Re-sampling Remotely Sensed Data to Improve National and Regional Mapping of Forest Conditions with Confidential Field Data

Raymond L. Czaplewski

USDA Forest Service

Forest Service Research and Development (R&D) and State and Private Forestry Deputy Areas, in partnership with the National Forest System Remote Sensing Applications Center (RSAC), built a 250-m resolution (6.25-ha pixel) dataset for the entire USA. It assembles multi-seasonal hyperspectral MODIS data and derivatives, Landsat derivatives (i.e., summary statistics for the 70 or so Landsat pixels within each 6.25-ha pixel), and other national geospatial datasets (e.g., climate, soils, ecoregions). All datasets are re-sampled to the same 250-m grid used for standard MODIS derivatives. There are approximately 100 layers in this "stack" of full-coverage geospatial data.

The unique value of this stack comes from integration with field data from the R&D Forest Inventory and Analysis (FIA) program. Field data from each of approximately 160,000 forested plots are associated with the corresponding pixel in the stack of geospatial data. Each field plot represents approximately 1-ha. FIA data are conveniently available to train empirical classification and regression models that use geospatial data as predictor variables, and field measurements as the response variables.

The Forest Service has two short-term objectives: (1) develop a forest biomass map for the USA; and (2) develop a detailed tree-species composition map as input to national forest risk mapping. Longer-term objectives might include a continuous development of additional geospatial data and model predictions using a shared geospatial data infrastructure. All products recognize the random errors in predicting detailed field conditions at the pixel level with remotely sensed data.

The first limitation of the stack comes from the need to keep locations of FIA field plots confidential, which is legislated by the Food Security Act of 1985 (as amended by PL 106-113). Since the stack could be used locate the 6.25-ha pixel containing the plurality of a FIA field plot, public release of the stack is prohibited by Public Law.

The second limitation of the stack is the substantial scale-mismatch between the 0.40-ha FIA field data and the 6.25-ha geospatial pixel data; field measurements only characterize a small portion of a 6.25-ha pixel, and it is likely that the field data do not realistically represent the field conditions of many 6.25-ha pixels. The field plot is essentially a sample size of n=1 for its surrounding pixel.

One partial solution to the second limitation is re-sampling to a smaller pixel size, such as a 150-m (2.25-ha) pixel. Although not a completely satisfactory solution to the scale-mismatch, this pixel size improves the match between the scales of field data and remotely sensed data, while considering the reality of miss-registration caused by errors in measuring the exact position of each field plot (e.g., random GPS errors). For example, assume a positional error of 30-m in a random direction; the full 1-ha FIA field plot could exist anywhere within a 2-ha circular area. Data volumes with a 2.25-ha pixel increase modestly relative to a 6.25-ha MODIS pixel, but data volumes remain much smaller relative to a 0.1-ha Landsat pixel database.

The third limitation of the stack is miss-registration error caused by a re-sampling grid that is misaligned on the field plot locations. For example, a 1-ha field plot could straddle up to four contiguous 2.25-ha pixels. The current practice picks the single pixel from the full-coverage re-sampling grid that covers the largest portion of the field plot, even though part of the plot, sometimes the majority of the plot, could be outside of that pixel.

One solution to the third limitation is to resample a 2.25-ha area centered on the best-known position of each field plot. This area would rarely, if ever, be identical to the closest 2.25-ha pixel on the full-coverage re-sampling grid. This kind of re-sampling would be independently repeated for each field plot in the training dataset. Statistical prediction models would be fit with the geospatial predictor variables from 2.25-ha pixels centered on each field plot. The resulting prediction models would then be applied to the full-coverage re-sampled grid with 2.25-ha pixels.

The proposed solution to the third limitation has an extremely beneficial consequence. It might allow the Forest Service to offer its stack of remotely sensed data and corresponding field data, without violating plot confidentiality legislation. This is a partial solution to the first limitation of the stack. It is very unlikely that any of the four 2.25-ha pixels on the full-coverage re-sampling grid would have exactly the same multivariate geospatial data, or "fingerprint", as the one 2.25-ha pixel that is centered on the field plot and that straddles those four pixels. The untested expectation is that the remotely sensed fingerprint of a field plot is unique within the immediate vicinity of that plot, and perhaps unique within the full stack of remotely sensed data. Inclusion of re-sampled 0.10-ha Landsat pixels would make the match even more unlikely. This might be sufficient to adequately obscure the position of a field plot within the stack. If this expectation proves valid, unpopular methods such "fuzz-and-swap" and adding random noise to geospatial predictor variables are unnecessary to protect plot confidentiality. The limitation of this solution is that these training data would only apply to the national stack of geospatial data created by the Forest Service.

The hope is that the Forest Service will (1) build a new full-coverage stack of remotely sensed and other geospatial data based on re-sampling to 150-m (2.25-ha) pixel size; (2) build a new training dataset where geospatial data are re-sampled to a 2.25-ha area centered on the position of each field plot and matched with detailed field data; and (3) offer the two integrated national datasets for public use. This would be possible if the solutions described above protect confidentiality of FIA plot locations.

A resampling solution for public distribution of exact FIA training data with 150-m pixel data

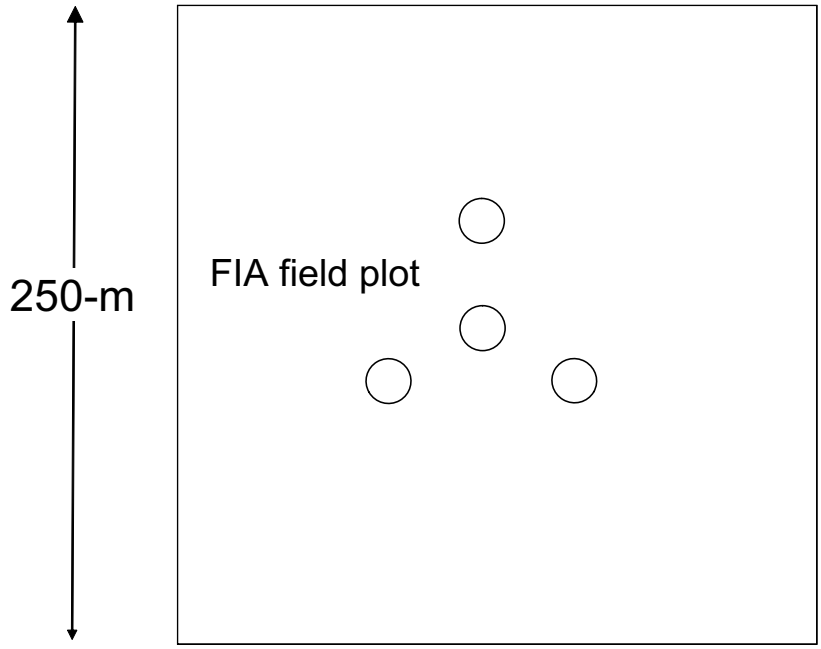
FHTET Workshop: *EVALUATION OF QUANTITATIVE TECHNIQUES FOR DERIVING NATIONAL SCALE DATA FOR ASSESSING AND MAPPING RISK*

Ray Czaplewski
USDA Forest Service
Rocky Mountain Research Station
Fort Collins, CO

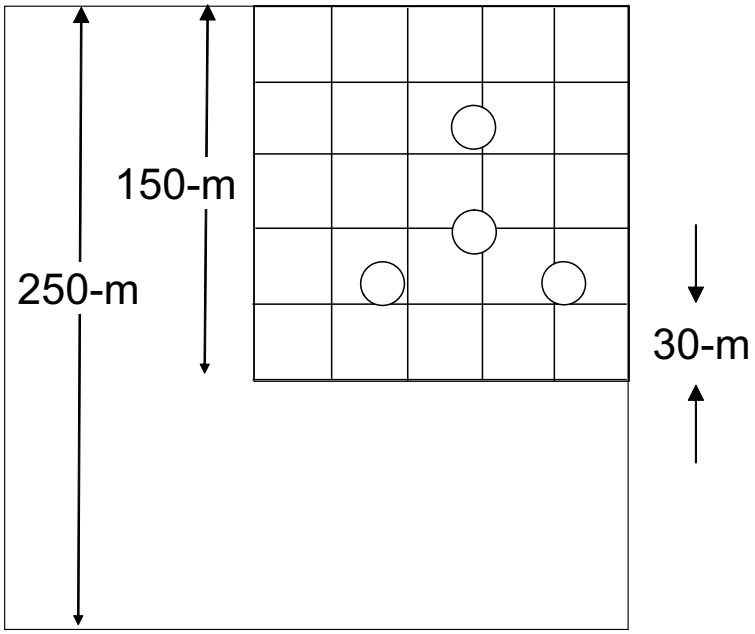
Forest Service national pixel database for use with FIA field plots

- FIA/RSAC "stack" of remotely sensed data already available based on 250-m data
  - Multi-season time-series MODIS composites (phenology) resampled to common 250-m grid
  - 30-m (Landsat) MRLC-NLCD land cover data and DEM data resampled to 250-m MODIS grid
  - Coarse-scale geospatial data (e.g., STATSGO soil polygons, 1-km climate data, ecoregions) resampled to same 250-m MODIS grid
  - Data volume well suited for national mapping

Problem 1: Scale mis-match between 1-acre FIA plot and 250-m (16-acre) MODIS pixel

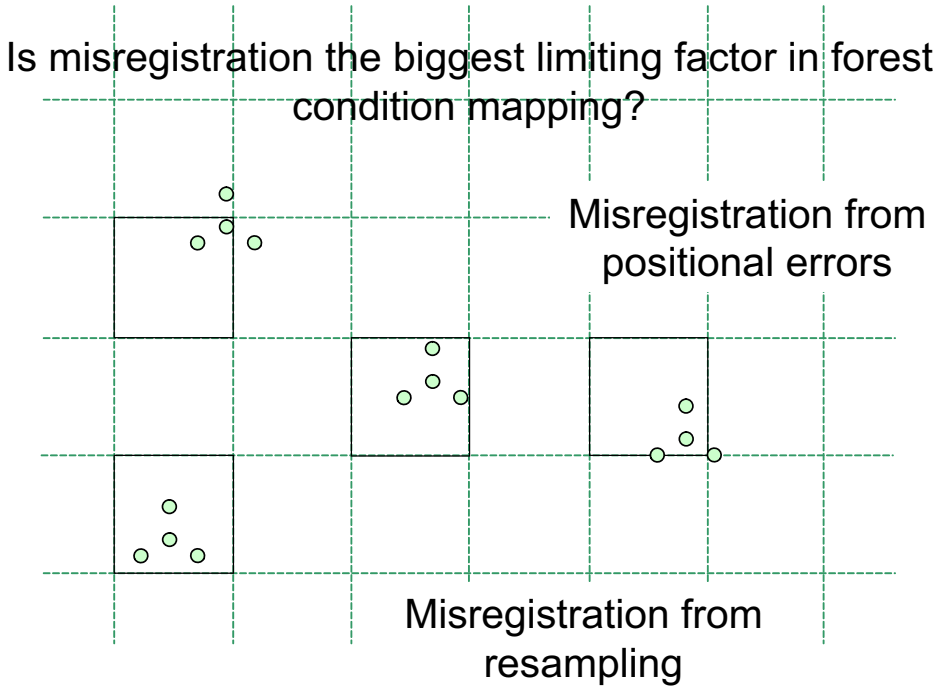


Proposal 1: Build national "stack" of geospatial data resampled to 150-m (5-acre) scale



Proposal 1: Build national "stack" of geospatial data resampled to 150-m (5-acre) scale

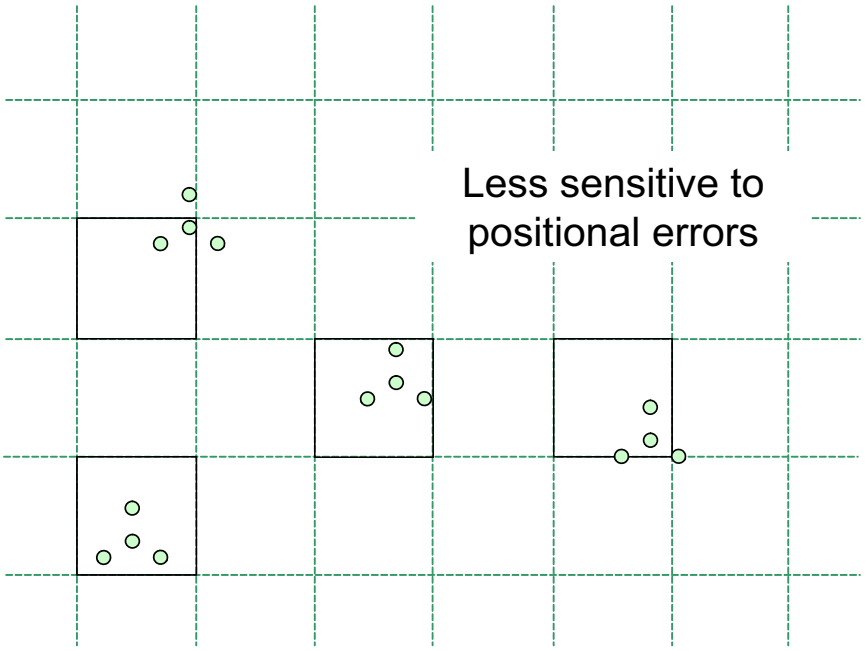
- Reduce scale mis-match between field plot and pixel size
- Realistically accommodate registration error between field data and pixel data



Training sites randomly located relative to pixel boundaries

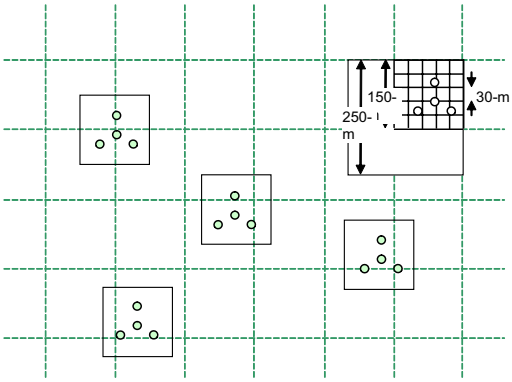
Proposal 2: Resample 150-m pixel data separately for training plots

- Improve registration between pixel data and training plots



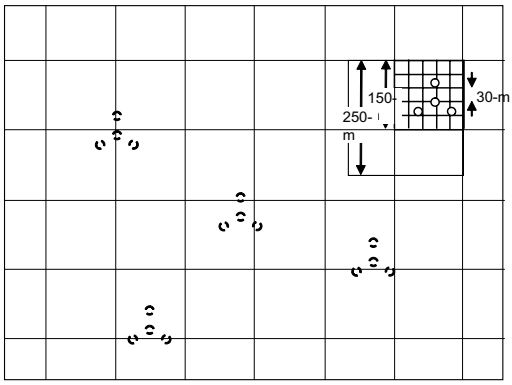
Independently resample remotely sensed data centered on each training site

A. Re-sampling frame for training data and model construction



Independently center a 150-m pixel on best known geospatial location of FIA field plot to minimize effects of misregisration between remotely sensed and field data

B. Re-sampling frame for wall-to-wall mapping with models from step A



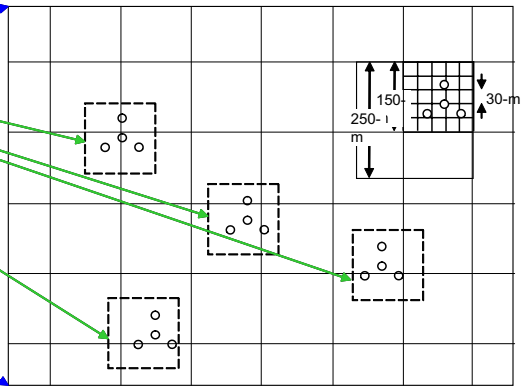
Apply models built in from training pixels in (**A**) to wall-to-wall coverage of geospatial data, acknowledging that no single pixel grid will ever simultaneously match well with field plots

Potential advantages

1. The 150-m pixel size would more realistically match the scale of the field data, plus a little slop to acknowledge inevitable mis-regisration error
2. The 150-m pixel size would reduce data volumes to (30x30)/(25x25)=0.04 times smaller than the size of corresponding Landsat files, and only (250x250)/(150x150)=1.67 times larger than corresponding MODIS files.

Unexpected advantage

3. Any plot in training dataset in (A) could be very difficult to find in image file (B) because of differences caused by re-sampling.



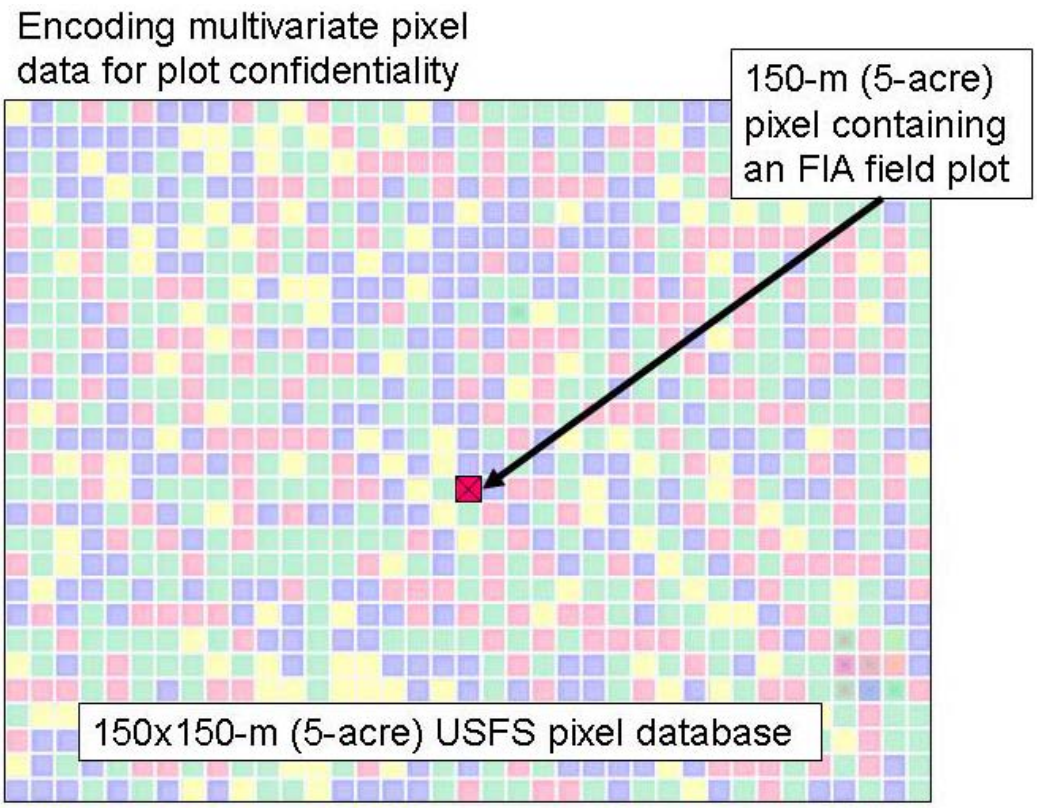
Therefore, datasets (A) and (B) potentially could be made publicly available without risking disclosure of exact plot locations. This might eliminate need for “fuzz and swap”.

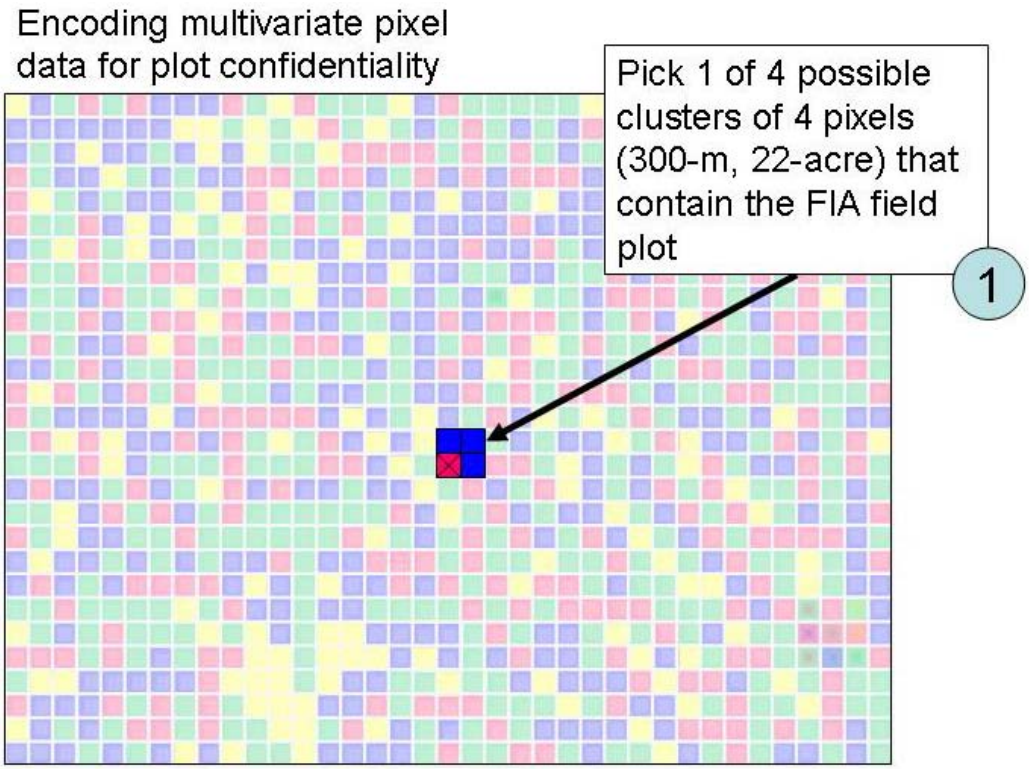
Encoding to further protect plot confidentiality

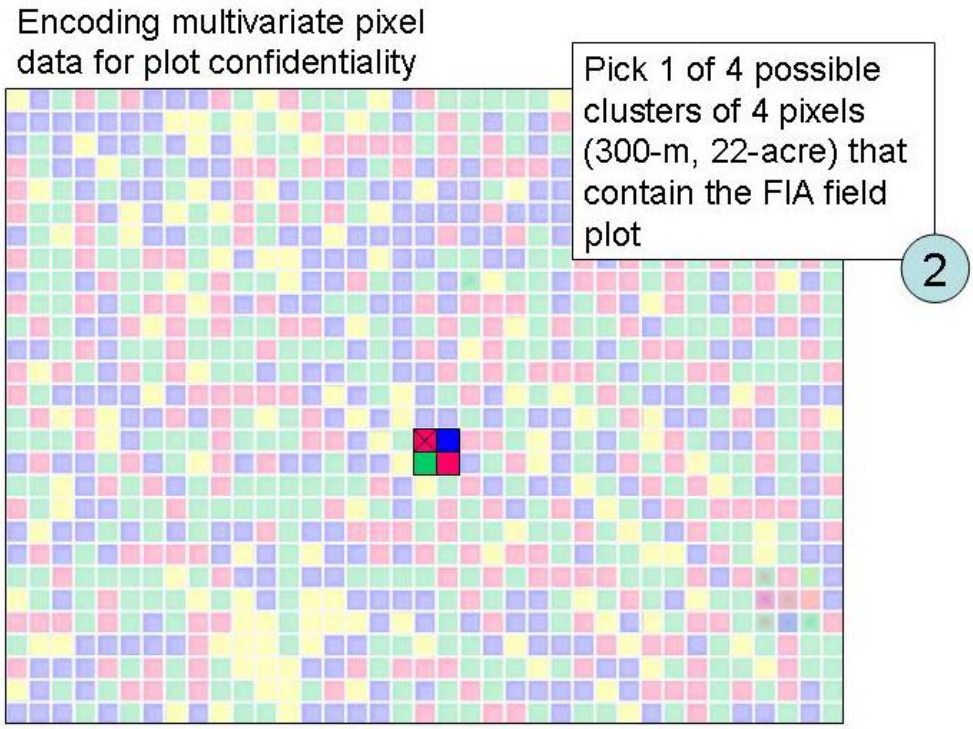
Worst case: a 5-acre pixel near an FIA training plot has a very rare spectral signature that might be located in a full image

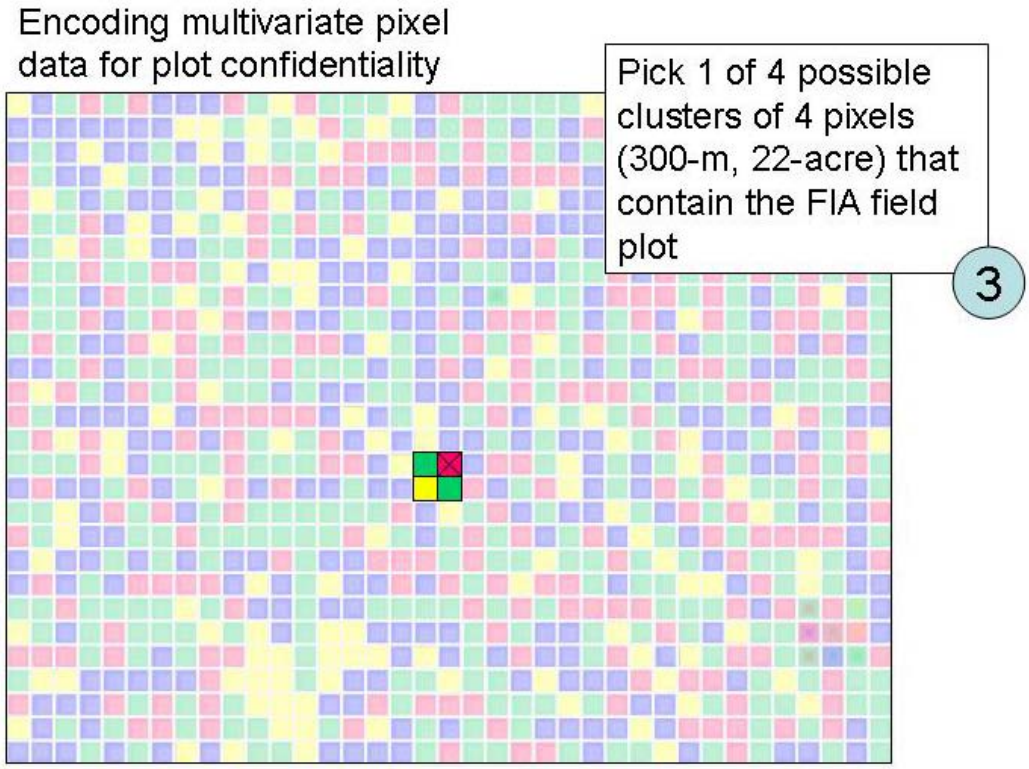
Option 1: Add random error to spectral data to pixel in training data (degrades models predicting forest conditions)

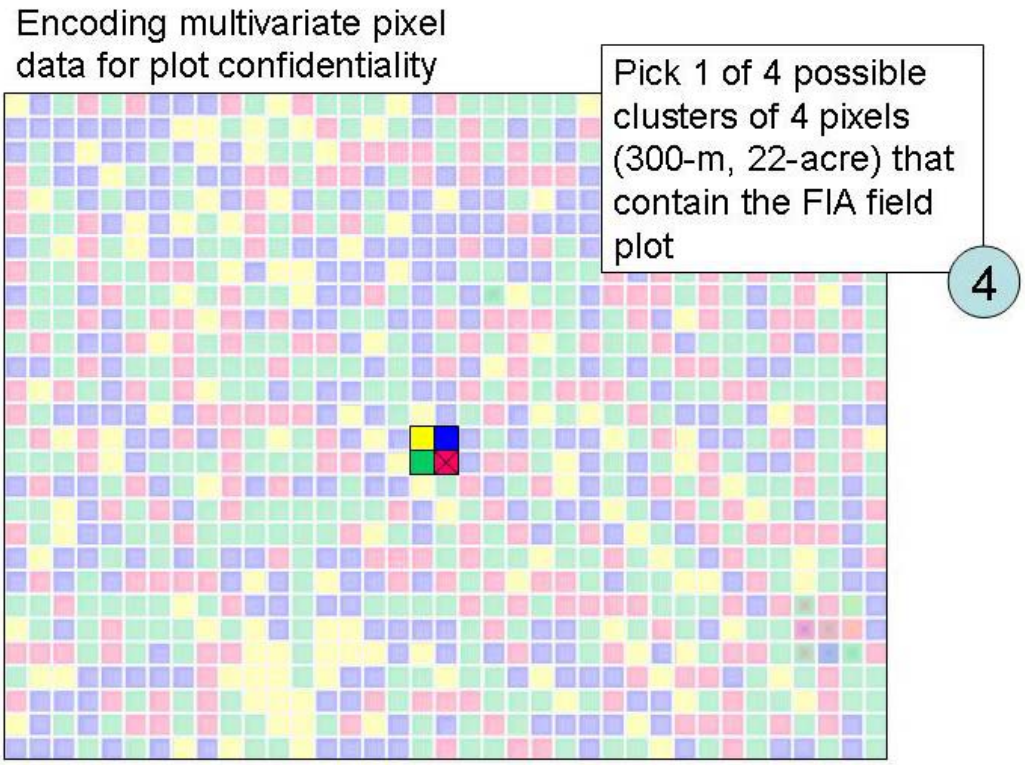
Option 2: Add randomness to a relatively small number of pixels, but not to spectral data for a training pixel

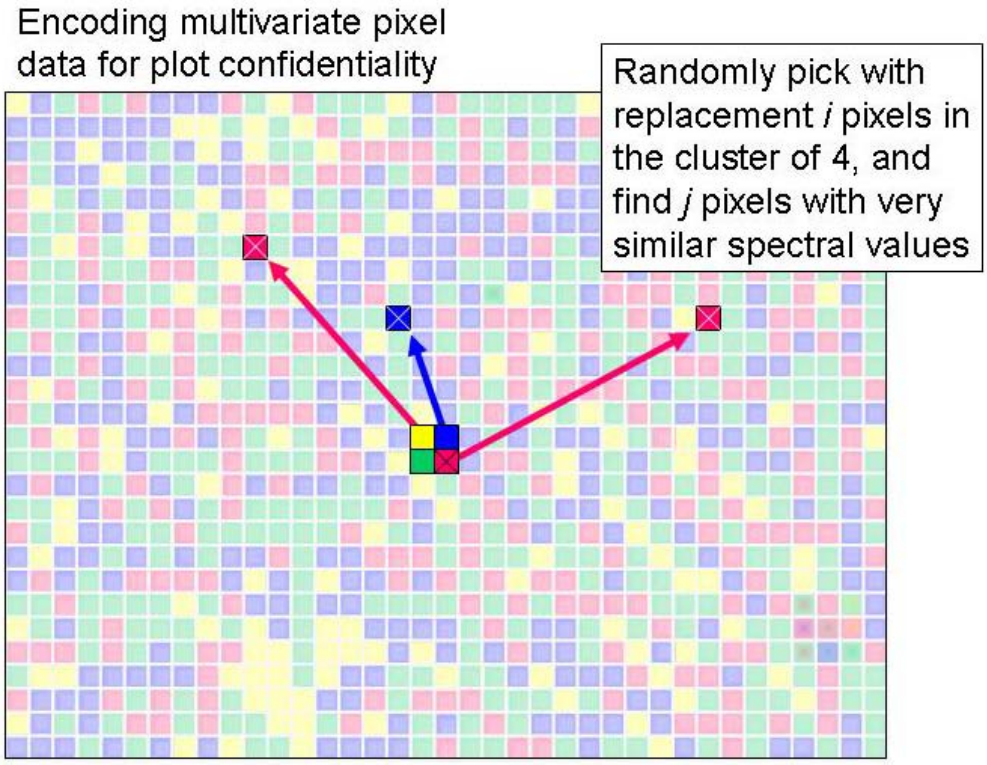

Encoding multivariate pixel data for plot confidentiality
150-m (5-acre) pixel containing an FIA field plot
150x150-m (5-acre) USFS pixel database

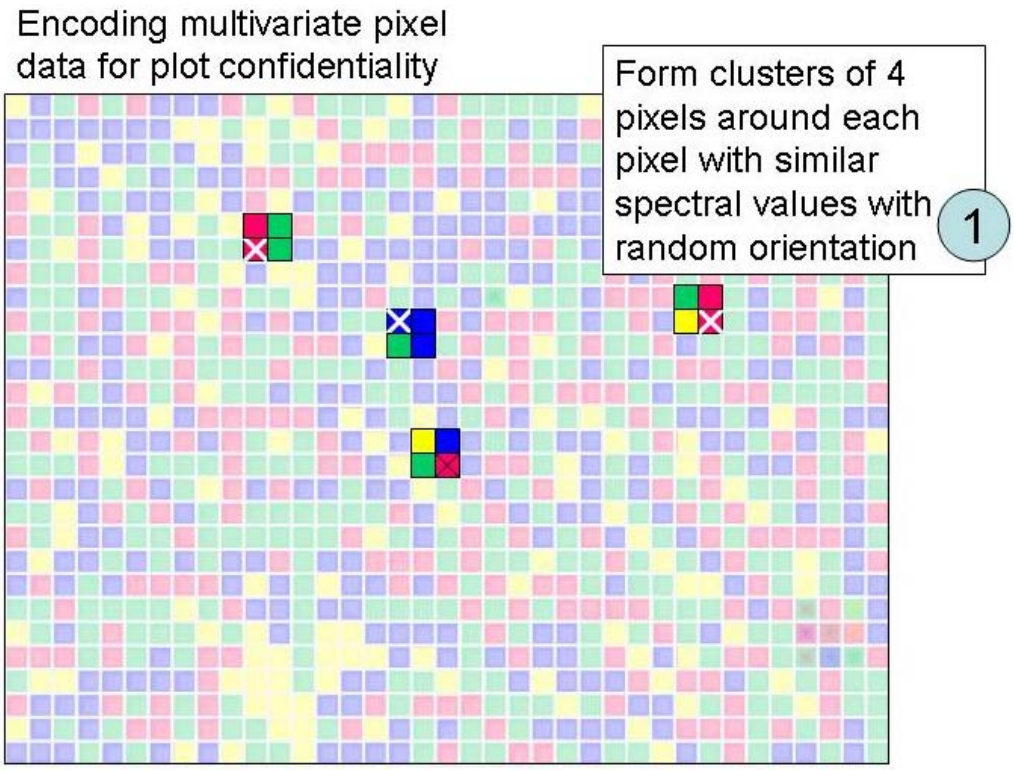

Encoding multivariate pixel data for plot confidentiality
Pick 1 of 4 possible clusters of 4 pixels (300-m, 22-acre) that contain the FIA field plot
1

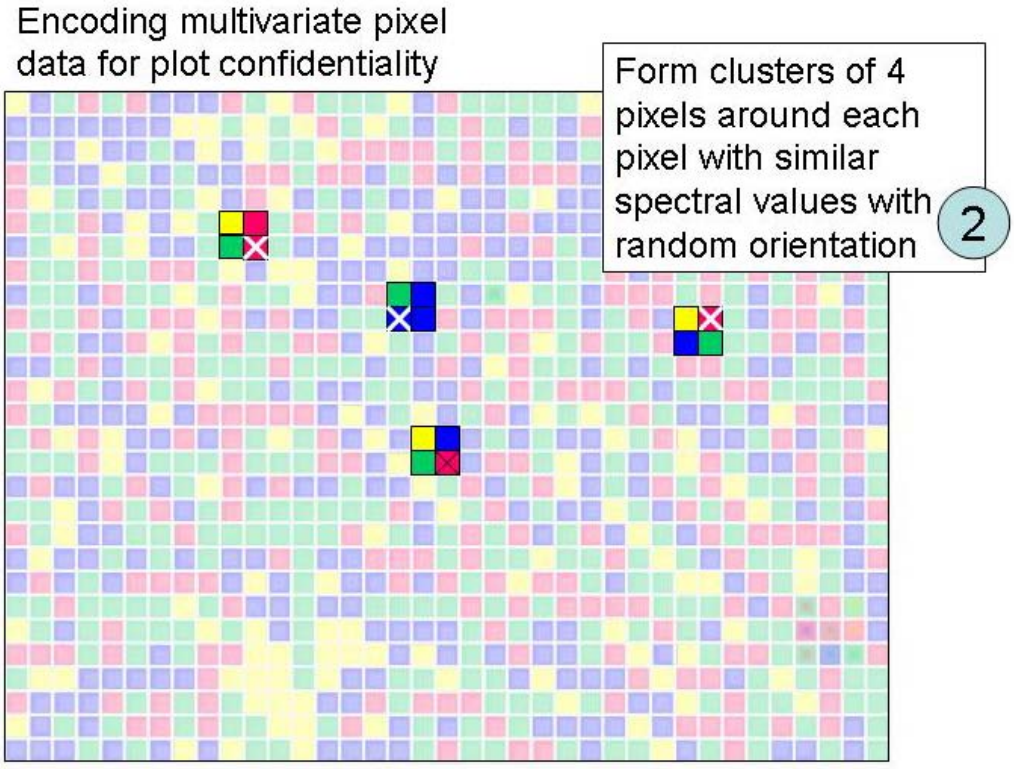

Encoding multivariate pixel data for plot confidentiality
Pick 1 of 4 possible clusters of 4 pixels (300-m, 22-acre) that contain the FIA field plot
2

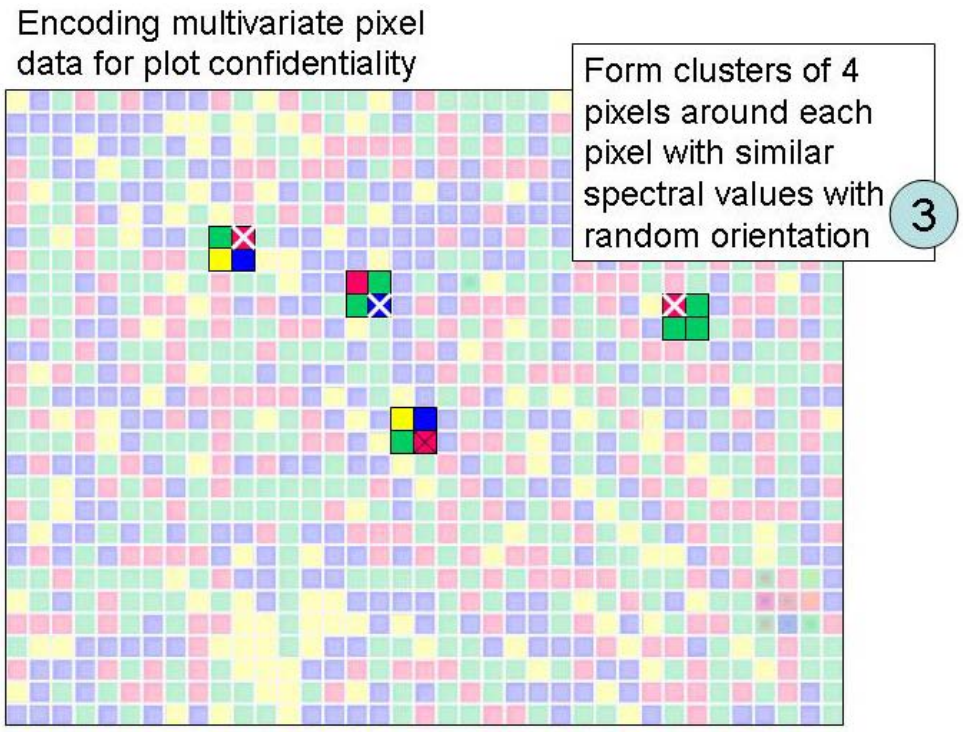

Encoding multivariate pixel data for plot confidentiality
Pick 1 of 4 possible clusters of 4 pixels (300-m, 22-acre) that contain the FIA field plot
3

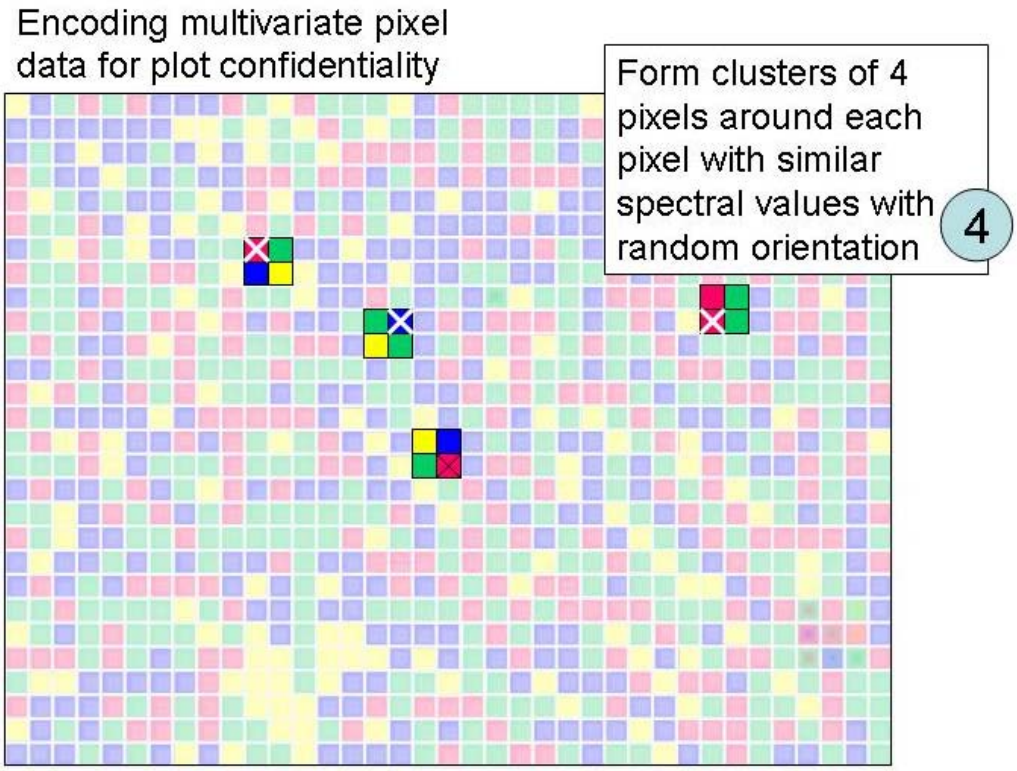

Encoding multivariate pixel data for plot confidentiality
Pick 1 of 4 possible clusters of 4 pixels (300-m, 22-acre) that contain the FIA field plot
4

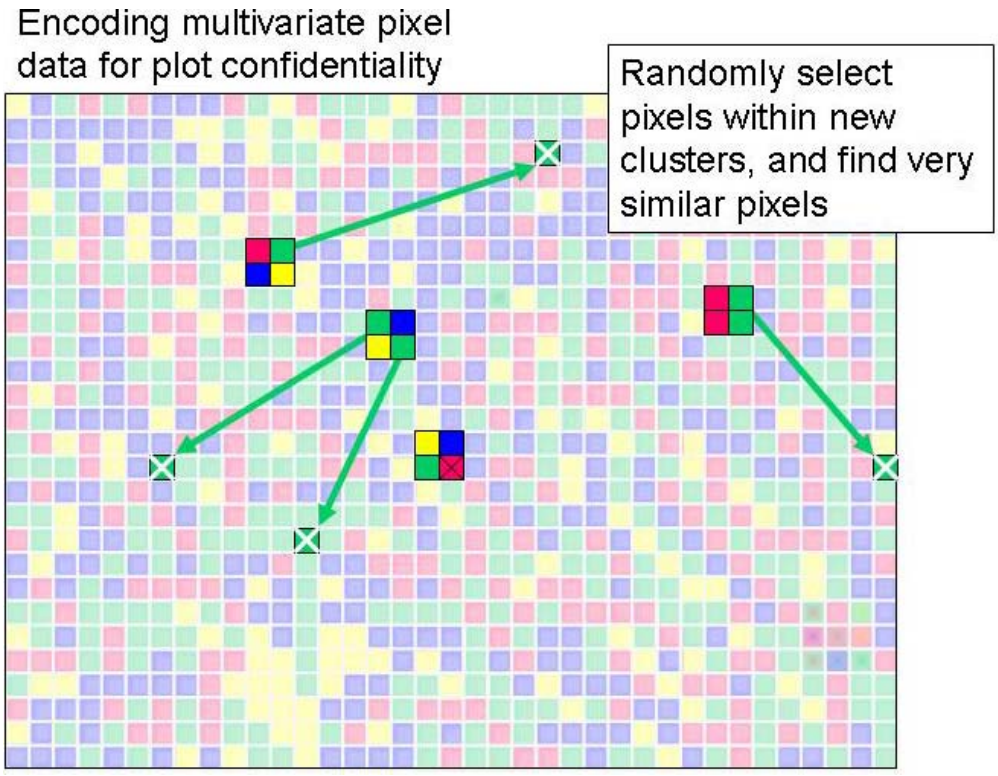

Encoding multivariate pixel data for plot confidentiality
Randomly pick with replacement *i* pixels in the cluster of 4, and find *j* pixels with very similar spectral values

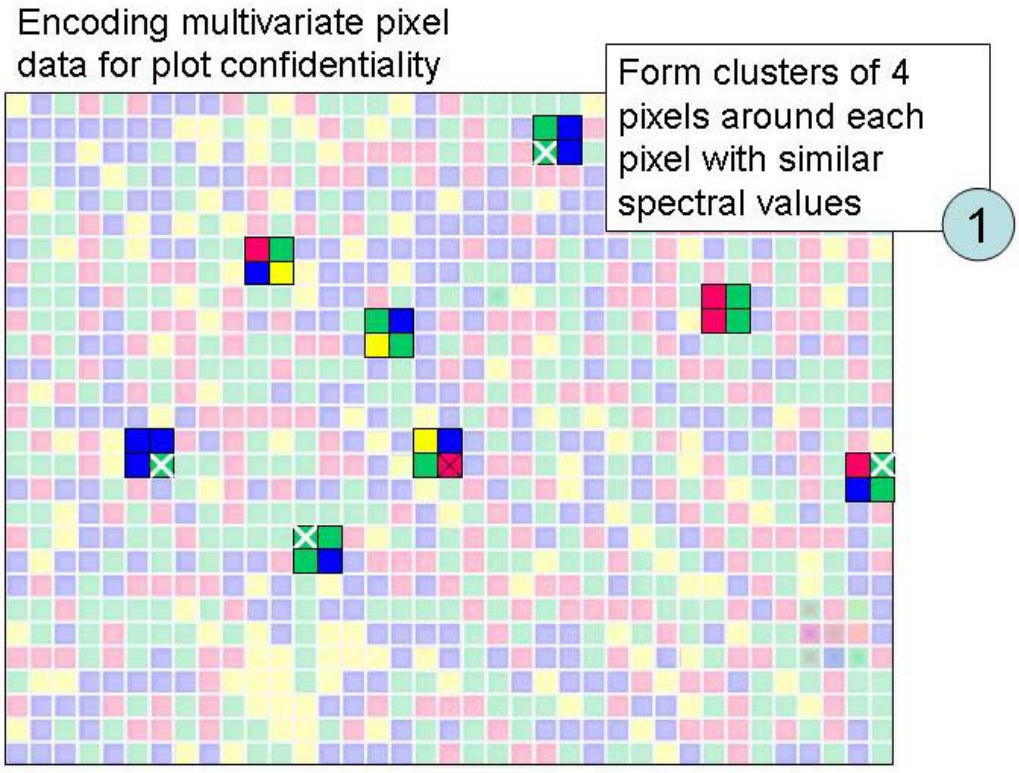

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
1

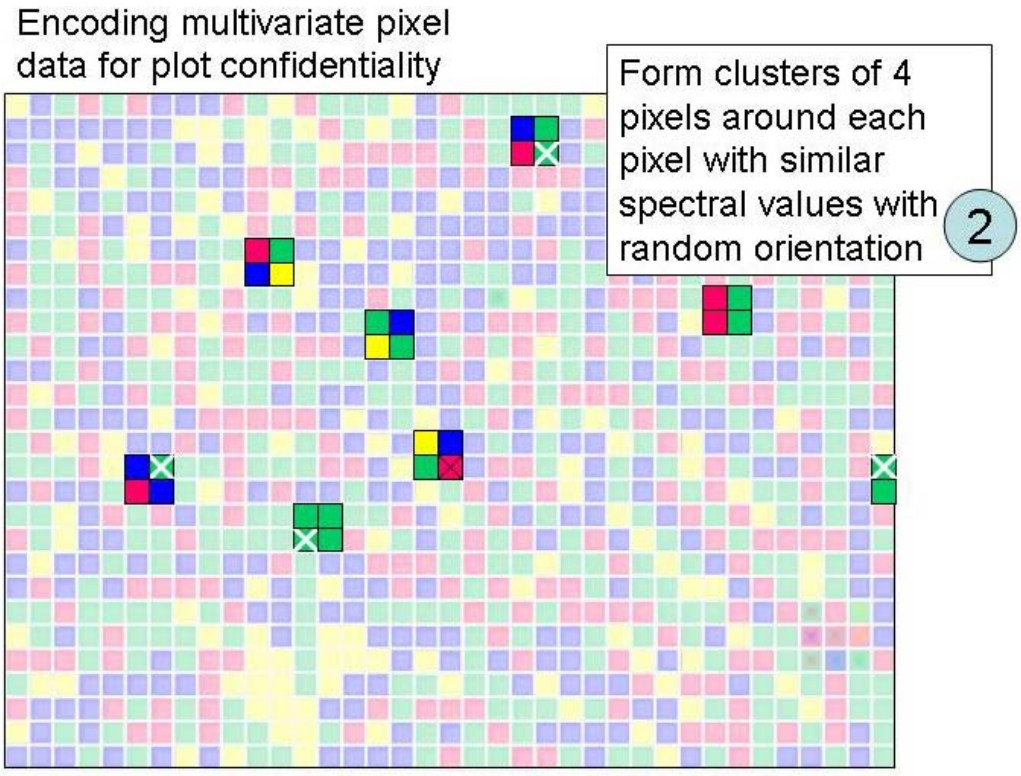

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
2

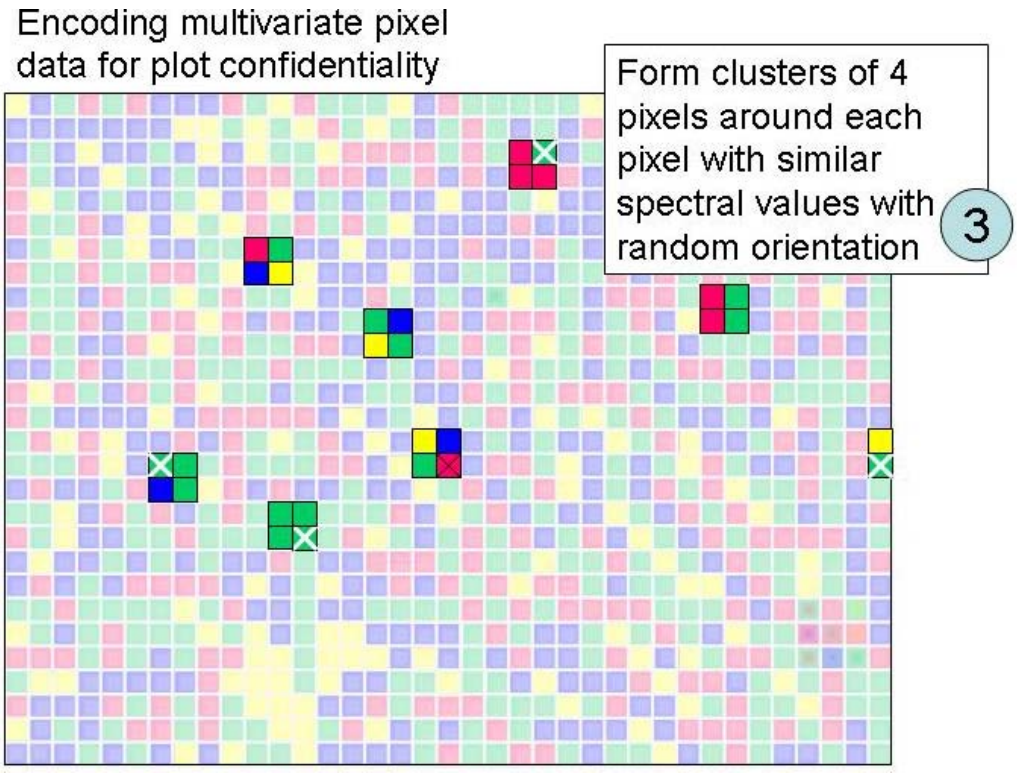

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
3

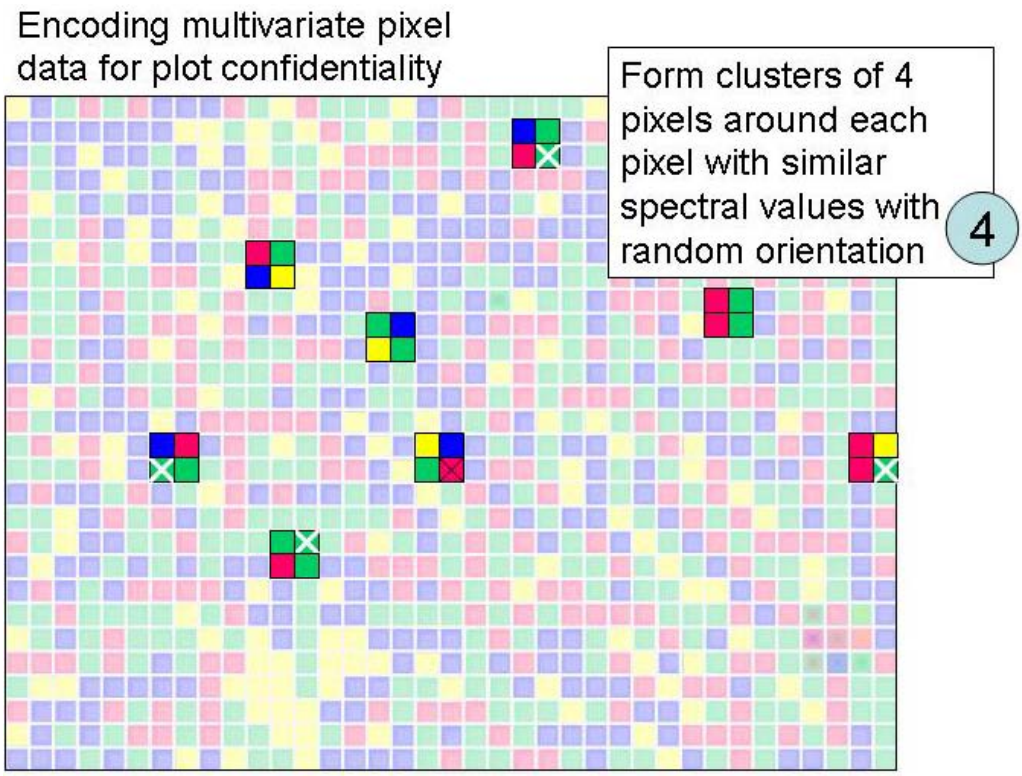

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
4

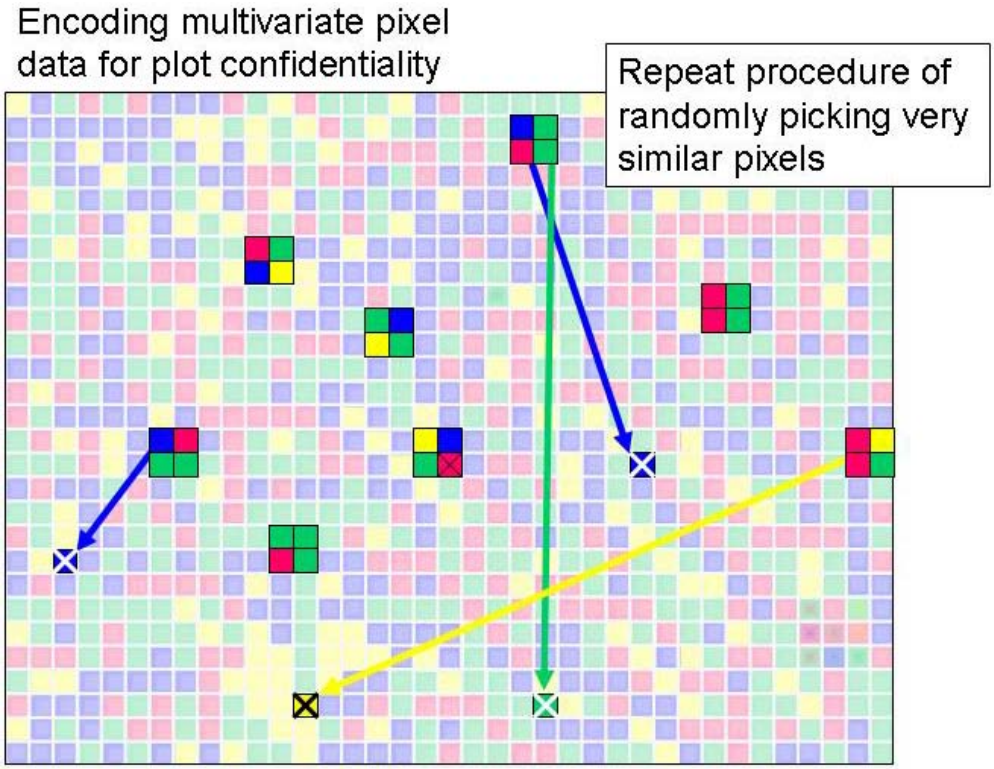

Encoding multivariate pixel data for plot confidentiality
Randomly select pixels within new clusters, and find very similar pixels

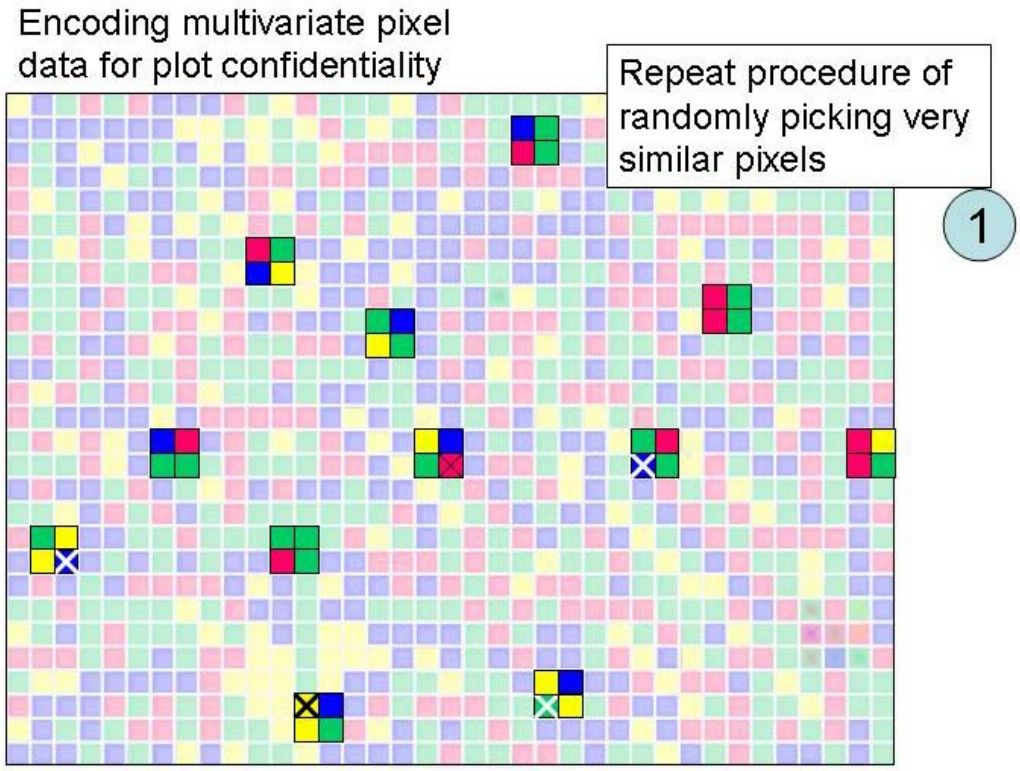

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values
1


Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
2

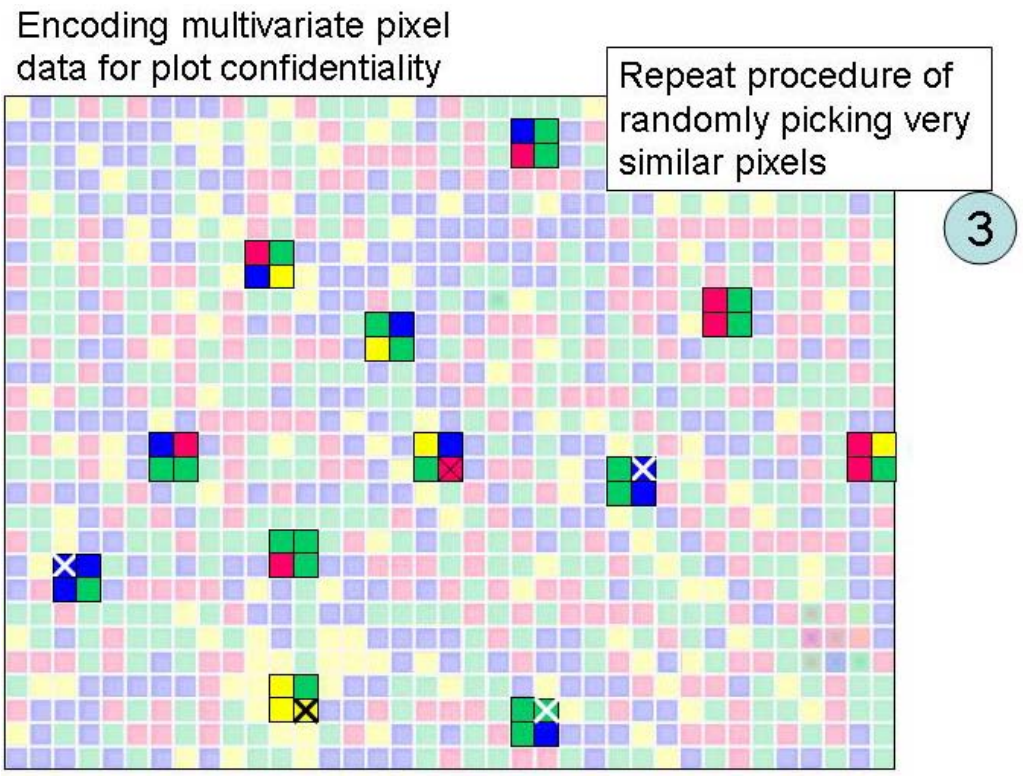

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
3

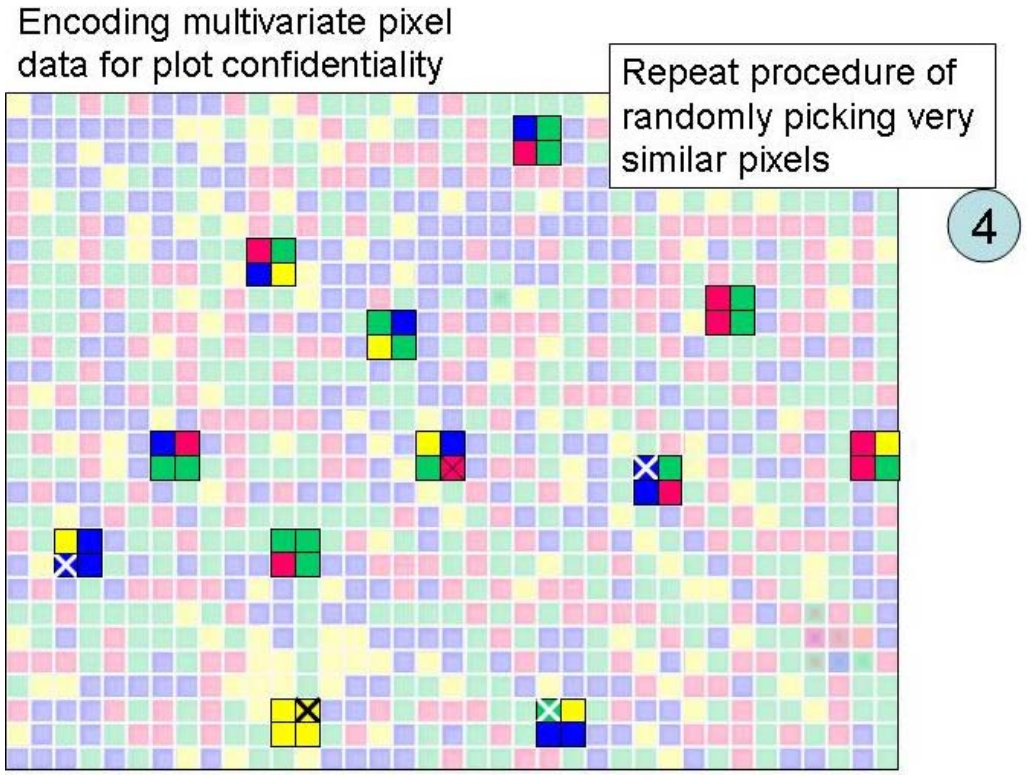

Encoding multivariate pixel data for plot confidentiality
Form clusters of 4 pixels around each pixel with similar spectral values with random orientation
4

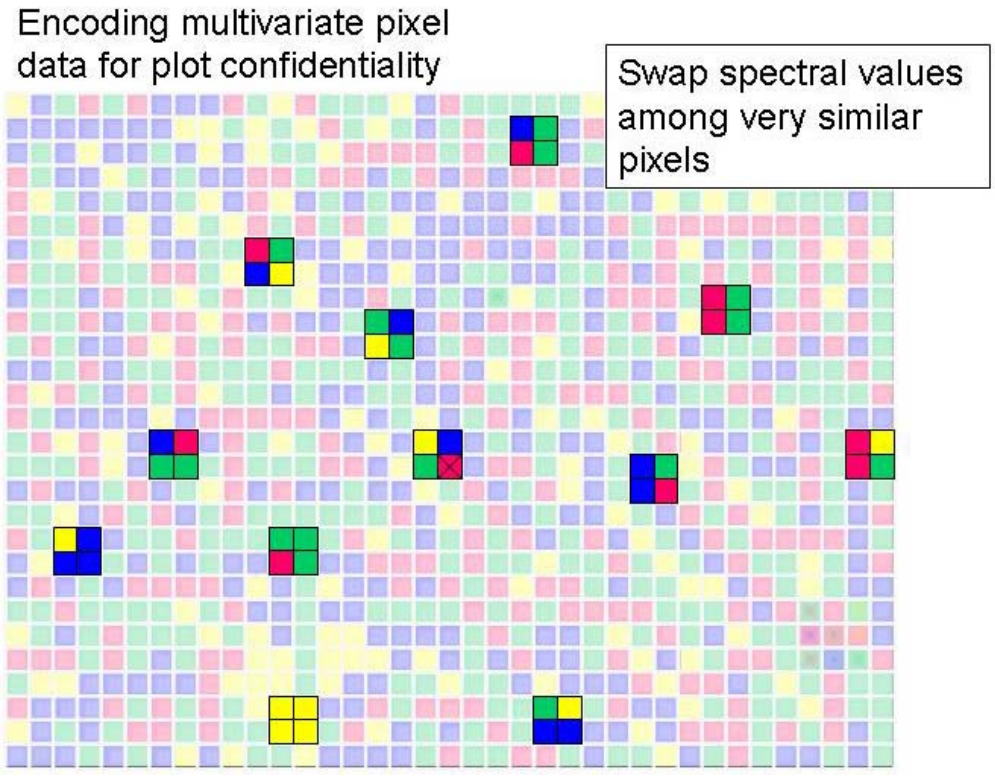

Encoding multivariate pixel data for plot confidentiality
Repeat procedure of randomly picking very similar pixels


Encoding multivariate pixel data for plot confidentiality
Repeat procedure of randomly picking very similar pixels
1

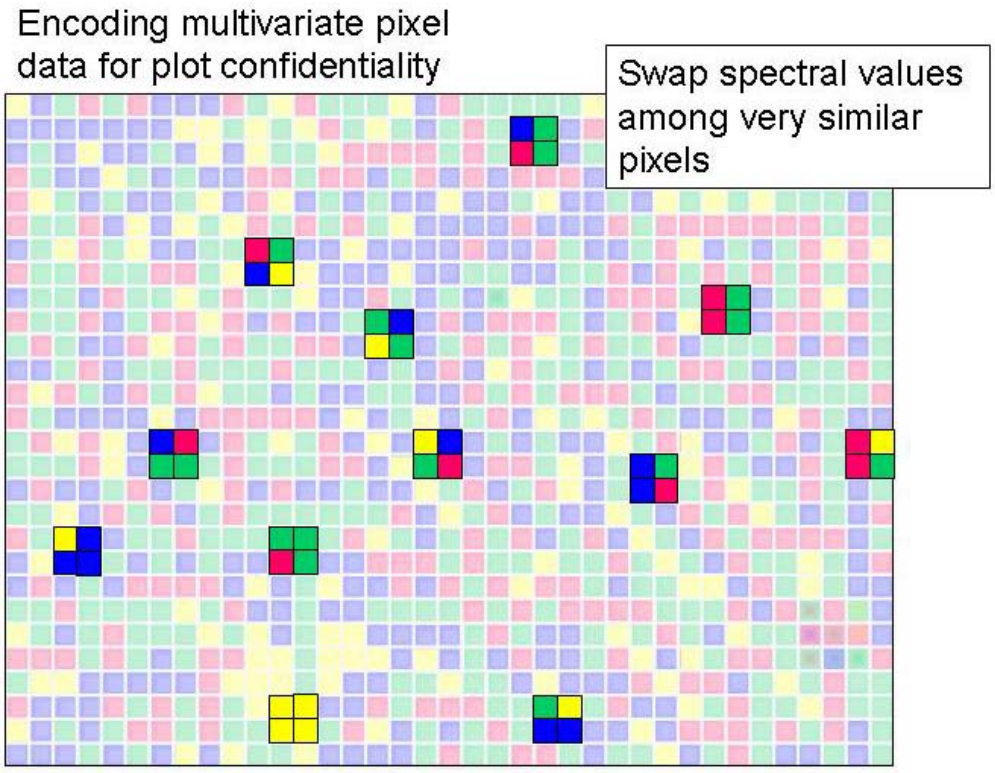

Encoding multivariate pixel data for plot confidentiality
Repeat procedure of randomly picking very similar pixels
2

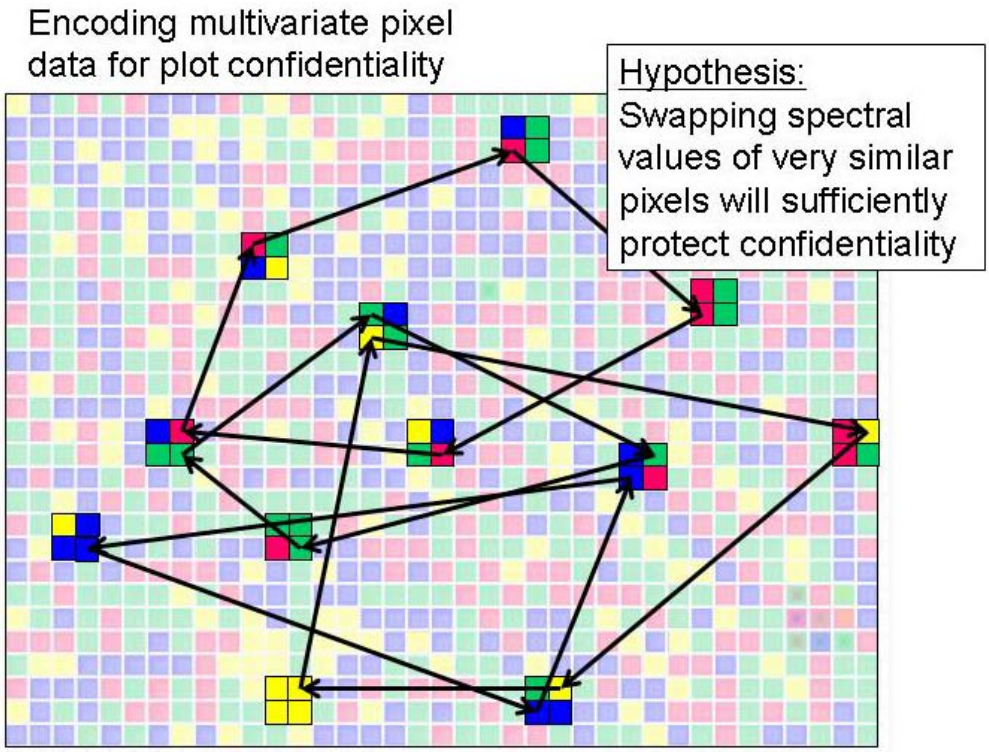

Encoding multivariate pixel data for plot confidentiality
Repeat procedure of randomly picking very similar pixels
3

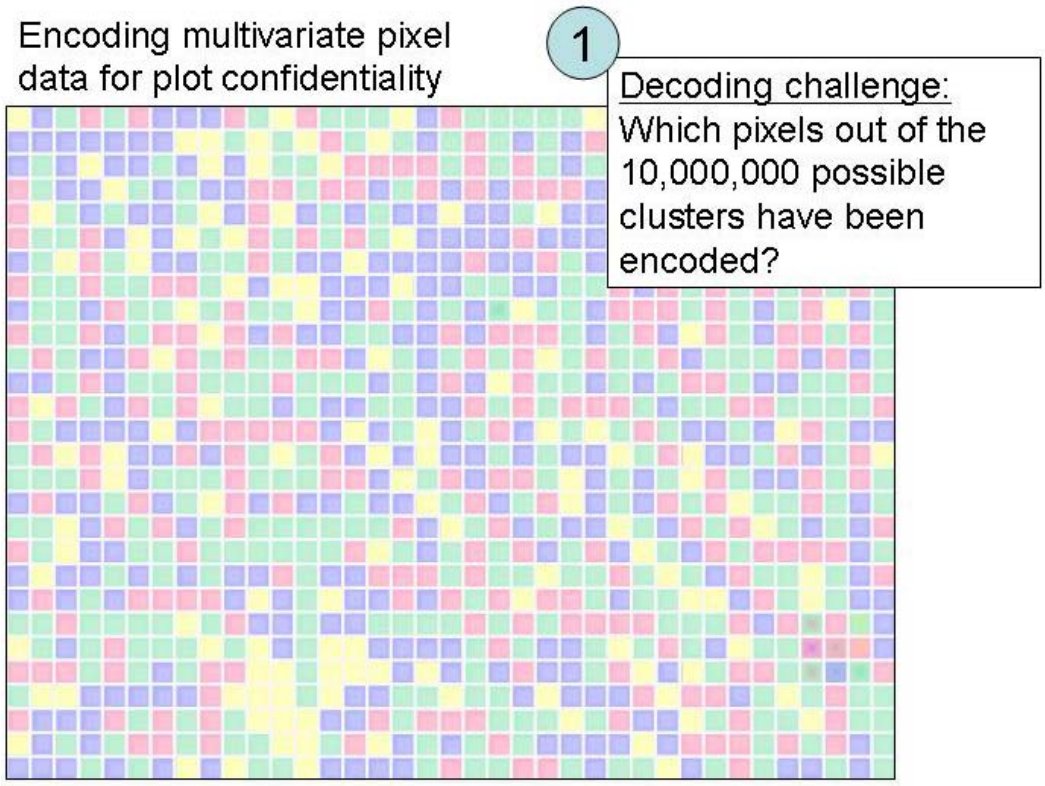

Encoding multivariate pixel data for plot confidentiality
Repeat procedure of randomly picking very similar pixels
4

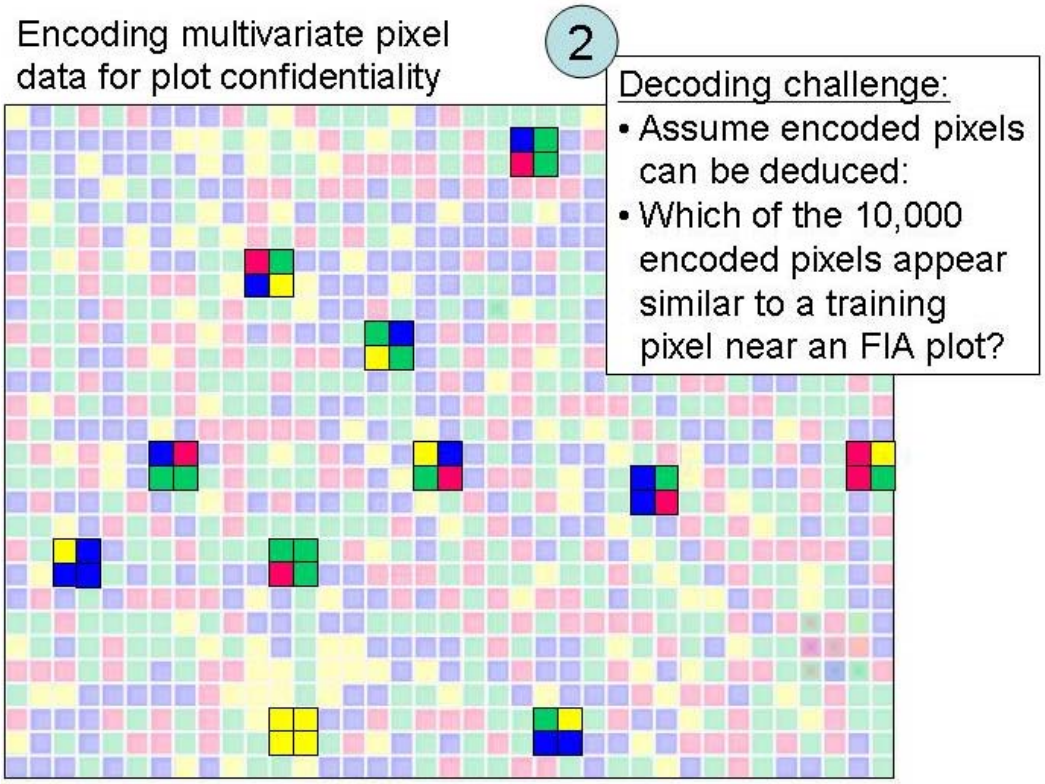


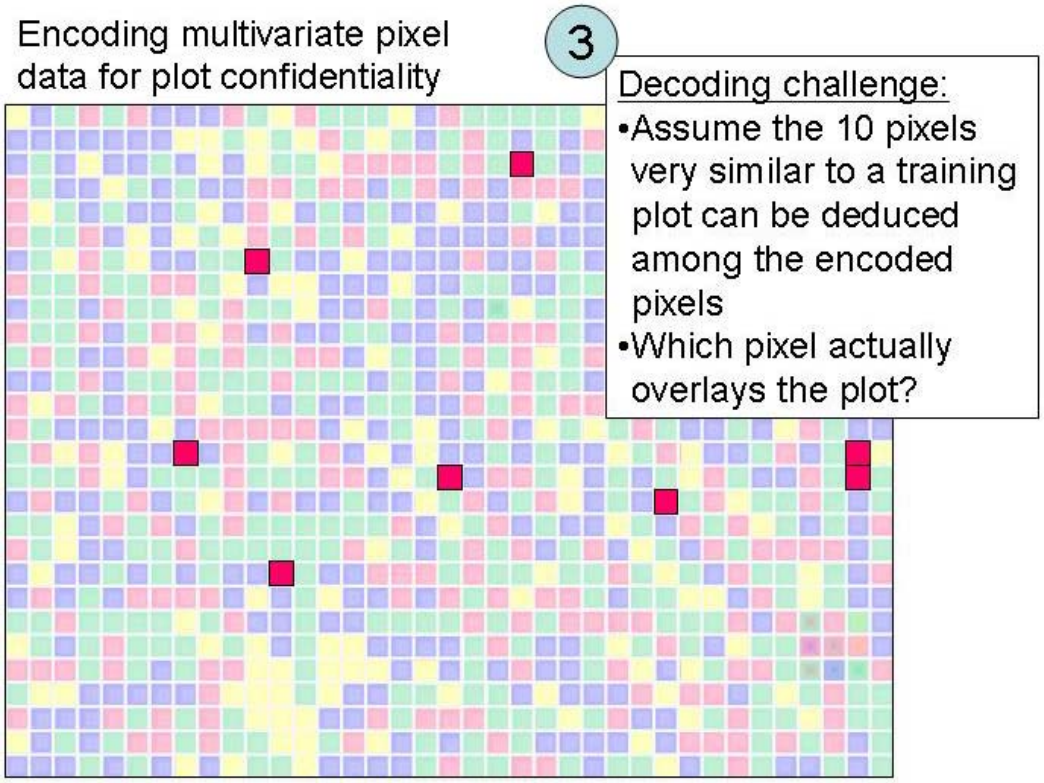
Encoding multivariate pixel data for plot confidentiality
Swap spectral values among very similar pixels

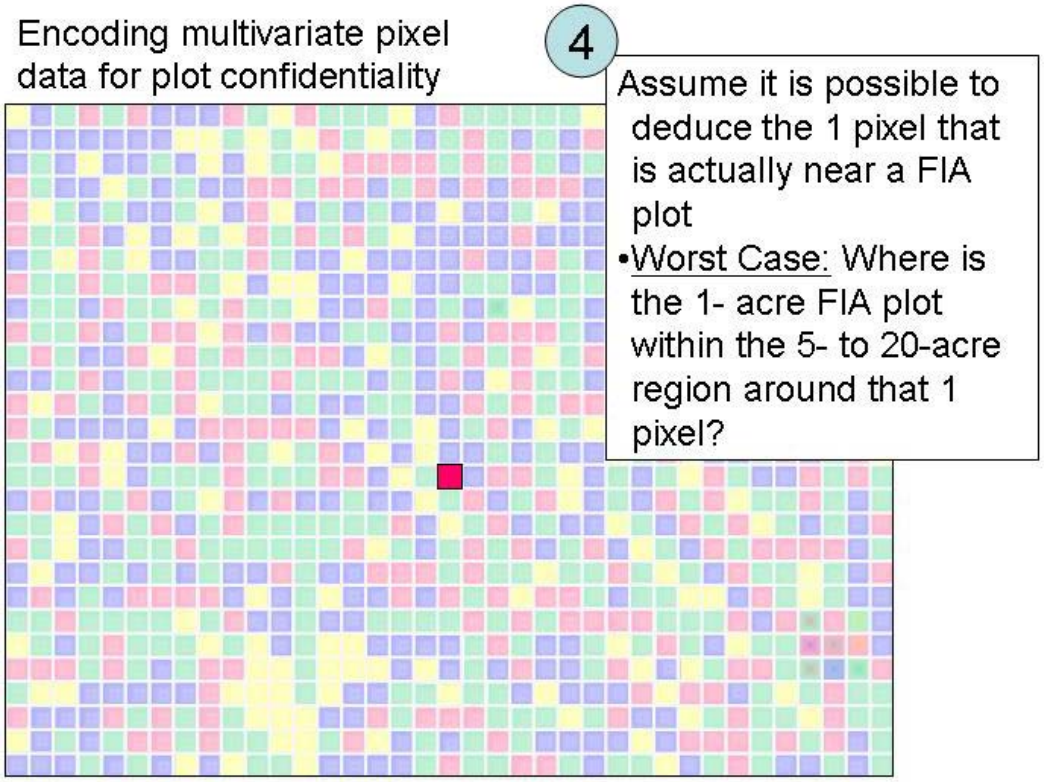

Encoding multivariate pixel data for plot confidentiality
Swap spectral values among very similar pixels


Encoding multivariate pixel data for plot confidentiality
Hypothesis:
Swapping spectral values of very similar pixels will sufficiently protect confidentiality


Encoding multivariate pixel data for plot confidentiality
1
Decoding challenge:
Which pixels out of the 10,000,000 possible clusters have been encoded?


Encoding multivariate pixel data for plot confidentiality
2
Decoding challenge:
• Assume encoded pixels can be deduced:
• Which of the 10,000 encoded pixels appear similar to a training pixel near an FIA plot?


Encoding multivariate pixel data for plot confidentiality
3
Decoding challenge:
•Assume the 10 pixels very similar to a training plot can be deduced among the encoded pixels
•Which pixel actually overlays the plot?


Encoding multivariate pixel data for plot confidentiality
4
Assume it is possible to deduce the 1 pixel that is actually near a FIA plot
•Worst Case: Where is the 1- acre FIA plot within the 5- to 20-acre region around that 1 pixel?