David M. Johnson
Geographer

United States Department of Agriculture
National Agricultural Statistics Service
Research and Development Division
Spatial Analysis Research Section

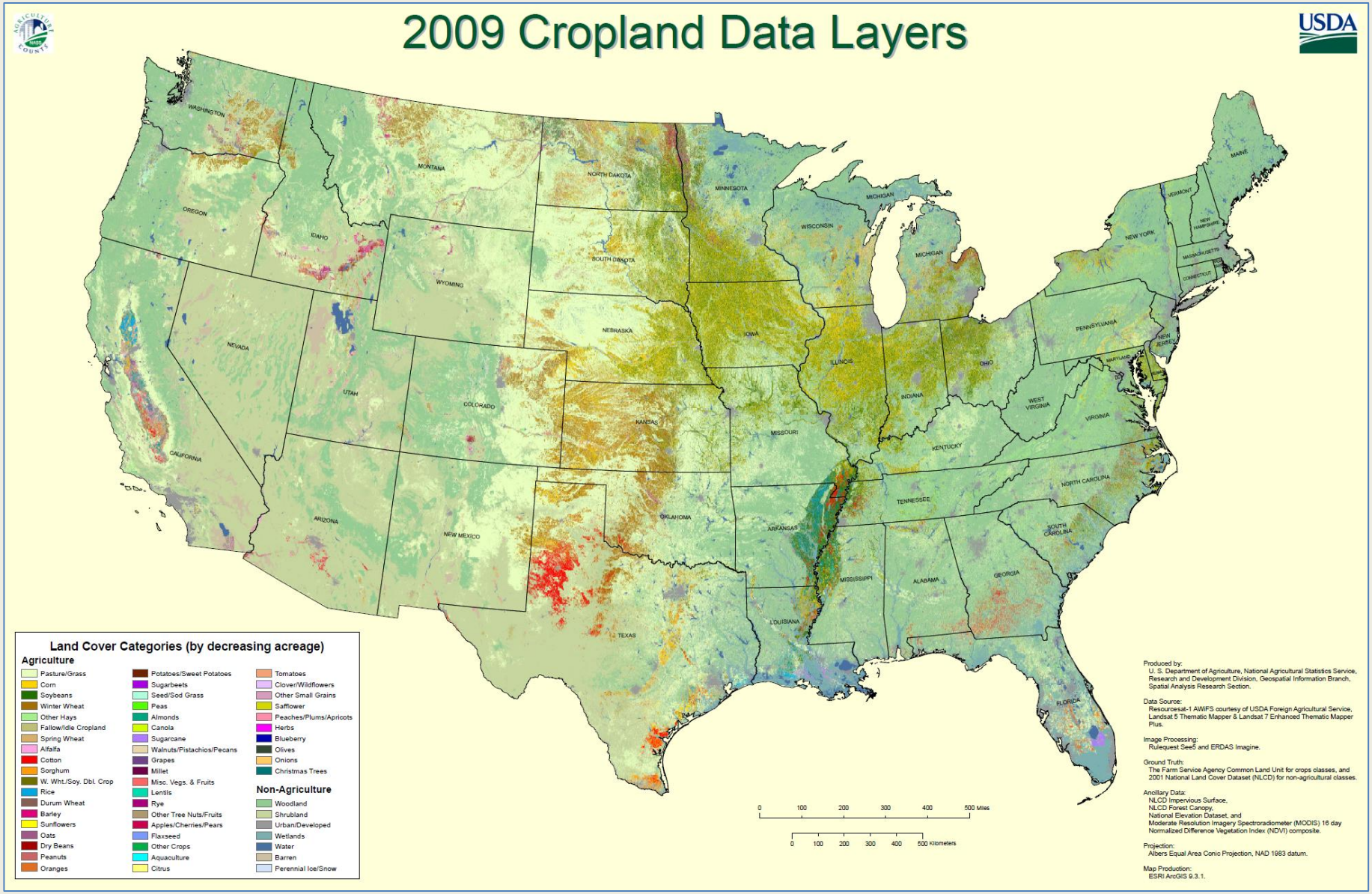
Cropland Classification Accuracy as a Function of Training Data Accuracy

Association of American Geographers, 2010 Annual Meeting

Study Overview

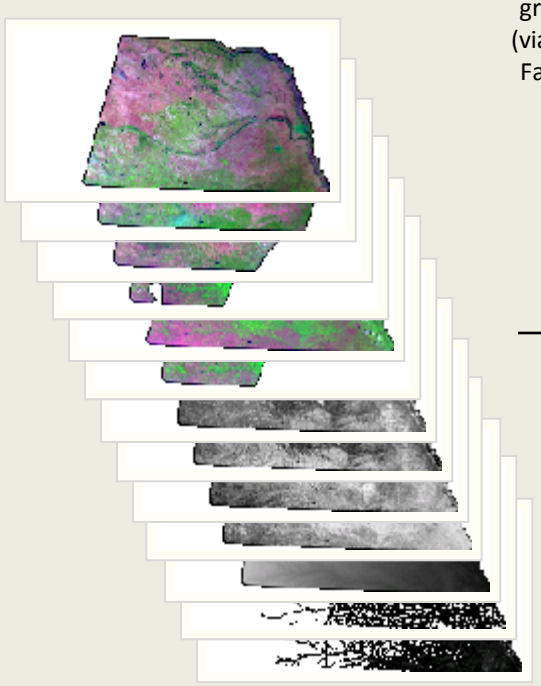
- Want to understand how potential errors in training data impact decision-tree based land cover classification
  - Especially tailored to mapping efforts within NASS
  - Primarily in regions dominated by common commodity crops
- Hypothesis : Classification accuracy decreases as training data accuracy decreases
  - By how much?
  - Is there a threshold?
  - What's the relationship?
  - Is it linear?
  - Are there scenarios where it improves the outcome?
- Chose 3 states to test these questions
  - Iowa
  - Idaho
  - North Dakota

Operational land cover mapping within NASS



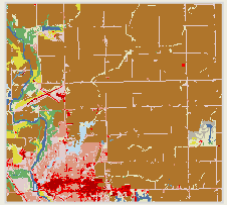
Classification Methodology Overview

1) "Stack" AWiFS, TM, MODIS, and ancillary data layers within a raster GIS
   - 56 m grid cells, Albers Conic Equal Area projection, common extent by state
   - some compromised imagery (from clouds, haze, data gaps, etc.) is acceptable
2) Sample spatially from stack within known ground truth from FSA (ag. categories) and NLCD (non-ag. categories)
   - a heavy sample rate (100s of thousands) at the pixel level is employed
3) "Data-mine" samples using Boosted Classification Tree Analysis to derive best fitting decision rules
   - implemented with Rulequest See5.0, interfaced with ERDAS Imagine with the "NLCD Mapping Tool"
4) Create land cover map by applying derived decision rules back to input data stack



Imagery stack
(independent data)

Agricultural ground truth (via the USDA Farm Service Agency)

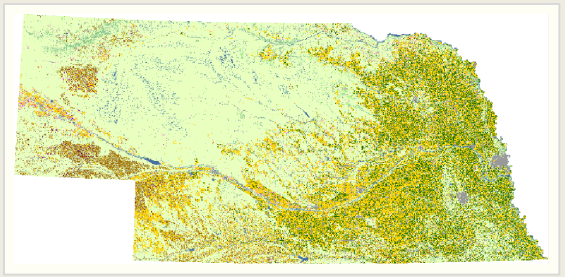


Non-agricultural ground truth (using the National Land Cover Dataset as a proxy)

(dependent data)

Manages and visualizes datasets

ERDAS IMAGINE®



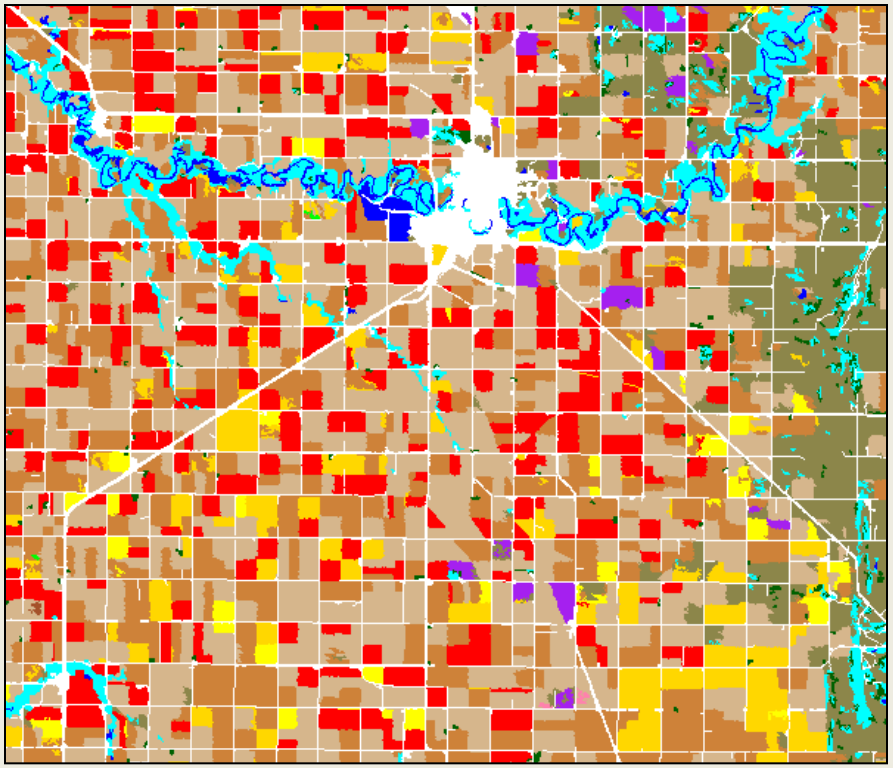
Output "Cropland Data Layer"

Rulequest See5.0

Derives decision tree-based classification rules

Generated rule set

Example Classification Subset

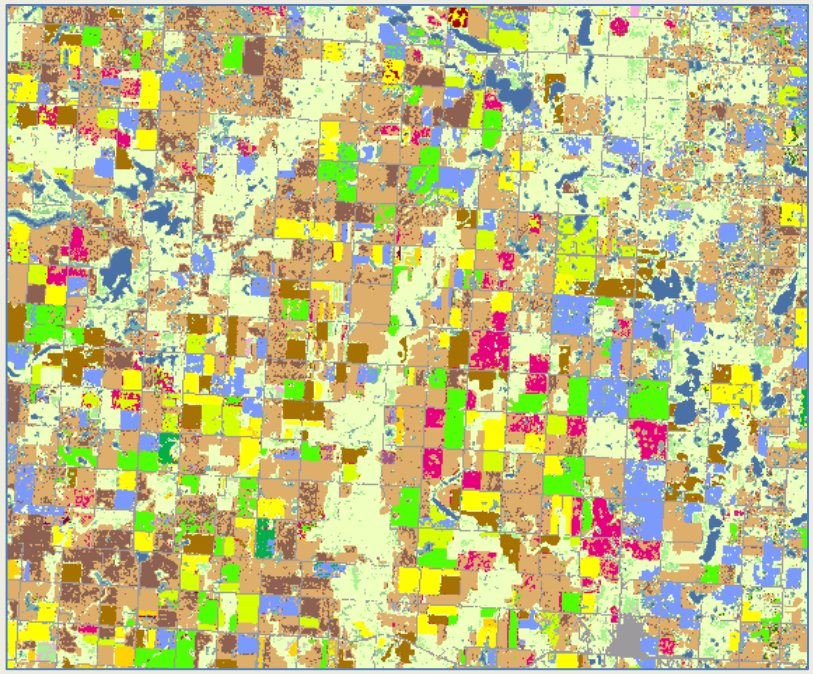


CDL Classification
(red = sugar beets, brown = soybeans
tan = spring wheat, gold = corn,
yellow = sunflowers)

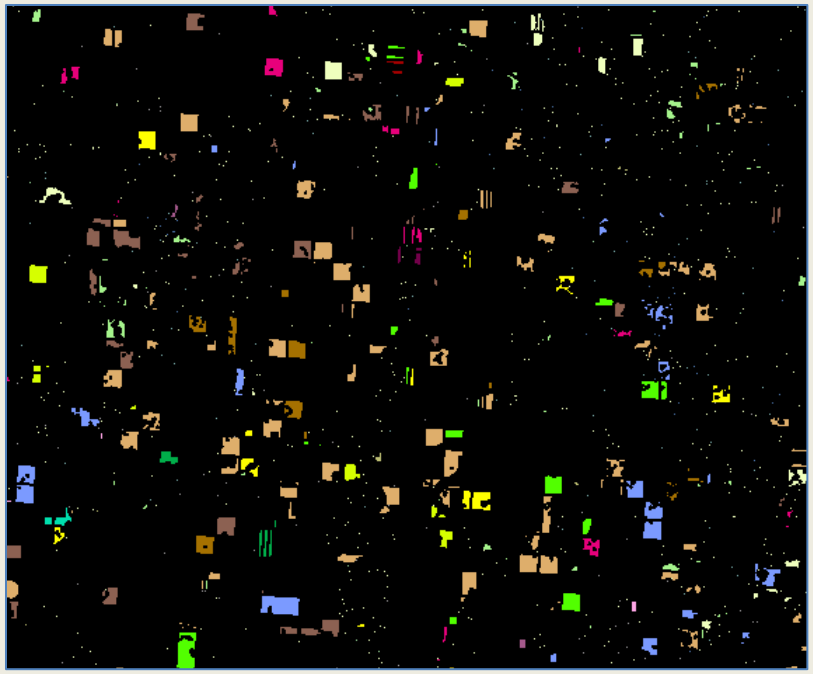
Resourcesat-1 AWiFS, 6 July 2007
(red =SWIR band,
green=NIR band,
blue=red band)

Accuracy Assessment

Each classification tested against independent set of ground truth data to determine overall and within class accuracies



Example classification subset



Example validation subset

Degradation methodology

Column with land cover category value

DG	DH	DI	DJ	DK	DL	DM	DN
133	22	3474	8460	434	0	0	1
132	21	3474	8460	436	0	0	1
133	24	4020	9158	435	0	0	1
138	24	2939	9249	440	0	0	1
133	20	3711	9183	439	0	0	1
134	22	3711	9183	439	0	0	1
133	23	3324	9359	437	0	0	1
134	30	4205	9011	440	0	0	1
140	28	3759	9042	438	0	0	1
134	25	3770	9252	436	0	0	1
135	24	2864	9244	439	0	0	1
132	23	3324	9359	437	0	0	1
134	24	3759	9042	437	0	0	1
134	23	4205	9011	438	0	0	1
134	24	3433	9104	435	0	0	1
134	23	3372	8987	439	0	0	1
133	23	3372	8987	440	0	0	1

Original sample file with no known errors (dozens of columns, hundreds of thousands of rows in reality)

Etc.

Altered sample files with X'th row scrambled

Every row

DG	DH	DI	DJ	DK	DL	DM	DN
133	22	3474	8460	434	0	0	1
132	21	3474	8460	436	0	0	5
133	24	4020	9158	435	0	0	181
138	24	2939	9249	440	0	0	1
133	20	3711	9183	439	0	0	5
134	22	3711	9183	439	0	0	1
133	23	3324	9359	437	0	0	1
134	30	4205	9011	440	0	0	1
140	28	3759	9042	438	0	0	1
134	25	3770	9252	436	0	0	5
135	24	2864	9244	439	0	0	1
132	23	3324	9359	437	0	0	5
134	24	3759	9042	437	0	0	1
134	23	4205	9011	438	0	0	1
134	24	3433	9104	435	0	0	1
134	23	3372	8987	439	0	0	1
133	23	3372	8987	440	0	0	5

Every other row

DG	DH	DI	DJ	DK	DL	DM	DN
133	22	3474	8460	434	0	0	1
132	21	3474	8460	436	0	0	5
133	24	4020	9158	435	0	0	1
138	24	2939	9249	440	0	0	1
133	20	3711	9183	439	0	0	1
134	22	3711	9183	439	0	0	1
133	23	3324	9359	437	0	0	1
134	30	4205	9011	440	0	0	1
140	28	3759	9042	438	0	0	1
134	25	3770	9252	436	0	0	5
135	24	2864	9244	439	0	0	1
132	23	3324	9359	437	0	0	5
134	24	3759	9042	437	0	0	1
134	23	4205	9011	438	0	0	1
134	24	3433	9104	435	0	0	1
134	23	3372	8987	439	0	0	1
133	23	3372	8987	440	0	0	1

Every third row

DG	DH	DI	DJ	DK	DL	DM	DN
133	22	3474	8460	434	0	0	1
132	21	3474	8460	436	0	0	1
133	24	4020	9158	435	0	0	181
138	24	2939	9249	440	0	0	1
133	20	3711	9183	439	0	0	1
134	22	3711	9183	439	0	0	1
133	23	3324	9359	437	0	0	1
134	30	4205	9011	440	0	0	1
140	28	3759	9042	438	0	0	1
134	25	3770	9252	436	0	0	1
135	24	2864	9244	439	0	0	1
132	23	3324	9359	437	0	0	5
134	24	3759	9042	437	0	0	1
134	23	4205	9011	438	0	0	1
134	24	3433	9104	435	0	0	1
134	23	3372	8987	439	0	0	1
133	23	3372	8987	440	0	0	1

Every forth row

DG	DH	DI	DJ	DK	DL	DM	DN
133	22	3474	8460	434	0	0	1
132	21	3474	8460	436	0	0	5
133	24	4020	9158	435	0	0	1
138	24	2939	9249	440	0	0	1
133	20	3711	9183	439	0	0	1
134	22	3711	9183	439	0	0	1
133	23	3324	9359	437	0	0	1
134	30	4205	9011	440	0	0	1
140	28	3759	9042	438	0	0	1

Run classifier

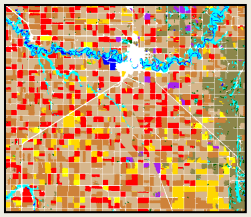
Rulequest See5.0

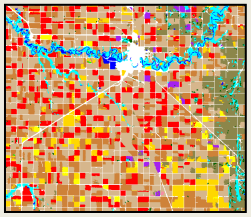
Rulequest See5.0

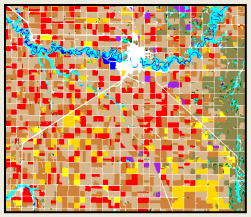
Rulequest See5.0

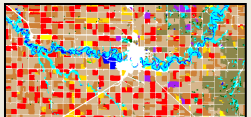
Rulequest See5.0

Output land cover map

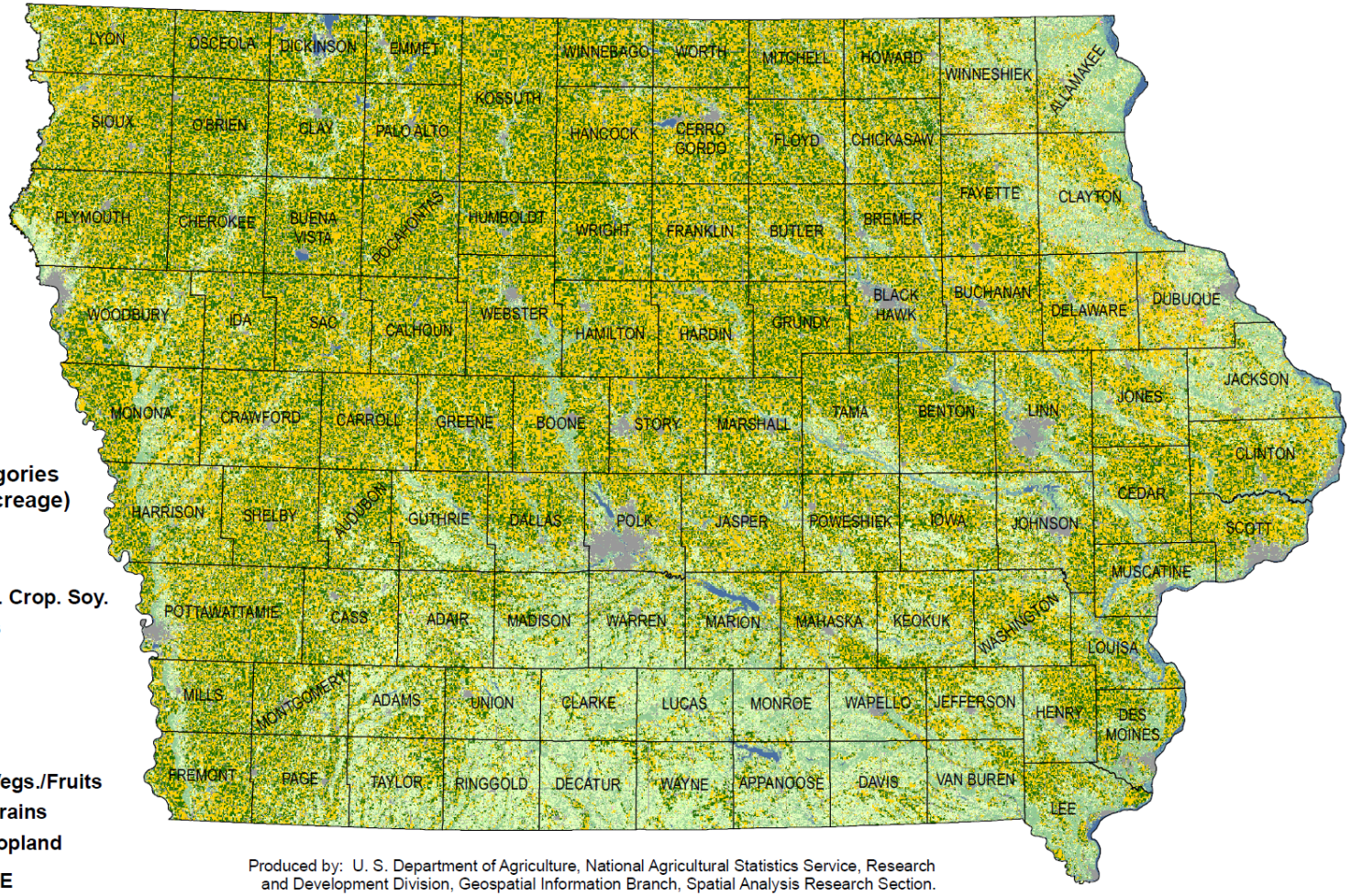








2009 Iowa Cropland Data Layer



Land Cover Categories
(by decreasing acreage)

AGRICULTURE

- Corn
- Soybeans/Dbl. Crop. Soy.
- Pasture/Grass
- Other Hays
- Alfalfa
- Oats
- Winter Wheat
- Other Crops/Vegs./Fruits
- Other Small Grains
- Fallow/Idle Cropland

NON-AGRICULTURE

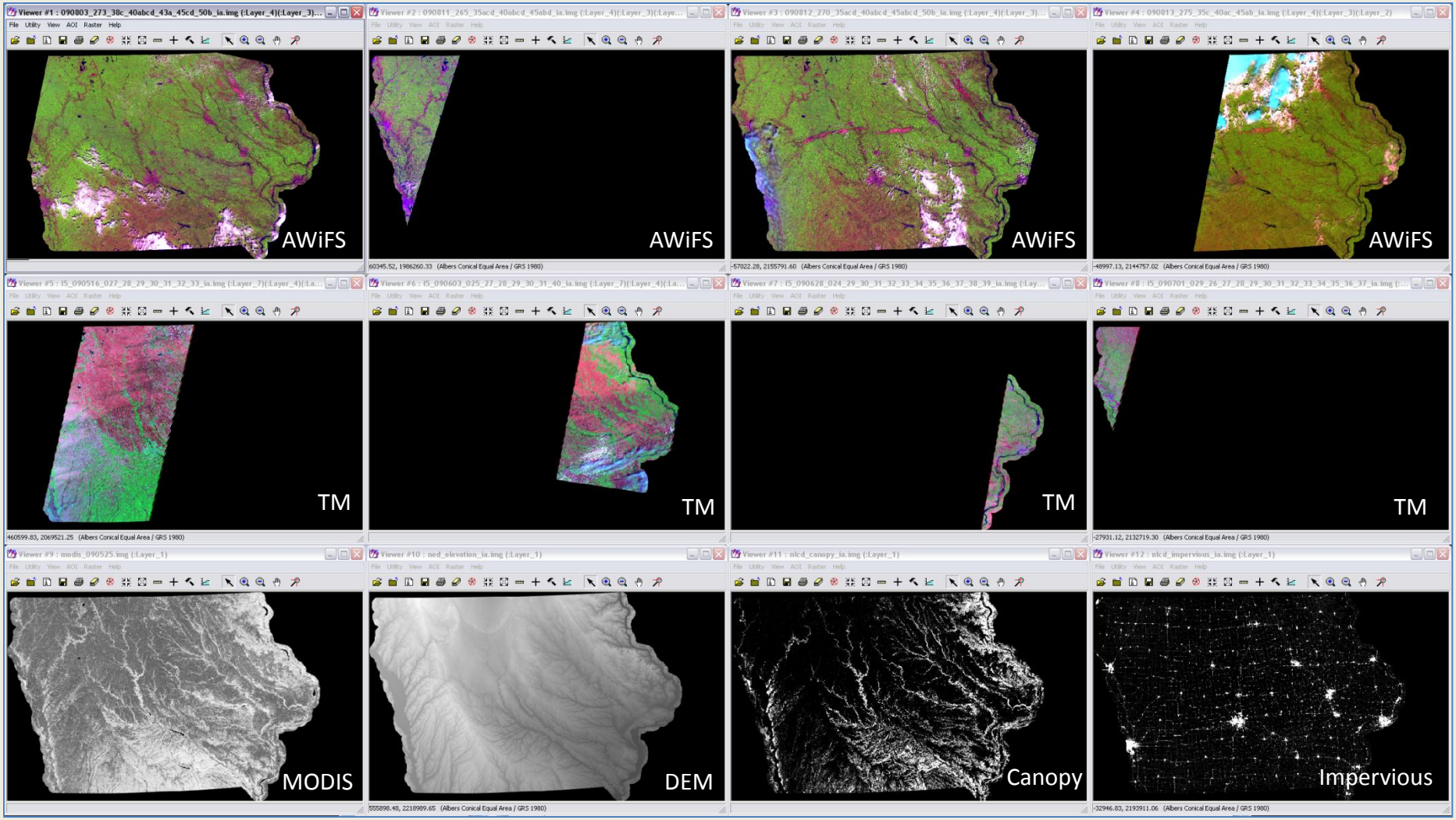
- Urban/Developed
- Woodland
- Wetlands
- Water
- Barren
- Shrubland

Produced by: U. S. Department of Agriculture, National Agricultural Statistics Service, Research and Development Division, Geospatial Information Branch, Spatial Analysis Research Section.
Data Source: Resourcesat-1 AWiFS courtesy of USDA Foreign Agricultural Service, Landsat 5 Thematic Mapper.
Image Processing: Rulequest See5 and ERDAS Imagine.
Ground Truth: The Farm Service Agency Common Land Unit for crops classes, and 2001 National Land Cover Dataset (NLCD) for non-agricultural classes.
Ancillary Data: NLCD Impervious Surface, NLCD Forest Canopy, National Elevation Dataset, and Moderate Resolution Imagery Spectroradiometer (MODIS) 16 day Normalized Difference Vegetation Index (NDVI) composite.
Projection: UTM zone 15, WGS84 datum.
Map Production: ESRI ArcGIS 9.3.

N

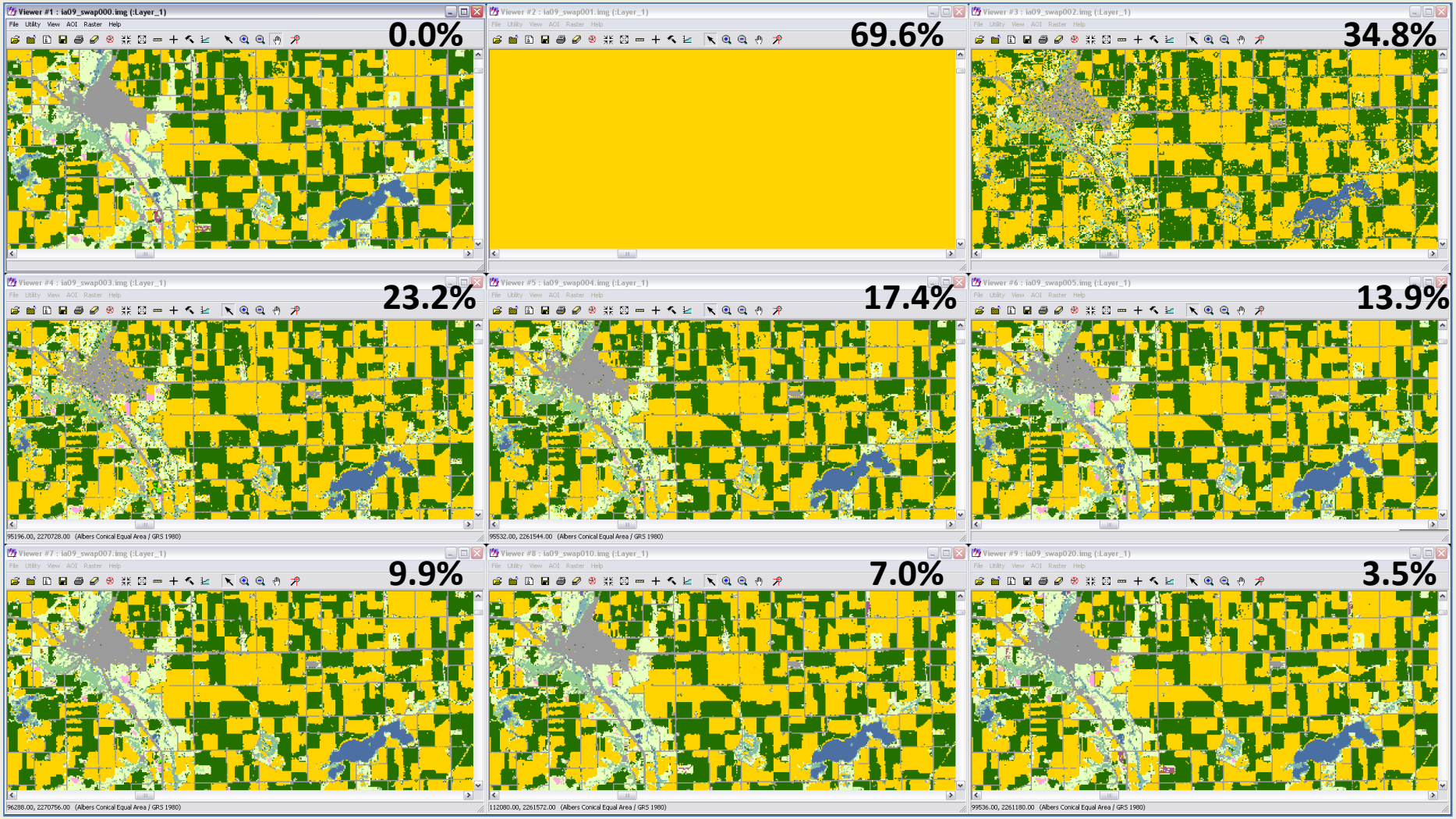
0 10 20 30 40 50 Miles

Iowa ‘09 CDL input layer examples



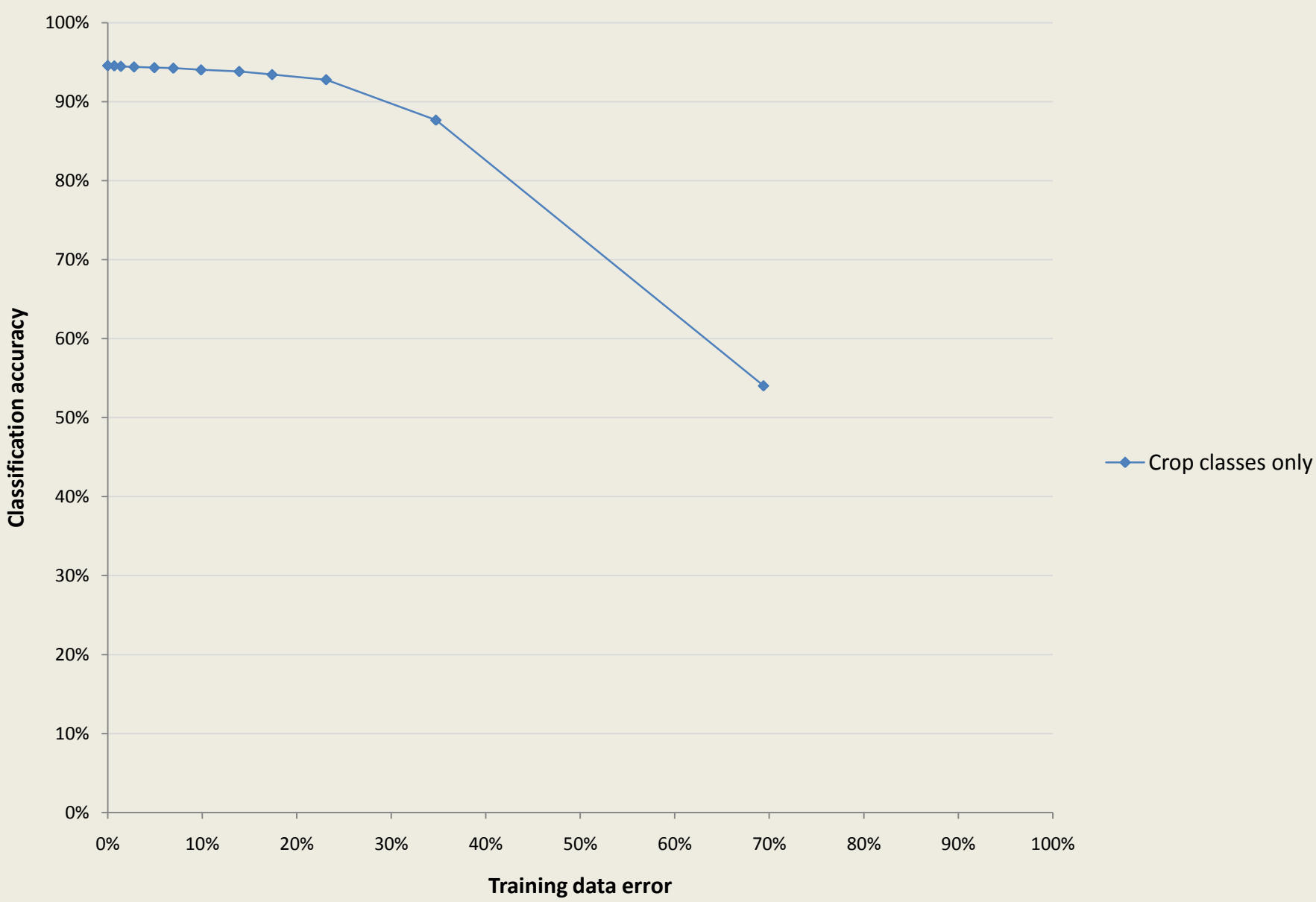
Scenes of data actually used: 10 AWiFS, 10 TM, 2 MODIS NDVI, DEM, Canopy, and Impervious
(dates ranged from 1 April ‘09 – 8 August ‘09)

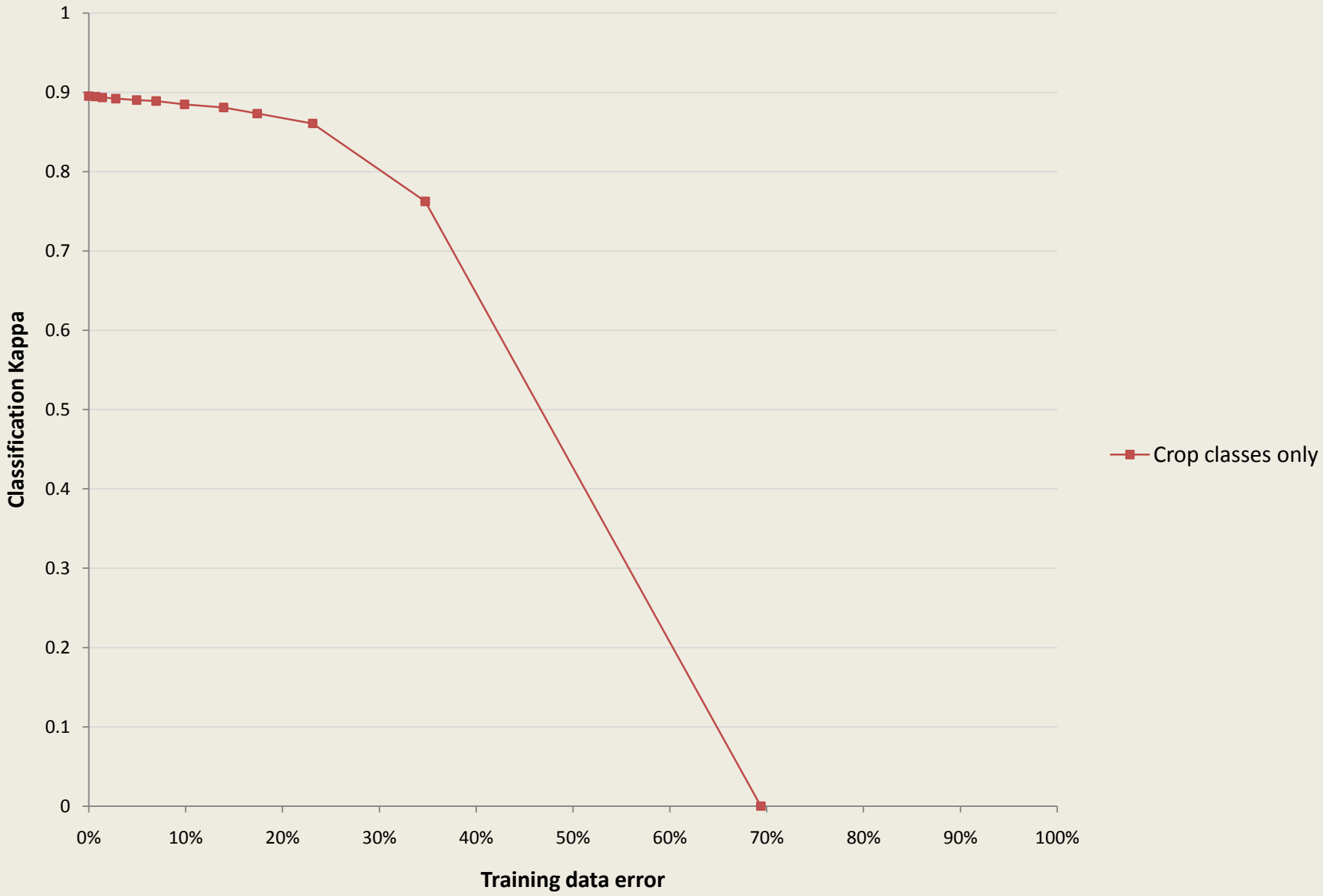
Iowa classifications with training data error %

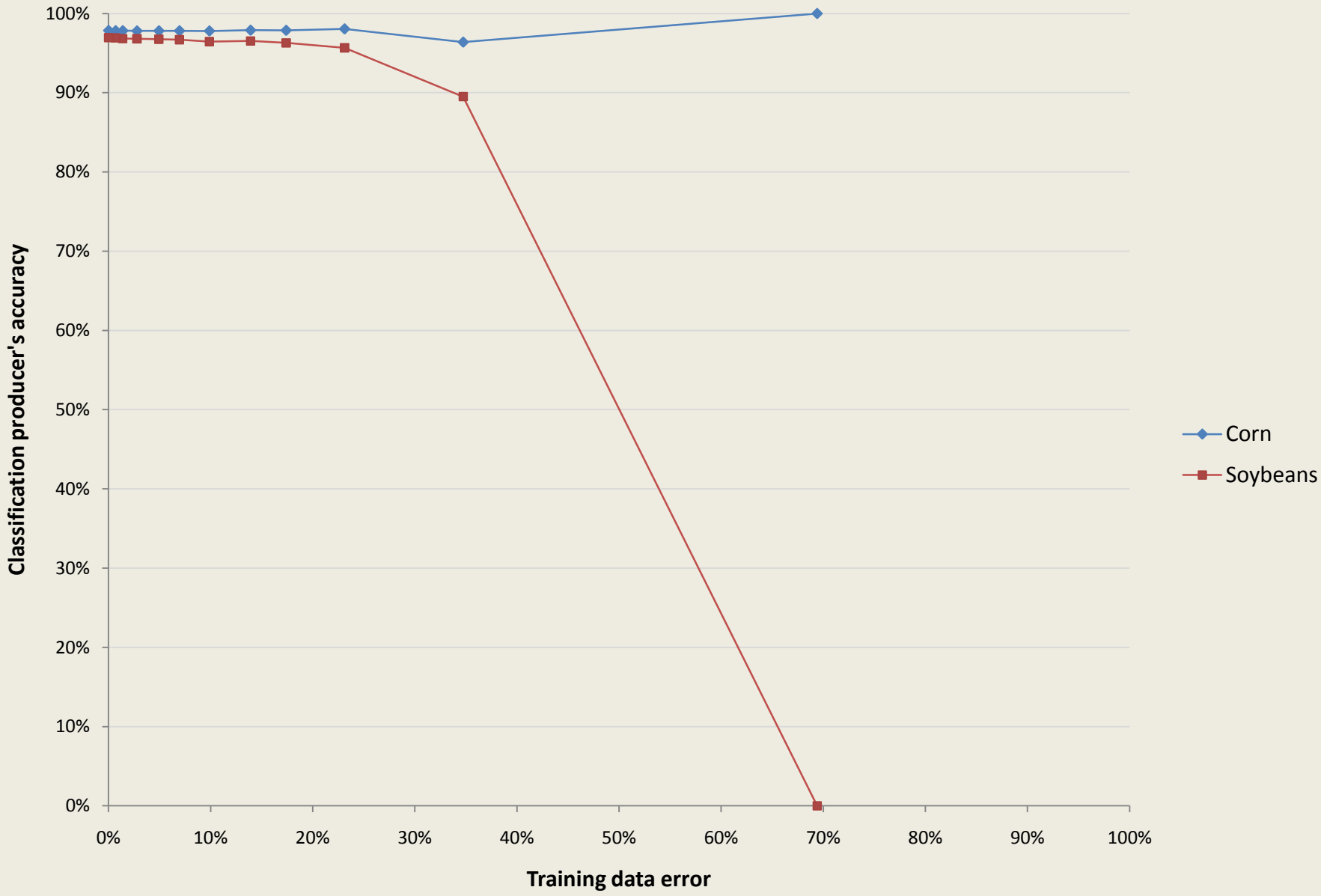


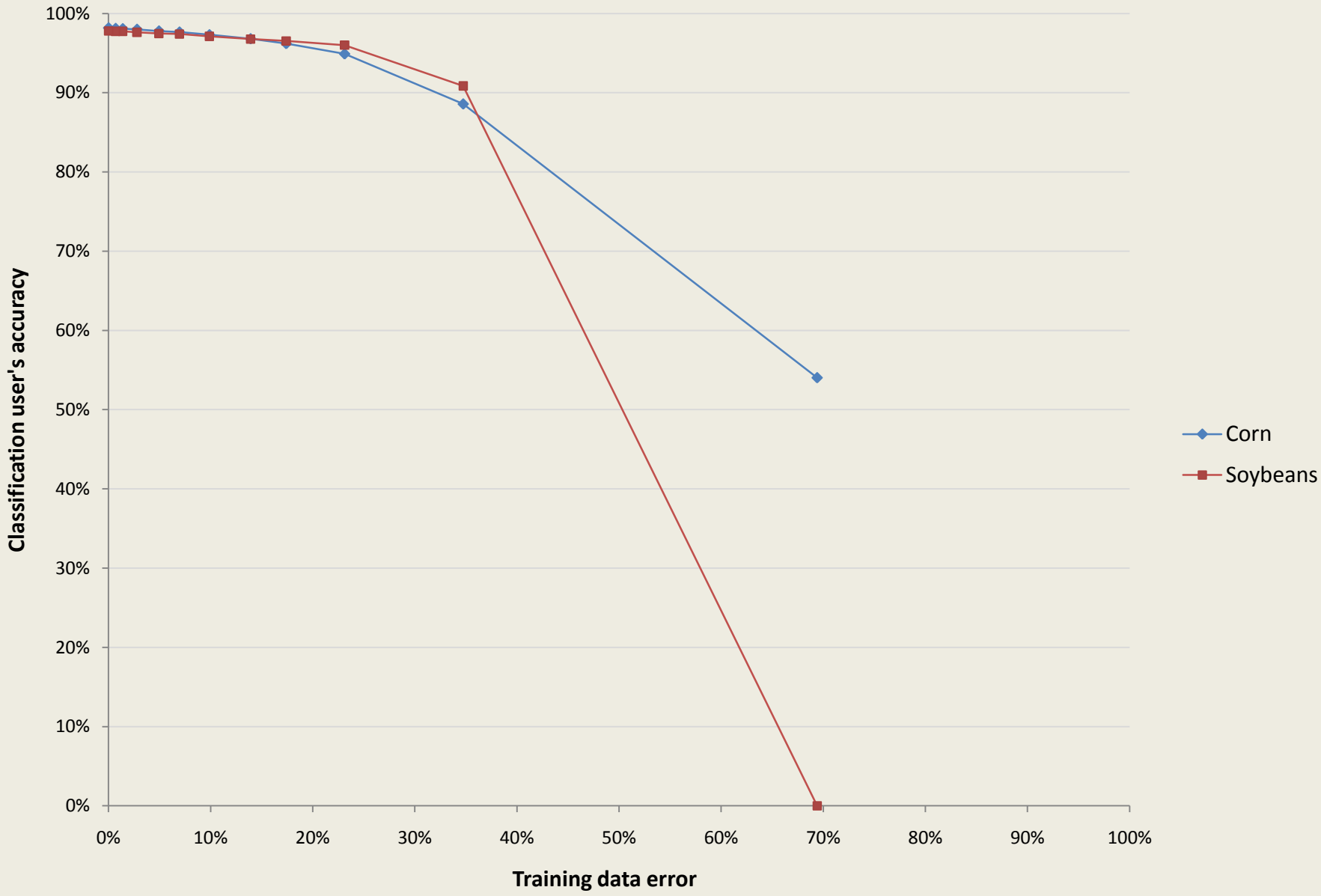
gold = corn, dark green = soybeans

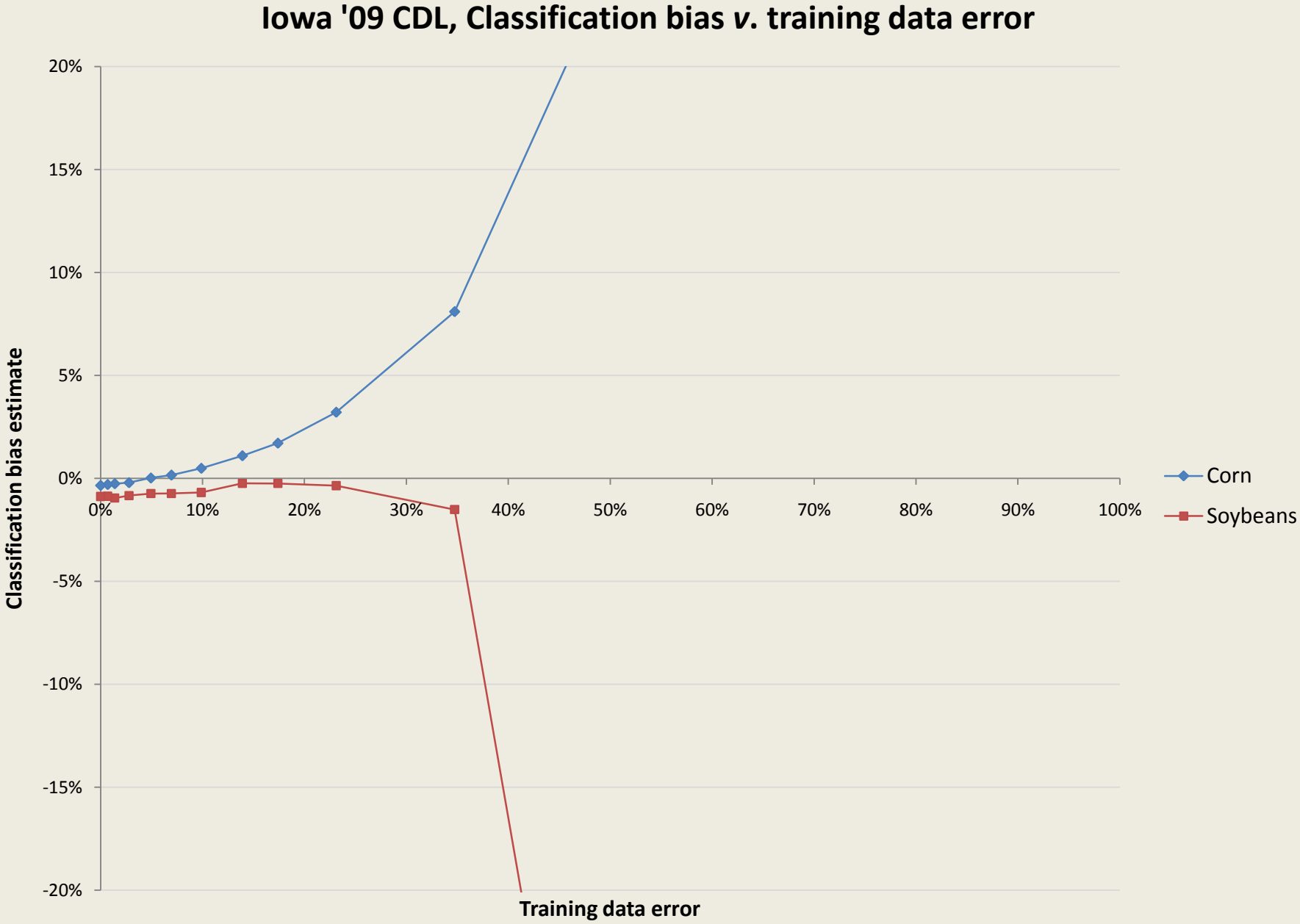
Total scene has 46,474,682 pixels, 755,116 (1.6%) chosen for training


Iowa '09 CDL, Classification accuracy v. training data error
Classification accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Crop classes only

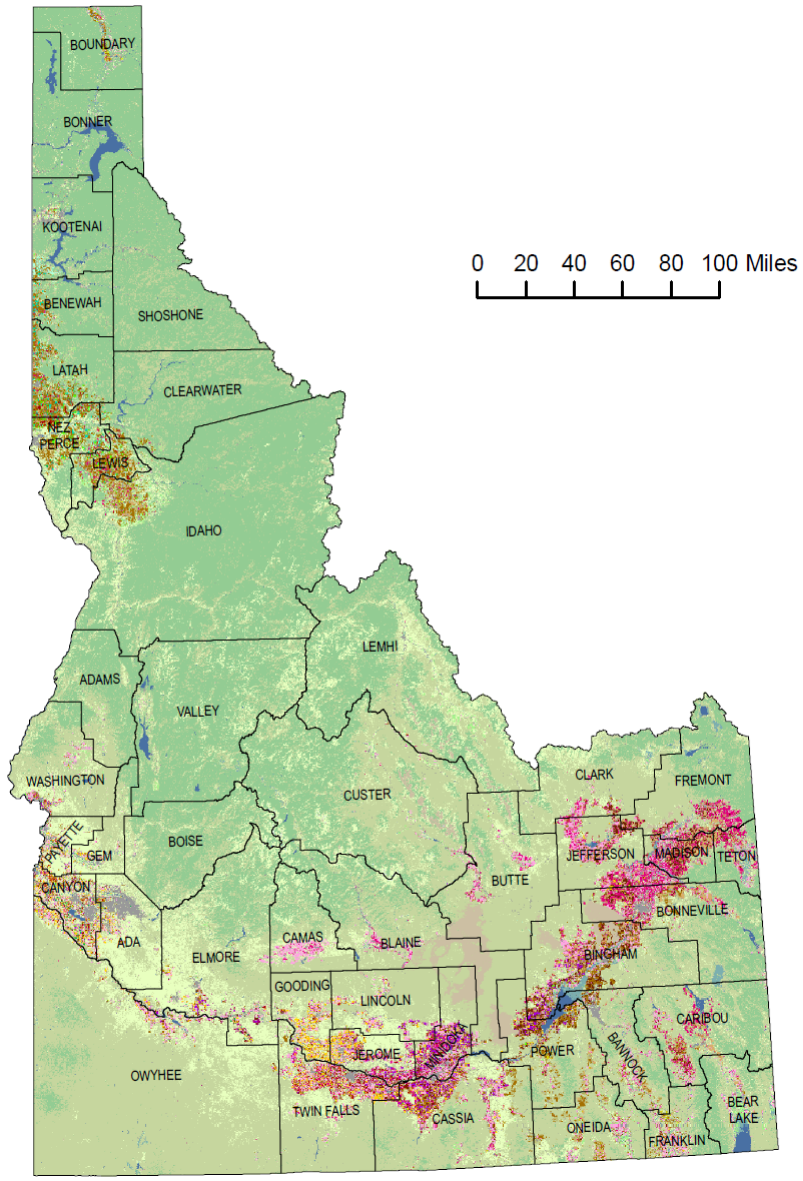

Iowa '09 CDL, Classification Kappa v. training data error
Classification Kappa
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Crop classes only


Iowa '09 CDL, Classification producer's accuracy v. training data error
Classification producer's accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Corn
Soybeans


Iowa '09 CDL, Classification user's accuracy v. training data error
Classification user's accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Corn
Soybeans


Iowa '09 CDL, Classification bias v. training data error
Classification bias estimate
20%
15%
10%
5%
0%
-5%
-10%
-15%
-20%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Corn
Soybeans

2009 Idaho Cropland Data Layer



Land Cover Categories (by decreasing acreage)

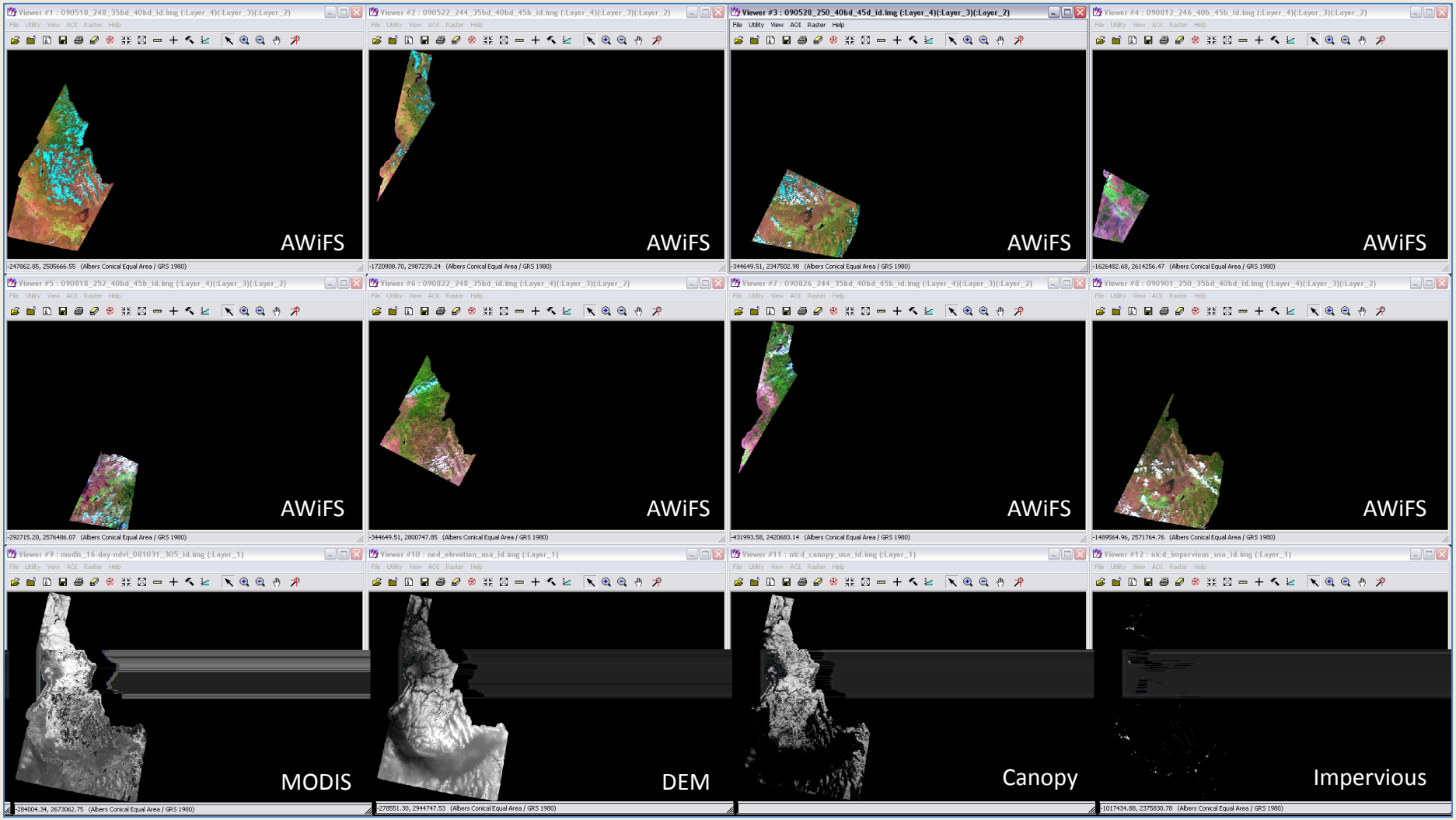
AGRICULTURE

- Pasture/Grass
- Alfalfa
- Winter Wheat
- Spring Wheat
- Barley
- Other Hays
- Potatoes
- Fallow/Idle Cropland
- Corn
- Sugarbeets
- Dry Beans
- Peas
- Lentils
- Seed/Sod Grass
- Oats
- Other Crops/Herbs
- Misc. Vegs. & Fruits/Tree Fruits
- Other Small Grains
- Canola
- Safflower
- Rye

NON-AGRICULTURE

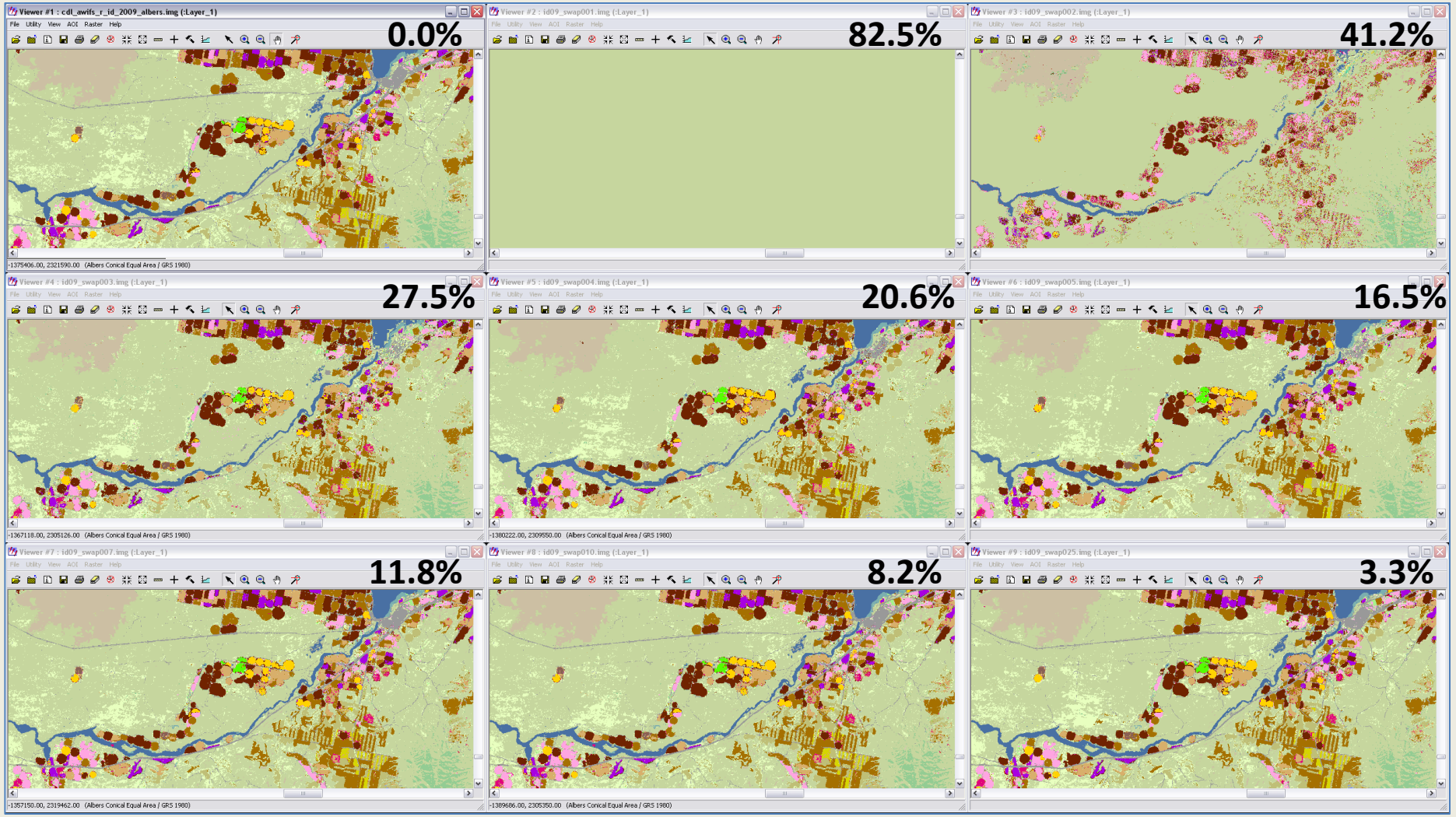
- Shrubland
- Woodland
- Urban/Developed
- Barren
- Water
- Wetlands
- Perennial Ice/Snow

Idaho ‘09 CDL input layer examples

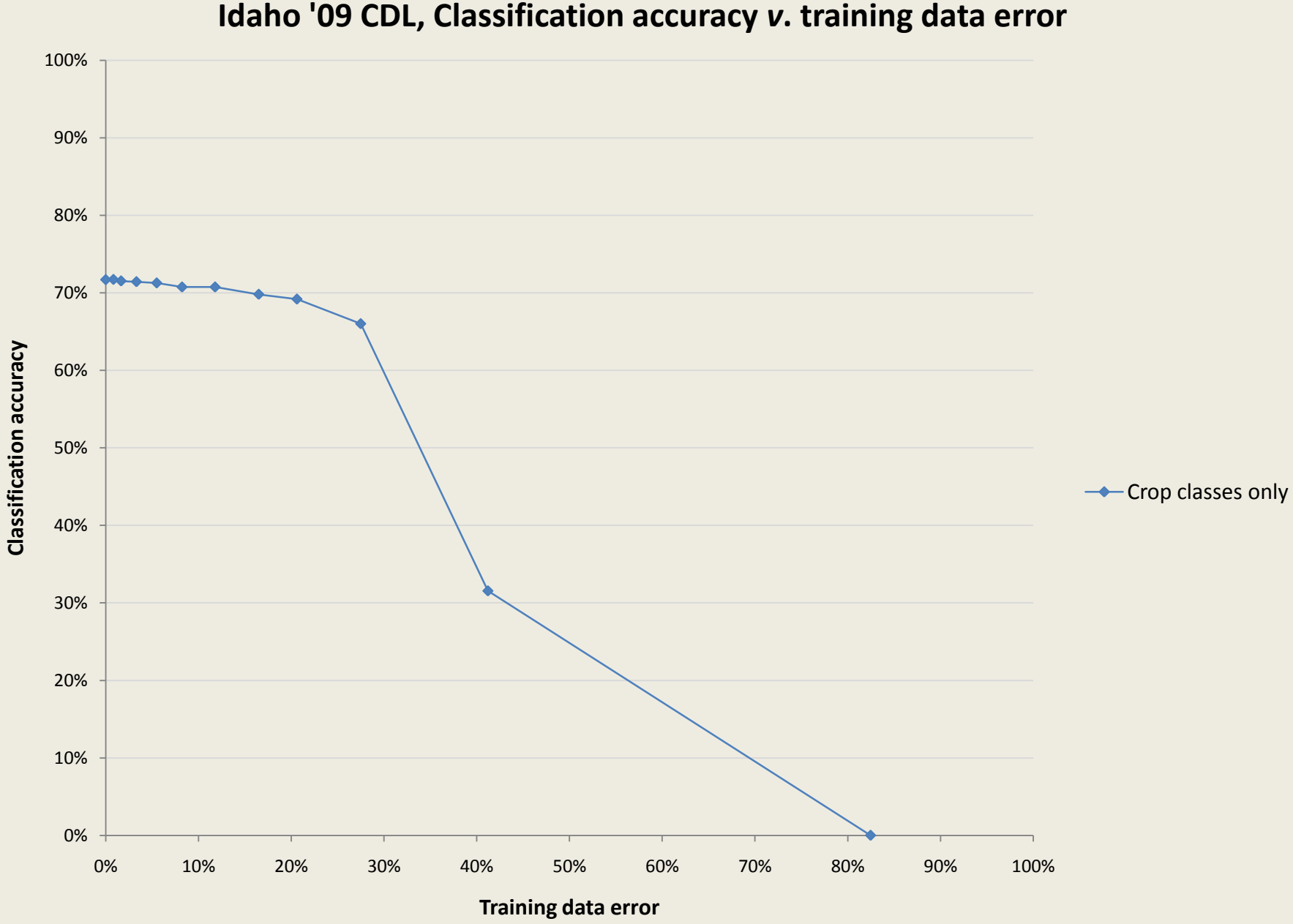


Scenes of data actually used: 15 AWiFS, 7 MODIS NDVI, DEM, Canopy, and Impervious
(dates ranged from 29 September ‘08 – 1 September ‘09)

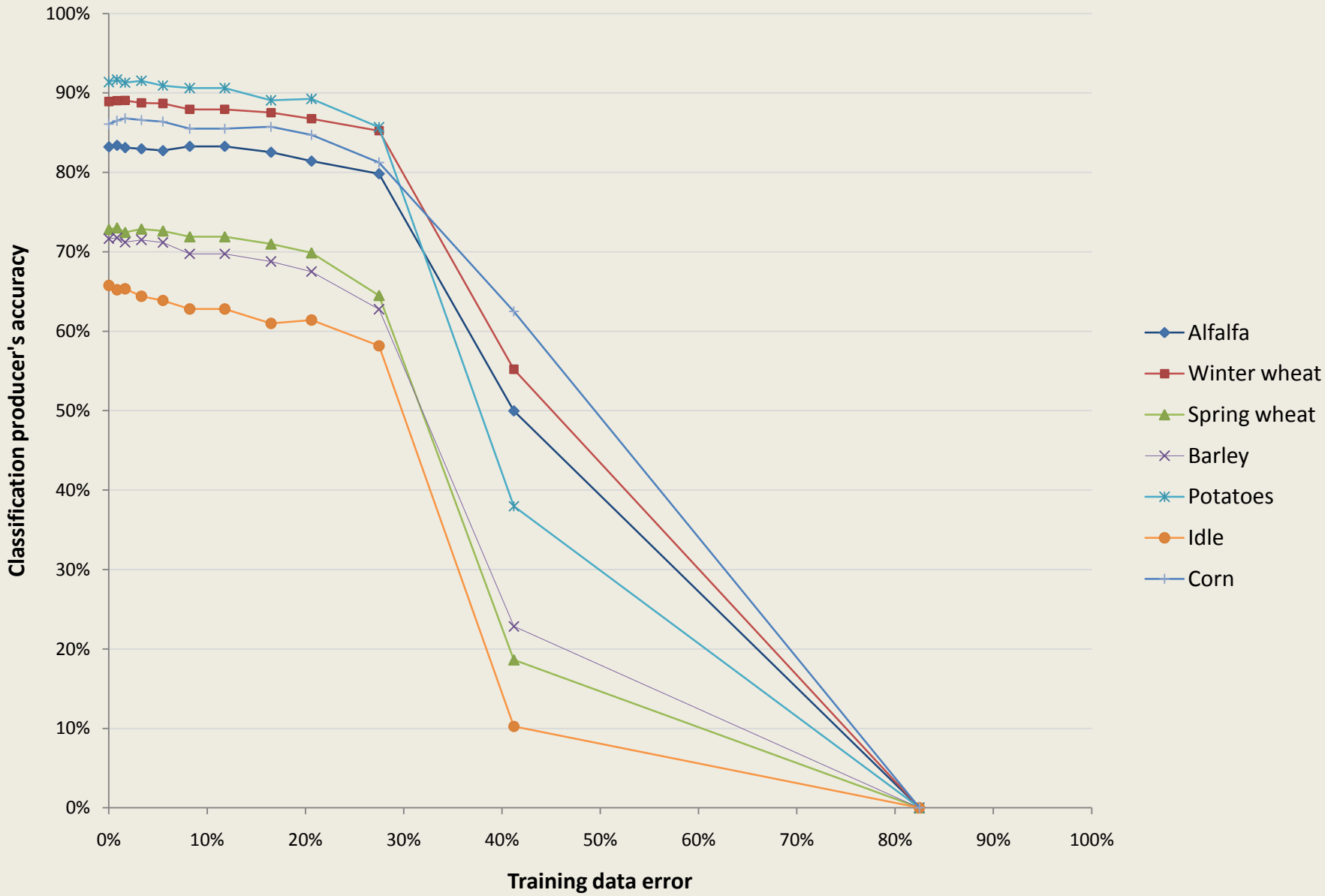
Idaho classifications with training data error %

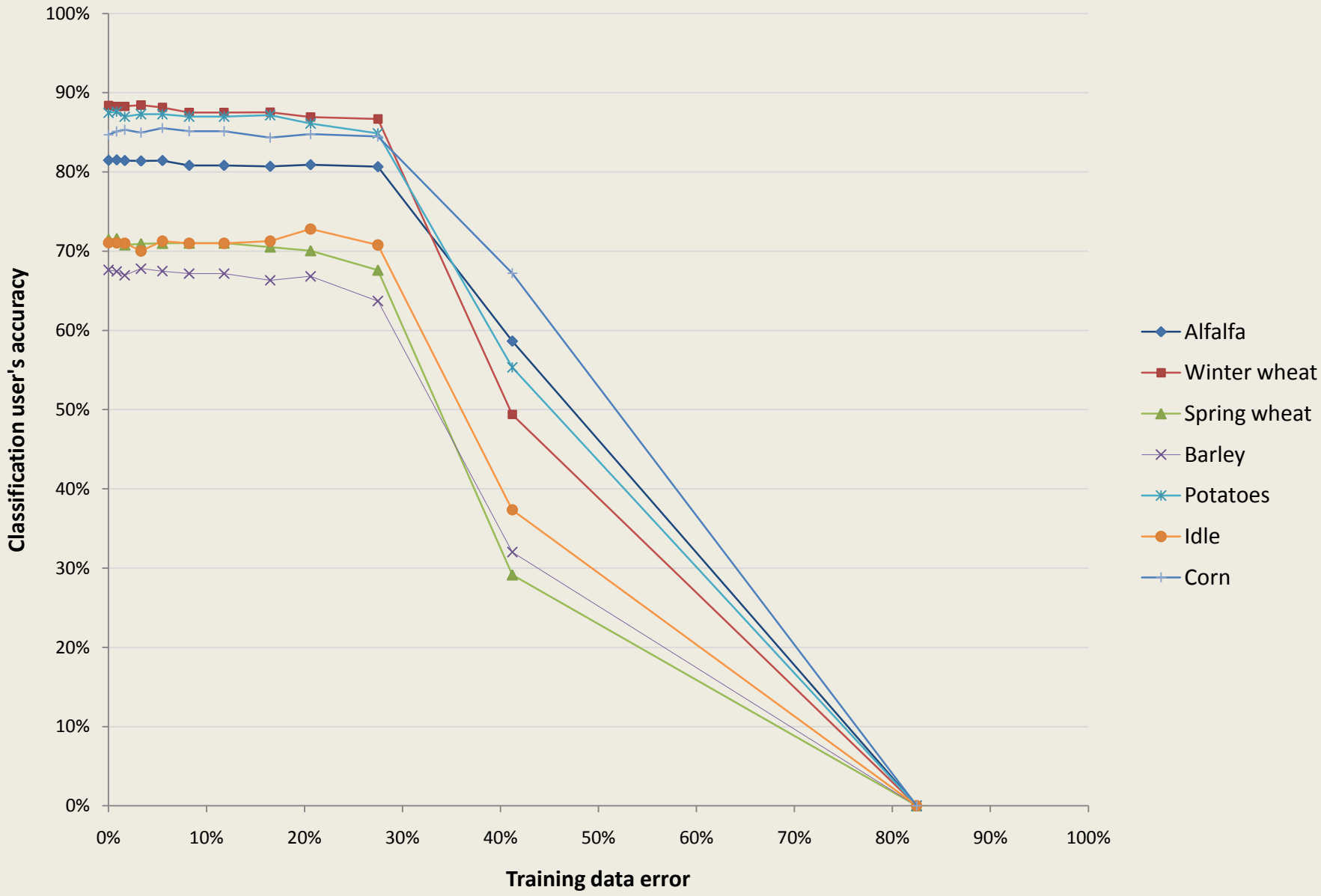


Total scene has 69,018,509 pixels, 891,793 (1.3%) chosen for training


Idaho '09 CDL, Classification accuracy v. training data error
Classification accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Crop classes only

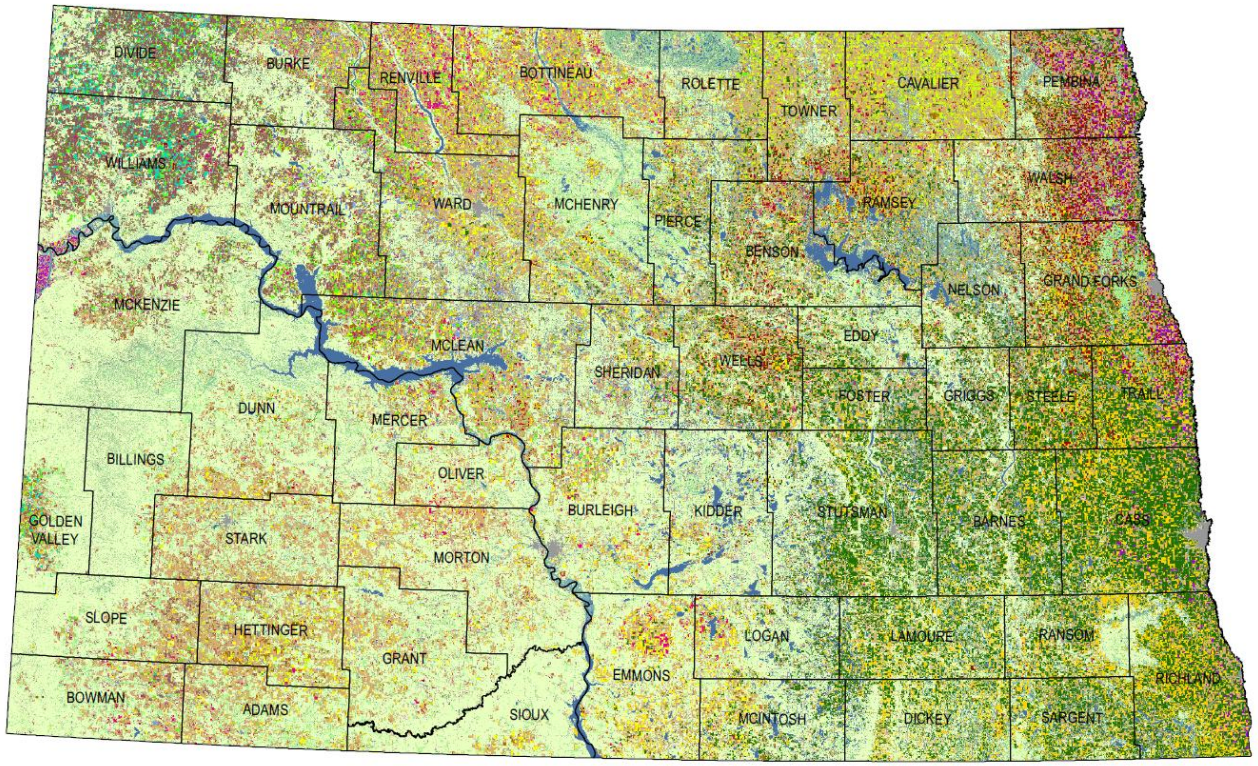

Idaho '09 CDL, Classification Kappa v. training data error
Classification Kappa
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Crop classes only


Idaho '09 CDL, Classification producer's accuracy v. training data error
Classification producer's accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Alfalfa
Winter wheat
Spring wheat
Barley
Potatoes
Idle
Corn


Idaho '09 CDL, Classification user's accuracy v. training data error
Classification user's accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Alfalfa
Winter wheat
Spring wheat
Barley
Potatoes
Idle
Corn


Idaho '09 CDL, Classification bias v. training data error
Classification bias estimate
20%
15%
10%
5%
0%
-5%
-10%
-15%
-20%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Alfalfa
Winter wheat
Spring wheat
Barley
Potatoes
Idle
Corn

2009 North Dakota Cropland Data Layer



0 20 40 60 80 100 Miles

Land Cover Categories (by decreasing acreage)

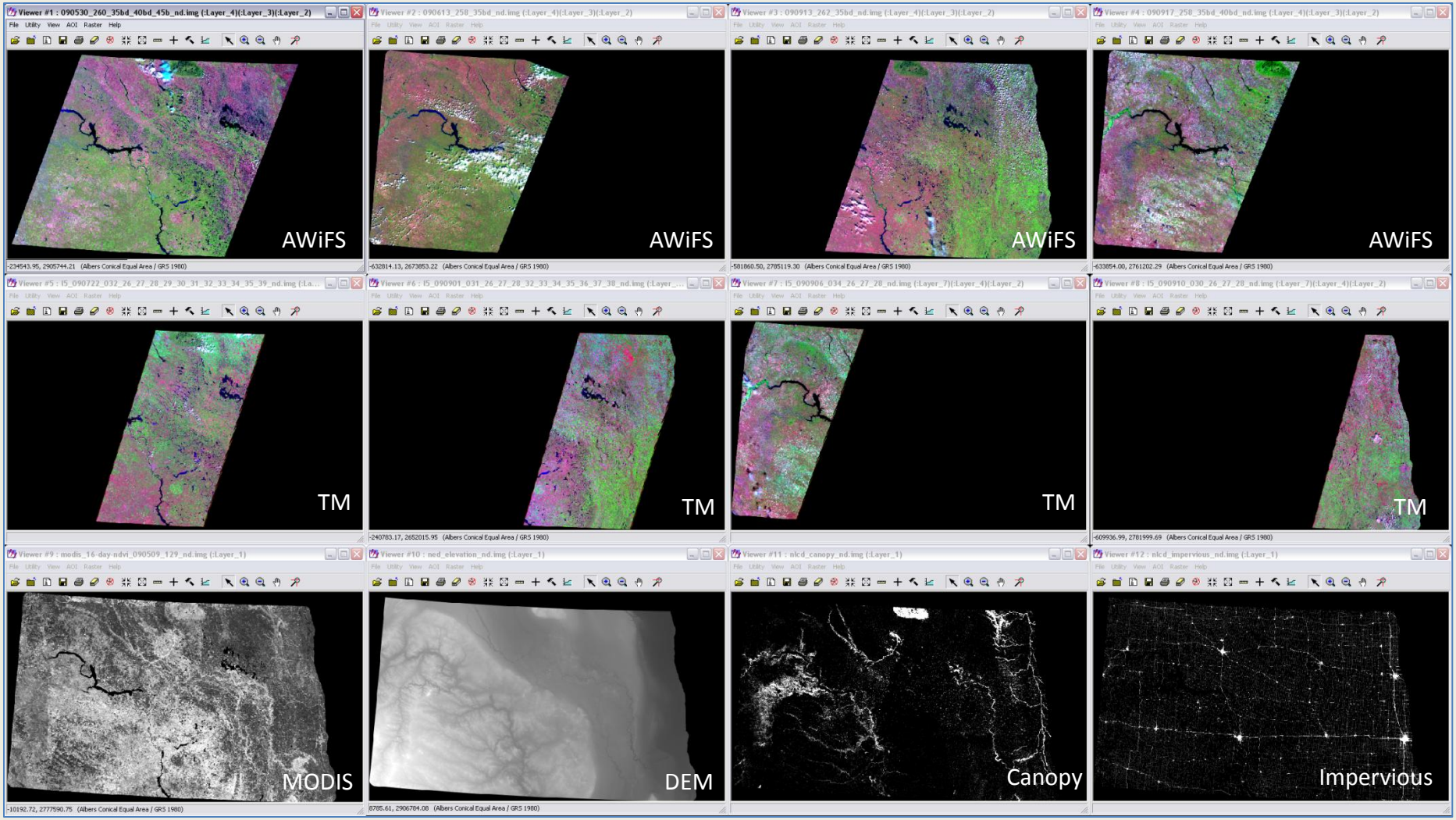
AGRICULTURE

- Pasture/Grass
- Spring Wheat
- Soybeans
- Other Hays
- Corn
- Durum Wheat
- Canola
- Sunflowers
- Dry Beans
- Barley
- Winter Wheat
- Peas
- Alfalfa
- Fallow/Idle Cropland
- Flaxseed
- Sugarbeets
- Lentils
- Oats
- Potatoes
- Other Crops/Vegs./Fruits
- Millet
- Safflower
- Sorghum
- Rye
- Seed/Sod Grass

NON-AGRICULTURE

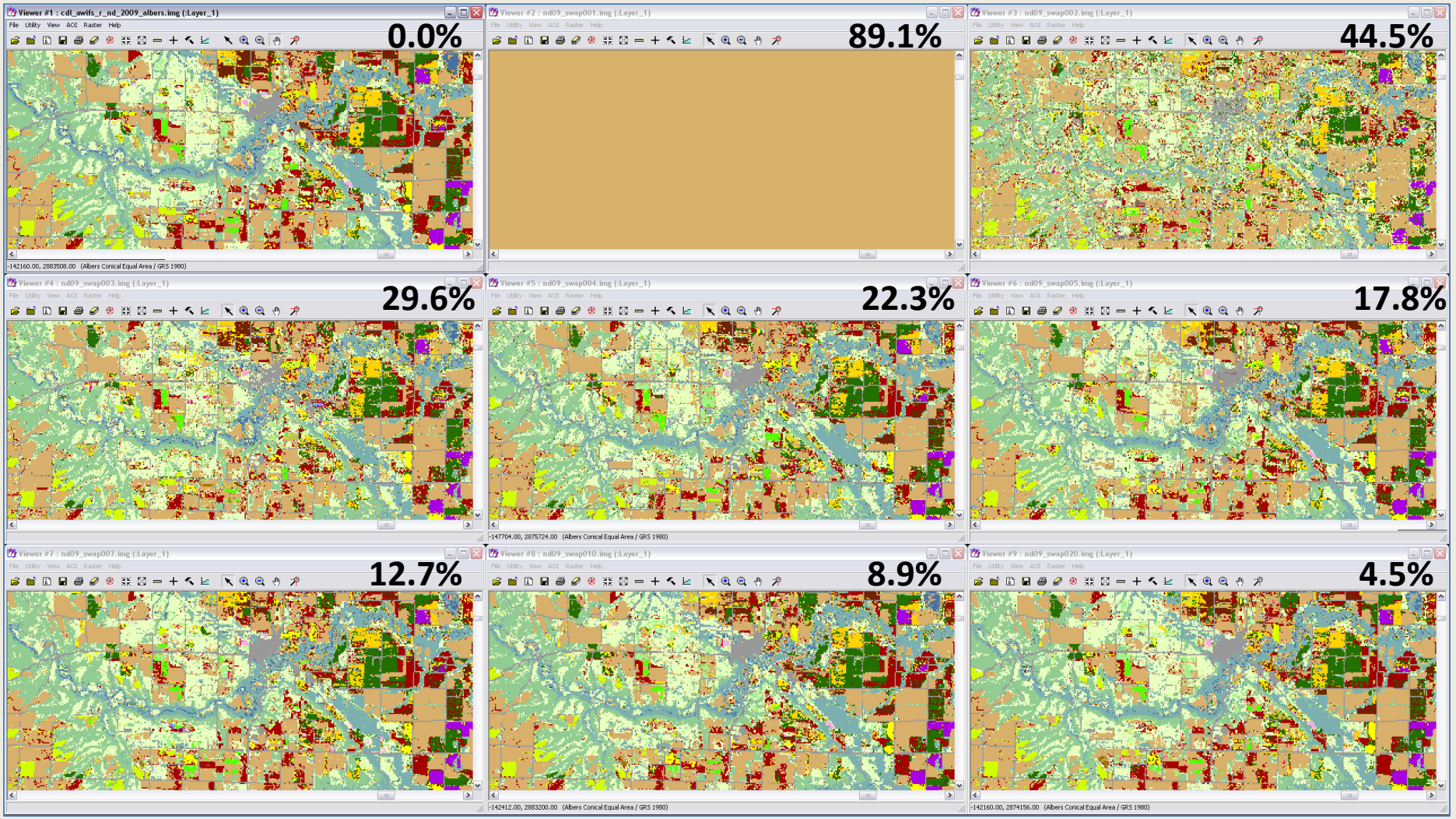
- Urban/Developed
- Wetlands
- Water
- Woodland
- Shrubland
- Barren

North Dakota ‘09 CDL input layer examples

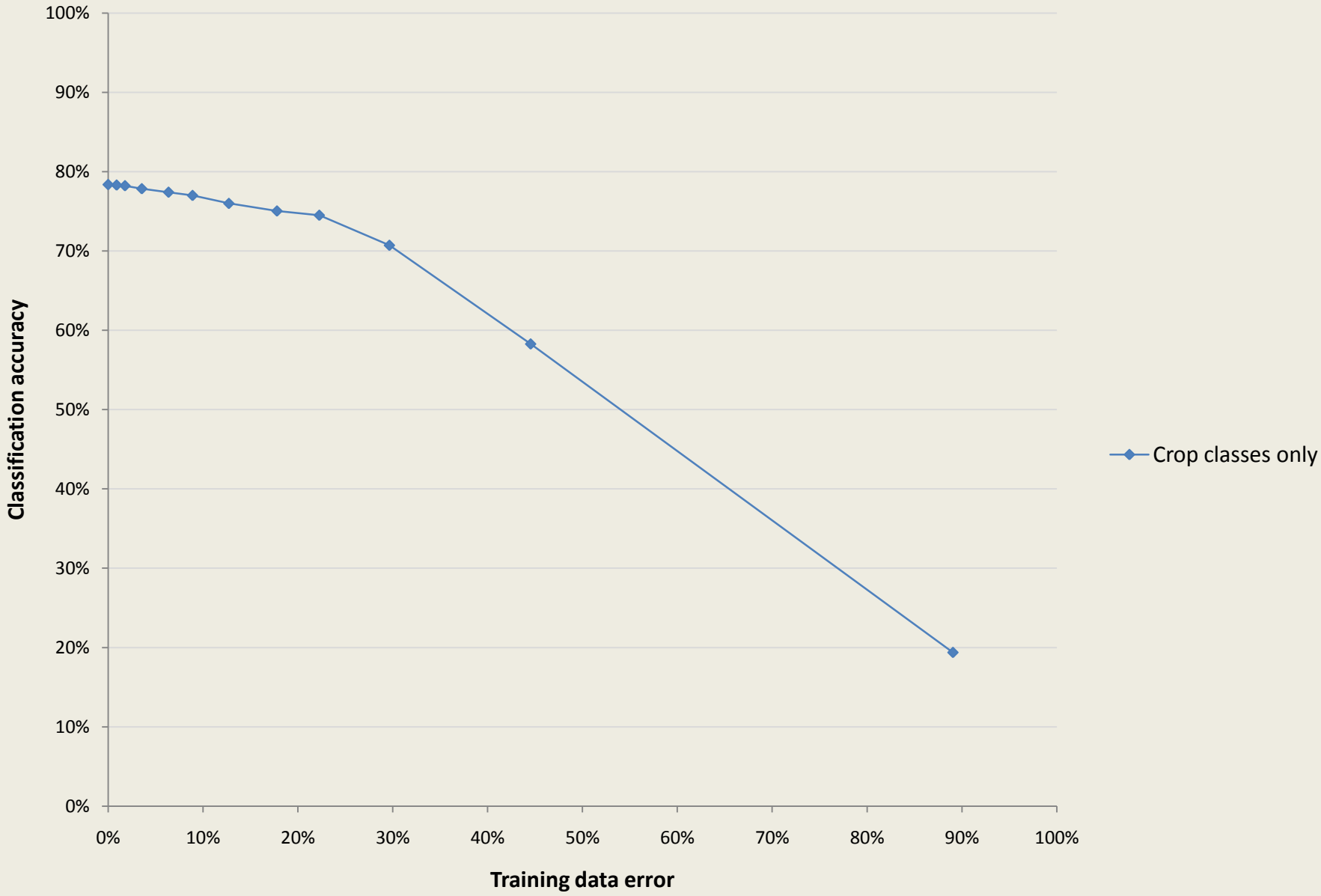


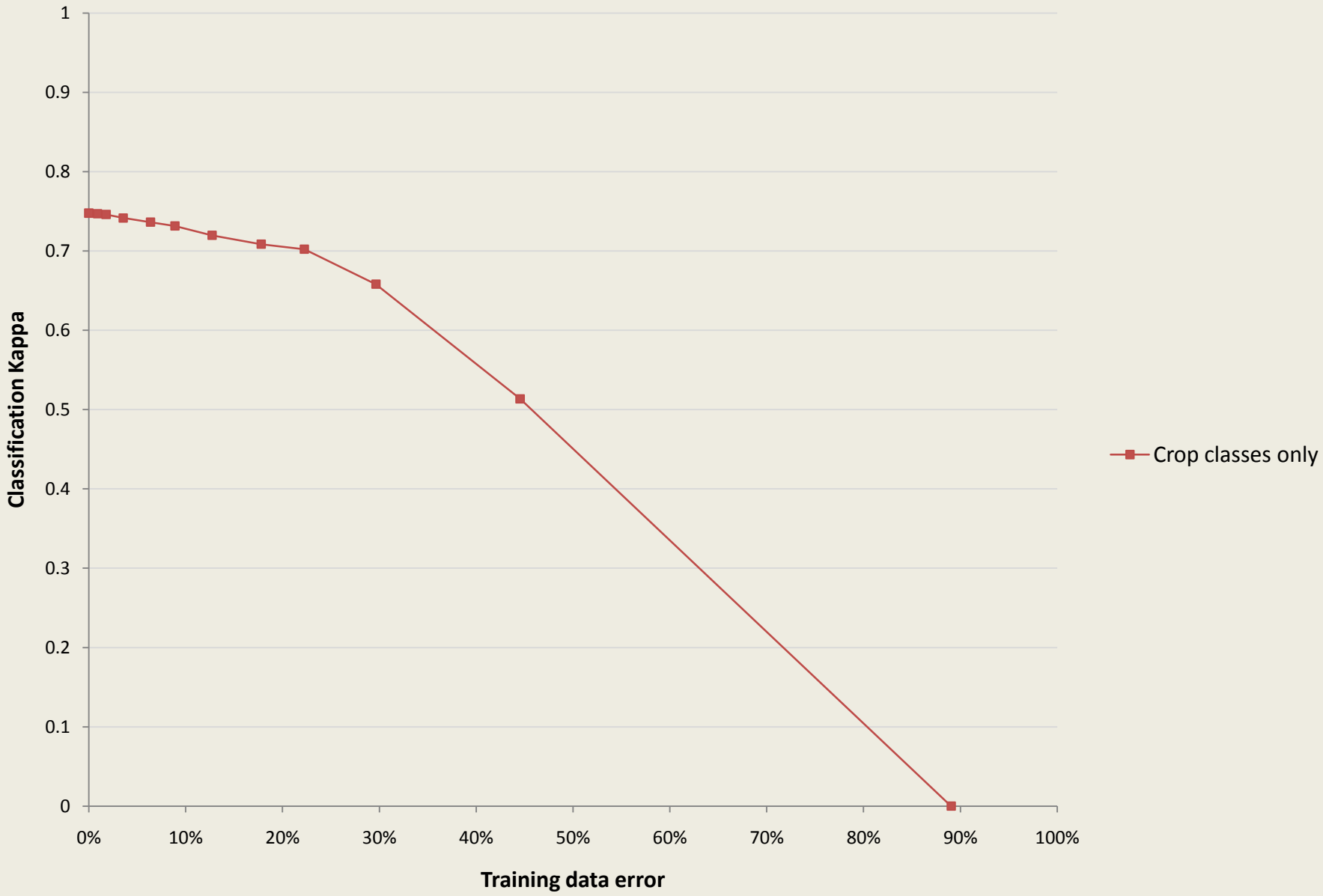
Scenes of data actually used: 14 AWiFS, 13 TM, 1 MODIS NDVI, DEM, Canopy, and Impervious (dates ranged from 6 May ‘09 – 17 September ‘09)

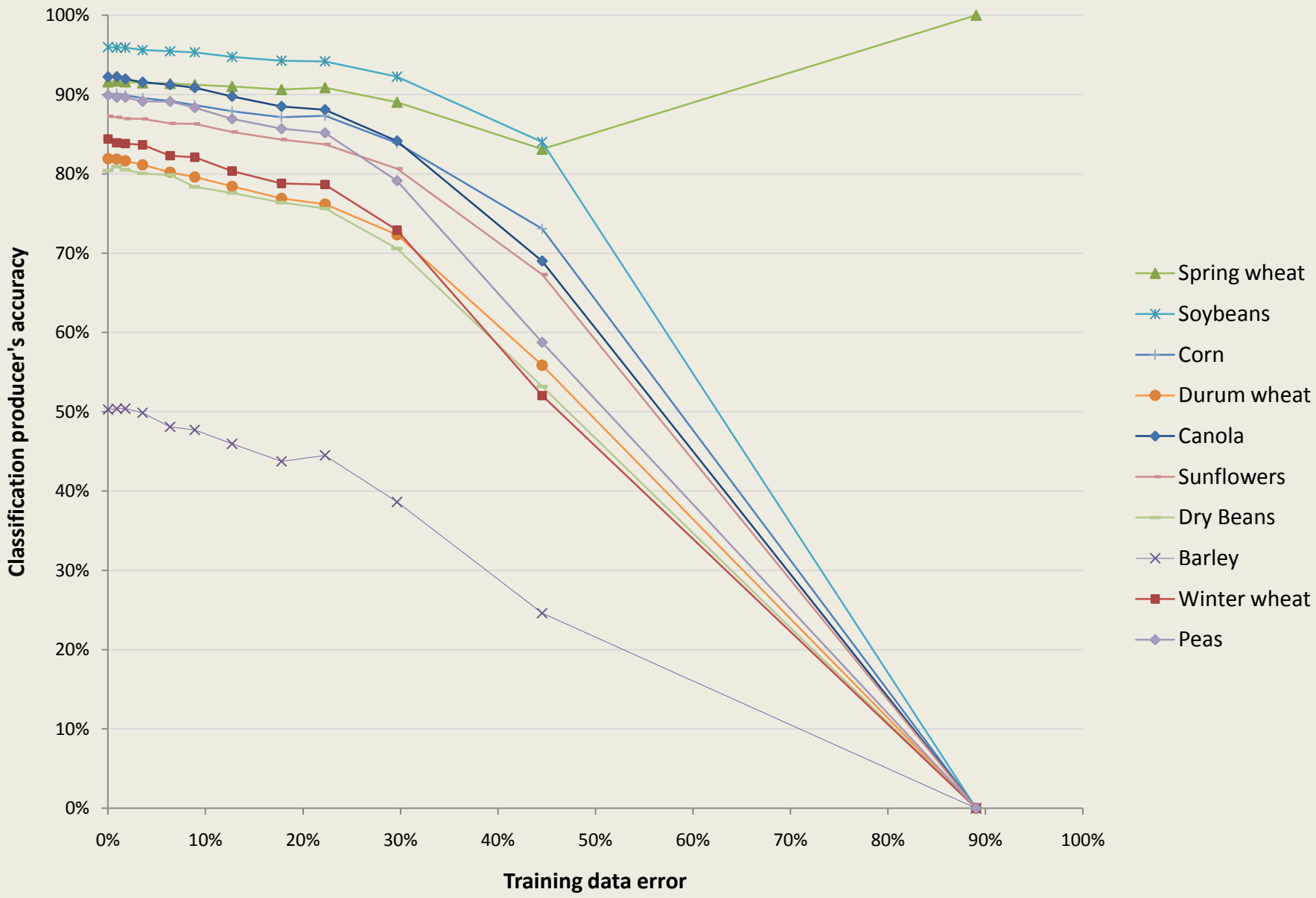
North Dakota classifications with training data error %

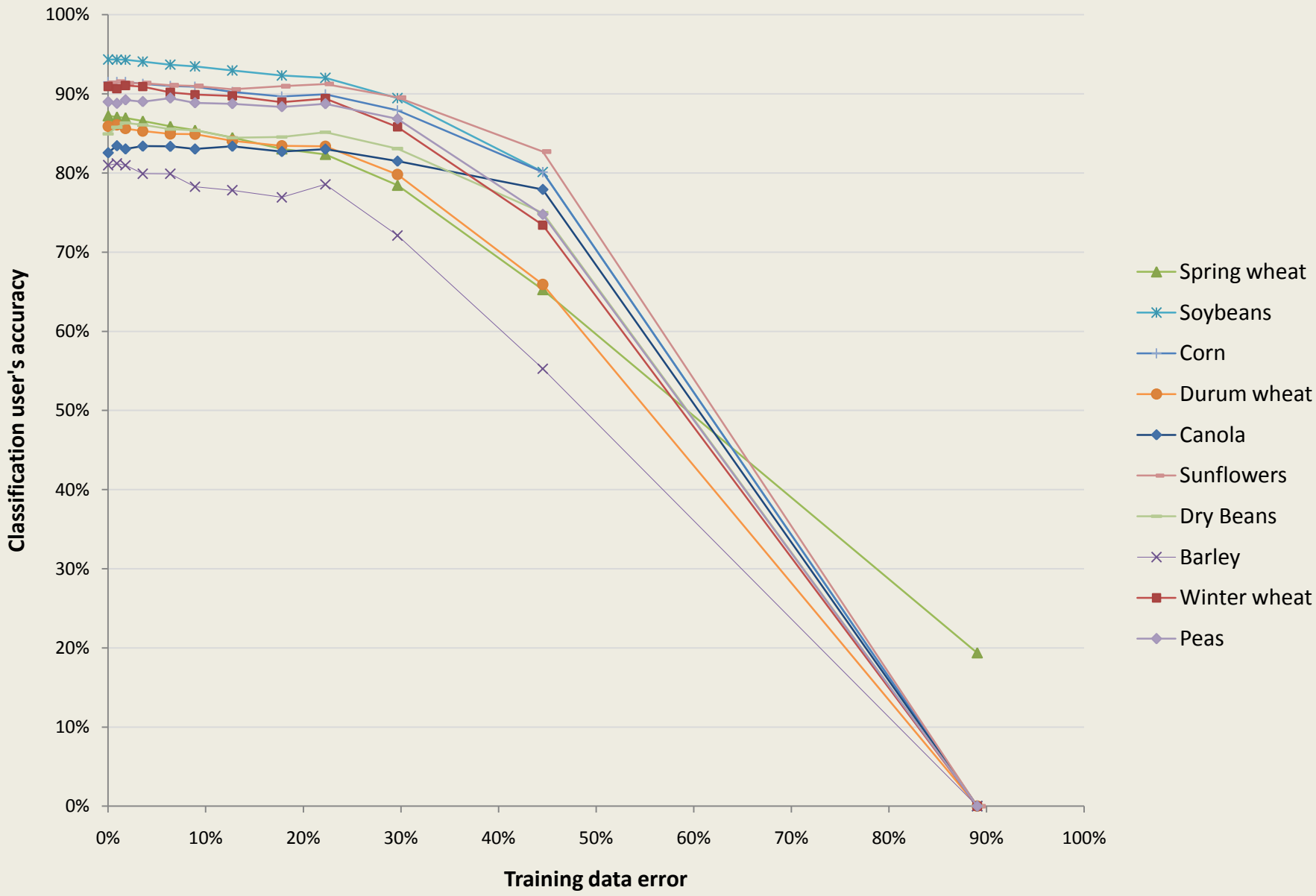


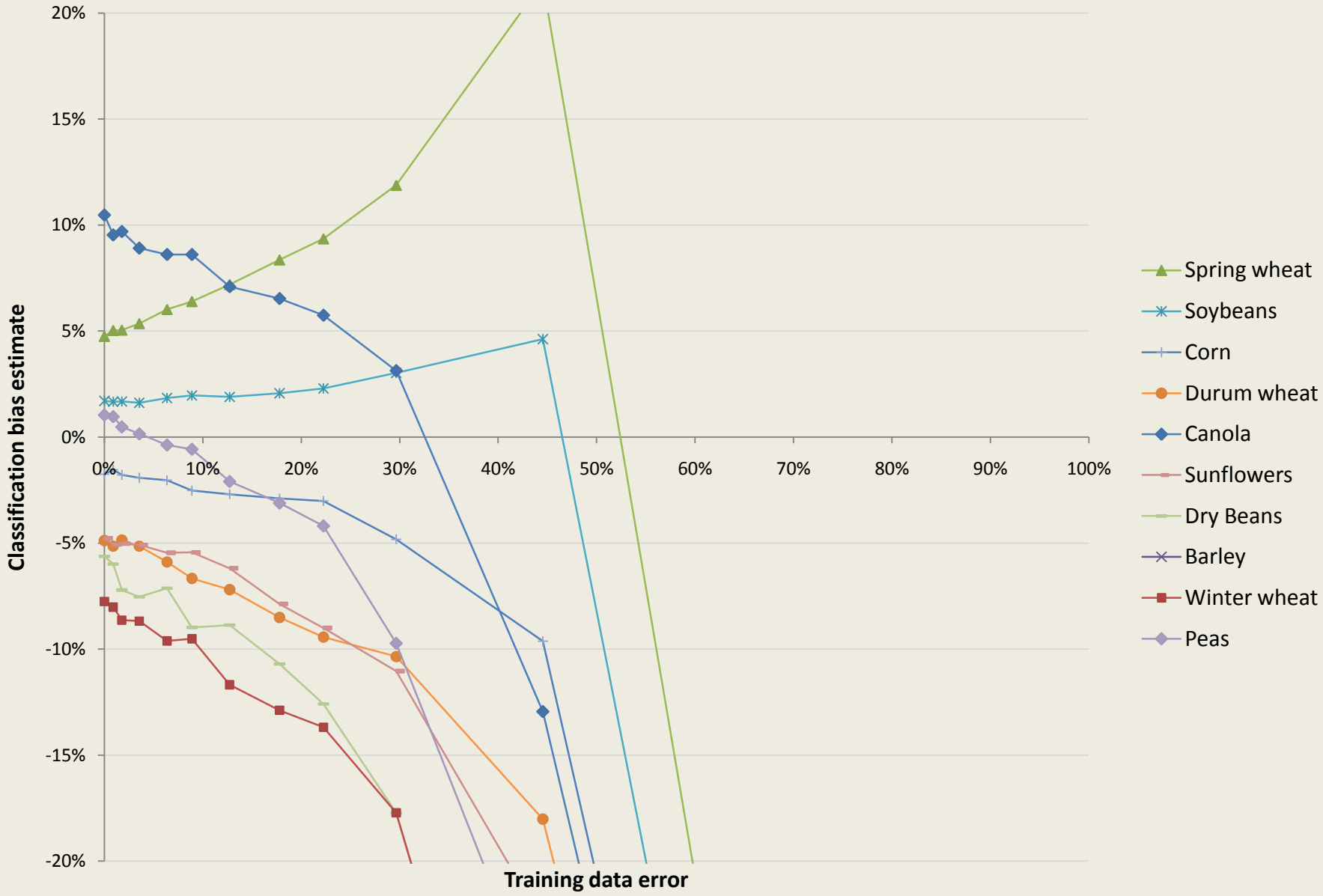
Total scene has 58,388,946 pixels, 737,633 (1.3%) chosen for training

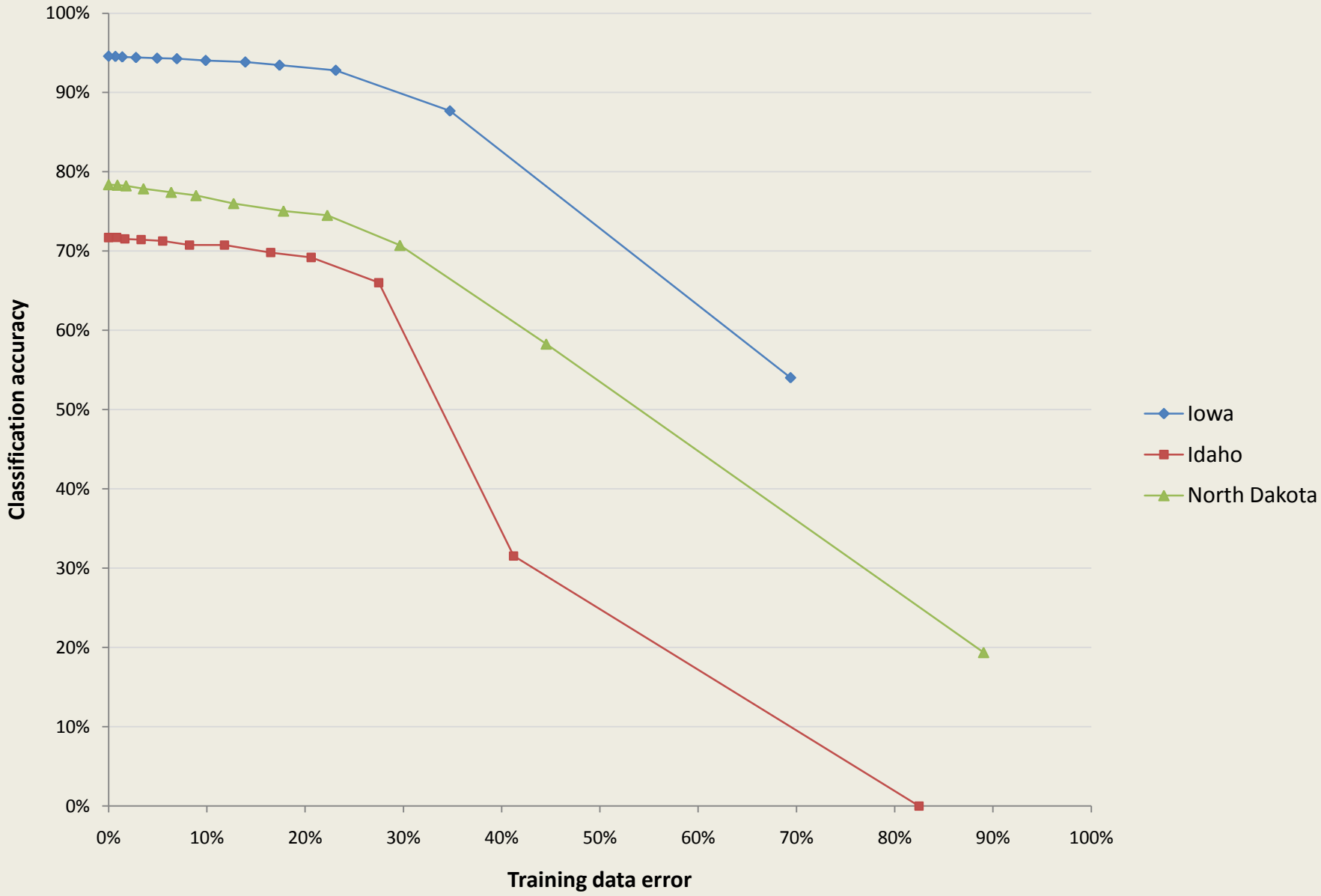

North Dakota '09 CDL, Classification accuracy v. training data error
Classification accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Crop classes only

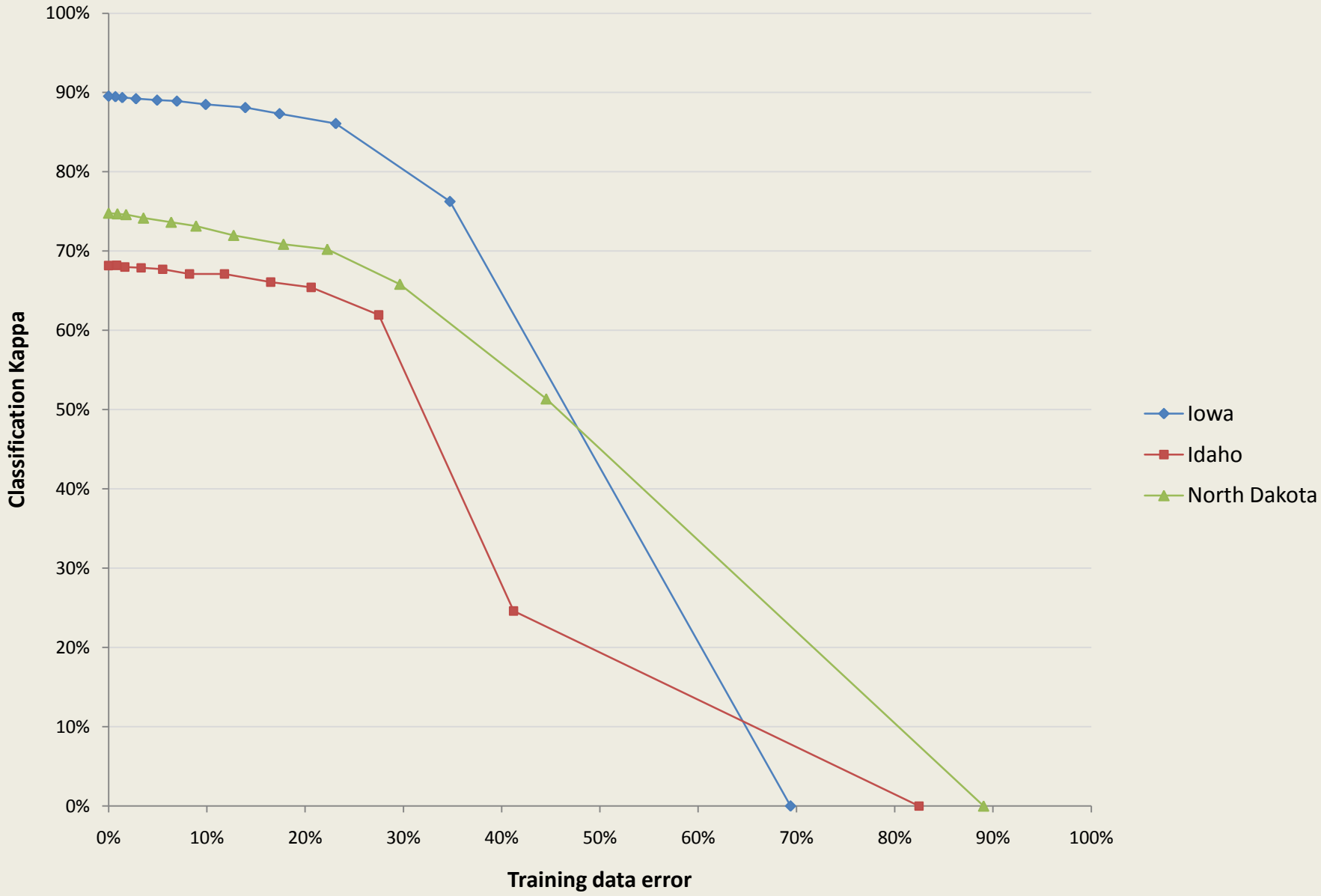

North Dakota '09 CDL, Classification Kappa v. training data error
Classification Kappa
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Crop classes only

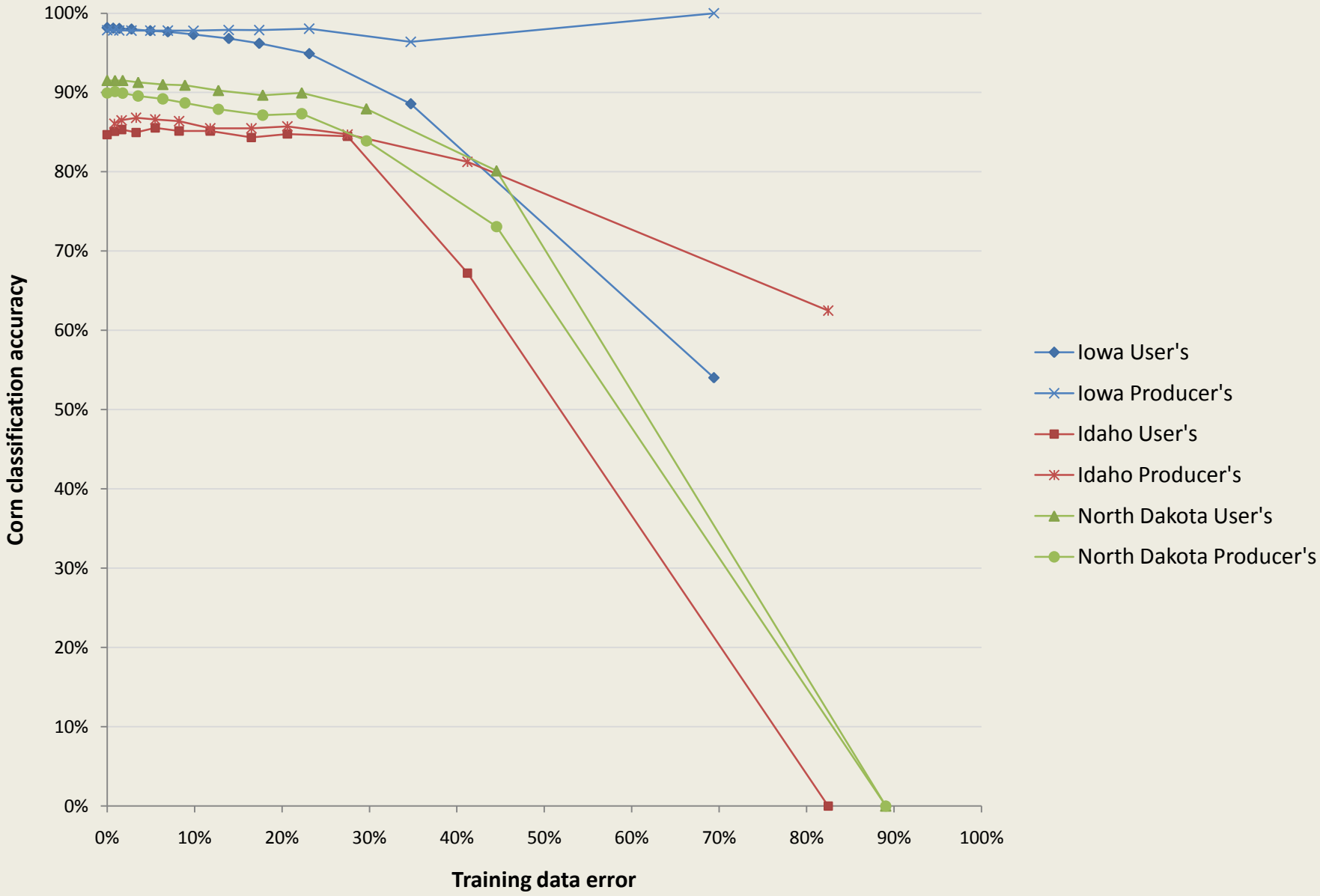

North Dakota '09 CDL, Classification producer's accuracy v. training data error
Classification producer's accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Spring wheat
Soybeans
Corn
Durum wheat
Canola
Sunflowers
Dry Beans
Barley
Winter wheat
Peas


North Dakota '09 CDL, Classification user's accuracy v. training data error
Classification user's accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Spring wheat
Soybeans
Corn
Durum wheat
Canola
Sunflowers
Dry Beans
Barley
Winter wheat
Peas


North Dakota '09 CDL, Classification bias v. training data error
Classification bias estimate
20%
15%
10%
5%
0%
-5%
-10%
-15%
-20%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Spring wheat
Soybeans
Corn
Durum wheat
Canola
Sunflowers
Dry Beans
Barley
Winter wheat
Peas


'09 CDL, Cropland classification accuracy v. training data error
Classification accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Iowa
Idaho
North Dakota


'09 CDL, Cropland classification Kappa v. training data error
Classification Kappa
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Iowa
Idaho
North Dakota


'09 CDL, Corn classification accuracy v. training data error
Corn classification accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Iowa User's
Iowa Producer's
Idaho User's
Idaho Producer's
North Dakota User's
North Dakota Producer's


'09 CDL, Corn classification bias v. training data error
Classification bias estimate
20%
15%
10%
5%
0%
-5%
-10%
-15%
-20%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Training data error
Iowa
Idaho
North Dakota

Conclusions

- Degradation of training data…..
  - degrades the classification.
  - has relatively modest impacts on the classification until more than roughly 25% of training data is in error (then it falls rapidly, and thus is not linear).
  - hurts the classification more when lots of classes are present.
  - never improves a classification.
  - impacts differently the areal bias of categories within the classification.

dave_johnson@nass.usda.gov
www.nass.usda.gov