

## Using New and Emerging Technologies through the Lens of Improving Official Statistics

**Instructions:** Click on the link to access each author's presentation.

**Organiser:** Linda J. Young

**Chair:** Maria Luiza Guerra de Toledo

**Discussant:** Denise Silva

### Participants:

**Linda J. Young:** Using New Technologies to Leverage Alternative Data in the Production of Official Statistics

**José Hernández:** Use of Satellite imagery to validate statistical data on agricultural activity from different sources

**Raul Emilio Ospina Villalobos:** Use of geospatial technologies to enhance the generation of official statistics in Colombia



# Using New Technologies to Leverage Alternative Data in the Production of Official Statistics

Linda J. Young

USDA National Agricultural Statistics Service (NASS)

May 17, 2024



# Outline

- Motivation for using all (survey and non-survey) data
- Alternative (non-survey data)
- List building
- Data collection
- Editing
- Estimation
- Final thoughts

**The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.**

# Why Turn to Non-Survey Data?

- Increasing demands for more official statistics
  - More often
  - Finer geospatial scales
  - Increasing response burden
- Decreasing list coverage
- Declining response rates

**Question: What can be done to alleviate these concerns?**

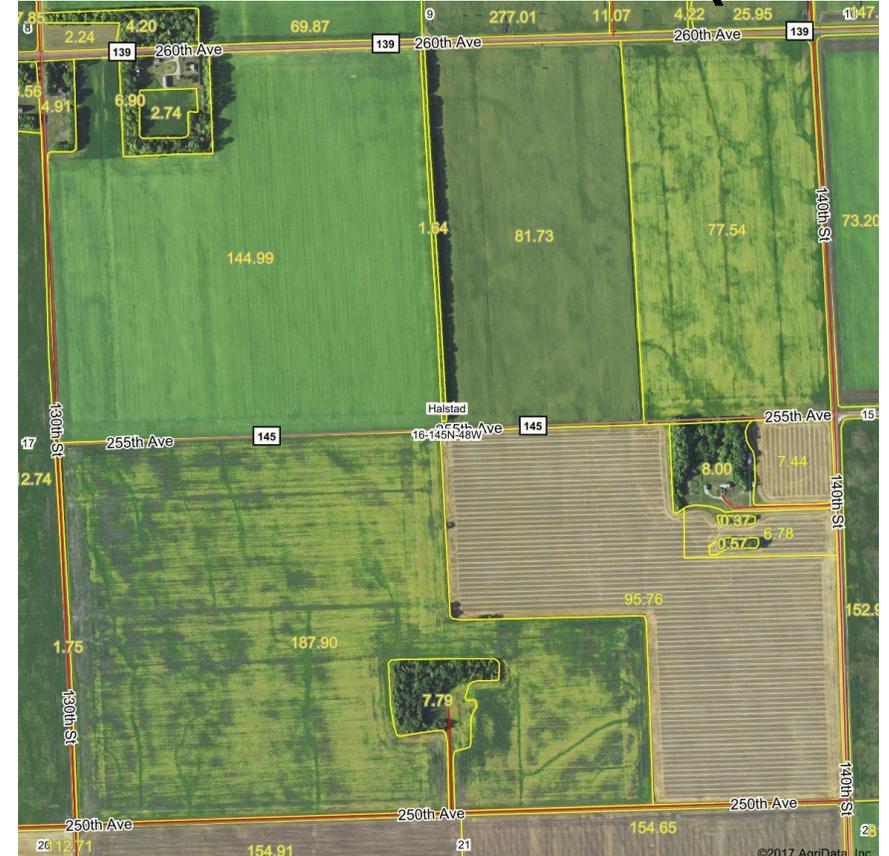
# Alternative (Non-survey) Data



# Farm Service Agency (FSA) Form FSA-578

- Completed by all producers participating in a USDA program for that crop season
- Information for each Common Land Unit
  - Crops
  - Acreage
  - Irrigation
- Variable coverage for crops and states, but high in major corn states
- Provides lower bound for acreages planted to a crop within a county

## Common Land Units (CLUs)

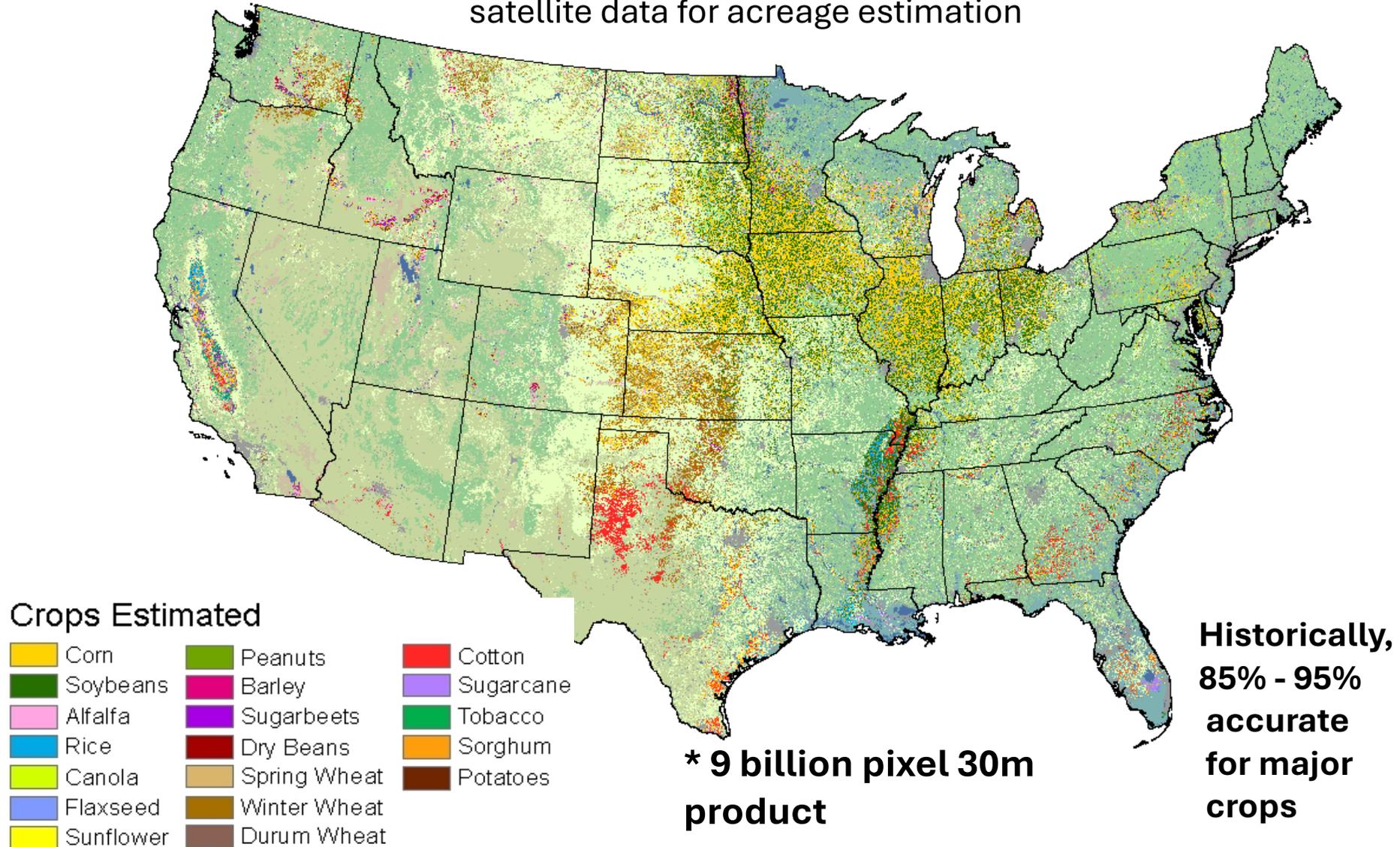


[https://www.agridatainc.com/Home/Products/Mapping%20Features/Land%20Resource%20Intelligence/FSA%20Field%20Boundaries%20\(CLU\)](https://www.agridatainc.com/Home/Products/Mapping%20Features/Land%20Resource%20Intelligence/FSA%20Field%20Boundaries%20(CLU))

# Cropland Data Layer (CDL)

Annual national coverage since 2008

A raster, crop-specific, land cover data set produced using satellite data for acreage estimation



# Predictive Cropland Data Layers and Entropy Layers

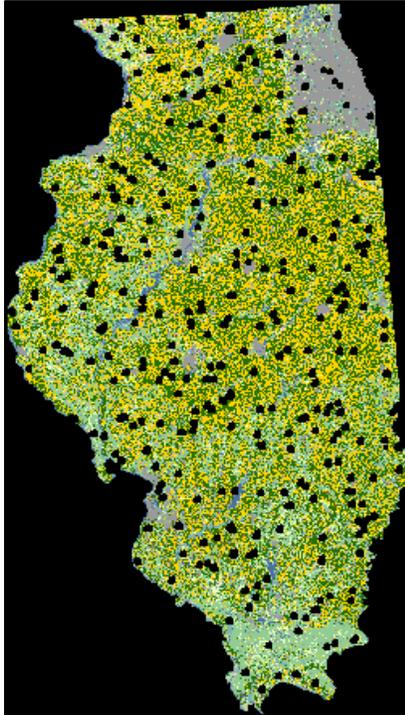
Land Cover Categories  
(by decreasing acreage)

## AGRICULTURE

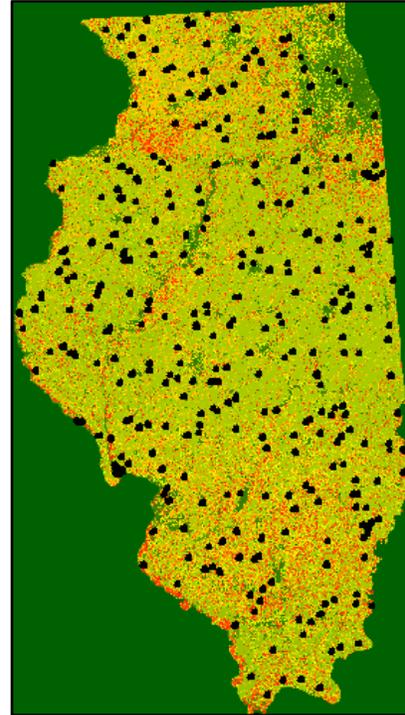
- Soybeans
- Corn
- Grass/Pasture
- Winter Wheat
- Dbi Crop WinWht/Soybeans
- Alfalfa
- Other Hay/Non Alfalfa
- Fallow/Idle Cropland
- Other Crops

## NON-AGRICULTURE\*

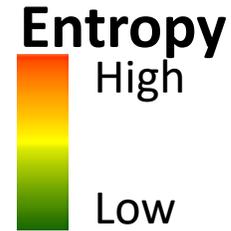
- Developed/Open Space
- Mixed Forest
- Developed/Low Intensity
- Deciduous Forest
- Woody Wetlands
- Developed/Medium Intensity



Illinois (2021)  
PCDL and  
Segments



Illinois (2021)  
Entropy Layer



**PCDL based on**  
High-Order Markov Chains

**Entropy layer based on**  
normalized Shannon entropy  
from the predictive distribution

## Accuracies in IL

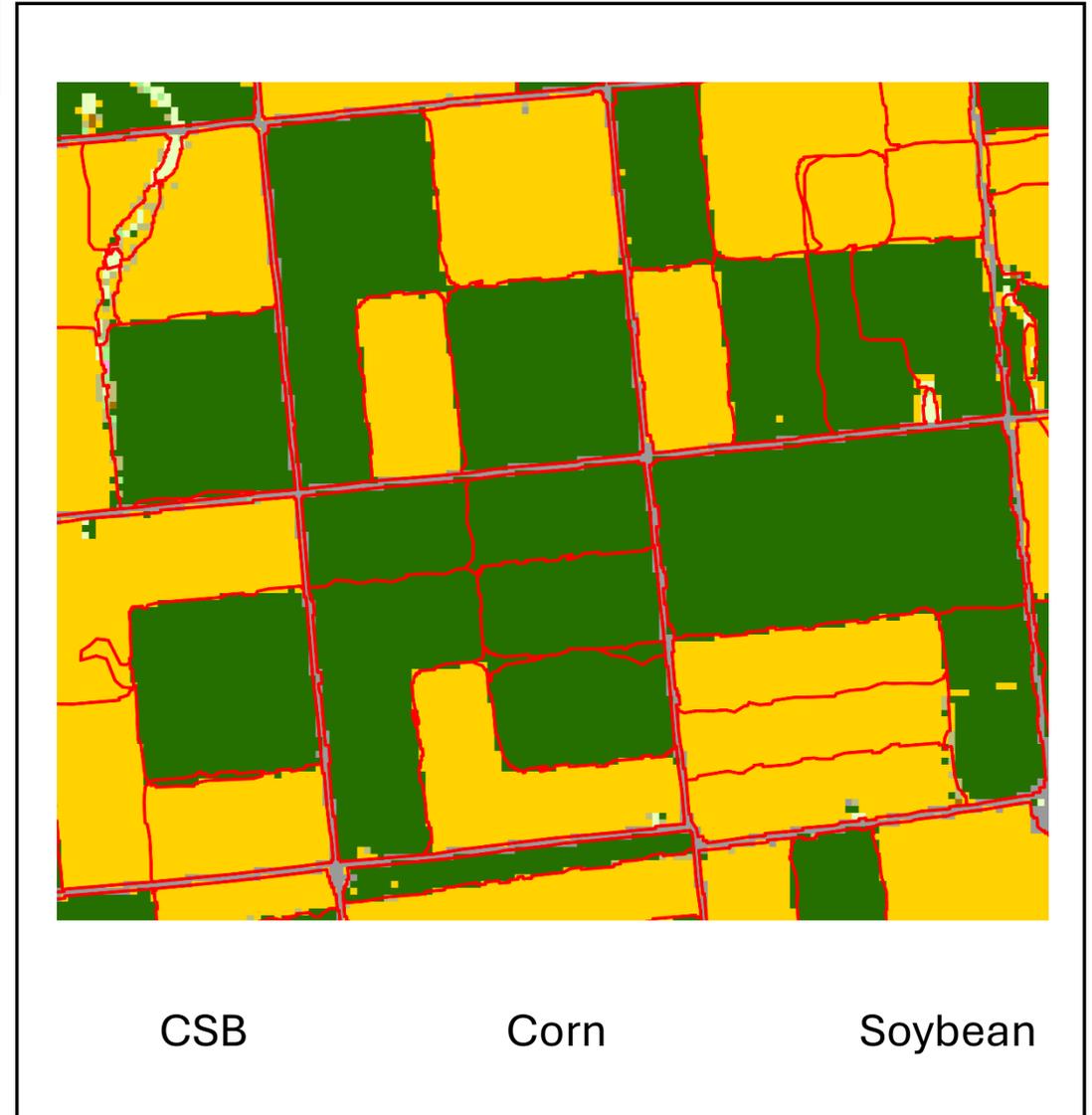
F1-score for corn: 81.5%

F1-score for soybeans: 80.5%

# Crop Sequence Boundaries (CSBs)

## An agricultural field managed over time

- Uses historic Cropland Data Layers
  - Based on 8-year historic panels
  - Uses U.S. Census TIGER roads & rails features
- Created in Google Earth Engine (GEE) and ArcGIS
- Data products correspond with CDL availability
  - Contiguous U.S. 2008-2023
- Product is in both polygon and raster (grid/pixel) file
- Joint effort with USDA Economic Research Agency



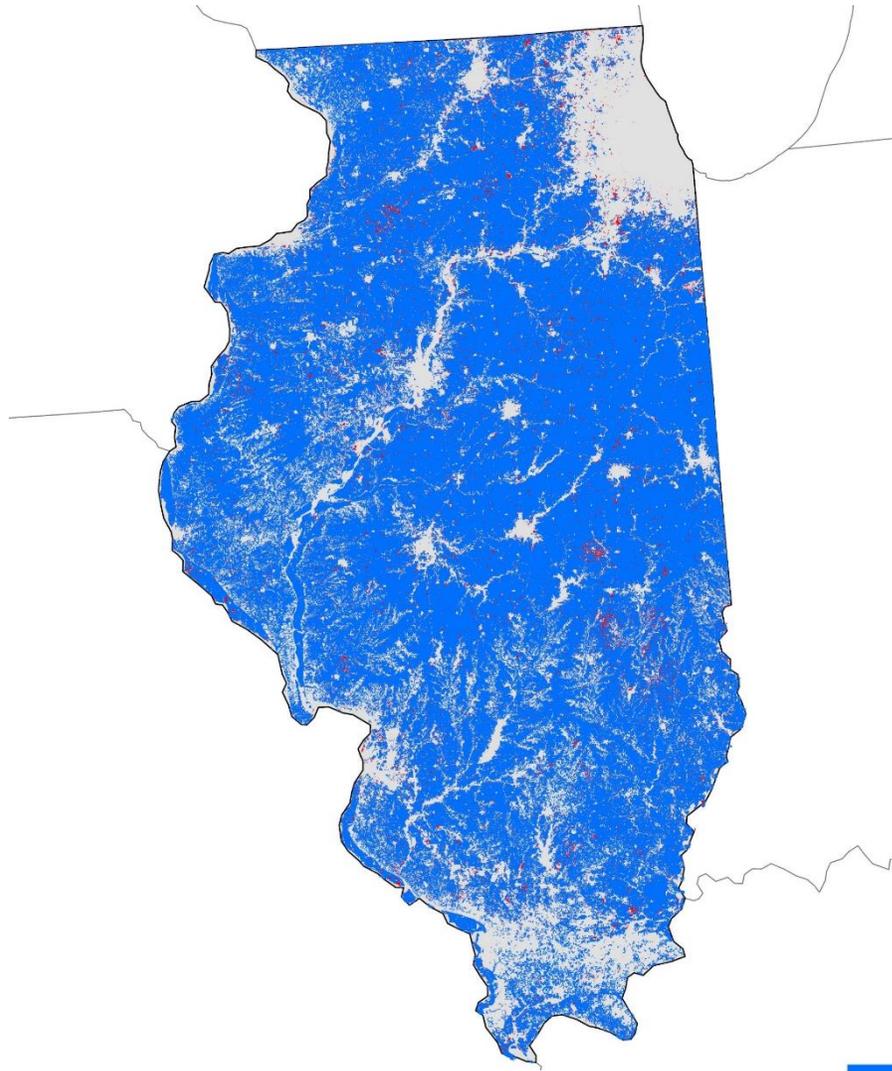
# Applications Leveraging All Data



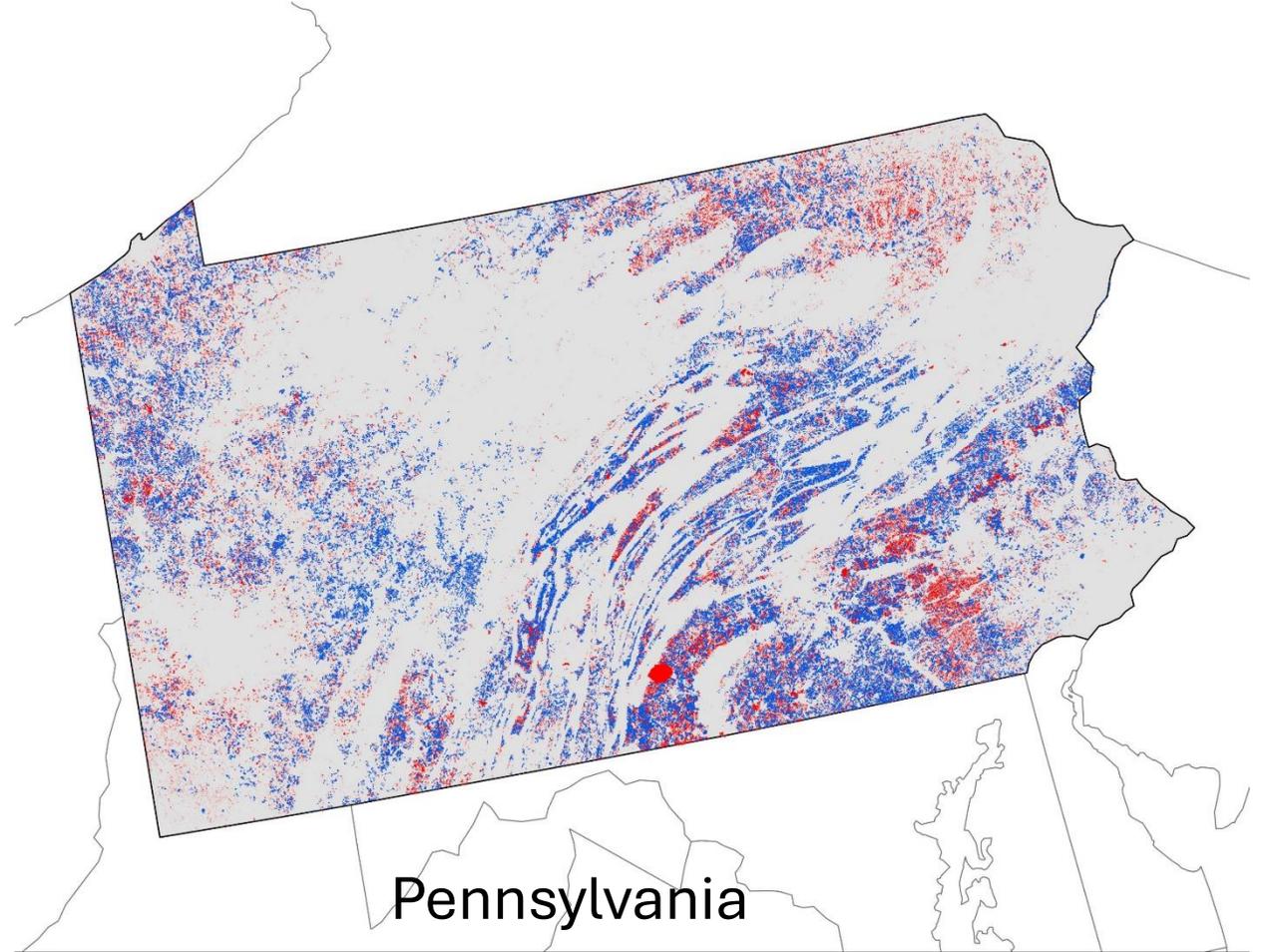
# Leveraging All Data to Identify List Frame Undercoverage

- FSA data have been used to identify farms for the NASS list frame
- Challenge: accounting for non-FSA farms
- Approach
  - Overlay the CSBs on the most recent Cropland Data Layer
  - Identify all CSBs associated with cropland
  - Identify the CSBs with cropland that do not have FSA data
  - Assess the farm status of all CSBs with cropland, not on the NASS list frame, and without FSA data
- Results vary by state
- Identifying livestock operations more challenging
  - Few USDA programs related to livestock → Limited FSA data
  - Small to mid-size operations difficult to identify using satellite imagery

# Identifying Farms Not on the NASS List Frame



Illinois



Pennsylvania



# Using Non-Survey Data to Complete Surveys

June Area Survey (JAS) is conducted annually in June

**Frame:** All land in U.S. provides a complete frame assuming accurate screening

**Sample Unit:** A segment, which is typically a 1-square mile area of ~640 acres (~259 hectares)

Segments divided into tracts, representing unique operations

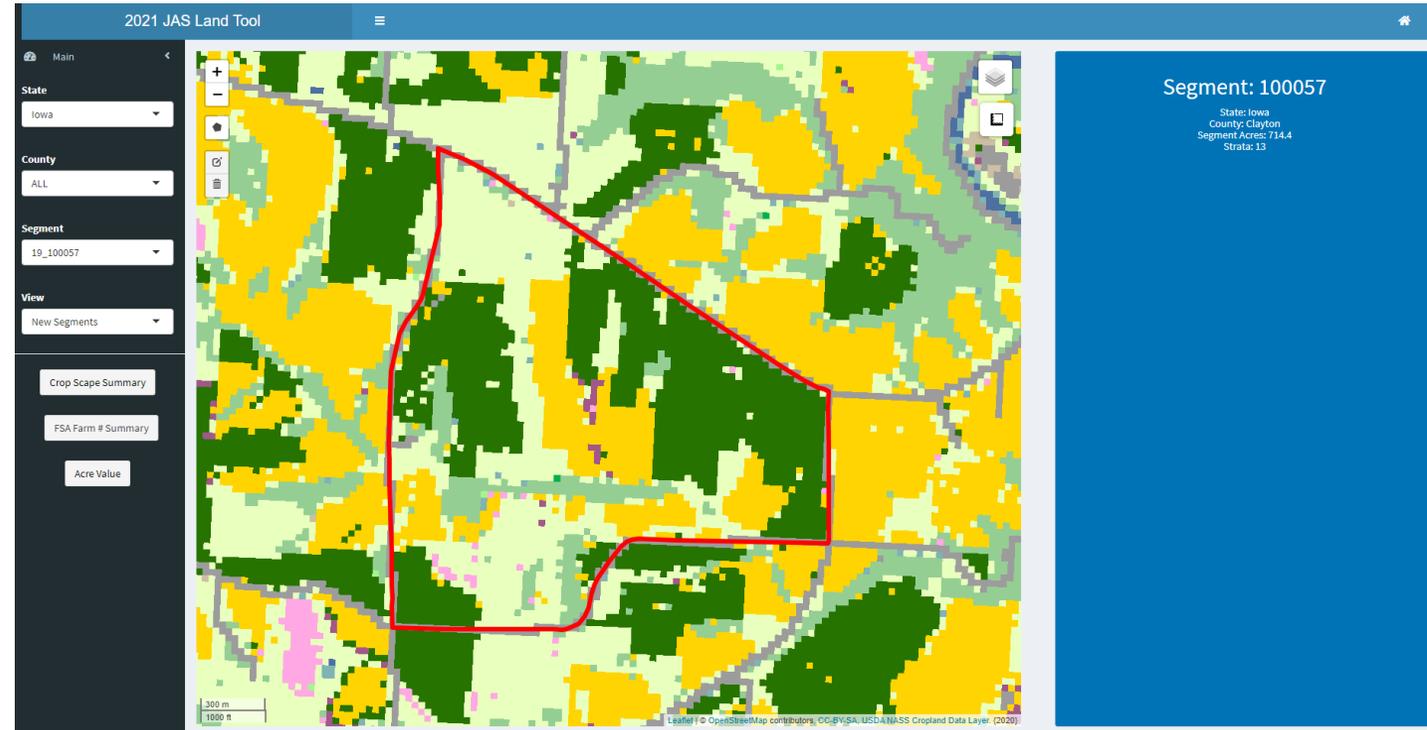
**Design:** Stratified Random Sample of segments, strata based on percent cultivated (>50%, 15%-50%, < 15%)

20% of the sample enters each year and remains for 5 years



# Tract-Level Information Required

- Nonresponse: tract-level data imputed
- June Area Tool
  - Historical CDLs
  - Historical FSA Data
  - Predictive CDLs (beginning in 2021)
- Predictions for current season
  - Predictive CDL
  - Modeled CSB prediction
- If the two predictions agree, imputation tends to be accurate
- Imputation will be automated for these tracts beginning June 2024



# Leveraging Survey and Non-Survey Data for Estimation

- Modeling at an aggregated level of geography
  - Examples: county or state
  - Combine multiple estimates and covariates to produce estimate
- Modeling at the unit level
  - Requires linkage of survey and non-survey data
- Goal: estimate acres planted to corn
  - Pre-season
  - In-season
  - Post-season

# Estimating Planted Acreage: Corn

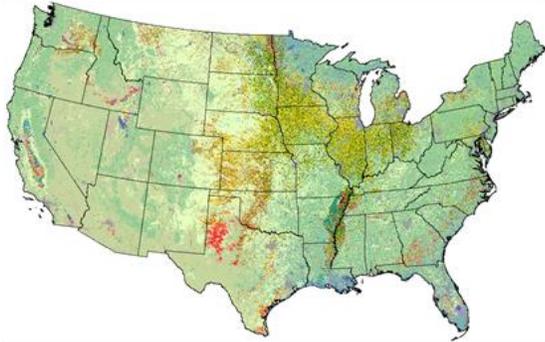
## Agricultural Survey

- Conducted quarterly (March, June, September, December)

## County Agricultural Survey

- Additional data collected in December
- December surveys provide foundation for county estimates
  - **Planted acreages**
  - Harvested acreages
  - Production
  - Yield

# Wealth of Non-Survey Data



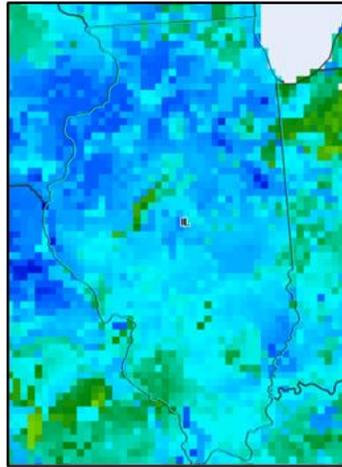
Cropland Data Layers (CDL)



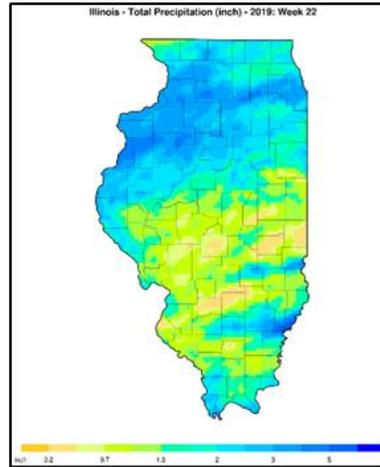
FSA Common Land Unit  
and 578 data



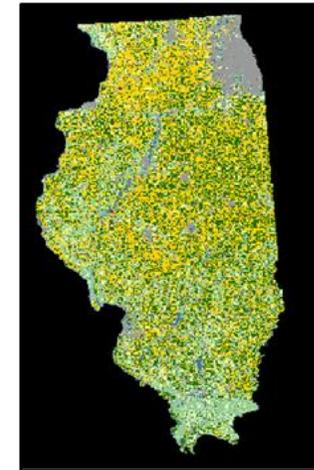
Crop Sequence Boundaries  
("Fields")



Soil Moisture Data



Precipitation Data



Early Season CDLs

# Ready to Link Survey and Non-survey Data?



- Non-survey data are geospatially referenced
- Survey data are collected at the **farm level**
  - Multiple fields in most farms
  - A farm may be in multiple counties or states
  - May be able to determine acreage of corn for a set of fields
  - BUT, cannot determine which particular fields are to be planted to corn

# Estimating Planted Acreage: Corn

- Three Bayesian hierarchical models used to combine information at the county level
  - **Planted acreage**
  - Harvested acreage, which must be no greater than planted acreage
  - Yield—production estimated by  $(\text{yield}) \cdot (\text{harvested acreage})$
- Challenges
  - County estimates must sum to state estimate
  - Honoring the bounds obtained from administrative data
  - Rounding
- Moved into production in 2021 for 2020 Growing Season

# Leveraging All Useful (Survey and Non-Survey) Data

- FSA and NASS have different definitions of a farm
- NASS list frame is not fully geo-referenced
- Surveys
  - Generally, not designed to provide estimates lower than a state
  - Information at farm level does not provide field-level data
- Integration into existing production process
  - Flow of survey and non-survey data
  - Analysis methods
  - Review processes

# Final Thoughts

- NASS conducts over 400 surveys annually to produce over 450 reports each year
  - Respondent burden is high, especially for large producers
  - Response rates decreasing
  - List frame coverage decreasing
- Leveraging all data has had an impact on production processes
- Challenges to leveraging all useful data (survey and non-survey)
  - Access is often challenging
  - Record-level versus higher level of geography
  - Survey design
  - Major effort underway to modernize processes

**Progress is being made!**

# Selected References

Abernethy, J., P. Beeson, C. Boryan, K. Hunt, L. Sartore. Preseason crop type prediction using crop sequence boundaries. *Computers and Electronics in Agriculture*. In Press.

Boryan C, Yang Z, Mueller R, Craig M. (2011) Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto Int.* 26(5):341–58.

Chen, L., Nandram, B., Cruze, N.B. (2022) Hierarchical Bayesian model with inequality constraints for county estimates. *Journal of Official Statistics* 38(3):709–732. <https://doi.org/10.2478/jos-2022-0032>.

Chen, L., Nandram, B., Cruze, N.B. (2021a) Hierarchical Bayesian model with inequality constraints for US county estimates. (*Journal of Official Statistics* accepted.)

Chen, L., Cruze, N.B., Nandram, B. (2021b) Preserving acreage relationships in small area agricultural models. (In preparation.)

Chen, L., Cruze, N. B., and Young, L. J. (2022). Model-based Estimates for Farm Labor Quantities. *Stats* 5(3): 738-754. <https://doi.org/10.3390/stats5030043>.

Cruze, N.B., Chen, L., Guindin, N. and Nandram, B. (2020). Dancing distributions: developing a better understanding of county-level crop yield from posterior summaries. In *JSM Proceedings, Section on Government Statistics*. American Statistical Association, Alexandria, VA. 2262-2272.

Cruze, N.B., Erciulescu, A.L., Nandram, B., Barboza, W.J., Young, L.J. (2019) Producing official county-Level agricultural estimates in the United States: needs and challenges. *Statistical Science*. 34(2), 301-316. <https://doi.org/10.1214/18-STS687>

Erciulescu, A.L., Cruze, N.B., Nandram, B. (2018). Benchmarking a triplet of official statistics. *Environmental and Ecological Statistics*, 25(4), 523-547. <https://doi.org/10.1007/s10651-018-0416-4>

Erciulescu, A.L., Cruze, N.B., Nandram, B. (2019). Model-based county-level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society, Series A*, 182, 283-303. <https://doi.org/10.1111/rssa.12390>

# Selected References

Erciulescu, A.L., Cruze, N.B., Nandram, B. (2020) Statistical challenges in combining survey and auxiliary data to produce official estimates. *Journal of Official Statistics*. 36(1), 63-88. <https://doi.org/10.2478/jos-2020-0004>.

Nandram, B., N.B. Cruze, A.L. Erciulescu and L. Chen. (2022). Bayesian Small Area Models under Inequality Constraints with Benchmarking and Double Shrinkage. Research Report RDD-22-02, National Agricultural Statistics Service, USDA.

Available at:

[https://www.nass.usda.gov/Education\\_and\\_Outreach/Reports,\\_Presentations\\_and\\_Conferences/reports/ResearchReport\\_constraintmodel.pdf](https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/ResearchReport_constraintmodel.pdf).

Nandram, B., Erciulescu, A.L. and Cruze, N.B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 45, No.

2. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00004-eng.htm>

National Academies of Sciences, Engineering, and Medicine (2017). *Improving Crop Estimates by Incorporating Multiple Data Sources*. National Academies Press. <https://doi.org/10.17226/24892>.

Sartore, L., C. Boryan, P. Willis. (2022) Developing entropies of Predictive Cropland Data Layers for crop survey imputation. *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 1404–1407.

<https://doi.org/10.1109/IGARSS46834.2022.9884059>.

Sartore, L., Boryan, C., Dau, A., & Willis, P. (2023). An Assessment of Crop-Specific Land Cover Predictions Using High-Order Markov Chains and Deep Neural Networks. *Journal of Data Science* 21(2): 333-353. Available at:

<https://doi.org/10.6339/23-JDS1098>,

Young, L.J. and L. Chen. (2022) Using Small Area Estimation to Produce Official Statistics. *Stats* 5(3): 881-897.

<https://doi.org/10.3390/stats5030051>.





**Thank you!**

**Linda.J.Young@usda.gov**





# Use of Satellite imagery to validate statistical data on agricultural activity from different sources

IAOS-ISI 2024 Mexico Conference  
Improving Decision-Making for All



Mexico City, May 15th -17th, 2024

# Content

1. Validation of land area with no apparent agricultural activity to strengthen the geographic coverage of the Census of Agriculture (CA) 2022.
2. Identification of crops as support the validation of results of the National Agricultural Survey (ENA).
3. Validation of crops obtained through Administrative Records (AR).

**1. Validation of land area with no apparent agricultural activity to strengthen the geographic coverage of the Census of Agriculture (CA) 2022**



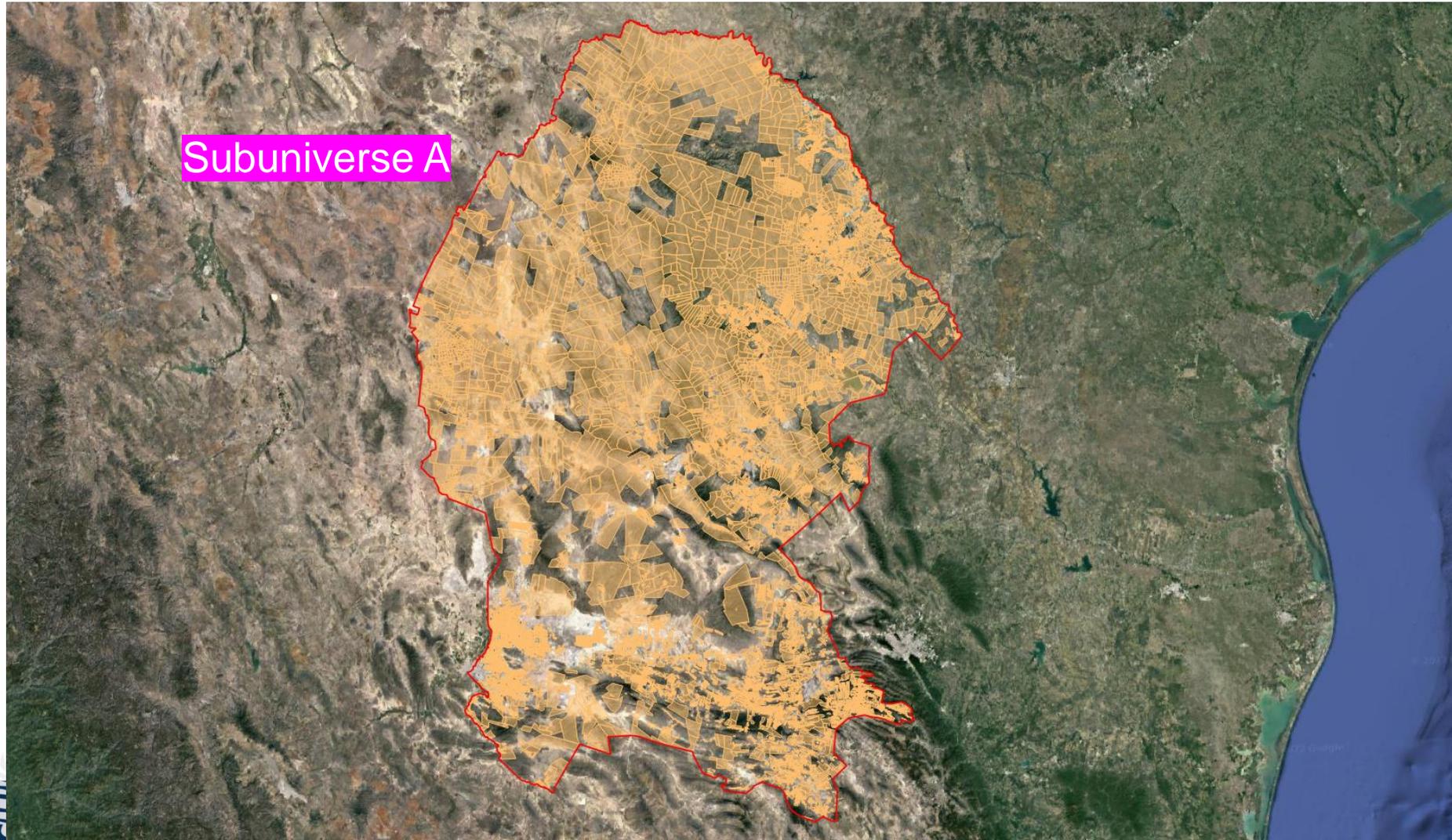
# Initial Universe of the CA2022

One of the great challenges of any census is to ensure coverage of the entire population under study.

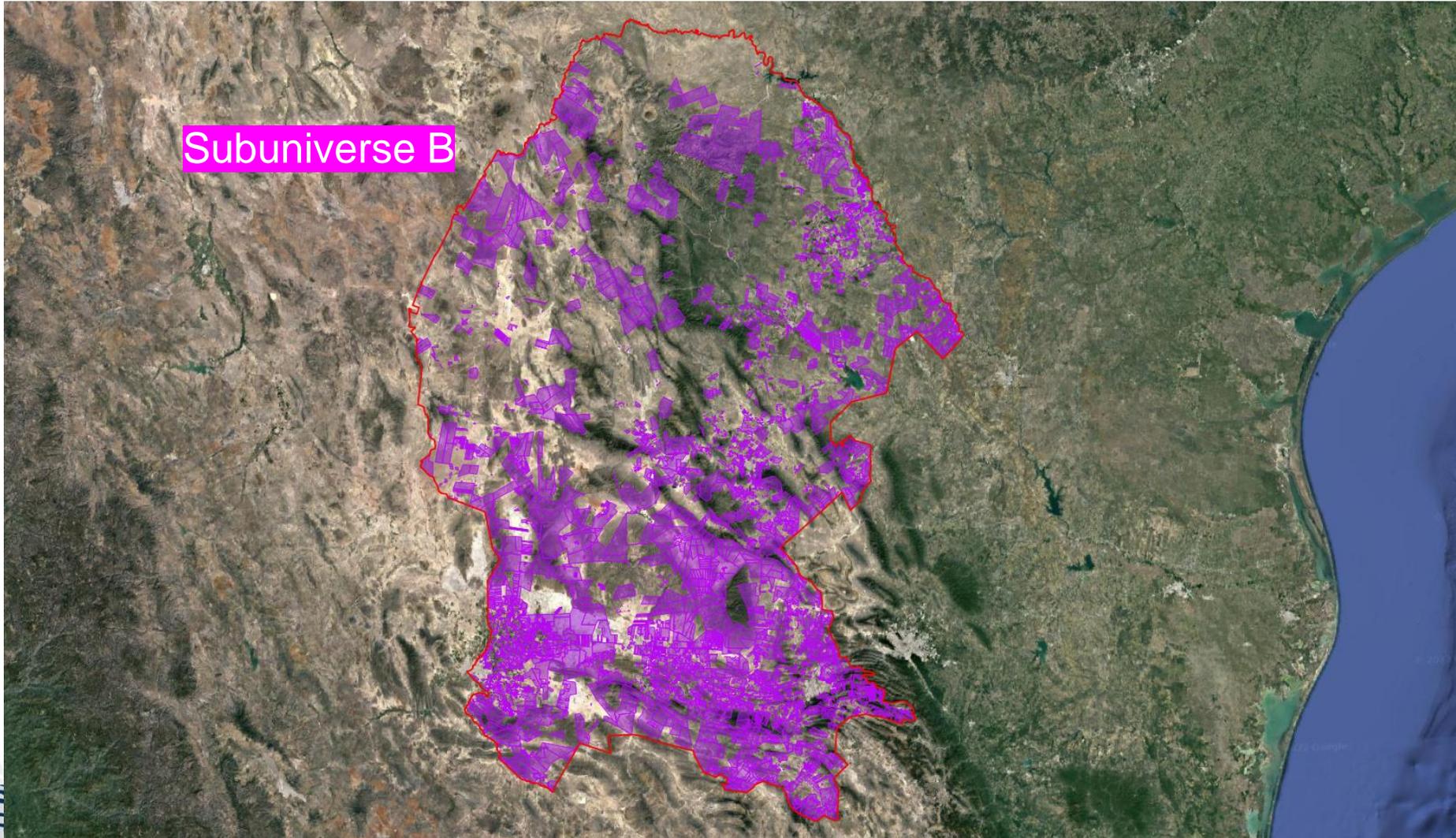
To address this challenge, The census was planned based on the Master Framework for Agricultural Statistics (MMESAGRO), which included:

- **The directory of agriculture, livestock and forestry producers.** Where each producer has associated the land plots in which he carries out his activities.
- **The mosaic of all land plots with and without agricultural activity** (in shape files), separated into two large subuniverses:
  - Subuniverse A: Land plots with agricultural activity associated with a producer.
  - Subuniverse B: Land plots without apparent agricultural activity not associated with a producer or directory.

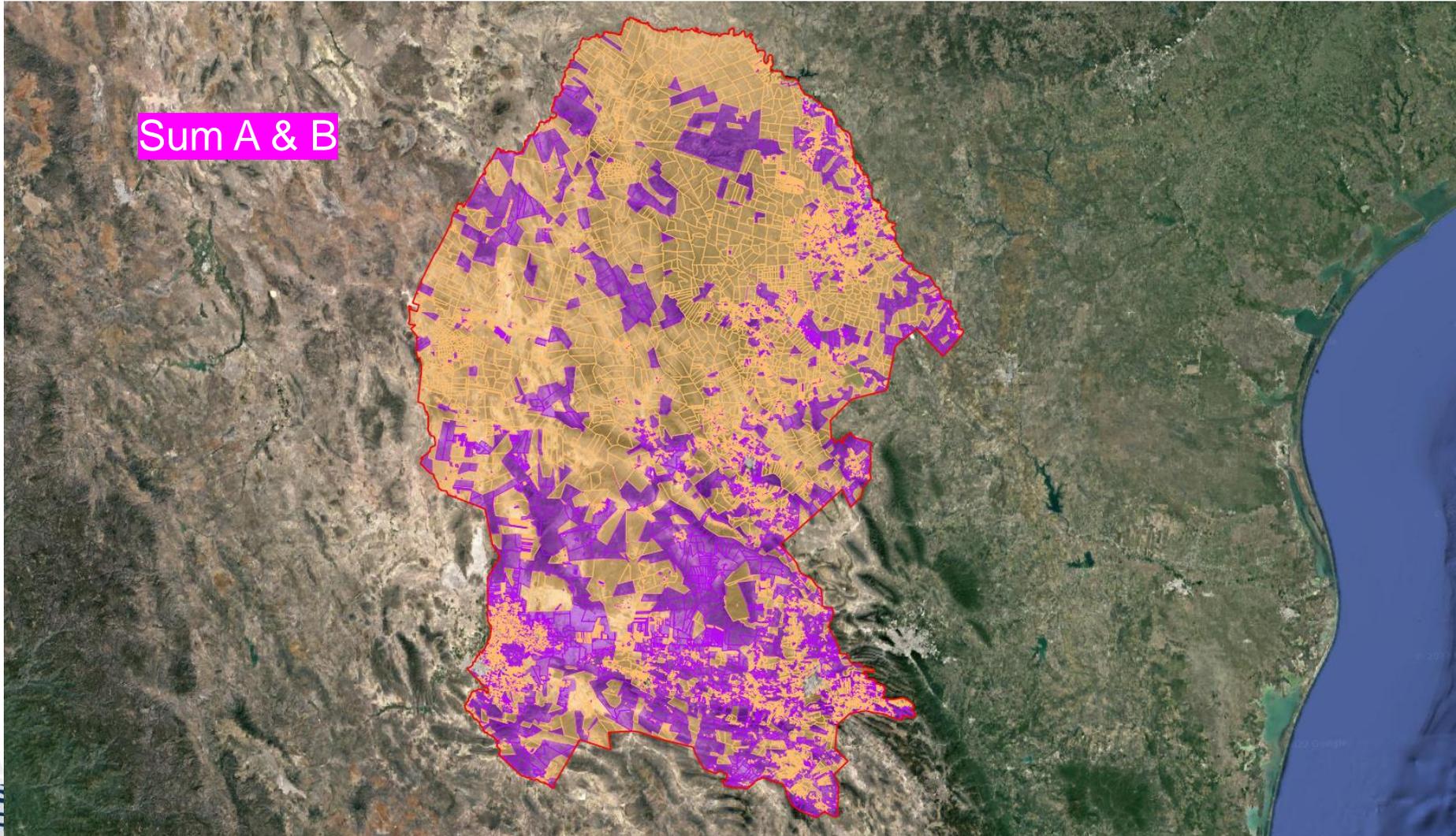
# Land plots with agricultural activity associated with a producer



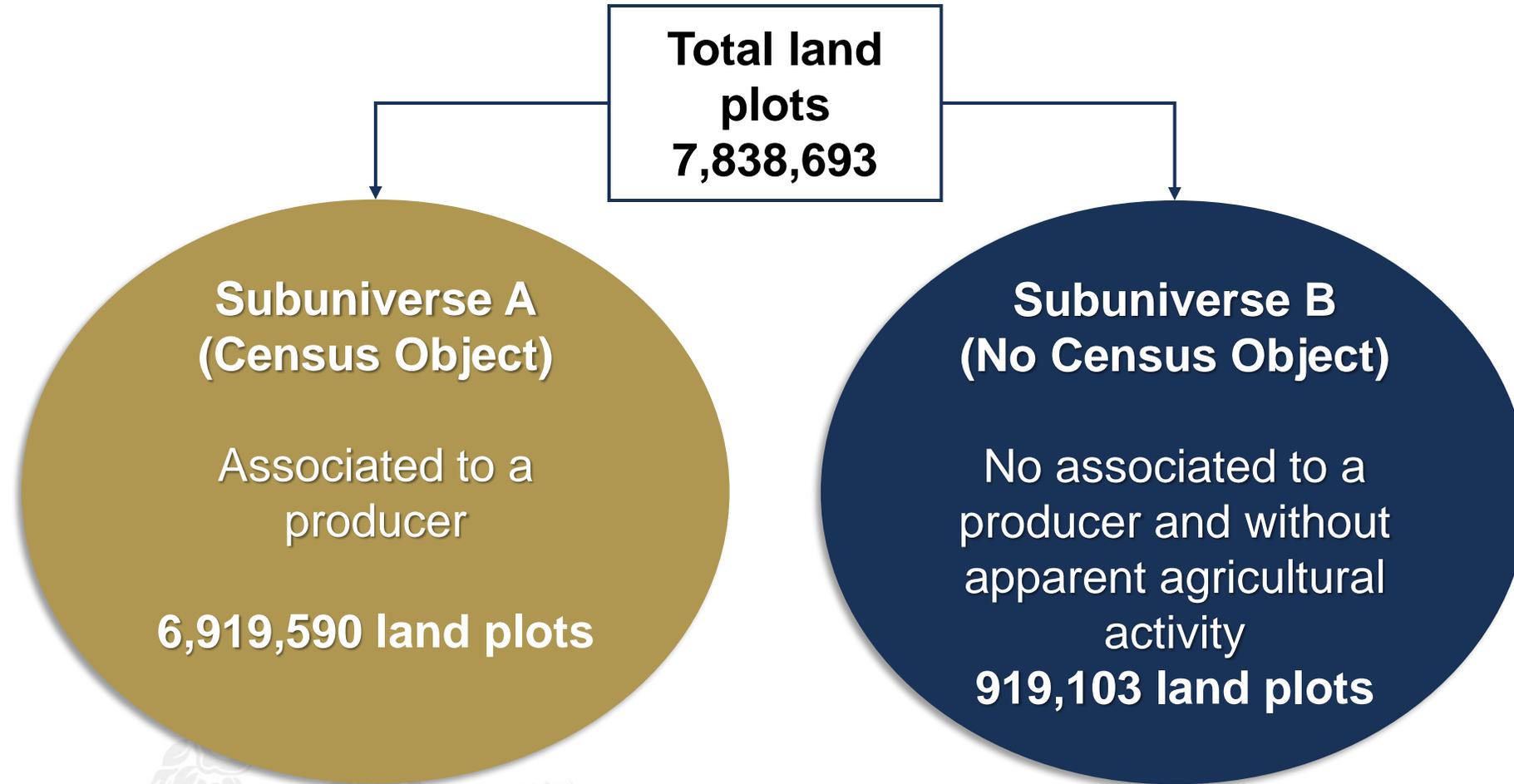
Land plots without apparent agricultural activity not associated with a producer or directory 



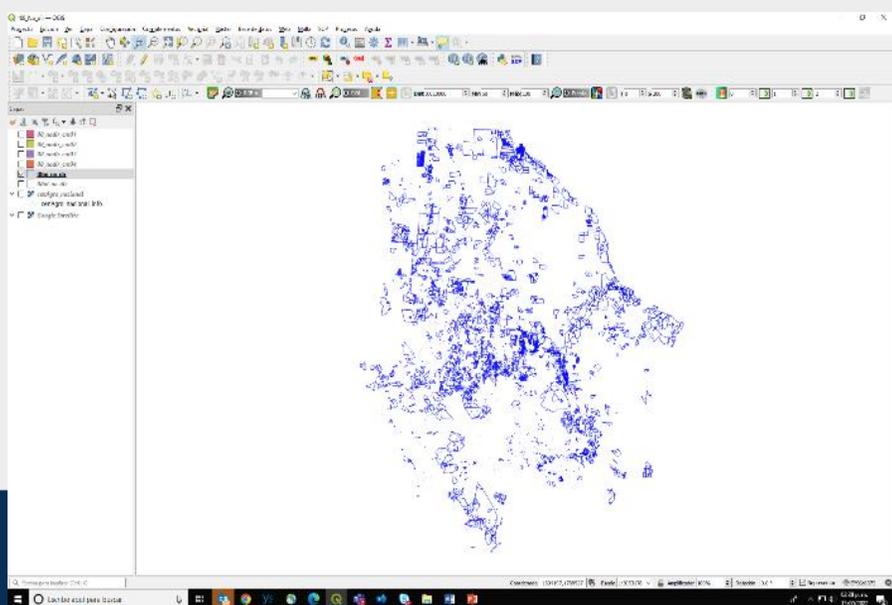
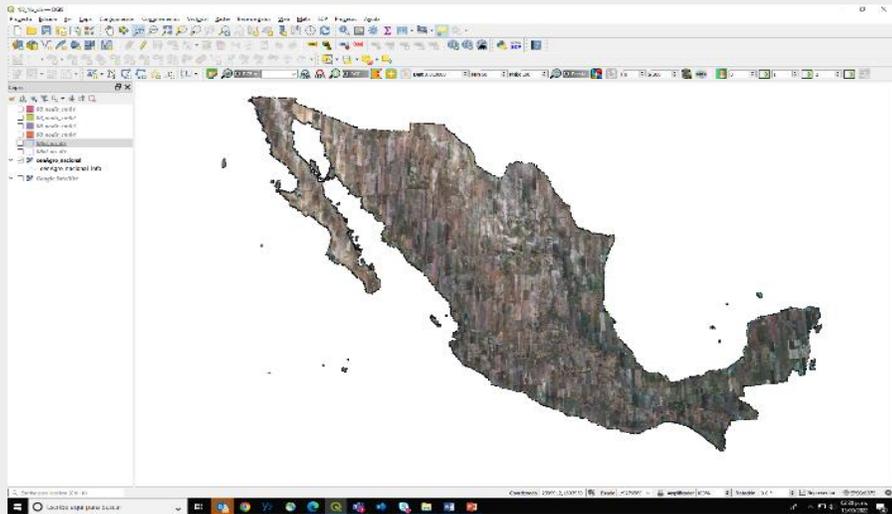
Complete mosaic of all the land plots, adding the two subuniverses:



# Coverage of the CA2022



In order to support census coverage, the situation of all the land plots of this Subuniverse was validated using satellite images, to verify that they had no agricultural activity.

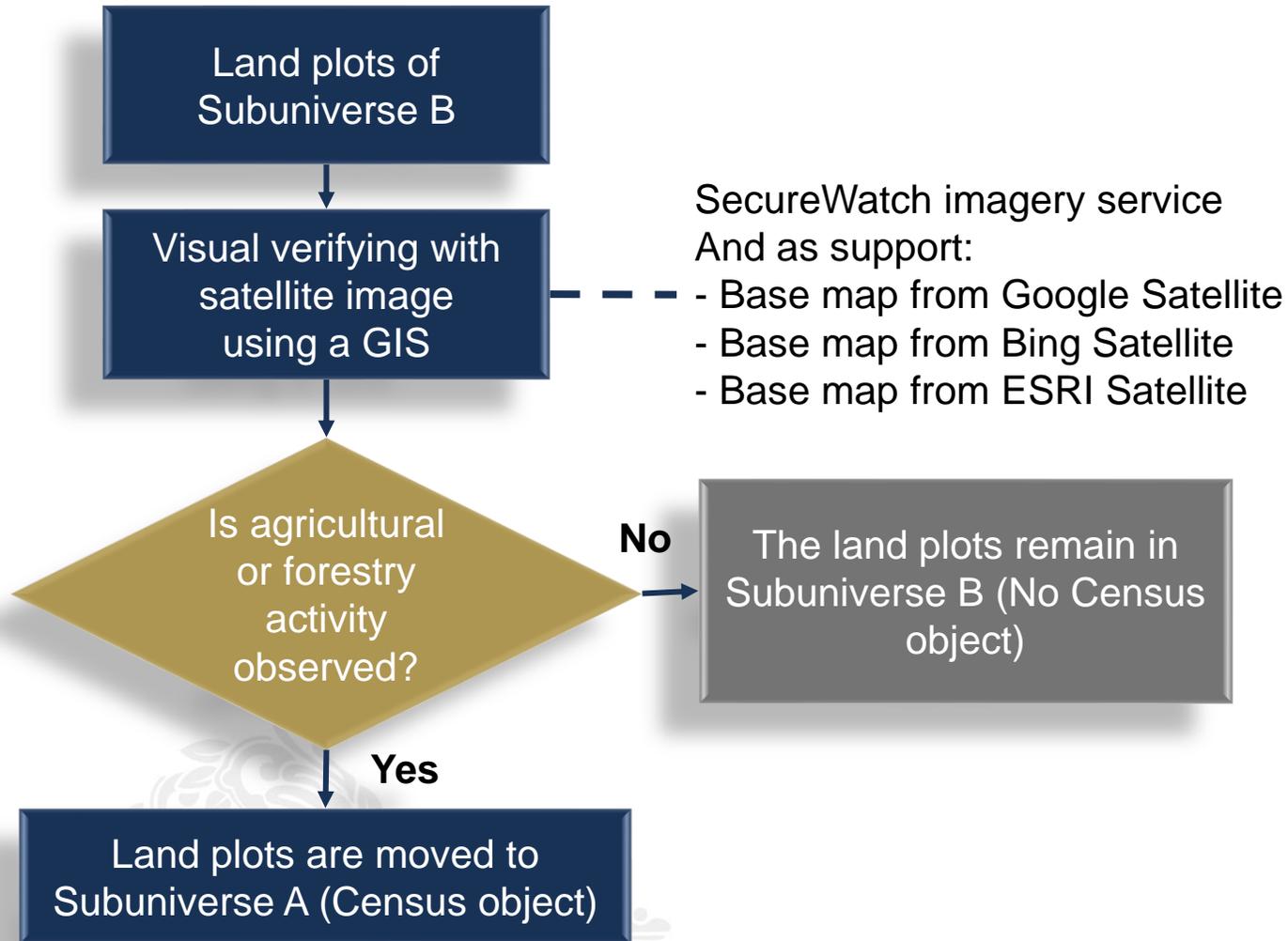


# Inputs Used

- GIS software.
- SecureWatch image service, with a spatial resolution of 1.0 m in urban areas, and 1.5 m in rural areas.
- Other services of satellite images available.
- File of all land plots in shape format.
- Layers with supporting elements (localities, water bodies, roads, among others).



# Procedure to verify land plots of Subuniverse B



# How to identify a land plots with agricultural activity

Land plots with agriculture activity



Land plots with livestock activity



Cropland Idle (abandoned or resting for recovering nutrients)



# How to identify a land without agricultural activity

Shrub land



Land with forest or jungle



Water bodies



Land plots in dessert



Land plots with facilities or infrastructure



Urban areas



# Results

- Derived from the review, it was confirmed that most of the land plots in subuniverse B had no agricultural or forestry activity.
- However, approximately 8.1 million of ha were identified partially or completely with some type of agricultural or forestry activity and were moved from subuniverse B to subuniverse A, this means that, they were incorporated into the census coverage.



**2. Identification of crops as support  
the validation of results of the  
National Agricultural Survey (ENA)  
2019.**



# Validation of avocado in the National Agricultural Survey 2019 (ENA19)

In Mexico there are crops cultivated in certain regions of the country, that are conditioned by specific characteristics of climate, soil, height above sea level, etc.

Such is the case of avocado, which is grown in a specific region in the west of the country, where four states account for 91% of total production.

In recent years, the statistical exercises carried out have reflected a high rate of non-response by the producers of that region, which has forced to look for alternatives to obtain information that allows to know the reality in the production of this important crop in the region.



# Identification of avocado with satellite images

## Objective

Obtain through the support of satellite images, the planted area with avocado in the main producing states and municipalities, to validate the data of this crop obtained in the National Agricultural Survey (ENA) 2019.

## Work universe

State	Number of Municipalities
Michoacan	30
State of Mexico	11
Jalisco	13
Nayarit	5
<b>Total</b>	<b>59</b>



# Identification of avocado with satellite images

## Procedure

1. Identify the crops in the area of interest to detect possible confusion with avocado.

### Sowing progress (Ministry of Agriculture)

#### Situation as at November 30th, 2019

Municipality	Crop	Sown/Planted area (ha)
Salvador Escalante	<b>Avocado</b>	<b>16,293</b>
	Peach	29
	Asparagus	28
	Raspberry	12
	Guava	8
	Zarzamora	203
<b>Total municipality Salvador Escalante</b>		<b>16,573</b>

### Google Earth



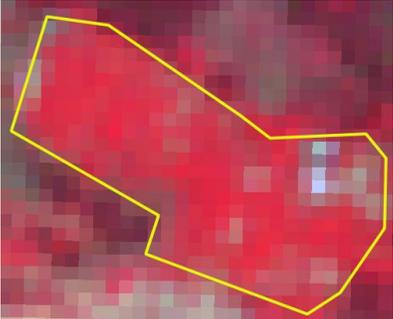
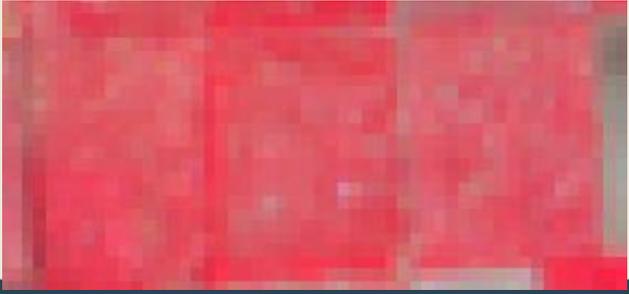
### Sentinel-2 2019 false infra-red (IR) color



# Identification of avocado with satellite images

## Procedure

2. Supported by free use mosaics of high-resolution satellite imagery to identify and photo interpret the avocado trees.
3. Obtain the satellite image coverage Sentinel-2 of the area of interest, with take-over date consistent with the Survey period
4. Corroborate the presence of avocado using Sentinel-2.
5. Digitize the area planted with avocado.
6. Calculate area in hectares.

Google Earth	Sentinel-2 2019 false IR color
	
	
	

# Identification of avocado with satellite images

## Results

Estimated data by identifying avocado with satellite images		Data published ENA19*	
Area (hectares)	Production (tons)	Area (hectares)	Production (tons)
212,106.84	2,309,178.79	213,422.12	2,013,590.93

\* Expansion factors of the ENA 2019 were calibrated using the information of avocado obtained by satellite images.

### **3. Validation of crops obtained through Administrative Records (AR).**



# Validation of crops obtained through Administrative Registers

INEGI has a master framework, consisting of a national mosaic of rural land plots (polygons), as well as a national directory of producers.

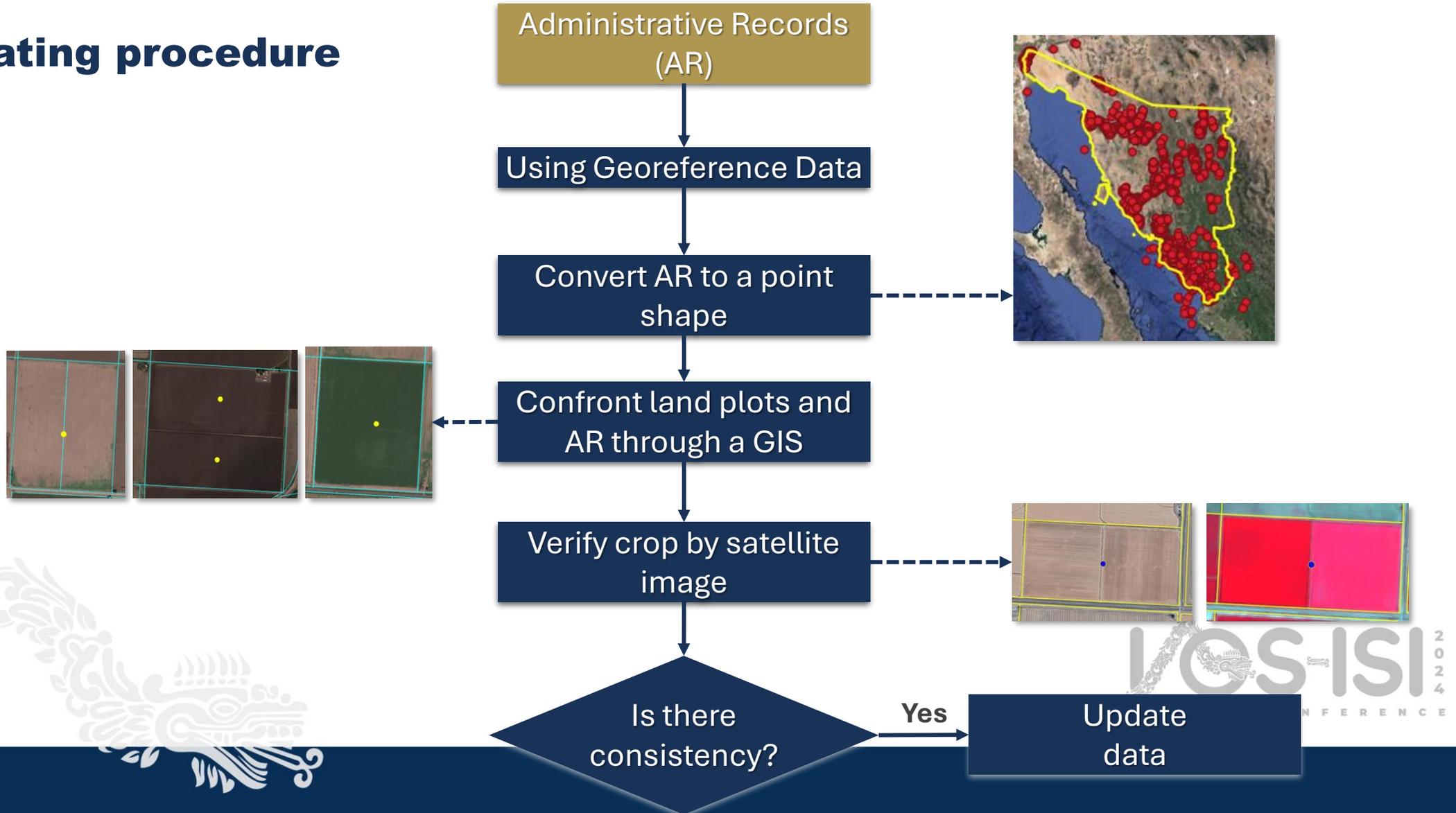
In order to find alternatives for updating this framework, the possibility of using administrative registers was explored.

A diagnosis was made of the existence and availability of georeferenced administrative registers, mainly agricultural.

In 2020 a project to update the framework was carried out through the use of agricultural administrative registers.

# Validation of crops obtained through Administrative Records

## Updating procedure



# Importance of satellite imagery in updating the Framework with AR

## Verify the activity and crop reported by AR

### Using satellite Images

1. **Google satellite.** Analysis of the land plot in high resolution.
2. **Sentinel-2.** Analysis of the land plots in the temporality according to the date of update of the AR to confirm the presence or absence of some crop either cyclic or perennial.

AR = wheat

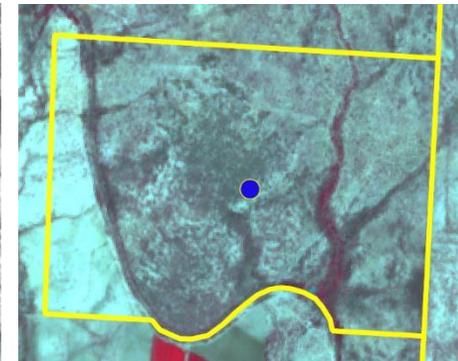
Google Satellite



Sentinel-2 real color



Sentinel-2 false IR color



AR = wheat

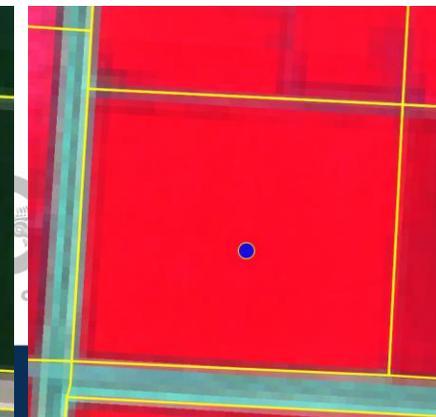
Google Satellite



Sentinel-2 real color



Sentinel-2 false IR color



### Legend

● Location of AR

□ Land plot of the Master Framework



# Importance of satellite imagery in updating the Framework with AR

## Verify the activity and crop reported by AR

AR = corn



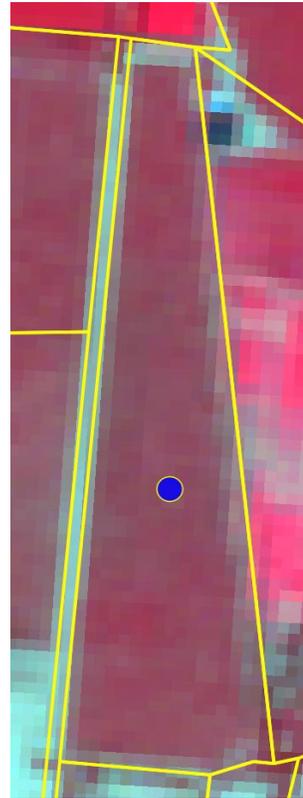
Google  
Satellite



Sentinel-2  
real color



Sentinel-2  
false IR color

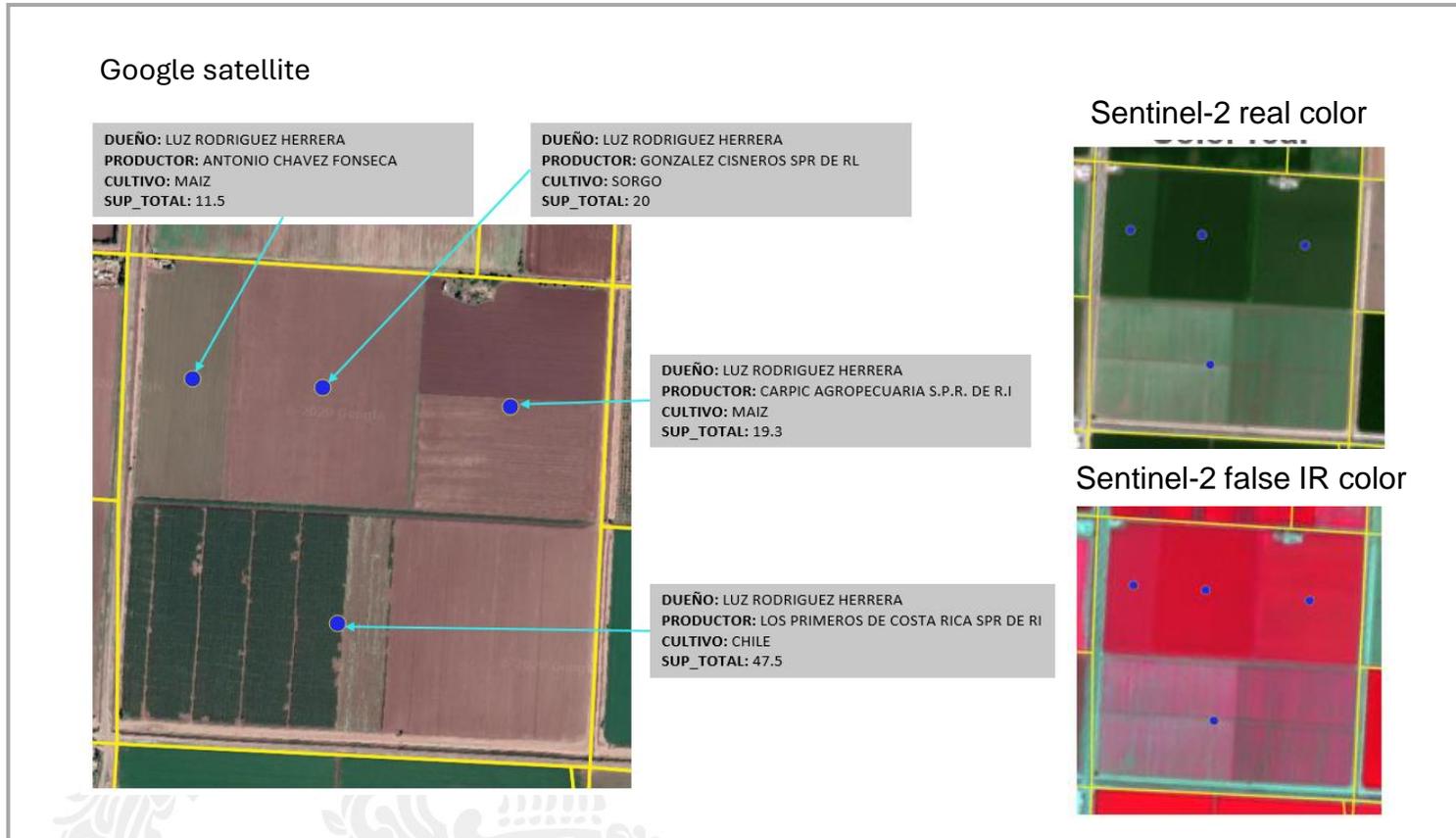


This activity added quality to update the Framework with information obtained from the AR.

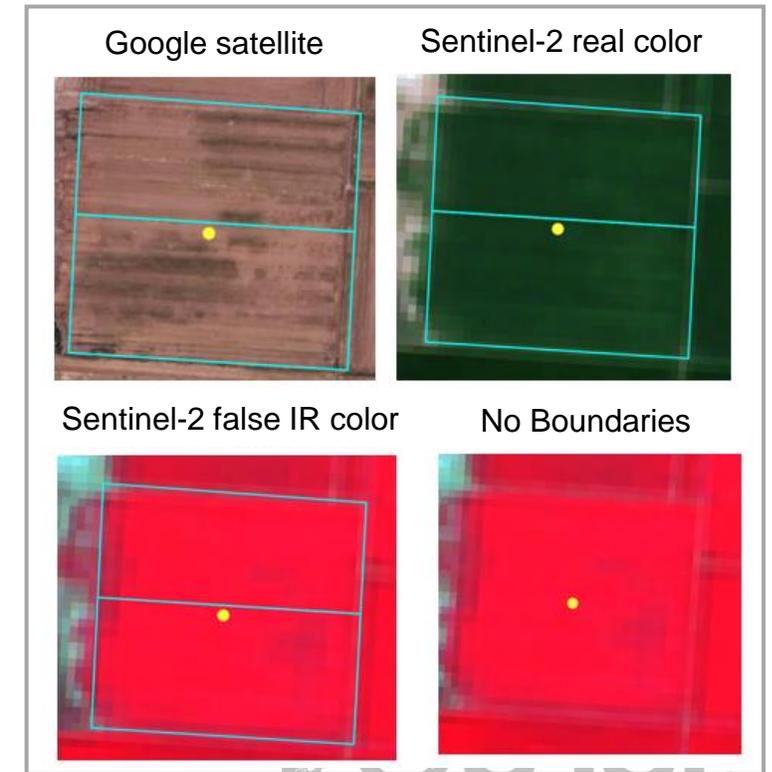
# Importance of satellite imagery in updating the Framework with AR

## Features in the satellite image

### Relation of several AR (points) to one land plot



### Relationship of an AR with two lands plots



These elements facilitated the interpretation of the information reported by the AR and thus identify the correspondence with the land of the Framework.

# Importance of satellite imagery in updating the Framework with AR

## Measure areas, using geographic features of the satellite image



This activity, besides to adding quality, made it possible to better interpret the information reported by the AR and thus identify the correspondence with the land plots of the Framework.

# Validation of crops obtained through Administrative Registers

## Results

**Total land plots  
updated:  
140,455**

Information updated	
Producer	34,114
Main activity	15,978
*Land plot information	71,700
Crops	129,712
Forest species	602
Livestock species	414
Confirmed the same information	3,972

\* Rights over land, tenure, land use and water availability.



# Thank you

[jose.hernandez@inegi.org.mx](mailto:jose.hernandez@inegi.org.mx)





# **New and emerging technologies through the lens of improving official statistics**

**Use of geospatial technologies to enhance the  
generation of official statistics in Colombia**



# Importance of geospatial information

---

The possibility of referencing official statistics produced by a country to its territory is not only fundamental but also allows for added value within the entire statistical process. These statistics, besides occurring at a specific place and time, must enable spatial analysis to build capacities in national and territorial entities and spatially empower them.

## Current information demands require comprehensive data to:

- Measure and monitor the framework of global indicators for sustainable development goals.
- Make comparisons at different scales or levels, including local, subnational, regional, and global.
- Facilitate data exchange between institutions.
- Provide more detailed information for smaller areas.
- Integrate with new data sources and utilize them in statistical generation.

---

## Integrating statistical information with geographic data



- Provides spatial analysis tools for statistical data
- Provides a more comprehensive understanding of the territory through studies and statistical operations, aiding decision-making.



# Frameworks for improving official statistics from a geospatial perspective

## Integration - Statistical and Geospatial Information

### ◆ Expectations

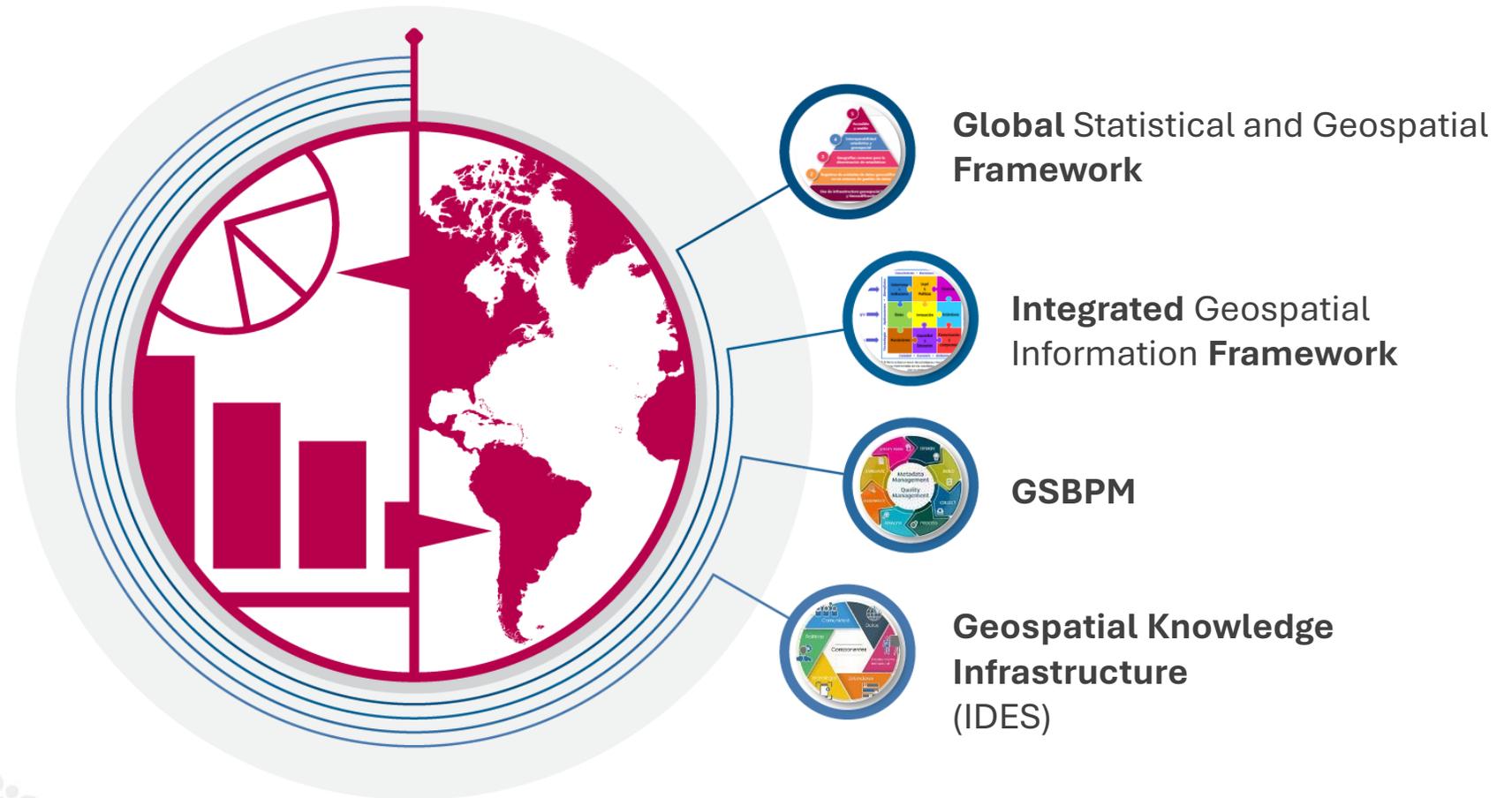
Applying the frameworks and implementing them in the processes

### ◆ Objectives

Improving the processes implemented in all statistical operations by increasing the use of new technologies.

### ◆ Purposes

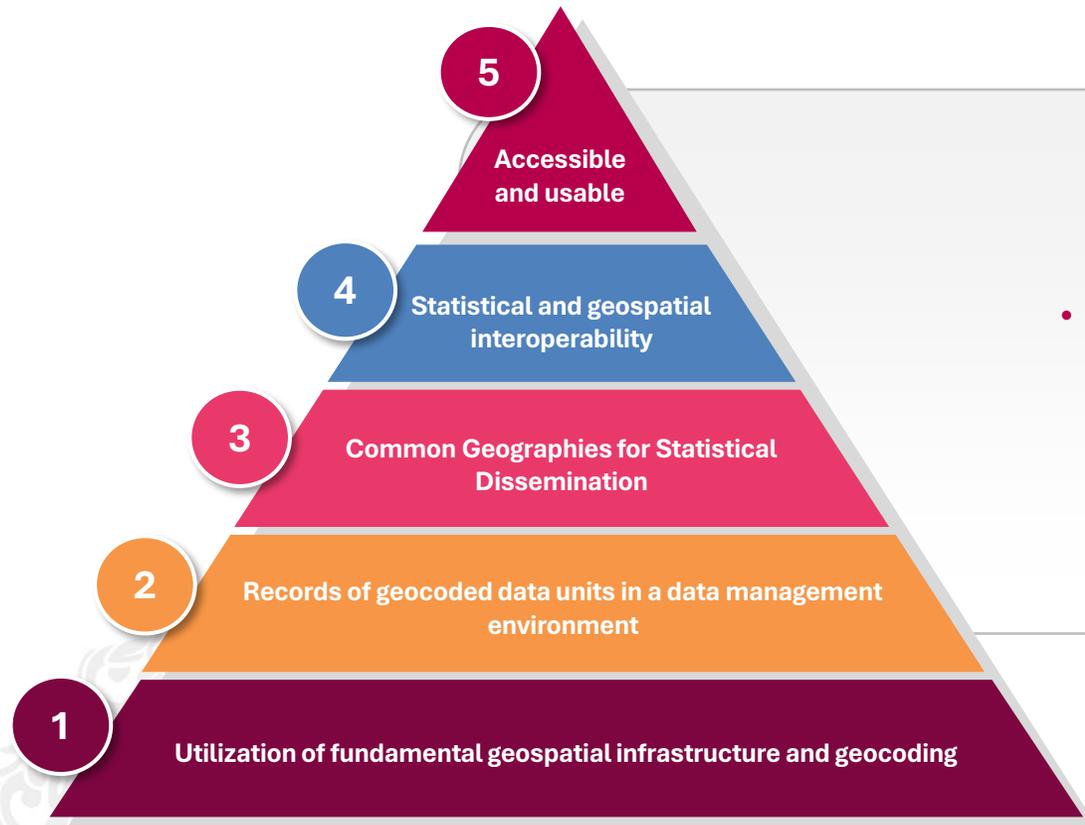
Generating value with high-quality data and statistics for public policy



# Guidelines for the appropriate use of geospatial information

## ① Global Statistical and Geospatial Framework: Five Principles

The main objective of the GSGF is to enhance the integration of geospatial and statistical information at the global level, assisting countries in improving the quality and accessibility of geospatial and statistical information.



- It enables the **integration** of a variety of data from both the statistical and geospatial communities.
- Through the application of its five principles and key supporting elements, it allows for the **production of statistical data with harmonized and standardized geospatial capabilities**.
- The resulting data can be integrated with other datasets to inform and facilitate **evidence-based decision-making**.

# Guidelines for the proper use of geospatial information

## ① Integrated Geospatial Information Framework

**Vision:** The efficient use of geospatial information by all countries to measure, monitor, and achieve sustainable social, economic, and environmental development, leaving no one behind.

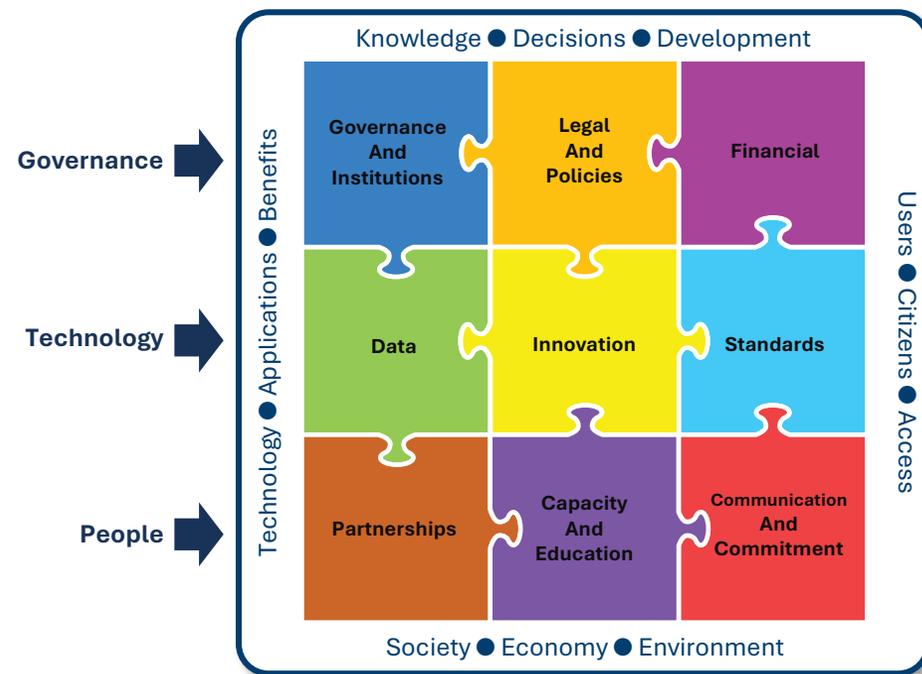
The framework is anchored in nine (9) strategic pathways and three (3) main **areas of influence**: governance, technology, and people.

The aim of these strategic pathways is to guide governments towards the implementation of integrated geospatial information systems in a manner that provides a vision for sustainable social, economic, and environmental development.

[https://ggim.un.org/IGIF/documents/PARTE\\_1\\_MARCO\\_%20ESTRATEGICO\\_GLOBAL.pdf](https://ggim.un.org/IGIF/documents/PARTE_1_MARCO_%20ESTRATEGICO_GLOBAL.pdf)

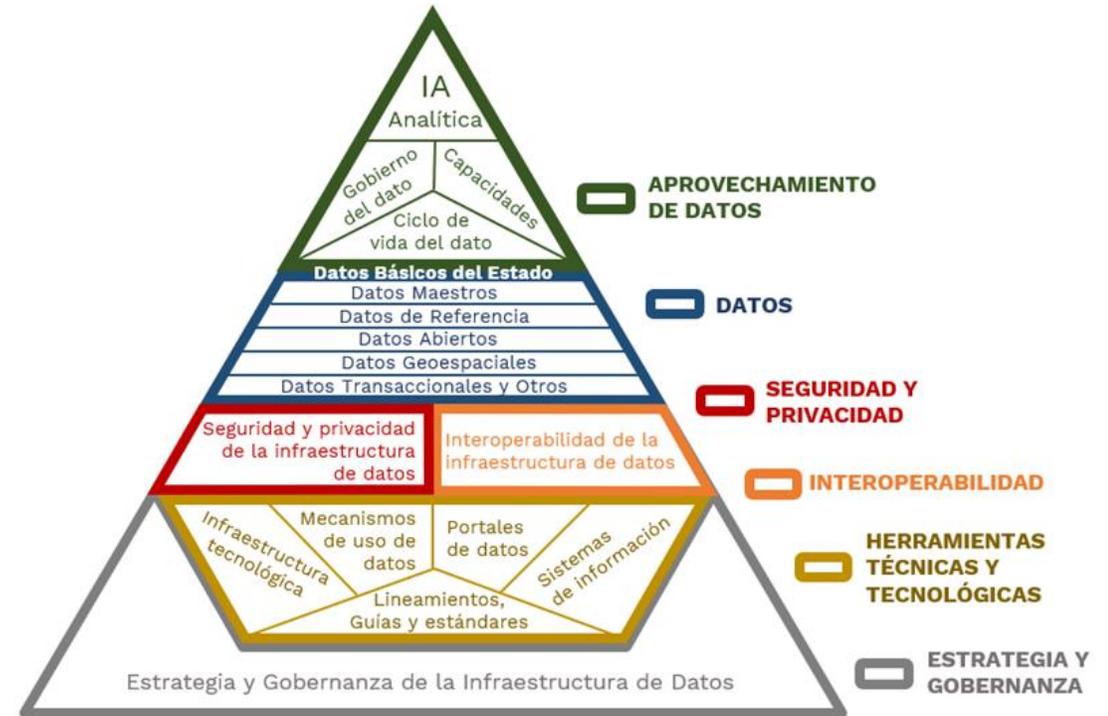
## Basic Principles:

- Strategic Enablement
- Transparent and Accountable
- Reliable, Accessible, and User-Friendly
- Collaboration and Cooperation
- Integrated Solution
- Sustainable and Valued
- Leadership and Commitment



# Colombian State Data Infrastructure

- 1. Governance Strategy of the State Data Infrastructure:** Policies, regulations, guidelines, and standards.
- 2. Technical and Technological Tools:** Instruments that facilitate the utilization of the infrastructure by various stakeholders.
- 3. Interoperability of the Infrastructure and 4. Data Security and Privacy:** Structural elements for the development and maintenance of the data infrastructure.
- 4. Data:** Central and most important asset of the Colombian State data infrastructure.
- 5. Utilization of the Data Infrastructure:** Ultimate objective of managing and implementing the infrastructure.



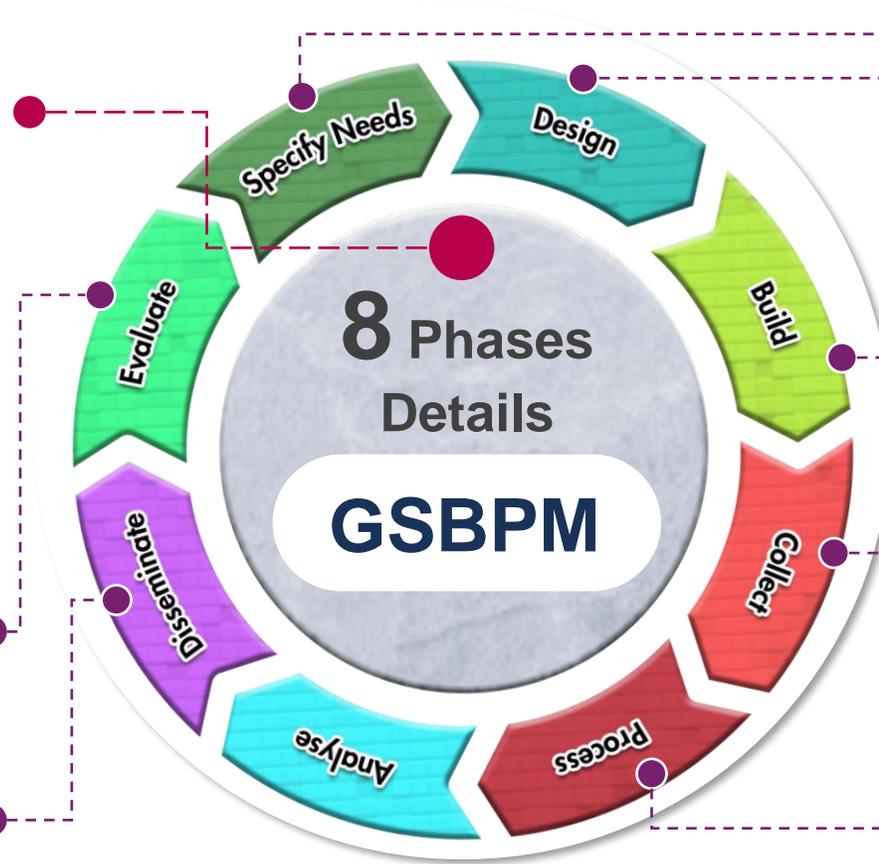
# For whom are the data integrated: Current and projected users.



# Use of geospatial data within the statistical process

## Address normalization and georeferencing

Applications for geocoding observation units



### CEED App

For the collection of the Building Census



### Cartographic Updates App

For registering cartographic updates in the field



### RUE App

For registering count data



### ENA App

App for collecting National Agricultural Survey



### Statistical Operations Monitoring Geoviewers

Georeferenced Tracking for Surveys.

### Housing variable dissemination geoviewer

For registering and visualizing the housing variable



### Geoviewer for counting economic units

Economic unit counting geoviewer



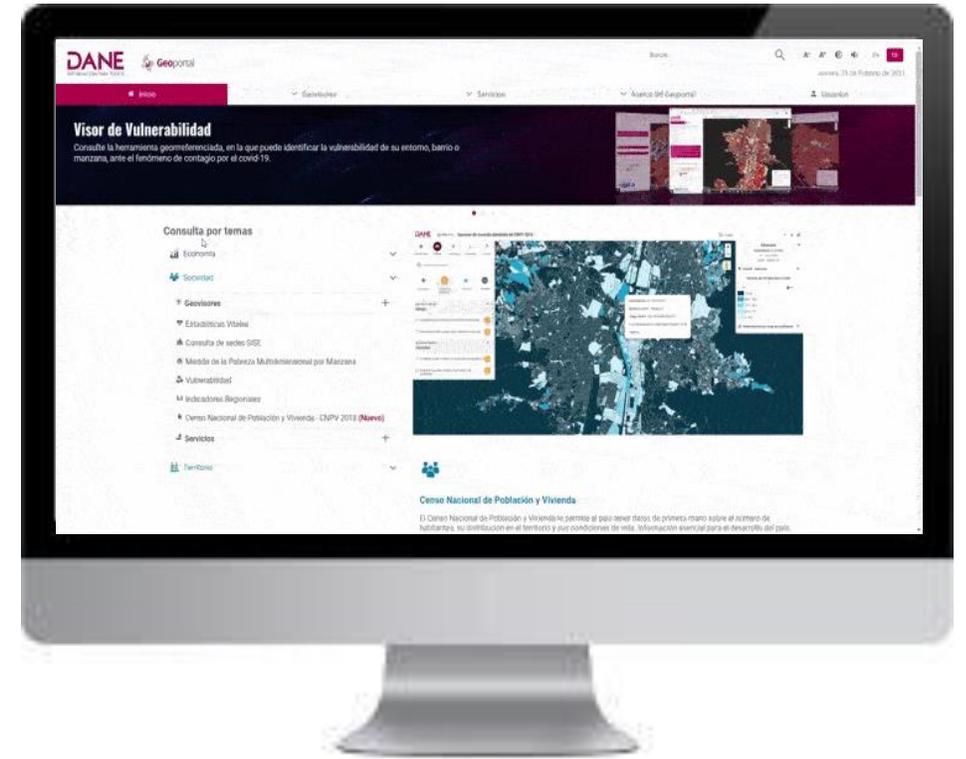
# Description of the Geoportal



The DANE Geoportal is the website serving as a gateway to statistical information resources produced by DANE that have territorial disaggregation.

It provides query tools such as Geoviewers, interactive maps, mobile applications, and data downloads, among which stand out the National Geostatistical Framework, the National Address Framework, and the Political-Administrative Division.

 <https://geoportal.dane.gov.co/>





Gracias



---

Thank you

