS counties, as many regular (6 mile by 2 mile) strata are formed and then the remaining sections or partial sections are formed into 12 section groups.

PLSS Counties with Varying PSU/Segment Sizes. Due to the heterogeneity in some irrigated land and homogeneity in forest, range, and barren land in the west, some strata were constructed with differing PSU/segment sizes of as small as 40 acres up to 640 acres, though only 3 sizes were used (40, 160, and 640) beyond some variation due to non-square county borders.

Non-PLSS Counties in Ohio & Southern States. In Ohio, Louisiana, and Arkansas, these areas a grid pattern was superimposed on the county maps and then sampled similarly to PLSS counties.

Non-PLSS Counties in the 13 Northeast States. The strata in the 13 northeastern states are areas of land two minutes of latitude by four minutes of longitude in size. The PSUs are rectangular areas of land 20 seconds of latitude by 30 seconds of longitude. The PSUs range in size from 96 acres in northern Maine to 113 acres in southern Virginia.

### Original PSU Sample Selection Methods

Within each PSU stratum, PSUs (segments) were selected either using a simple random sample without replacement for strata with equal sized PSUs, or for strata with some differing in PSU size, they were sampled with probability proportional to size. Initially, a 2, 3, and 4 percent samples were selected. This was done to facilitate choices in sample reduction in some PSUs before making the final sample choices.

- In the simplistic case of a stratum with 48 equal sized PSUs, a 2 percent sample would be the selection of one PSU, while a 4 percent sample would be the selection of 2 PSUs.
- Within a township, a 3 percent sample was also selected by selecting 2 PSUs in one of the three strata, and 1 PSU from the other two.
- Other schemes were employed for non-standard counties to choose 2, 3, and 4 percent samples.
- The final sample of 300,000 PSUs for the 48 coterminous States, Hawaii, and the Caribbean territories (Puerto Rico and the Virgin Islands) was determined by a fixed budget estimate and sample choices that would minimize variance of key variables.

### Sample Changes and Sub-Sampling over Time

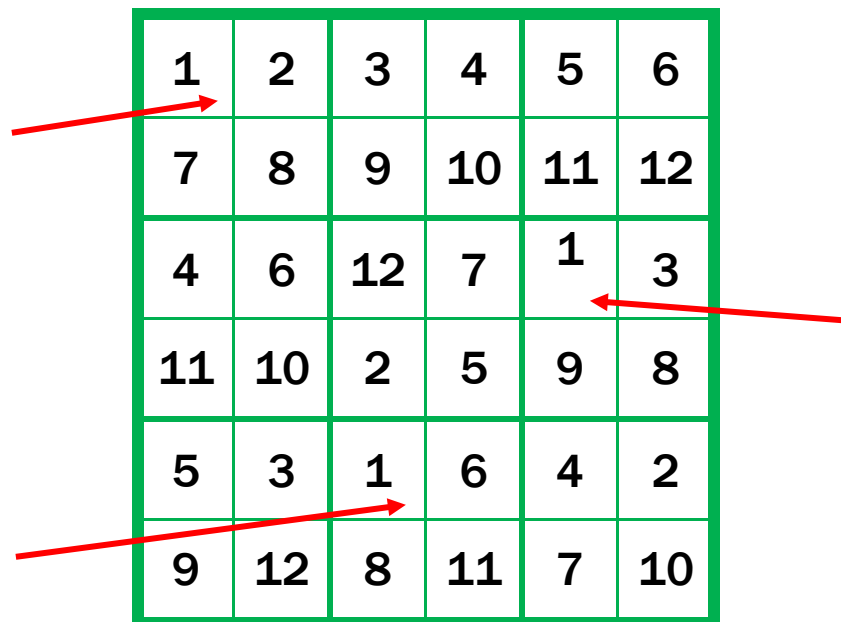
Between 1982 and 1992, the original sample of 320,000 PSUs in 1982, were reduced to 300,000 PSUs with some augmentation in selected counties where analysis showed a need for additional sample size.

### Selection of Sample Points in the Sample PSUs

The last step in selecting the sample was to locate three sample points within each PSU. There were exceptions-two points were selected from 40-acre PSUs and only one point was selected per PSU in Louisiana and northwestern Maine.

The procedure for selecting the points within a PSU was as follows:

1. A grid consisting of squares formed with three rows and three columns was superimposed on the PSU. Each square was subdivided into four equal blocks. The numbers 1 to 12 were assigned to the blocks in each row with a number appearing once in each row and once in each column. No adjoining blocks had the same number.
2. Two numbers between 1 and X were selected at random, where X is the width of the side of the PSU in feet. These two numbers determine the coordinates of sample point #1 in feet north and east from the PSU's southwest corner.
3. Points #2 and #3 were located in the blocks with the same label as the block for point #1. They were positioned in the same relative position within the blocks as point #1. Steps for selection of two sample points within a PSU were similar, except the PSU was divided into 4 blocks instead of 36.

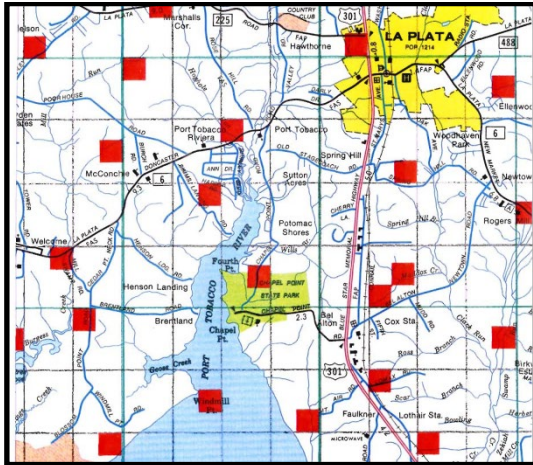


1	2	3	4	5	6
7	8	9	10	11	12
4	6	12	7	1	3
11	10	2	5	9	8
5	3	1	6	4	2
9	12	8	11	7	10

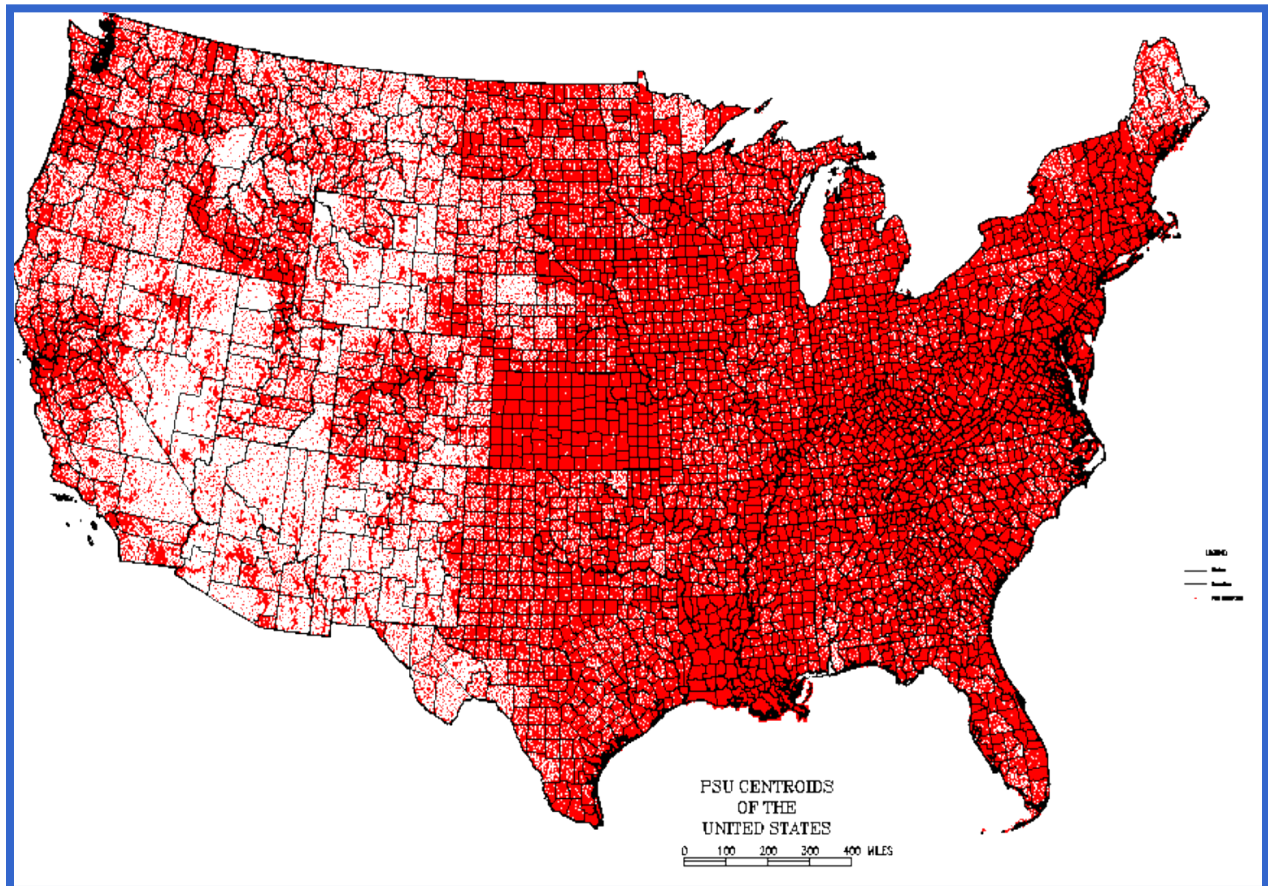


## Sample Results

The resulting PSU selection is a random area sample of segments in every county across the country. The localized idea is depicted below in the first illustration. Within each segment, the points are selected in a balanced fashion resulting in points as depicted in the second illustration.



The distribution of segments throughout the country is as follows:



## The 2000 and Later Annual Samples

Prior to 2000, in 1982, 1987, 1992, and 1997, the entire sample of about 300,000 segments and over 800,000 points were collected in one year. That presented a huge resource impact both in use of personnel and in the cost of the collection. In addition, such a large collection and the strain it puts on resources tends to have a negative impact on data quality. For all of those reasons, starting in 2000, the NRI Program changed to an annual sample approach. After extensive research documented in Breidt and Fuller (1999) a rotating panel design was shown to produce the best results, consisting of a fixed sample of “core” segments that are included in the collection every year, combined with a sample of “rotating” segments which rotate in and out of the annual sample over time.

The core sample of segments consists of just over 41,000 segments. To construct the core sample, segments were selected in every county using a stratified selection from the following strata:

- Wetland (contains one or more wetland point)
- CRP (contains one or more CRP points and no wetland points)
- Developed Land Change (not in above)
- Urban (Urban in segment, not in above)
- High Erosion (Not in above, but has high erosion cropland point)
- Cropland (not in above and has one or more cropland points)
- Pasture (not in above and has one or more pasture points)
- Range (not in above and has one or more range points)
- Forest (not in above and has one or more forest points)
- 100% Urban
- 100% Federal or Water
- Remainder

A similar approach is used annually to select the rotating panel of around 31,000 segments. Details of this entire process are provided in Fuller (2003).

The annual design was implemented as indicated from 2000 to 2003. After that, some variations were implemented by using some repeated rotation panels in their entirety.

## Appendix B

### Sampling Procedure for 2015-2016 National CEAP Survey

## 1 Frame

### 1.1 Eligible Land Cover/Uses

A point is included in the frame if the most recent collected land cover/use (LCU) satisfies one of the following conditions:

- $LCU > 0$  and  $LCU < 200$  and  $LCU \neq 7$
- $LCU$  in  $\{211, 212, 213\}$
- $LCU = 200$  and not range
- $LCU = 410$

The states included in the frame are the coterminous 48 states (not including DC). Aquaculture (171) is treated just like other cropland LCUs.

Points that are classified as urban or roads in 1997 and as 200-213 in the most recent collected year are not eligible. The last set of points removed contains 376 points, one of which is 213 in the most recent year and the rest of which are 200. These 376 points are removed because NRI editing procedures change the a collected LCU of 200-213 to urban.

Each point is classified into one of five mutually exclusive and exhaustive LCU groups. With “t” representing the most recent year, the LCU categories obtained from this program in item 2 are defined as follows:

- 1 =  $LCU(t)$  in 1-20 (high value specialty crops)
- 2 =  $LCU(t)$  in 141-144 and  $LCU(t-1)$ ,  $LCU(t-2)$ ,  $LCU(t-3)$  not in the set 11-116
- 3 =  $LCU(t)$  in 200-213 and  $LCU(t-1)$ ,  $LCU(t-2)$ ,  $LCU(t-3)$  not in the set 11-116
- 4 =  $LCU(t)$  in 21-116, 170, 171, 180, or  $LCU(t)$  in 141-144 and at least one of  $LCU(t-1)$ ,  $LCU(t-2)$ ,  $LCU(t-3)$  in the set 11-116, or  $LCU(t)$  in 200-213 and at least one of  $LCU(t-1)$ ,  $LCU(t-2)$ ,  $LCU(t-3)$  in the set 11-116
- 5 =  $LCU$  410 (CRP)

The combination of groups 1, 2, and 4 below approximates the NRI definition of cropland (cultivated and non-cultivated combined). Category 3 approximates the NRI definition of pasture. (A perfect classification of NRI points into broaduses is not possible with the data available because of the 2004 protocol change.) The five LCU categories are aggregated to the following three groups.

- Cropland (1): LCU category of 1, 2, 4

- Pasture (3): LCU category of 3
- CRP (4): LCU category of 5

The CRP category is labeled 4 for consistency with the original code of 410.

## 1.2 CEAP Regions

Ten CEAP production regions are defined for the 2015-2016 National survey. Figure 1 shows a map of the ten production regions. The specifications for the regions are outlined in Table 1.

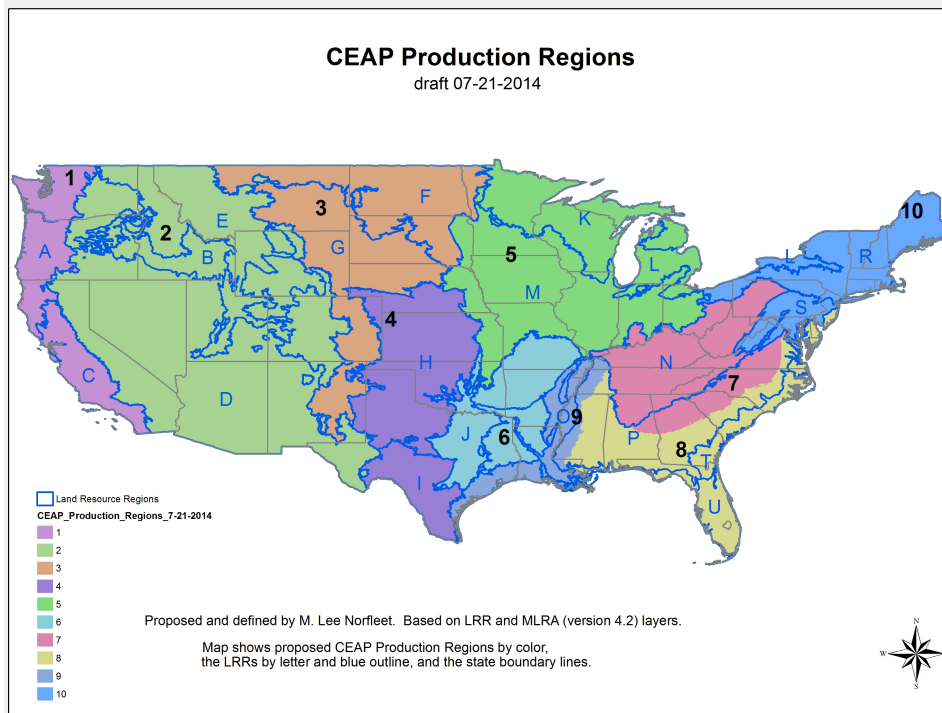


Figure 1: CEAP Production Regions

#	Working Name	LRR	Refinements	Concept or Rationale
1	Pacific Coast	A,C	None	High value irrigated crops. Water quantity issues
2	Irrigated West	B,D,E	None	Water quantity resource concern note the non Irr portions
3	Northern Plains Wheat Belt	F,G	Dryland wheat predominant Wind ero.	Dominated by Wheat and other small grains
4	Southern Plains Wheat/Cotton	H,I	Aquifer driven	Wheat cotton region, wind erosion and aquifer depletion
5	Corn Belt	K,L,M	Less 109 from LRR L	Corn soy dominated region
6	South Central Pasture and Crop	J,P, and western N	P west of MS River	Animal ag and manure on crop/pasture hay common
7	East Central pasture and Crop	eastern portion N	plus MLRA 136	Animal ag and manure on crop/pasture hay common
8	Southeast Mid Atlantic Coastal Plain	T, P, U	East of MLRA 134	Split from other for less rain intensity and coarser soil in general
9	Lower MS River and Texas Gulf	O, T and P	West with MLRA 134	Silty soils with cotton rice and cane. Intense rain
10	Northeast	R, S, + 101 from L		

Table 1: Description of 10 CEAP Production Regions.

Table 2 contains the average variance of APEX output variables from the 2003-2006 CEAP survey by CEAP region, where the average is across states for a particular region. For each APEX output variable, CEAP region 1 has the largest variance, which may not be surprising because this region is defined by high value irrigated crops in the Pacific Coast.

	7	8	6	9	1	2	3	4	10	5
Precips2	2.275	1.962	2.954	2.369	10.553	5.015	1.775	2.829	1.621	2.239
Runoffs2	2.335	2.055	2.324	2.702	8.379	2.419	0.409	0.508	1.525	1.223
Percs2	4.216	7.168	4.675	5.362	6.654	3.079	2.260	2.468	3.352	3.583
WaterYlds2	3.665	4.680	4.182	4.726	10.755	4.495	2.302	2.540	2.843	2.684
RUSLE2s2	0.826	0.837	1.596	2.711	0.498	0.144	0.036	0.222	0.436	0.463
Sediments2	3.199	1.887	1.590	4.205	6.466	3.086	0.130	0.226	2.142	1.207
WindEros2	0.019	0.099	0.616	0.988	0.211	4.399	3.117	5.034	0.066	0.571
TNlosss2	50.901	53.836	33.010	43.503	150.861	95.705	24.433	50.796	90.358	29.160
Nrunoffs2	2.859	1.385	1.782	2.555	9.442	1.840	0.468	0.244	1.072	1.165
Nseds2	13.379	4.194	5.187	10.978	13.102	9.335	9.317	10.536	10.415	8.410
PercNs2	25.675	30.894	19.855	23.172	80.382	51.439	11.385	28.414	47.680	15.609
Plosss2	6.196	3.013	2.016	4.879	13.544	8.850	2.865	4.424	7.907	3.086
solublePs2	3.476	2.143	0.885	2.915	6.348	0.680	0.346	0.175	4.393	1.458
sed_Ps2	3.809	1.464	1.635	3.321	9.037	8.602	2.771	4.387	4.466	2.222
C_starts2	8.168	10.352	7.571	12.101	49.092	11.152	8.347	8.854	11.516	25.907

Table 2: Average variance of APEX output variables by CEAP region.

### 1.3 Programs

Four programs construct the frame:

1. “extractlatestlcu3check1.R” obtains the most recent collected landuse and cropping history and checks the range indicator for points with latest LCU 200. The program also adds segment-specific information from PSU, SAMPLED.SEGMENT, and SDE including foundation selection probabilities, a core indicator, and location.
2. “addcolstoframecheckrev2.R” adds additional columns including indicators for whether the point was in the 03-06 CEAP or the 2013-2014 CEAP surveys. The program also adds an LCU category and obtains CEAP regions using an R overlay operation.
3. “mergeinceapregionsnewRev” obtains CEAP regions from an ArcGIS overlay operation. These CEAP regions are used for the sample selection.
4. “createcrdatRev” reorganizes the data by the CEAP regions obtained in step 3 instead of by state. This program also removes points in LA that are not in the pointgen, removes points that have been removed, and removes points that are classified as urban or roads in 1997 and as 200-213 in the most recent collected year. This programs also aggregates the five LCU categories into the three LCU groups.

## 2 Strata Definitions and Sample Sizes

### 2.1 Sample Sizes by CEAP Region

Let  $\hat{A}_h$  be the estimate of the area in CEAP-eligible categories in CEAP region  $h$  in the year 2010. The estimated areas are obtained from the 2010 pointgen. Due to topological errors in the CEAP region shapefile, not all NRI points are in a CEAP region. The total area corresponding to points classified in eligible LCUs that are not in a CEAP region is 83,800 acres. The total estimated area in CEAP-eligible categories is 480,436,800 acres. Because the area associated with missing CEAP regions is only 0.017% of the estimated area based on points located in a CEAP region polygon, the area with missing CEAP regions is ignored for the sample size calculation.

Define a target sample size for CEAP region  $h$  by  $n_h$ , the result of accumulating and rounding the  $\tilde{n}_h$ , where

$$\tilde{n}_h = \frac{N \hat{A}_h^{0.5}}{\sum_{h=1}^H \hat{A}_h^{0.5}}, \quad (1)$$

and  $N = 45,000$ . This yields the sample sizes in Table 3 below. The square root allocation is often used as a compromise between equal allocation, optimal for individual area estimates, and proportional allocation, optimal for the total of the regions combined (Bankier, 1988).

CEAP.Reg	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
$n_h$	2420	3113	3881	5733	5651	8701	4070	4120	3611	3700

Table 3: Target sample sizes by CEAP region.

The program to calculate the estimated areas and the target sample sizes is “estimatedareafrompgenoverlay1rev.R.”

### 2.1.1 Increasing Sample Size to Account for CRP

We obtain a revised sample size to account for differential eligibility rates across crop, pasture, and CRP. Assume the eligibility rate for CRP is 15%, and the eligibility rate for crop and pasture is 85%. Assume the response rate is 70% for all three domains. If the sample size is 45000, and the sample is drawn exclusively from pasture and cropland, then the expected realized sample size is,

$$\tilde{n} = 45000(0.85)(0.7) = 26675. \quad (2)$$

In the frame, approximately 6% of the segments contain at least one CRP point. If a sample of size 45000 is drawn and 6% of the points are CRP, then the expected realized sample size is

$$\tilde{n} = 45000(0.7)[0.06(0.15) + 0.94(0.85)] = 25452. \quad (3)$$

If the sample size is increased from 45000 to 47500, then the expected realized sample size is,

$$\tilde{n} = 47500(0.7)[0.06(0.15) + 0.94(0.85)] = 26866. \quad (4)$$

The CEAP region sample sizes for  $N = 47,500$  are provided in Table 4 below.

CEAP.Reg	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
$n_h$	2554	3286	4097	6051	5965	9185	4296	4349	3811	3906

Table 4: Target sample sizes by CEAP region.

### 2.1.2 Modifications for points in 2013 or 2014 CEAP surveys

It was decided that the data from the 2013 and 2014 CEAP surveys are to be used for points that are in the previous two regional survey samples and are selected for the national survey. Because this reduces the overall workload, cost constraints permit a larger total sample size. To determine the number of additional points to sample, the basic sampling procedure, described below, was implemented 30 times. The number of points in either the 2013 or 2014 sample was recorded in each implementation. The total sample size was increased by the median number of points obtained in a selected sample

that are also in either the 2013 or 2014 sample. The median was 1166, and the modified sample sizes are given in Table 5 below.

CEAP.Reg	1	10	2	3	4	5	6	7	8	9
$n_h$	2617	3366	4198	6199	6112	9410	4402	4456	3904	4002

Table 5: Modified CEAP region sample sizes.

## 2.2 Point Sample Sizes

The target sample size at the level of CEAP region and broaduse is converted to a target sample size for the intersection of CEAP region, sample class, LCU group, and state on the basis of the number of points in the frame. A point is classified into one of the three “LCU groups” defined in item 4 of Section 1 on the basis of the most recent collected land cover/use and associated cropping history.

We let  $N_{hbkl}$  be the number of points in the frame in CEAP region  $h$ , LCU group  $b$ , sample class  $k$ , and state  $\ell$ . The sample classes 8-12 are aggregated to a single sample class denoted by 0. The sample size  $n_h$  is split among the domains defined by intersections of  $(h, b, k, \ell)$  according to the  $N_{hbkl}$ . For CEAP regions other than regions 6 and 7, the allocation is in proportion to  $N_{hbkl}$ . For each of CEAP regions 6 and 7, the sample size is first distributed to LCU groups in proportion to the square root of  $N_{hb}$  and then to states and sample classes in proportion to  $N_{hbkl}$ . The allocation for regions 6 and 7 is square root instead of proportional at the first level because these are the only two regions where the estimated area in pasture exceeds the estimates area in cropland. (Estimates are based on the 2010 pointgen for the year 2010.) Specifically, let

$$\tilde{n}_{hbkl} = \frac{n_h N_{hbkl}}{\sum_{b,k,\ell} N_{hbkl}}, \quad h \neq 6, 7 \quad (5)$$

$$= \frac{n_h N_{hb}^{0.5} N_{hbkl}}{\sum_b N_{hb}^{0.5} \sum_{k,\ell} N_{hbkl}}, \quad h = 6, 7. \quad (6)$$

The target sample size  $n_{hbkl}$  is obtained by applying accumulate and round to the  $\tilde{n}_{hbkl}$  within a CEAP region.

### 2.2.1 Modifications to Point Sample Sizes for Regions 6 and 7

For the CRP category for regions 6 and 7, the target point sample size exceeds the number of available segments. The target point sample sizes were modified for regions 6 and 7 as follows. The modified  $n_{hbkl}$  is obtained by applying accumulate and round to  $\tilde{n}_{hbkl}$  defined,

$$\tilde{n}_{hbkl} = \frac{n_{hb}^* N_{hbkl}}{\sum_{k,\ell} N_{hbkl}}, \quad (7)$$



where  $c_6 = 165$ ,  $c_7 = 61$ ,

$$\begin{aligned} n_{hb}^* &= n_{hb}^{(0)} \left( 1 + \frac{c_h}{n_{h1}^{(0)} + n_{h2}^{(0)}} \right), \quad b = 1, 3 \\ &= n_{hb}^{(0)} - c_h, \quad b = 4, \end{aligned} \quad (8)$$

and

$$n_{hb}^{(0)} = \frac{n_h N_{hb}^{0.5}}{\sum_b N_{hb}^{0.5}}. \quad (9)$$

### 2.2.2 Resulting Frame Counts and Target Point Sample Sizes

	BU	SC	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
1	1	0	50	11	208	98	76	45	13	23	18	51
2	1	1	444	1716	890	3292	746	7352	126	611	2595	2425
3	1	2	10	83	641	1819	1519	5823	82	354	313	451
4	1	3	534	2227	562	156	489	3662	182	1640	1113	509
5	1	4	258	1607	443	322	564	5098	134	839	708	560
6	1	5	141	1092	3840	6185	4791	13606	338	1024	579	1324
7	1	6	3282	2457	7930	12060	12536	34364	1253	3370	3258	6078
8	1	7	60	49	110	66	9	140	82	185	74	32
9	3	0	47	42	90	50	35	84	136	190	130	77
10	3	1	343	1197	1117	531	110	2735	870	752	1605	821
11	3	2	10	26	283	503	903	1722	132	278	245	162
12	3	3	291	1187	340	70	169	1209	905	3010	855	303
13	3	4	147	813	194	61	142	1072	786	1375	423	275
14	3	5	27	209	180	293	384	1368	241	519	136	164
15	3	6	348	899	1660	1247	1113	4380	1933	2914	1010	809
16	3	7	256	367	746	356	294	1368	3830	2181	700	347
17	4	0	0	0	4	8	1	0	0	1	1	2
18	4	1	2	20	17	433	80	351	3	17	150	141
19	4	2	12	20	742	1278	1900	2224	36	172	320	263
20	4	3	0	18	9	0	9	59	1	17	22	6
21	4	4	0	4	3	10	19	54	1	18	20	12
22	4	5	0	12	239	369	354	377	1	19	35	45
23	4	6	3	21	247	493	539	427	12	59	81	75
24	4	7	0	0	4	2	0	8	1	1	2	2

Table 6: Frame point counts by CEAP region, LCU Group (BU), and sample class (SC).

	BU	SC	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
25	1	0	21	2	42	21	16	3	9	5	5	14
26	1	1	186	409	185	687	169	789	83	145	706	648
27	1	2	4	21	131	379	349	625	54	82	84	122
28	1	3	223	533	115	32	109	393	118	389	302	136
29	1	4	107	386	90	67	128	546	88	198	190	152
30	1	5	59	262	787	1292	1093	1463	223	243	155	355
31	1	6	1371	586	1625	2516	2862	3696	821	799	881	1626
32	1	7	24	11	24	14	3	15	53	43	19	9
33	3	0	20	11	18	10	9	9	43	41	34	22
34	3	1	142	285	229	110	26	296	286	147	436	220
35	3	2	5	5	58	105	204	186	43	57	66	44
36	3	3	122	283	68	16	39	133	298	603	231	82
37	3	4	61	192	38	13	33	117	258	276	115	74
38	3	5	12	51	38	59	87	147	78	103	39	44
39	3	6	145	216	338	261	254	471	633	586	276	219
40	3	7	107	88	152	75	66	147	1259	435	190	93
41	4	0	0	0	0	3	0	0	0	1	0	1
42	4	1	1	5	3	90	19	38	3	17	39	37
43	4	2	5	6	152	267	435	238	36	172	87	70
44	4	3	0	5	4	0	2	5	1	17	9	0
45	4	4	0	1	1	1	5	5	1	18	6	2
46	4	5	0	3	49	79	81	40	1	19	10	13
47	4	6	2	5	50	102	123	46	12	59	24	19
48	4	7	0	0	1	0	0	2	1	1	0	0

Table 7: Target point sample sizes by CEAP region, LCU Group (BU), and sample class (SC).

### 2.3 Segment Sample Sizes

The target sample size refers to a sample of points, but the selection procedure obtains a sample of segments. One point is to be selected from each sampled segment. The segments are classified into groups on the basis of the LCUs of the in-scope points in a segment. Table 8 gives the possible combinations. Table 9 gives the number of segments in each category  $S$  by CEAP region. No segment has more than three eligible points.

$S$	CRP	Crop	Pasture	Group
1	3	0	0	1
2	2	1	0	1
3	2	0	1	1
4	0	3	0	2
5	1	2	0	1
6	0	2	1	2
7	0	0	3	3
8	1	0	2	1
9	0	1	2	3
10	1	1	1	1
11	2	0	0	1
12	1	1	0	1
13	1	0	1	1
14	0	2	0	2
15	0	1	1	3
16	0	0	2	3
17	1	0	0	1
18	0	1	0	2
19	0	0	1	3

Table 8: Segment group based on number of points classified as crop, pasture, or CRP

$S$	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
1	2	1	194	444	460	176	5	9	35	38
2	1	3	46	111	112	250	3	8	10	35
3	2	1	22	50	43	124	2	8	12	7
4	983	774	2340	5155	4428	14016	209	536	815	2358
5	1	11	65	183	225	541	4	15	37	63
6	117	395	407	449	379	2791	176	476	216	232
7	87	100	368	228	240	644	893	563	283	251
8	1	1	10	16	28	117	5	10	15	11
9	56	242	206	146	167	1013	201	523	102	122
10	1	10	9	23	25	244	4	17	14	13
11		4	137	203	274	198	2	29	59	55
12	1	16	47	102	97	444	3	37	83	44
13		9	17	30	32	192	6	39	48	34
14	500	1388	2166	2404	2045	6711	218	871	1369	1287
15	143	844	451	351	289	2110	258	1164	448	225
16	185	478	611	308	336	1052	1576	1572	702	381
17	1	29	125	179	257	290	4	69	167	73
18	392	2217	1569	1728	1458	3895	318	1965	2312	753
19	461	1739	946	584	600	2181	2144	3616	1879	666

Table 9: Number of segments in each group  $S$  by CEAP region

### 3 Segment Sample Sizes

To define a procedure for selecting a sample of segments, let  $d$  denote the segment group. Let  $M_{hdk\ell}$  and  $m_{hdk\ell}$ , respectively, be the number of segments in the frame and sample for CEAP region  $h$ , aggregated sample class  $k$ , and CEAP region  $\ell$ , where

$$\tilde{m}_{hdk\ell} = \frac{m_h M_{hdk\ell}}{\sum_{d,k,\ell} M_{hdk\ell}}, \quad h \neq 6, 7 \quad (10)$$

$$= \frac{m_{hd}^* M_{hdk\ell}}{\sum_{k,\ell} M_{hdk\ell}}, \quad h = 6, 7, \quad (11)$$

where  $(m_{h1}^*, m_{h3}^*, m_{h4}^*) = (921, 3443, 38)$  for  $h = 6$ , and  $(m_{h1}^*, m_{h3}^*, m_{h4}^*) = (1763, 2452, 241)$  for  $h = 7$ . The target sample size  $m_{hdk\ell}$  is obtained by applying accumulate and round to the  $\tilde{m}_{hdk\ell}$  within a CEAP region.

	BU	SC	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
1	1	0	37	7	116	58	43	25	8	20	12	31
2	1	1	201	970	426	1440	346	3496	49	310	1774	1106
3	1	2	4	39	216	630	460	1800	26	131	145	143
4	1	3	267	1187	295	74	236	1750	88	860	613	223
5	1	4	125	858	220	151	254	2279	64	430	373	247
6	1	5	59	485	1538	2299	1827	5034	135	451	259	500
7	1	6	1273	1199	3616	5048	5138	12968	518	1546	1505	2365
8	1	7	26	29	55	36	6	61	33	100	31	15
9	3	0	30	39	68	30	24	63	118	150	95	63
10	3	1	229	892	602	282	64	1724	531	540	1192	471
11	3	2	2	13	116	213	380	547	54	143	130	65
12	3	3	214	912	240	44	120	846	596	2149	625	200
13	3	4	109	599	127	39	95	687	478	928	290	166
14	3	5	15	106	75	108	169	433	118	297	70	80
15	3	6	173	577	909	677	592	1934	1001	1738	597	397
16	3	7	160	265	445	224	188	766	2176	1493	415	203
17	4	0	0	0	2	5	1	0	0	1	1	2
18	4	1	2	19	13	267	42	294	2	17	125	92
19	4	2	6	18	388	617	1001	1555	24	130	220	176
20	4	3	0	17	5	0	4	49	1	14	18	6
21	4	4	0	4	2	7	14	48	1	17	19	10
22	4	5	0	11	125	183	194	279	1	14	28	34
23	4	6	2	16	133	260	297	345	8	47	68	51
24	4	7	0	0	4	2	0	6	1	1	1	2

Table 10: Frame segment counts by Group (BU), sample class (SC), and CEAP region (R1-R10)

	BU	SC	R1	R10	R2	R3	R4	R5	R6	R7	R8	R9
25	1	0	33	3	50	29	24	6	8	9	6	19
26	1	1	180	393	184	703	183	890	49	144	807	664
27	1	2	3	14	92	307	244	458	26	58	67	86
28	1	3	238	486	126	36	125	444	88	397	280	134
29	1	4	112	351	96	73	134	578	64	195	169	148
30	1	5	53	197	662	1123	972	1279	135	208	120	301
31	1	6	1135	488	1559	2466	2732	3298	518	708	684	1426
32	1	7	23	14	24	18	3	15	33	44	14	8
33	3	0	26	16	29	15	13	15	81	50	43	39
34	3	1	204	364	259	137	34	438	359	179	540	285
35	3	2	2	6	49	105	202	137	37	47	58	40
36	3	3	191	373	104	22	63	216	405	708	284	120
37	3	4	96	242	54	19	52	175	325	306	132	100
38	3	5	13	44	36	53	92	109	81	97	31	48
39	3	6	155	234	392	330	315	494	678	574	270	238
40	3	7	143	109	191	109	99	194	1477	491	187	123
41	4	0	0	0	0	2	0	0	0	1	0	1
42	4	1	2	8	6	132	24	77	2	17	56	54
43	4	2	6	6	170	301	533	399	24	130	99	106
44	4	3	0	6	1	0	2	13	1	14	6	4
45	4	4	0	2	1	3	6	14	1	17	8	6
46	4	5	0	4	52	90	101	72	1	14	12	19
47	4	6	2	6	58	126	159	88	8	47	31	32
48	4	7	0	0	3	0	0	1	1	1	0	1

Table 11: Segment sample sizes by Group (BU), sample class (SC), and CEAP region (R1-R10)

## 4 Selection of Segments and Points

This section describes the procedure to select a sample of segments and then select one point per segment. The procedure is guided by the following principles:

1. Attempt to attain geographic spread.
2. Retain points that provided usable data for the 2003-2006 CEAP survey..
3. Select one point per segment.
4. Avoid high variation in weights within groups defined above.
5. Avoid selecting core segments where possible.

## 4.1 Combination of One Per Stratum and Two Per Stratum Sampling

Each stratum (defined by the combination  $(h, b, k, \ell)$ ) is subdivided into smaller groups on the basis of geocode. A combination of one per stratum and two per stratum sampling is used to select a sample of PSUs from each group. This sampling procedure is intended to achieve geographic spread, similar to systematic sampling, but reduce the probability of undesirable samples that may arise if the geocode reflects a grid pattern in the underlying population. Selecting two per stratum for a sub-sample of the strata provides degrees of freedom for variance estimation. We first describe the general procedure and then explain an application to the 2015/2016 CEAP sample.

### 4.1.1 General Methodology

Let  $g$  denote the  $(h, b, k, \ell)$  used at the second stage of the allocation. Let  $N_g$  and  $n_g$  be the number of segments in the frame and sample, respectively, for group  $g$ . Let  $s_1 < \dots < s_{N_g}$  be the ranks of the geocodes of the elements in the frame for group  $g$ . (This maps the geocodes onto a line, where the segments in the frame are separated by equal distances.) Let  $N_g n_g^{-1} = k$ . Randomly select a starting point  $r \in [1, N_g]$ . For any segment  $j$  in the frame such that  $s_j < r$ , define  $s_j^* = s_j + s_{N_g}$ . For segments  $j$  ( $j = 1, \dots, s_{N_g}$ ) such that  $s_j \geq r$ , let  $s_j^* = s_j$ . Define stratum  $d$  by

$$U_d = \{j : r + k(d-1) \leq s_j^* < r + kd\}, \quad d = 1, \dots, n_g. \quad (12)$$

Randomly select  $d^*$  from  $d = 1, \dots, n_g$ .

- If  $d^* > 1$ , randomly select two elements from  $U_{d^*}$ . Randomly select 1 element from  $U_d$  for  $d \notin \{d^*, d^* - 1\}$ .
- If  $d^* = 1$ , randomly select two elements from  $U_{d^*}$ . Randomly select 1 element from  $U_d$  for  $d \notin \{d^*, d^* + 1\}$ .

This procedure is called the one-two sampling procedure below.

### 4.1.2 Application to CEAP

Let  $N_g$  and  $n_g$  be defined as above. Let  $N_{g,pc}$  be the number of segments in group  $g$  containing at least one prior CEAP point. Let  $N_{g,r}$  and  $N_{g,c}$ , respectively, be the number of rotation and core segments in group  $g$ . Consider several cases:

- Case 1:  $n_g - N_{g,pc} \leq 0$ . Use one-two sampling to select  $n_g$  of the  $N_{g,pc}$  segments in the sample for group  $g$ . Move to the next group.
- Case 2:  $m_g = n_g - N_{g,pc} > 0$  and  $N_{g,r} \geq m_g$ . Use one-two sampling to select  $m_g$  segments from the  $N_{g,r}$  rotation segments in group  $g$ .
- Case 3:  $m_g = n_g - N_{g,pc} > 0$  and  $N_{g,r} < m_g$ . Include all  $N_{g,r}$  segments in group  $g$ . If  $N_{g,c} \geq m_g - N_{g,r}$ , then use one-two sampling procedure to select a sample of size  $m_g - N_{g,r}$  from the  $N_{g,c}$  core segments. Otherwise, include all  $N_{g,c}$  segments, and print a warning message.

#### 4.1.3 Selection of Points within Segments

A simple procedure is used to select points within segments. If a segment contains a prior CEAP point, then one point is selected at random from the prior CEAP points in the segment. If the segment contains no prior CEAP points, one point is selected at random from the set of eligible points.

## 5 Summary of Realized Sample

Figure 2 compares realized point sizes (vertical axis) to target point sizes (horizontal axis) for intersections of CEAP regions, sample classes, states, and LCU groups. Each CEAP region is plotted separately. Table 12 contains realized and target point sample sizes at the level of CEAP region and LCU group. The total sample size is 48666, and 1228 points are in either the 2013 or 2014 CEAP surveys, leaving 47438 points that require data collection.



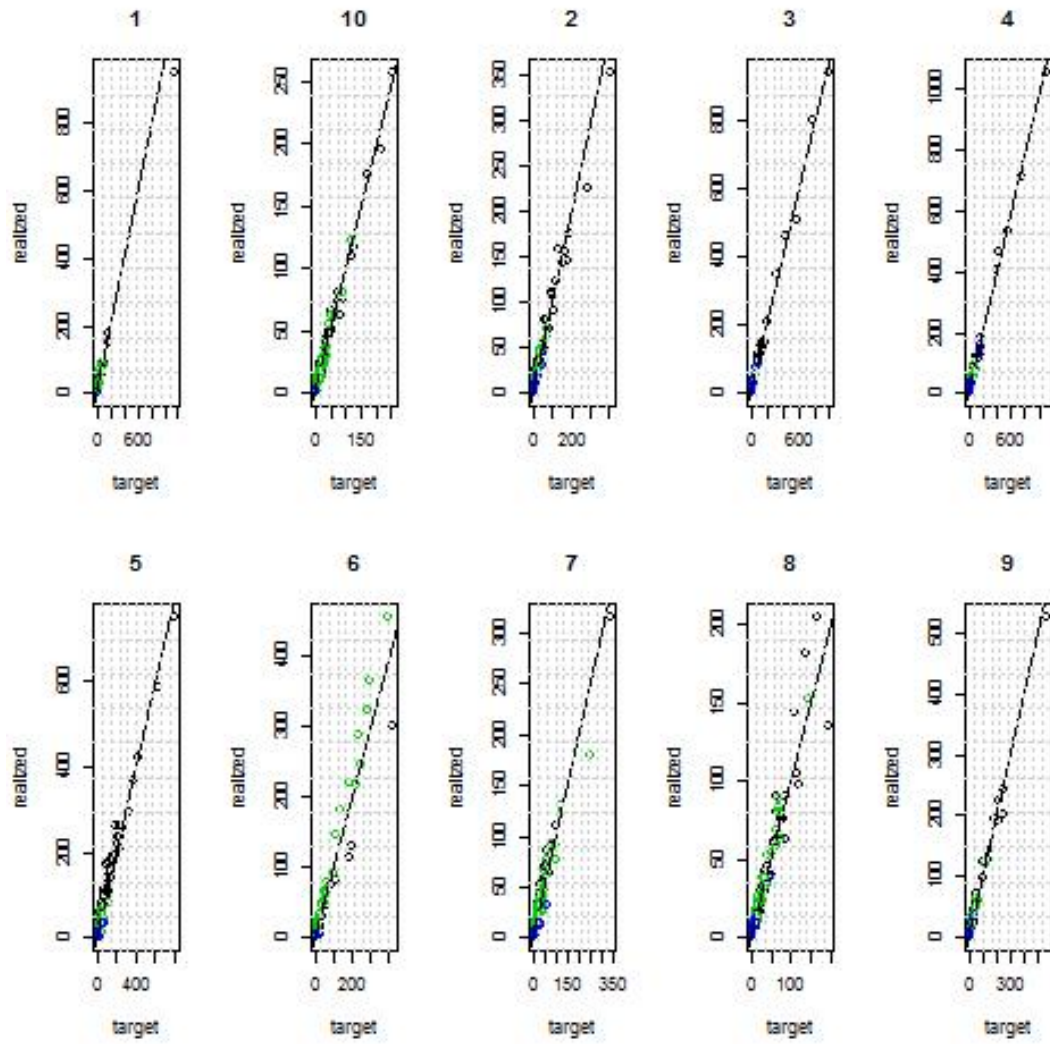


Figure 2: Realized (y-axis) and target (x-axis) point sample sizes for 10 CEAP regions. Blue = CRP, green = pasture, and black = cropland.

CEAP Region	LCU Group	Realized	Target
1	1	1837	1995
1	3	775	614
1	4	5	8
10	1	2208	2210
10	3	1140	1131
10	4	18	25
2	1	2986	2999
2	3	1001	939
2	4	211	260
3	1	4988	5008
3	3	712	649
3	4	499	542
4	1	4757	4729
4	3	774	718
4	4	581	665
5	1	8135	7530
5	3	1081	1506
5	4	194	374
6	1	1048	1449
6	3	3333	2898
6	4	21	55
7	1	2105	1904
7	3	2189	2248
7	4	162	304
8	1	2319	2342
8	3	1432	1387
8	4	153	175
9	1	2969	3062
9	3	889	798
9	4	144	142

Table 12: Realized and target point sample sizes for combinations of CEAP region and LCU group (1=crop, 3=pasture,4=CRP).

The core can be used in two cases. The first case is if a core segment is a prior CEAP segment. The second case is if there are not enough rotation segments to cover the target sample size. The table below summarizes the use of the core for the selected sample:

	CEAP Region	Prior CEAP	Count
2	1	0	200
4	1	1	83
6	10	0	4
8	10	1	337
10	2	0	1
12	2	1	375
14	3	0	5
16	3	1	665
18	4	0	2
20	4	1	861
22	5	0	1
24	5	1	2518
26	6	0	114
28	6	1	71
30	7	0	50
32	7	1	298
34	8	0	6
36	8	1	472
38	9	0	16
40	9	1	711

Table 13: Counts of points in selected sample that are in the core. Prior CEAP = 1 if point is a prior CEAP point and zero otherwise.

## 6 Division of Sample between Data Collection Years 2015 and 2016

- Begin by assigning the 2015 sample to be the subset of the selected sample with most recent observed year of 2008 or 2010-2012 that are not in the core.
- The remaining points in the sample are assigned to be sampled in 2016.
- Re-adjust the 2015 and 2016 sample sizes so that the sample sizes are approximately equal for the two years for each state. This is accomplished by completing the following two steps for each state. In the following, the notation  $[a]$  is the closest integer to  $a$ . If the original 2015 sample size is equal to  $n_{16} + k$ , where  $k > 0$ , then select  $[k/2]$  points from the 2015 sample for the state randomly and assign the  $[k/2]$  points to the 2016 sample. If the original 2016 sample size is equal to  $n_{15} + k$ , where  $k > 0$ , then let  $r = \min\{N_{16,03}, [k/2]\}$ , where  $N_{16,03}$  is the number of points in the 2016 sample for the state last observed in 2003. Select  $r$  points from the subset of the 2016 sample for the state that are not last observed in 2003, and randomly and assign the  $r$  points to the 2015 sample.

## 7 Programs to Compute Target Sample Sizes and Select Sample

1. `estimatedareafrompgenoverlay1rev.R`: estimated area of CEAP-eligible cropland by CEAP region
2. `SampleSelectCheck1.R`: Compute target sample sizes by CEAP region and point-level target sample sizes. Prepare data sets at the PSU level.
3. `SampleSelectCheck2.R`: Complete sample selection.
4. `evaluatesample1Rev.R`: Tabular comparisons and maps.

## Reference

- Bankier, M. D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42(3), 174-177.