New methods and data sources for official statistics

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

Chair: Tobias Thomas

Participants:

Alejandro Ruiz: Social and economic indicators from transactional banking data

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

Tomas Rudys: The use of scanner data in official statistics

Peter Knizat: Nowcasting industrial production index with high-frequeny highway toll data

Generation of Economic and Social Indicators from Banking Transactions.

Alejandro Ruiz
Researcher

The information contained in this presentation is not part of the official statistics of INEGI. Opinions and comments attributed solely to the researcher and do not necessarily reflect an institutional stance.

Challenge we face

- We, as NSO, face the challenge of collecting sensitive information, such as personal income or expenditure data.
- Income and expenditure data is important for public policy —well-being, labor market, fiscal policy—.
  - *National Survey of Household Income and Expenditure (ENIGH).*
  - *National Survey of Occupation and Employment (ENOE).*

However...

- Misreporting.
- Undercoverage.
- The data is not recalled properly.
- There is a growing demand for more disaggregated, timely, and frequent information.

Public-Private partnerships for leveraging privately held data

IOS-ISI 2024
MEXICO CONFERENCE

Bilateral agreements:

BANORTE, BBVA & SANTANDER

- Currently, at the state level, we can know how households are faring in terms of their economic well-being every two years. For the 2 469 municipalities , we can only access income information every five years, with no expenditure/consumption data.
- Now, for some subpopulation, we will have public, frequent, and quality municipality information on their economic well-being.
- There is no economic or in-kind compensation.

Transactional data sets

Payroll

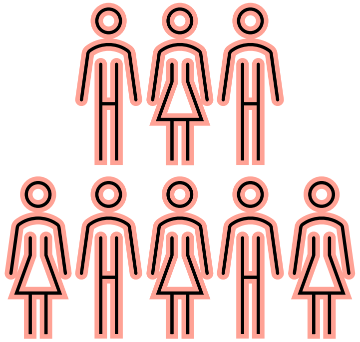
- Workers & retirees.
- Salaries, bonuses (Christmas bonuses & profit-sharing), severance pay.

Expenditure

- Debit and credit card transactions:
    1) Purchases.
    2) ATM cash withdrawals.
- On-line and In-person.

Sales

- Debit and credit card.
- On-line and In-person .



Demography & Geography

Statistics based on Payroll Transactions

Henceforth *payroll-disbursed income = income*

Payroll

Monthly data on 18 million clients

Statistics based on sex and age group:

- National,
- 32 states,
- 2 400 municipalities.

Statistics are calculated within the bank's servers

There is no transfer of personal information.

IOS-ISI 2024 MEXICO CONFERENCE

Who is represented in the data?

Official data sources:

41 million wage and salaried workers + retirees

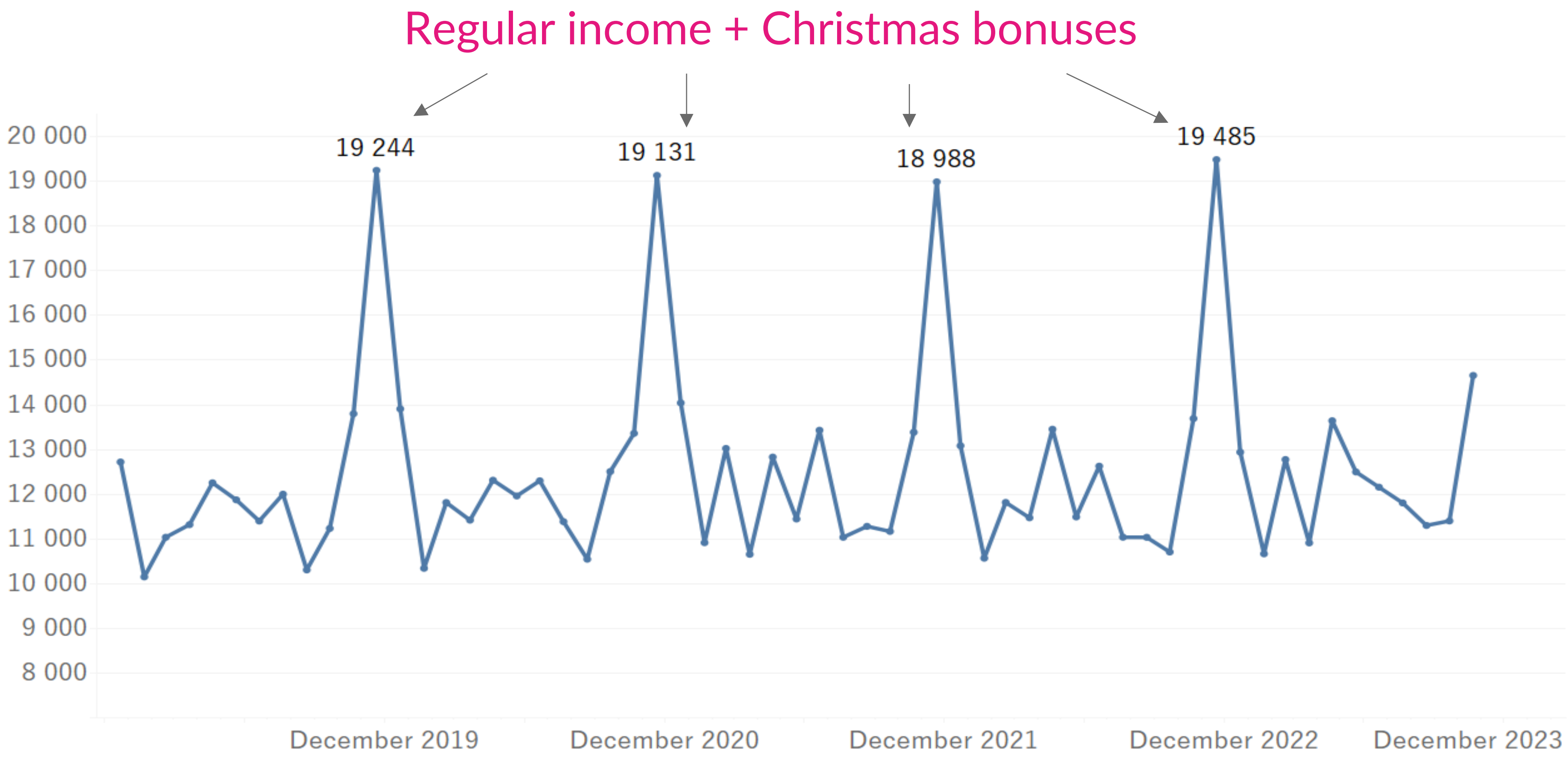
↓

24 million have a bank account.

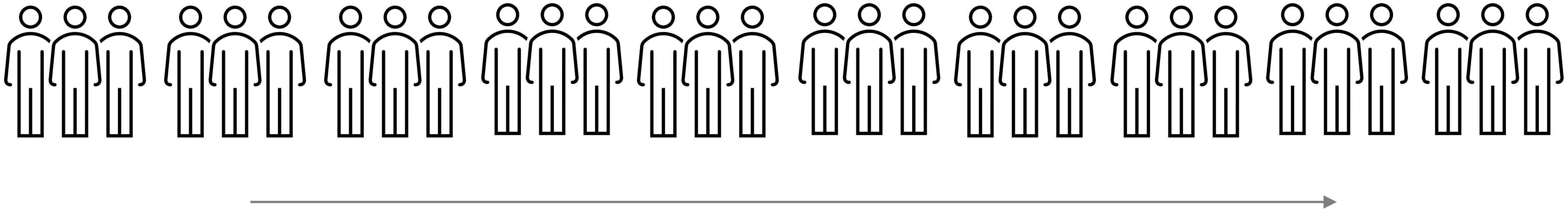
Most of them also have access to healthcare services (proxy for formal labor market ≈ half of the total labor market).

Monthly income

Average monthly income per bank-client = 15 000 (880 dollars, 1 dollar = 17 pesos)

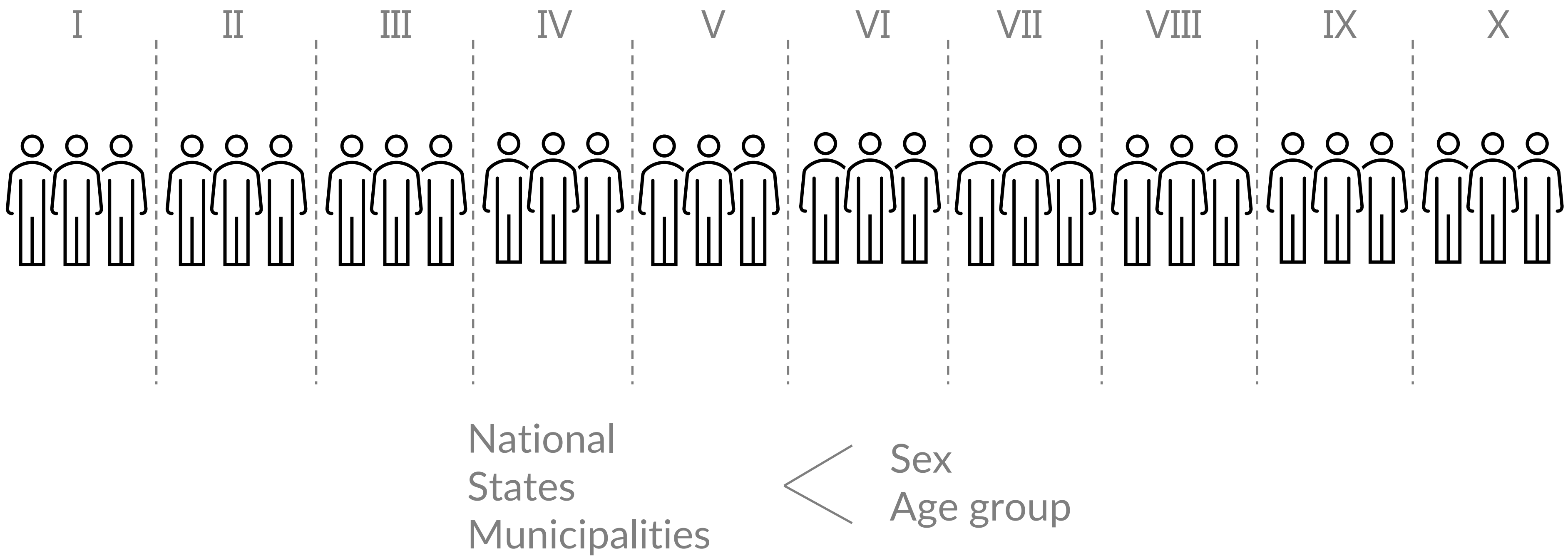


Decile Average



\- Monthly payroll dispersion +

Decile Average



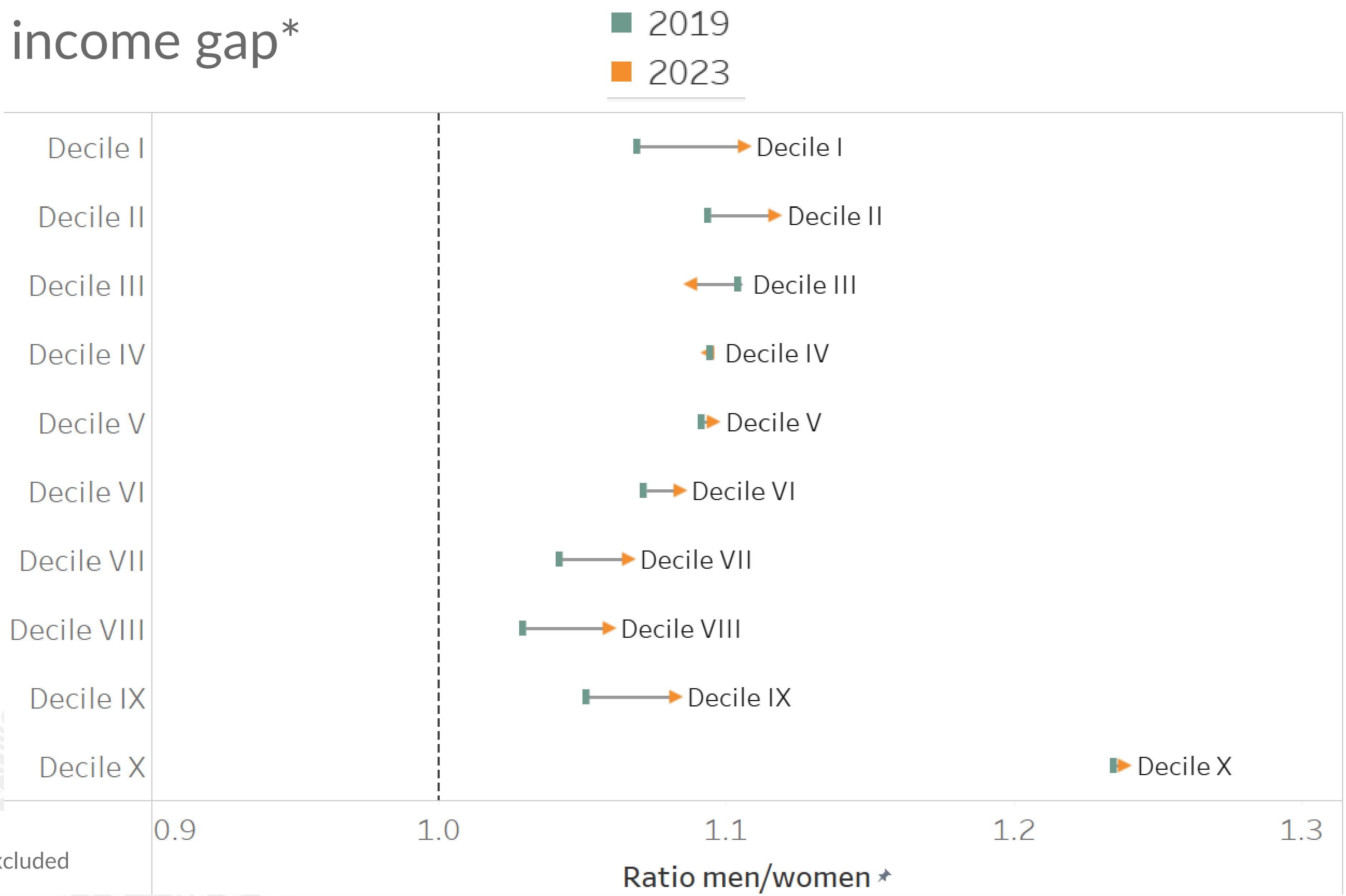
- This computational process is carried out on the bank's servers.
- INEGI receives the decile averages from each bank.
- The decile averages that would be made public result from a weighted average.

This data can contribute to the discussion of relevant topics:

1. Gender Income Gap.
2. Dynamics of the formal labor market by age group.
3. Poverty measurement.

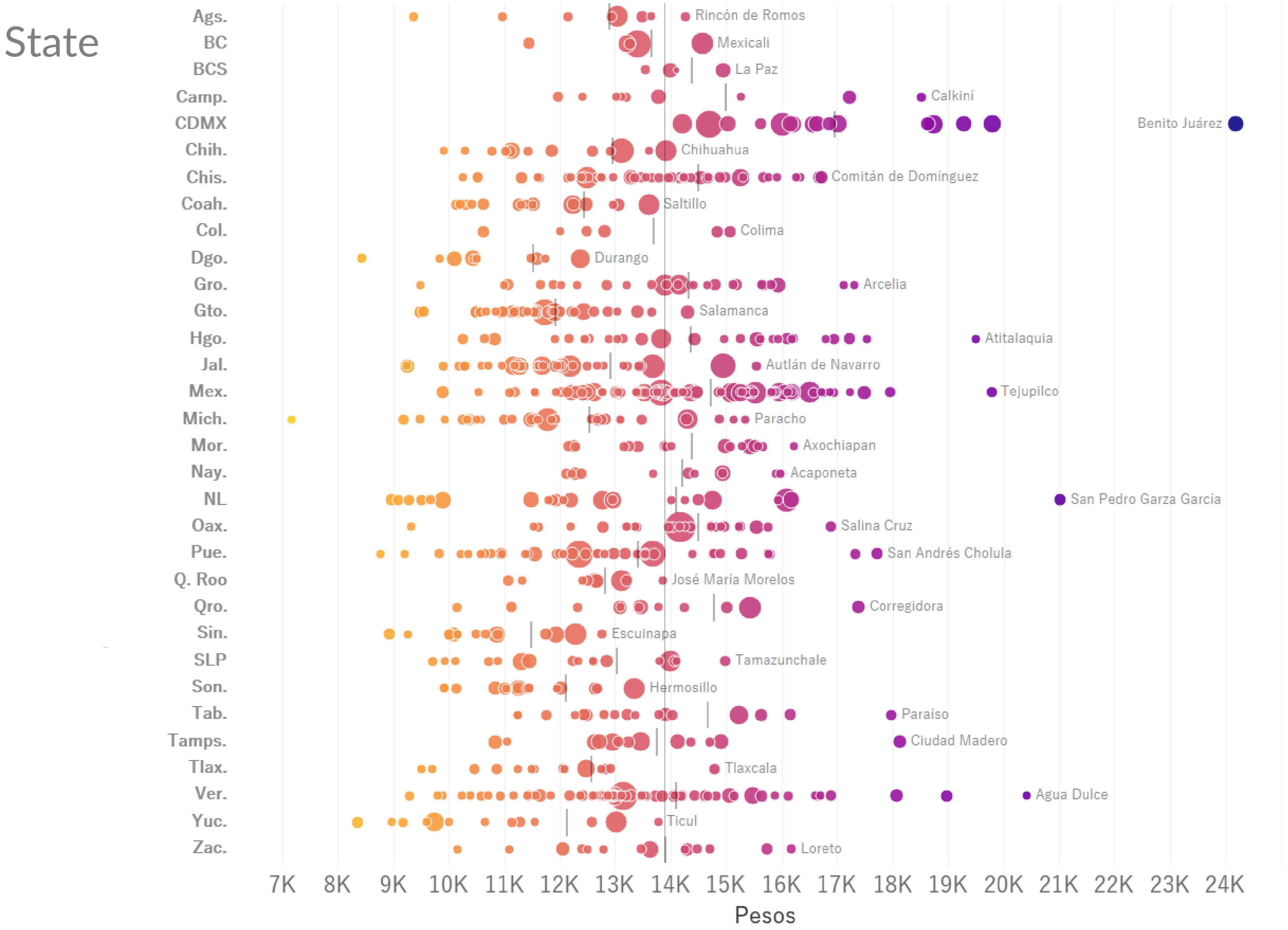
Gender gaps

Gender income gap*

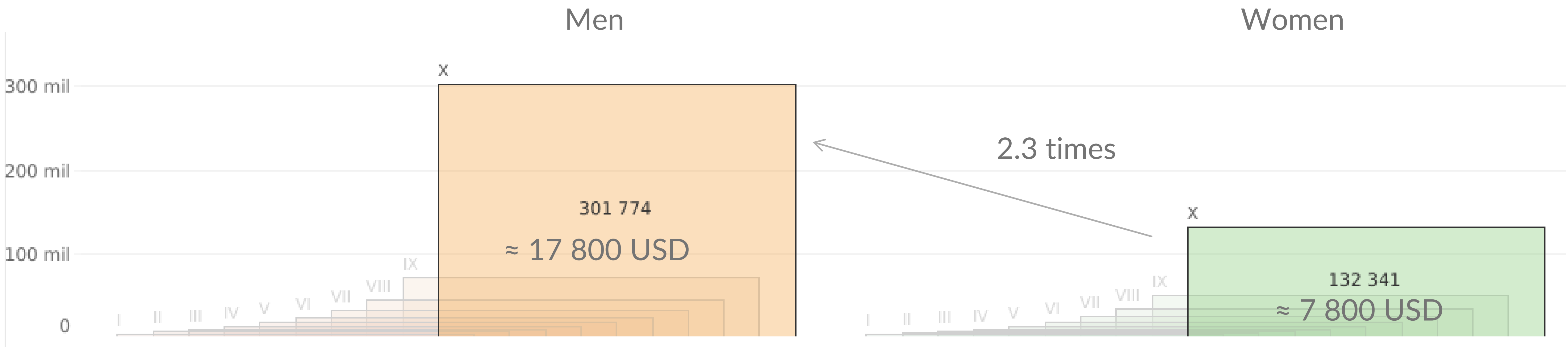


* December was excluded

Monthly average payroll for women, by municipality.



Gender gap in one of the richest municipality. Decile X.

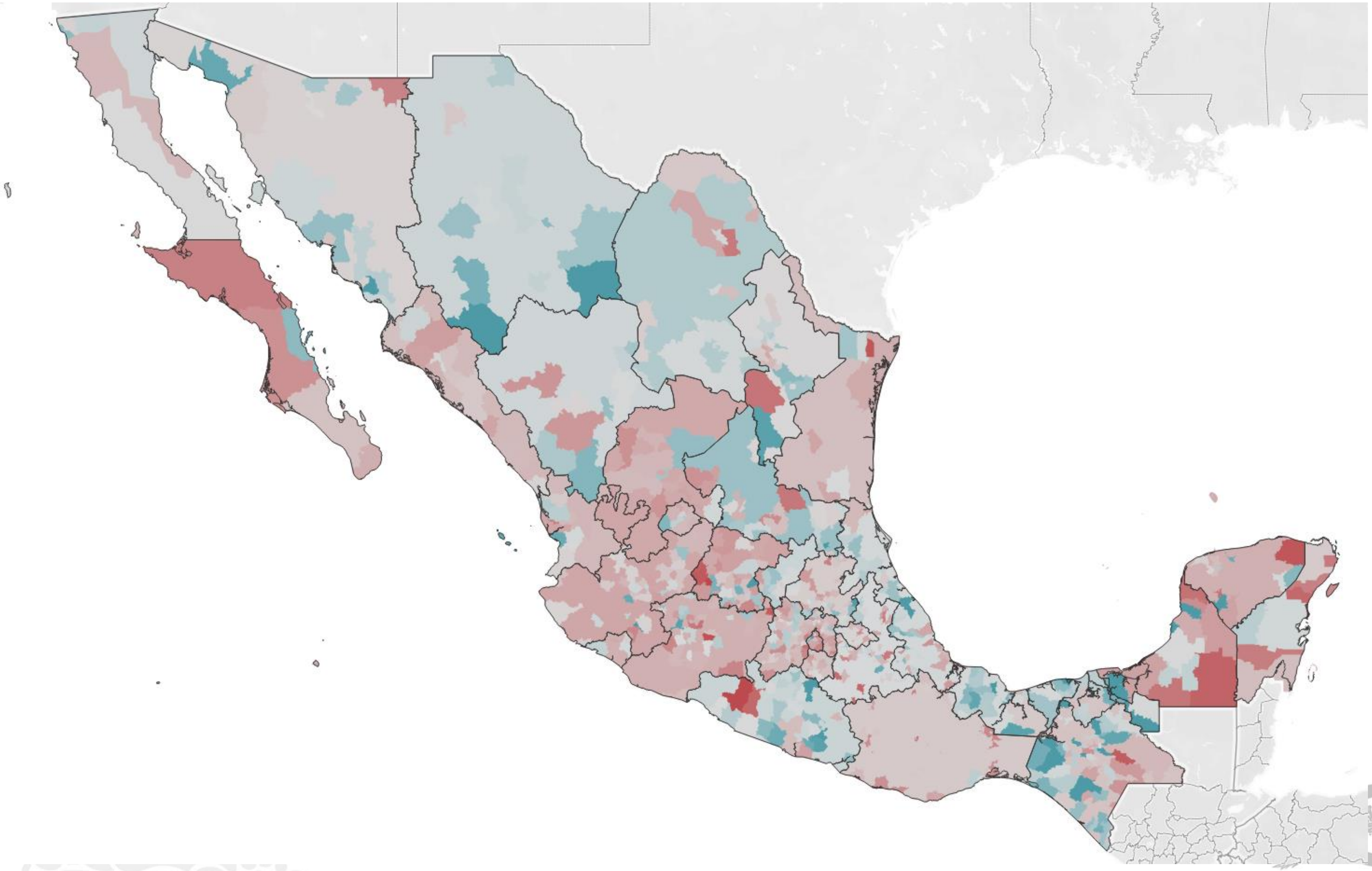


Dynamics by age group

Growth in number by age group:

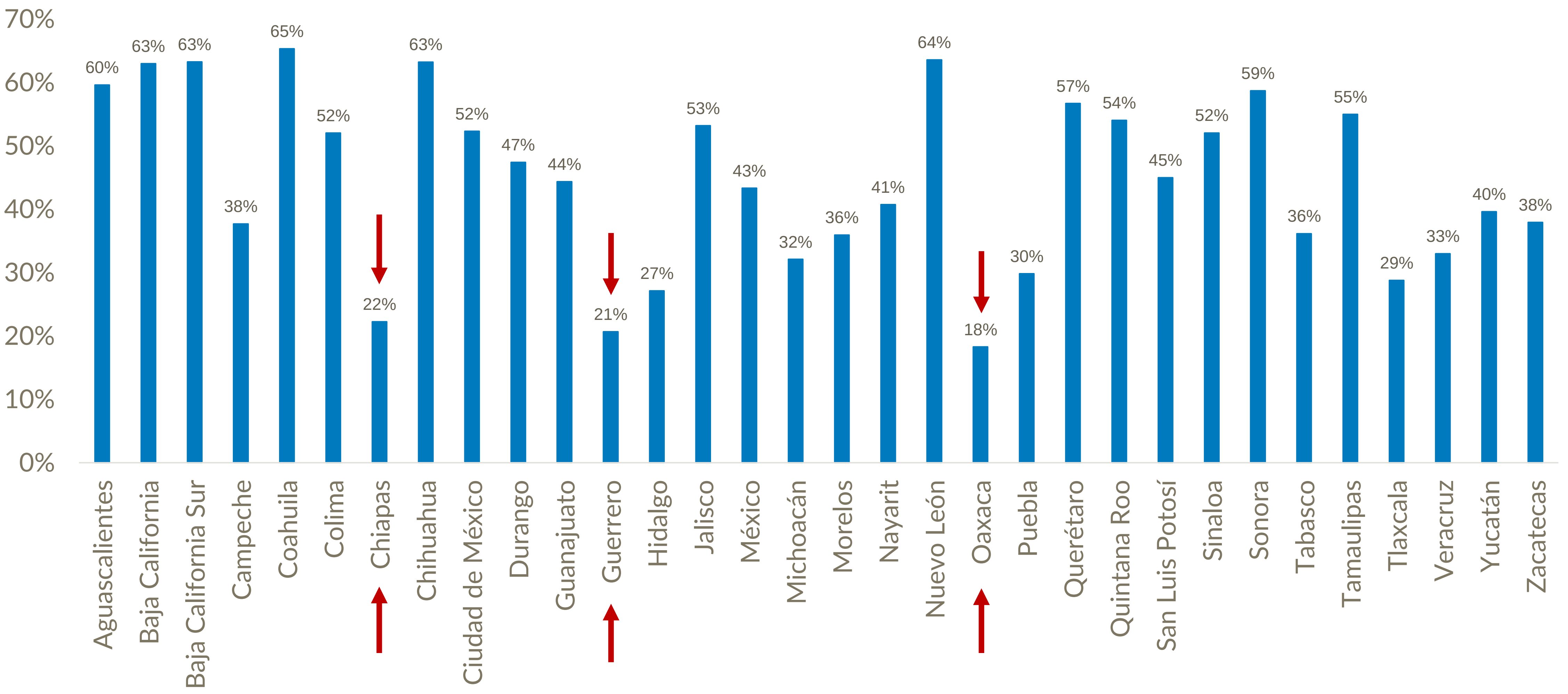
Age group	Bank clients	ENOE	
		Workers in formal sector + retirees	Workers in formal and informal sector + retirees
24 or younger	-6%	-5%	0%
25 a 34	4%	6%	5%
35 a 44	6%	4%	3%
45 a 54	10%	12%	8%
55 a 64	15%	15%	10%
65+	30%	23%	14%

Regions experiencing an increase or decrease of the youngest. 2022 vs 2019



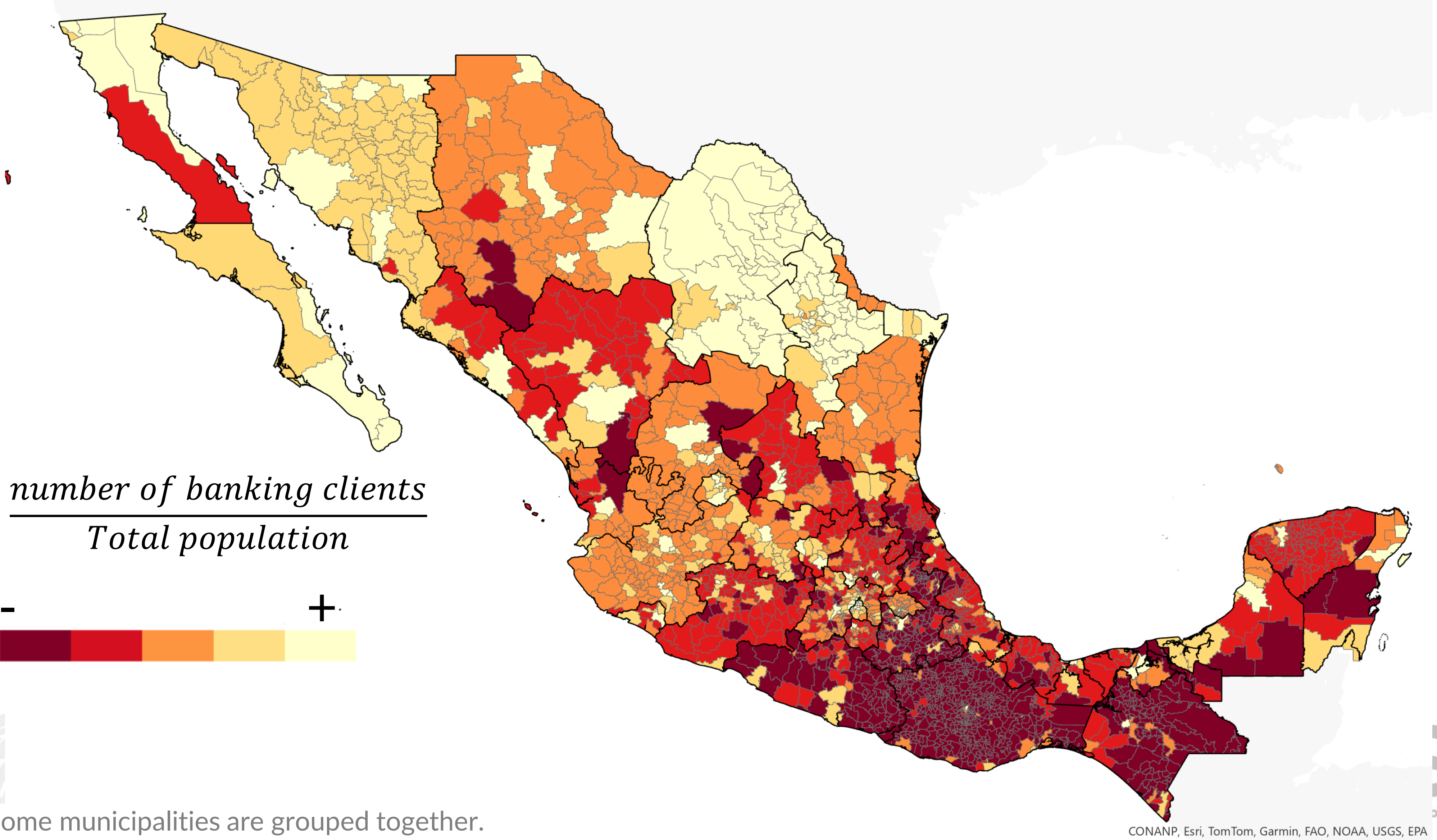
Poverty measurement

Formal sector share.



2024 CE

Coverage rate, 2020.*

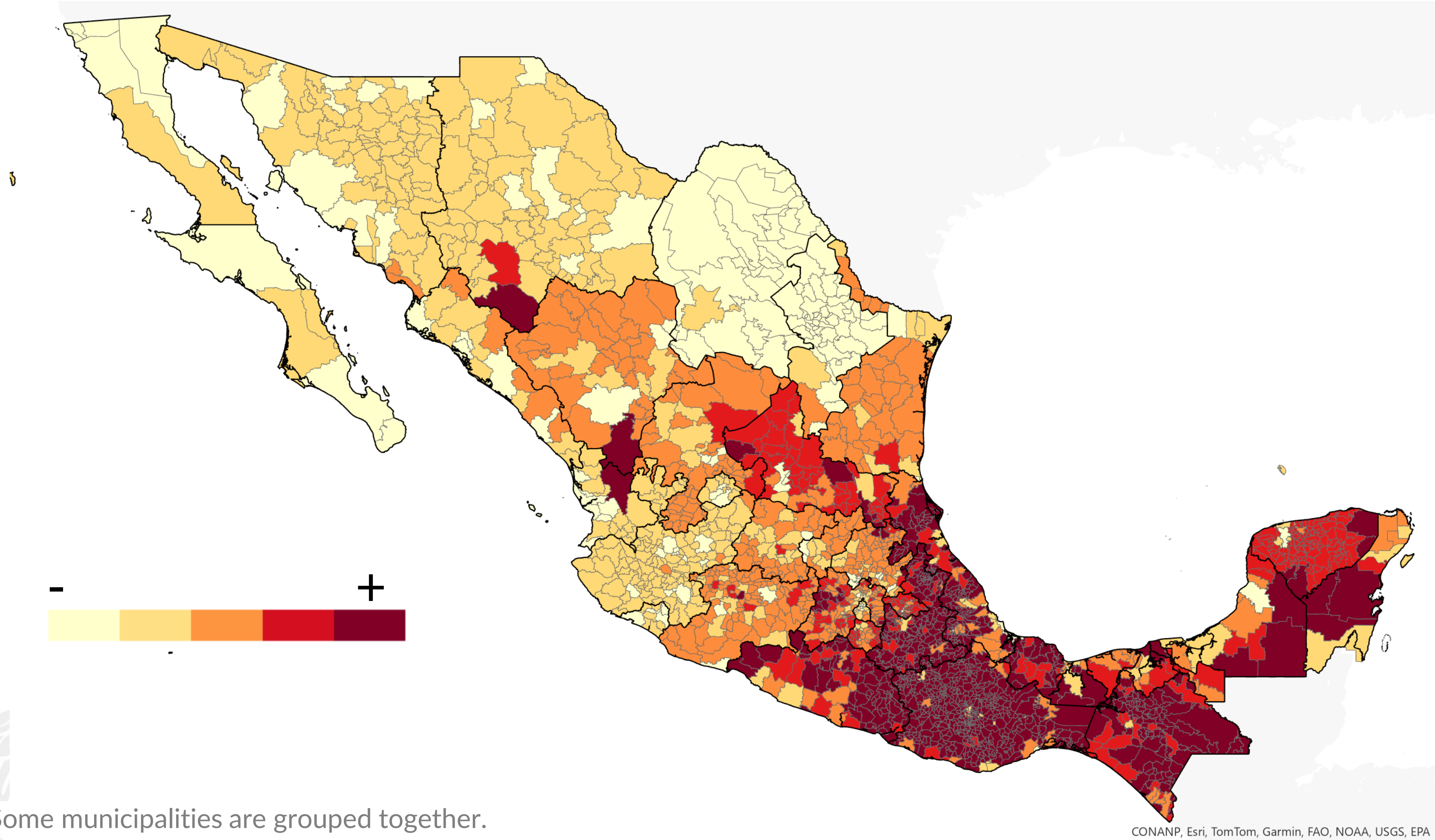


$$CR = \frac{number\ of\ banking\ clients}{Total\ population}$$

\- +

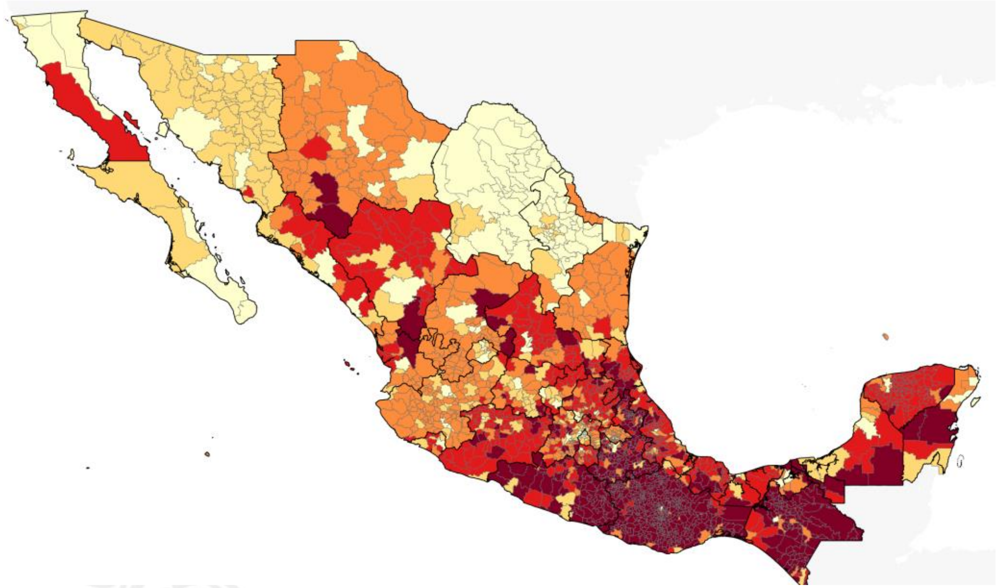
* Some municipalities are grouped together.

Poverty rates based on official data, 2020.*

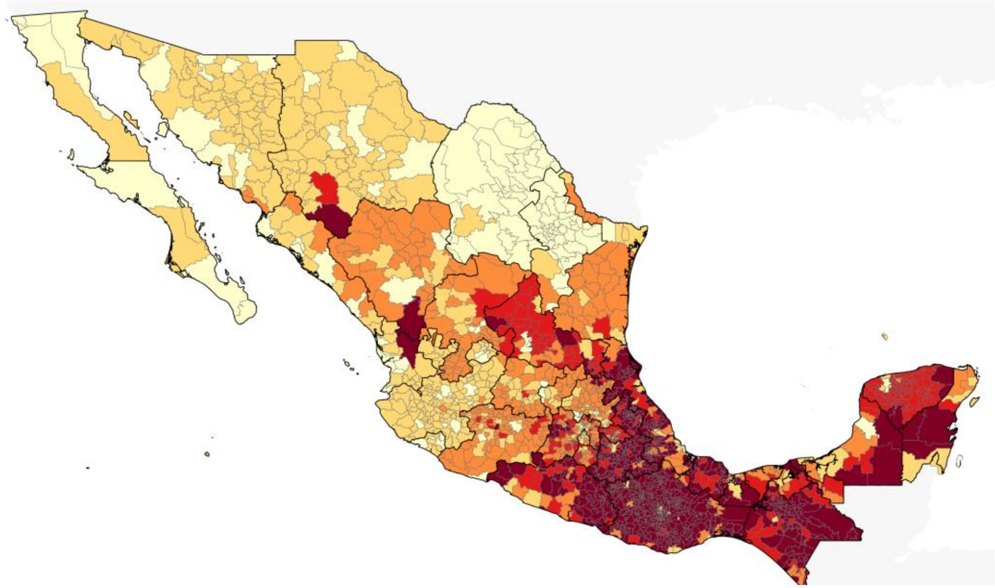


* Some municipalities are grouped together.

Coverage rate



Poverty rate



Correlation -0.8

R^2 0.79 (controlling by state)

MEXICO CONFERENCE

To wrap up

Monthly statistics on payroll dispersion:

- Dynamics in the number of people receiving payroll.
- Average payroll and average payroll by decile.

- National
- 31 States + CDMX
- More than 700 municipalities or regions.

What is next

- Publishing payroll information.
  - Talking to stakeholders.
- We will try to strength this project by:
  - Reinforcing the importance of this collaboration with the current banks = *Long-term relationship*.
  - Add more financial institutions.
- We will be working on expenditure data.

Thank you

jose.ruizs@inegi.org.mx

IAOS-ISI 2024
MEXICO CONFERENCE

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

INEGI International Statistical Institute IAOS IMPROVING OFFICIAL STATISTICS

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

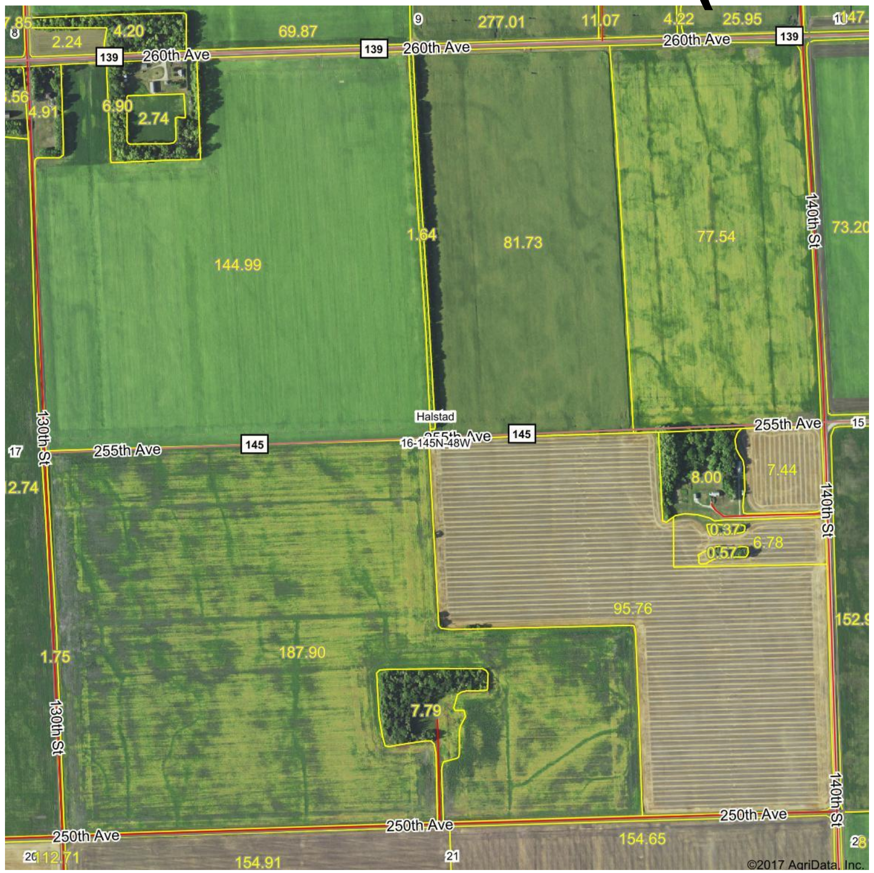
Question: What can be done to alleviate these concerns?

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

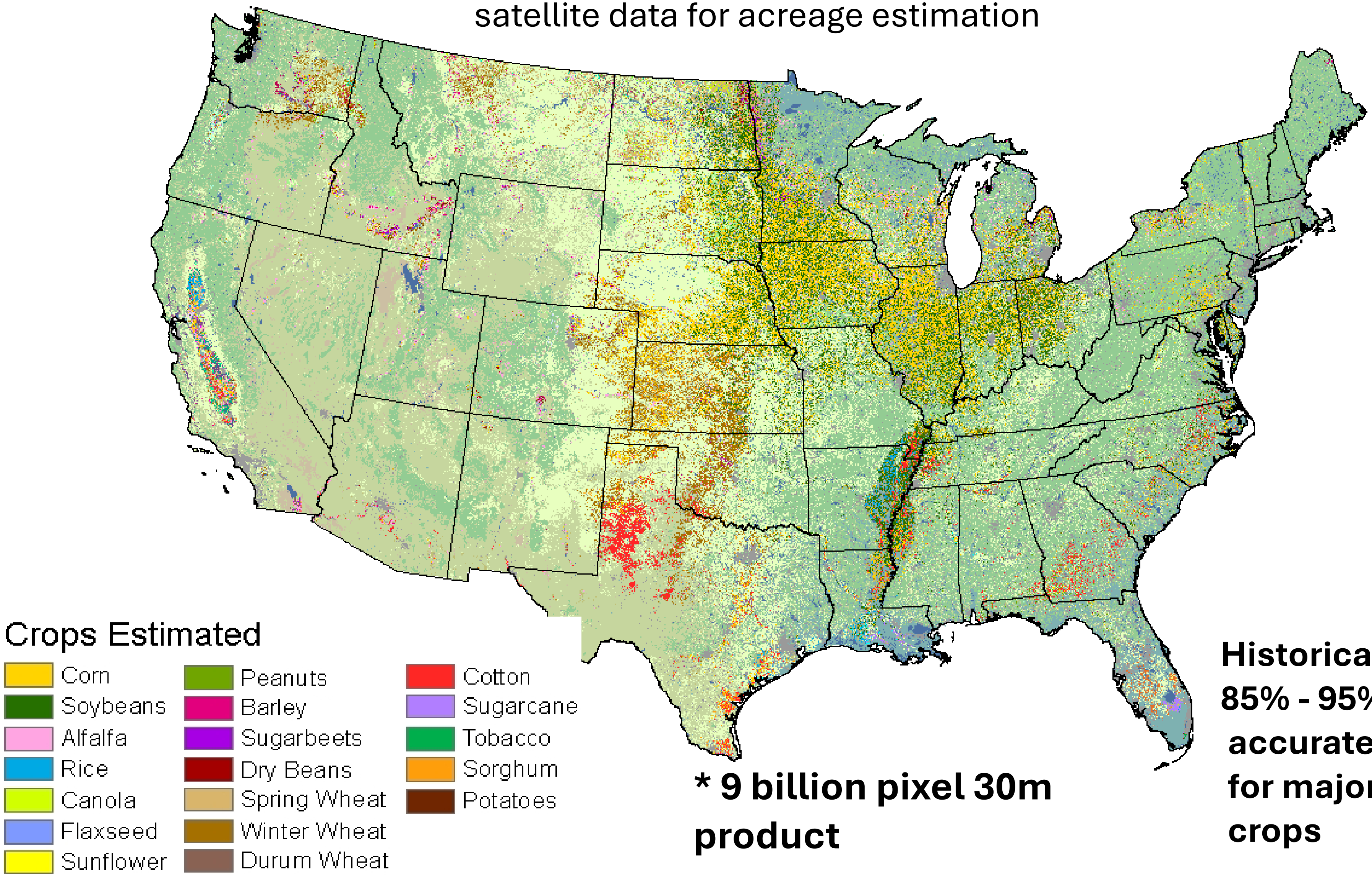
Common Land Units (CLUs)



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

Cropland Data Layer (CDL)

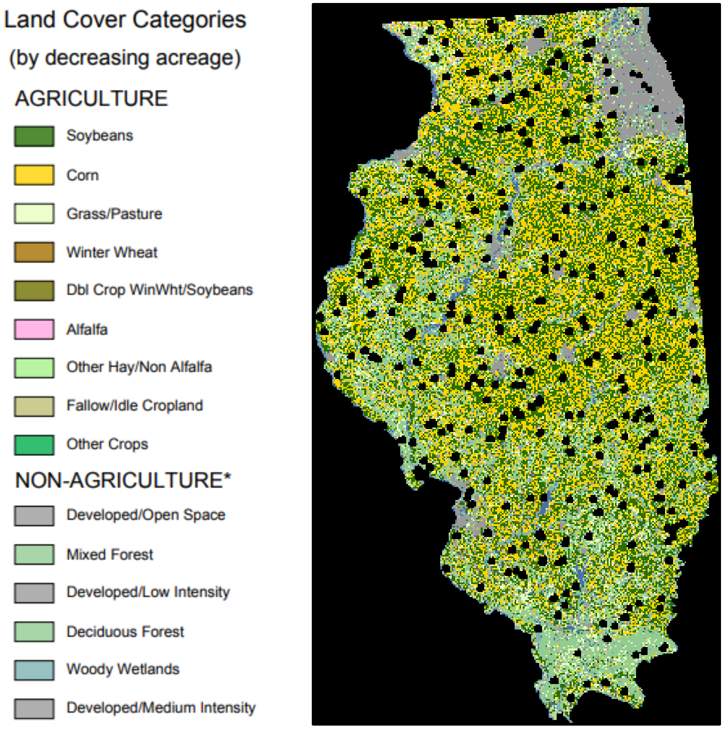
Annual national coverage since 2008
A raster, crop-specific, land cover data set produced using satellite data for acreage estimation



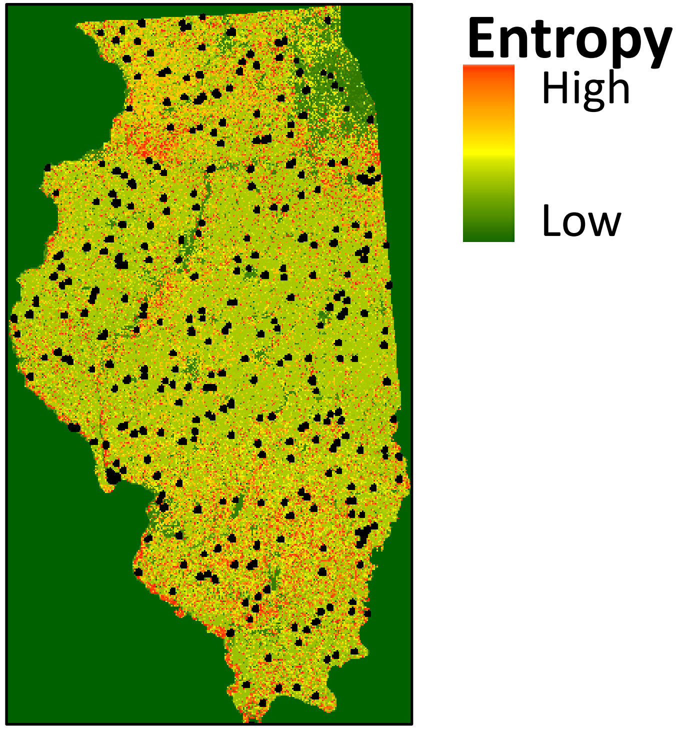
*** 9 billion pixel 30m product**

Historically, 85% - 95% accurate for major crops

Predictive Cropland Data Layers and Entropy Layers



Illinois (2021) PCDL and Segments



Illinois (2021) Entropy Layer

PCDL based on
High-Order Markov Chains

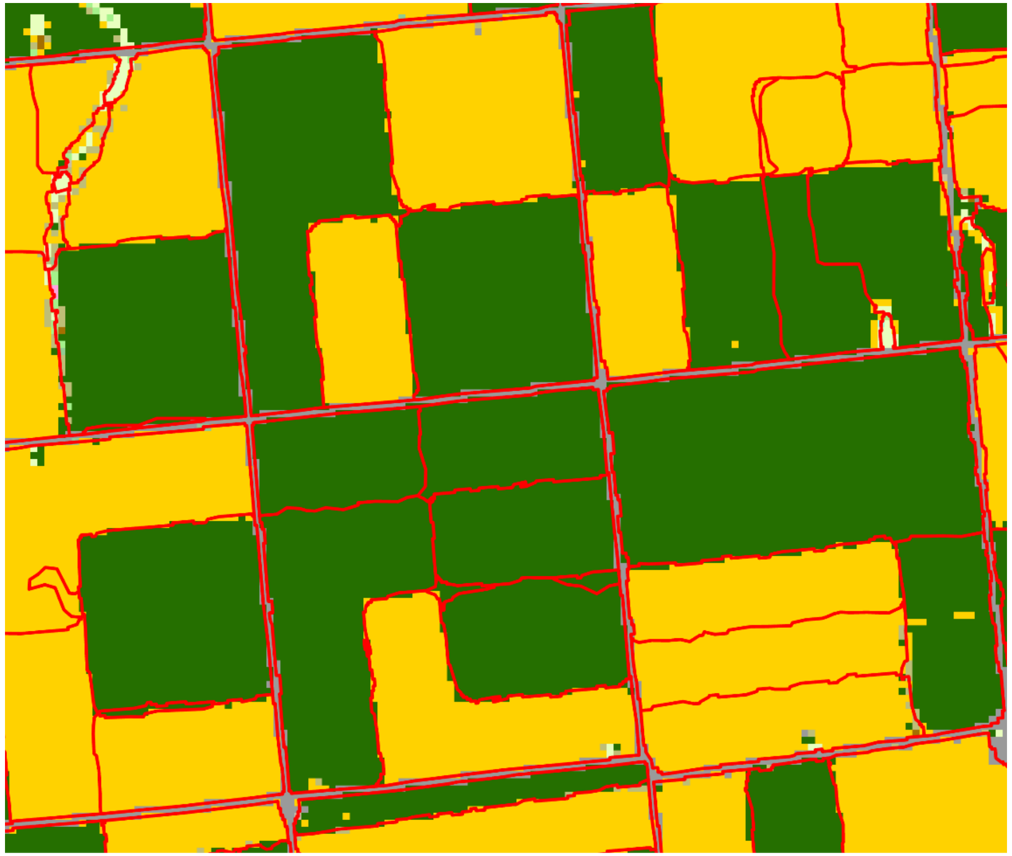
Entropy layer based on
normalized Shannon entropy from the predictive distribution

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

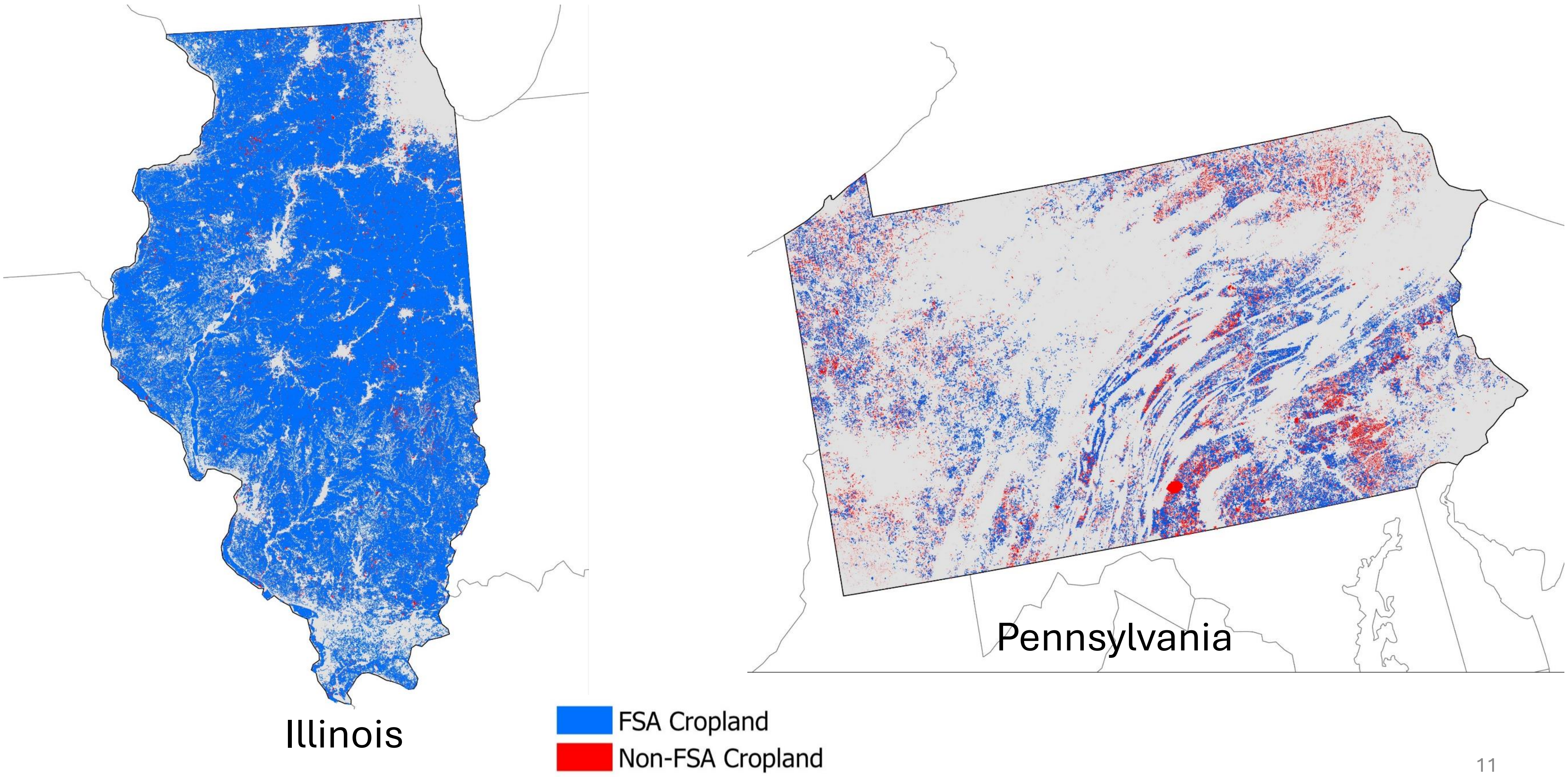


Applications Leveraging All Data

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



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)

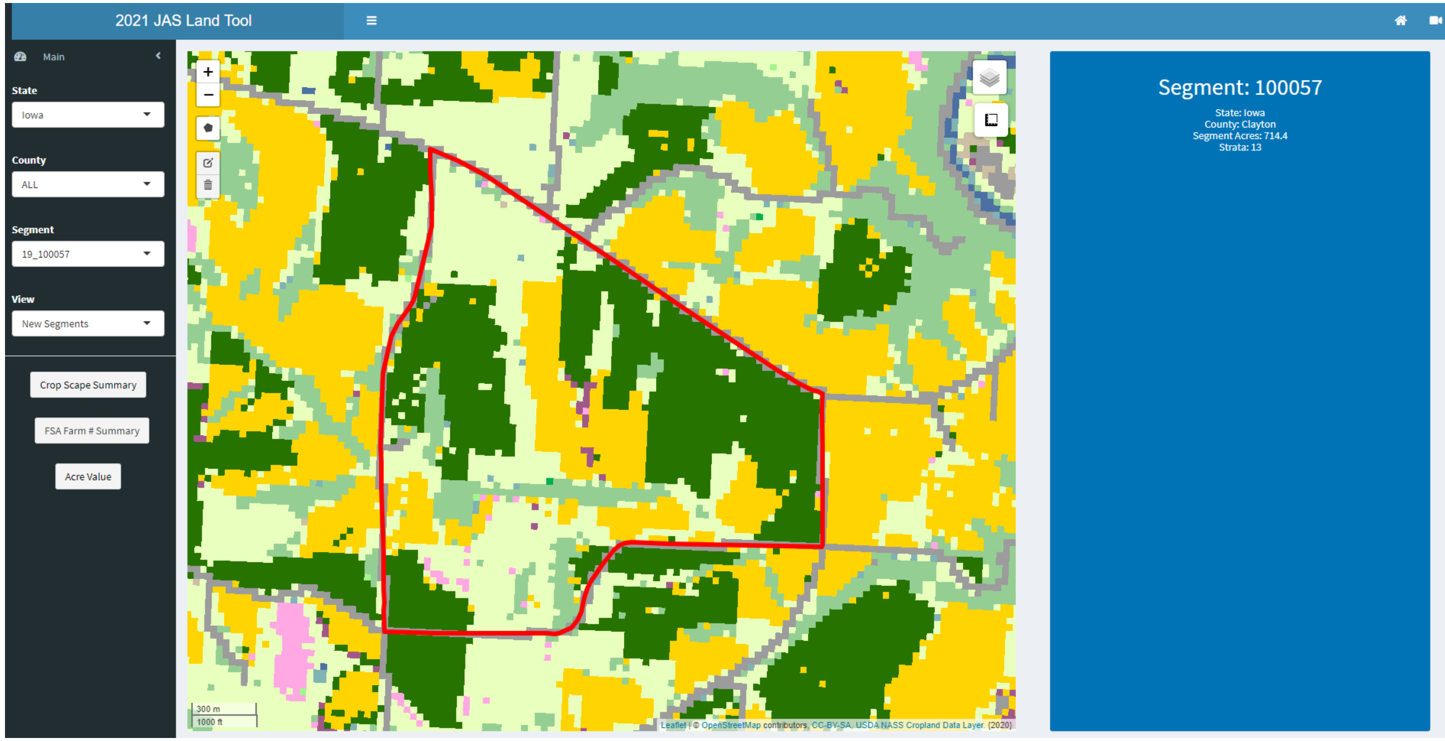
Segments divided into tracts, representing unique operations

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

IOS-ISI 2024
MEXICO CONFERENCE

Estimating Planted Acreage: Corn

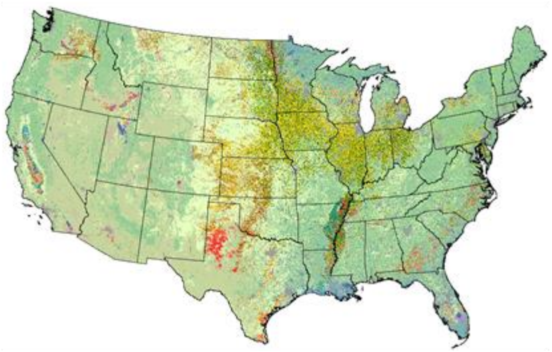
Agricultural Survey

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

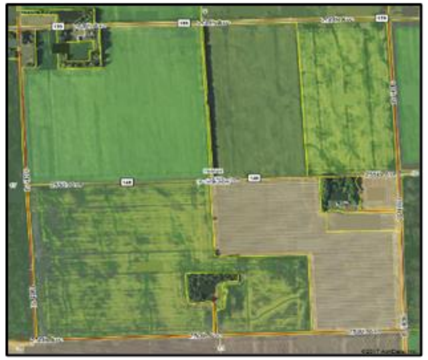
County Agricultural Survey

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

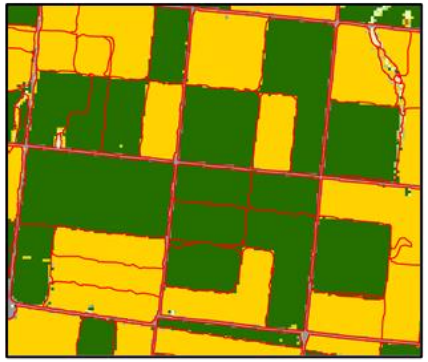
Wealth of Non-Survey Data



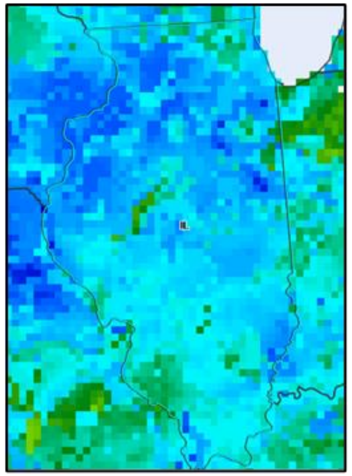
Cropland Data Layers (CDL)



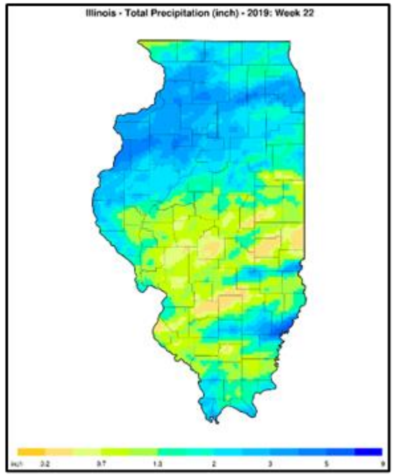
FSA Common Land Unit and 578 data



Crop Sequence Boundaries ("Fields")



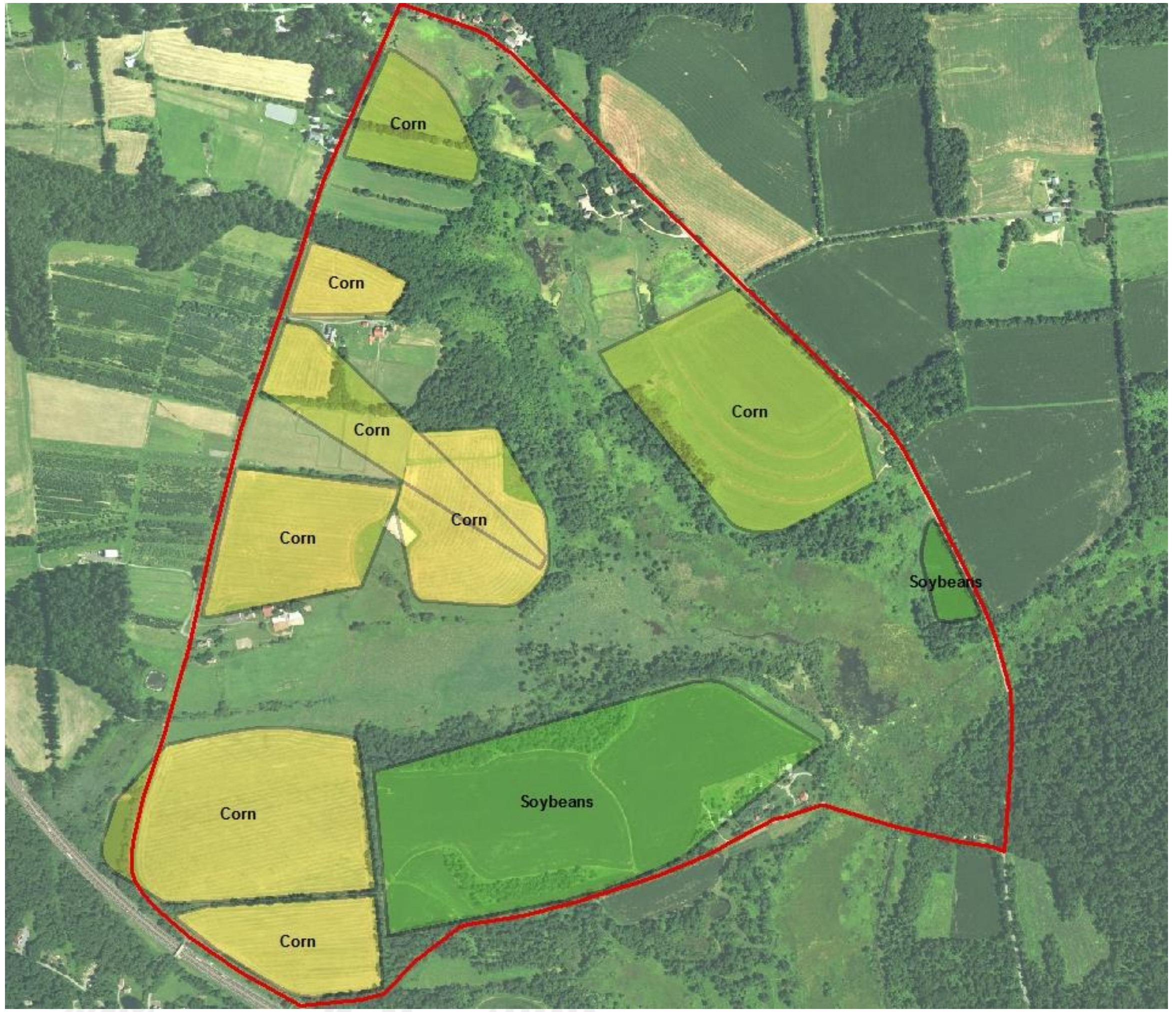
Soil Moisture Data



Precipitation Data

Early Season CDLs

Ready to Link Survey and Non-survey Data?



- Non-survey data are geo-spatially 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 (yield) · (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

IOS ISI 2024 MEXICO CONFERENCE

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.

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

International Statistical Institute

IMPROVING OFFICIAL STATISTICS

THE USE OF SCANNER DATA IN OFFICIAL STATISTICS

Tomas Rudys, State Data Agency (Statistics Lithuania)

IAOS-ISI 2024 MEXICO CONFERENCE

INEGI

International Statistical Institute

IAOS IMPROVING OFFICIAL STATISTICS

Outline

- General information
- Data acquisition process
- Health checks for raw data
- Classification
- Conclusions

Objective:

- Integration of scanner data received from retail trade companies/chains (private data owners) to produce price indices (particularly HICP)

Legislation:

- Private data owners **must provide** data for official statistics purposes **free of charge** according to national statistical law (Republic of Lithuania Law on Official Statistics and State Data Governance, articles 10, 13, 18)

No agreements:

- Order of DG of SL „ON THE PROVISION OF STATISTICAL DATA FOR THE STATISTICAL SURVEY OF CONSUMER PRICES " approves:
    - List of statistical indicators at item level (25-30 variables)
    - Information on survey
    - Respondent declaration

State Data Agency receives data from:

- 5 biggest retail trade chains (food products)
- 5 biggest retail trade chains (constructions, electronics)
- 5 biggest pharmacy chains

About data:

- **Periodicity:** daily of weekly (data providers can choose)
- **Aggregation:** at **item/product** (aggregated) or receipt (not aggregated) level
- **Transmission of data:** possible to choose different types (usually data providers are choose to send CSV files trough SFTP)

Amount of data:

- **3 chains**
- from 01-22 to 02-04 (**2 weeks**)
- Total rows at product level: ~ **13 mln.**

Data acquisition process:

1st Meeting with data owners → **2nd Meeting with data owners** → **3nd Meeting with data owners** → **Data transmission**

General information on:

- Explaining need, purpose
- Legislation
- Data confidentiality, IT security
- Draft data structures (list of indicators)
- IT and technical data transmission aspects

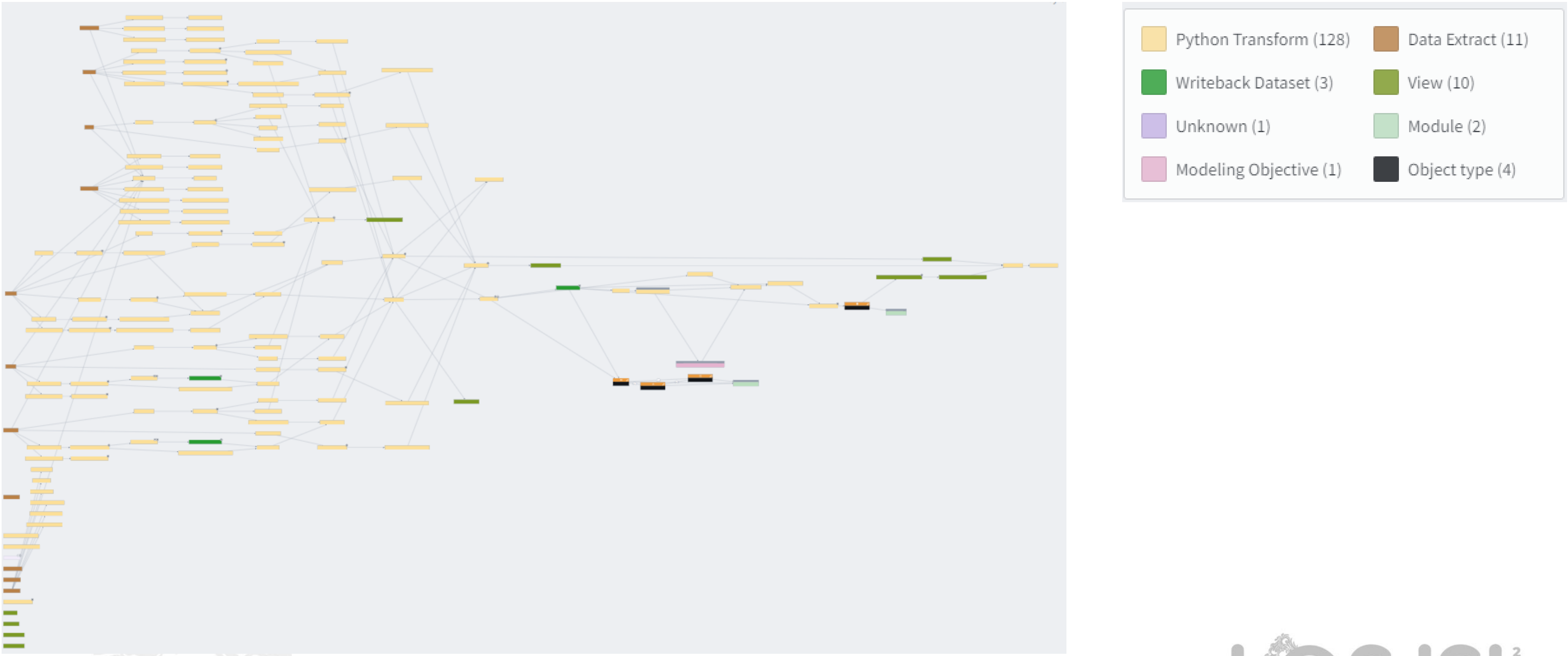
Detailed discussion on:

- Data structures (list of indicators)
- Data aggregation level
- Periodicity

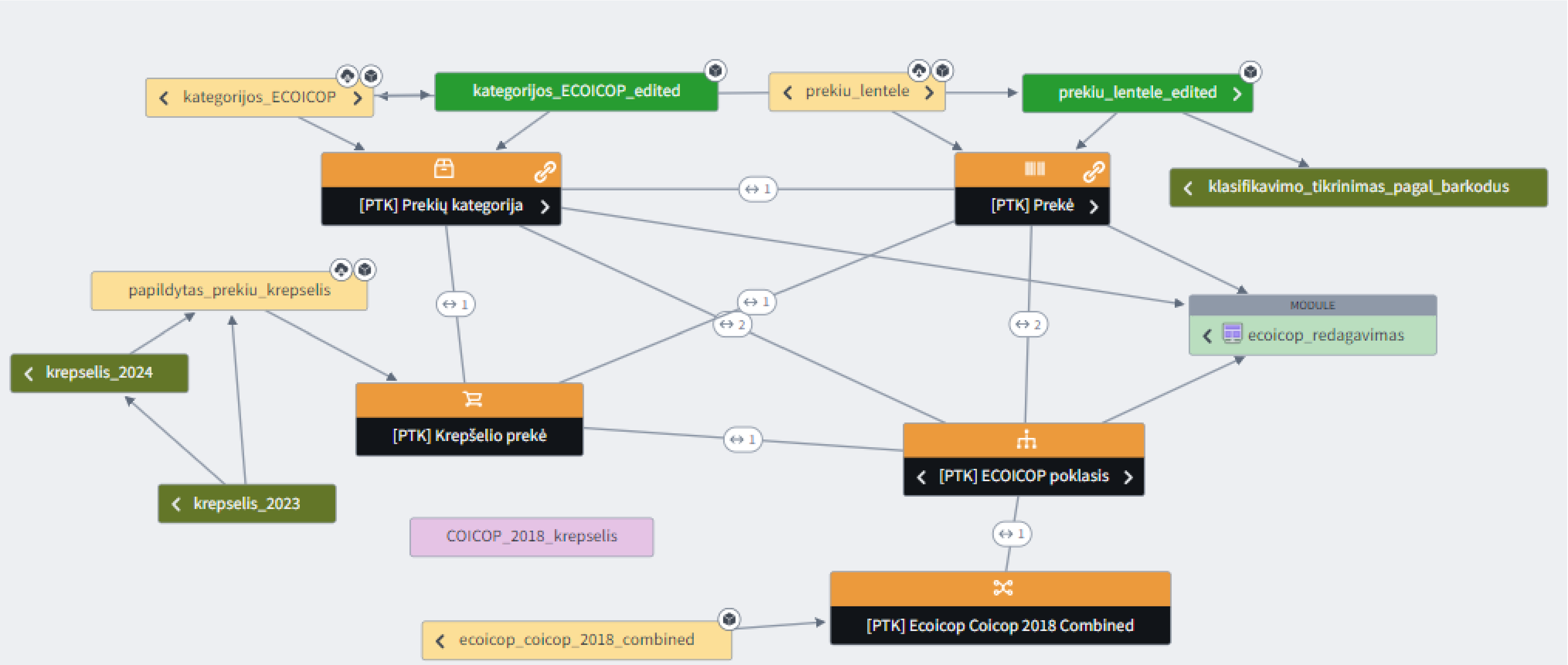
Detailed discussion on:

- Technical data transmission aspects
- We are flexible and offering different types of data transmission

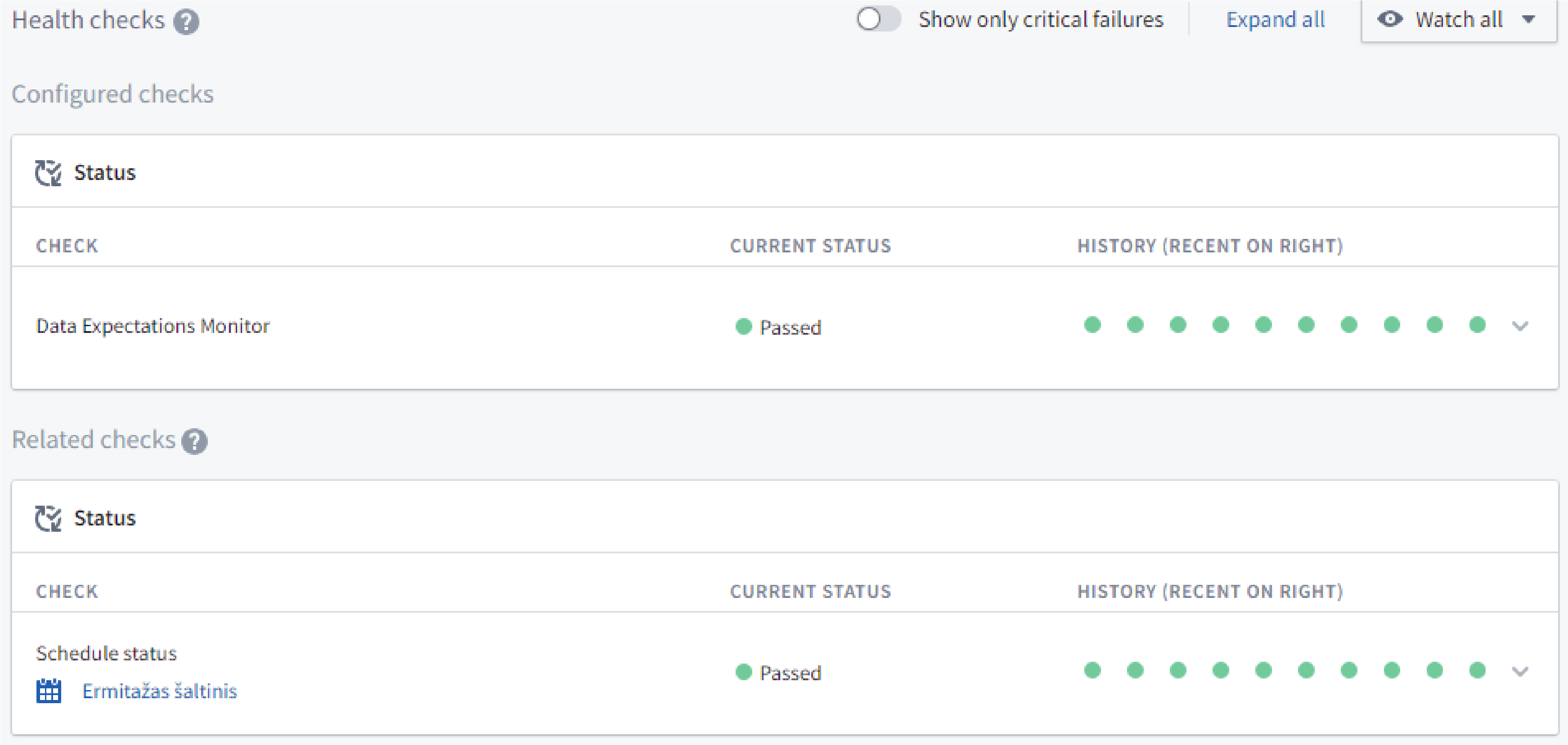
Data pipeline for automated data preparation



Data pipeline for automated data preparation



Health checks for raw data

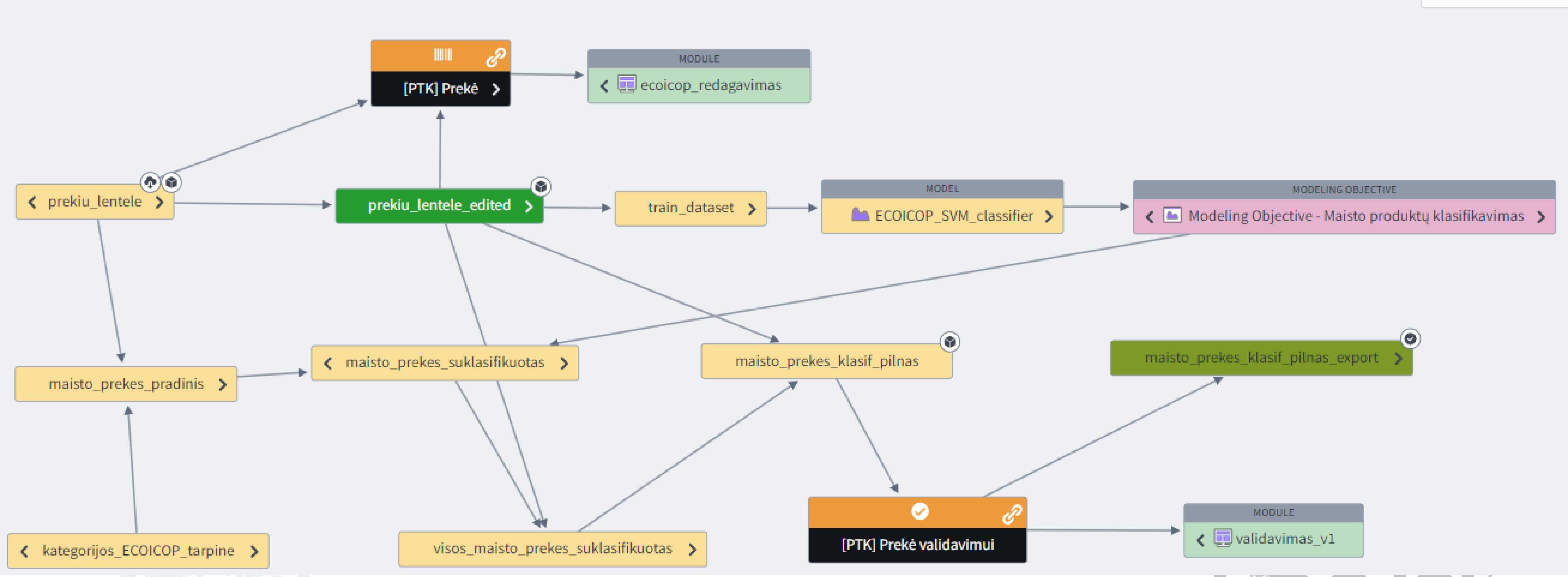


Health checks:

- Data integration to the platform passed
- Time since last updated
- Primary data validation passed
- Data freshness
- Corrupted files (null values)
- etc.

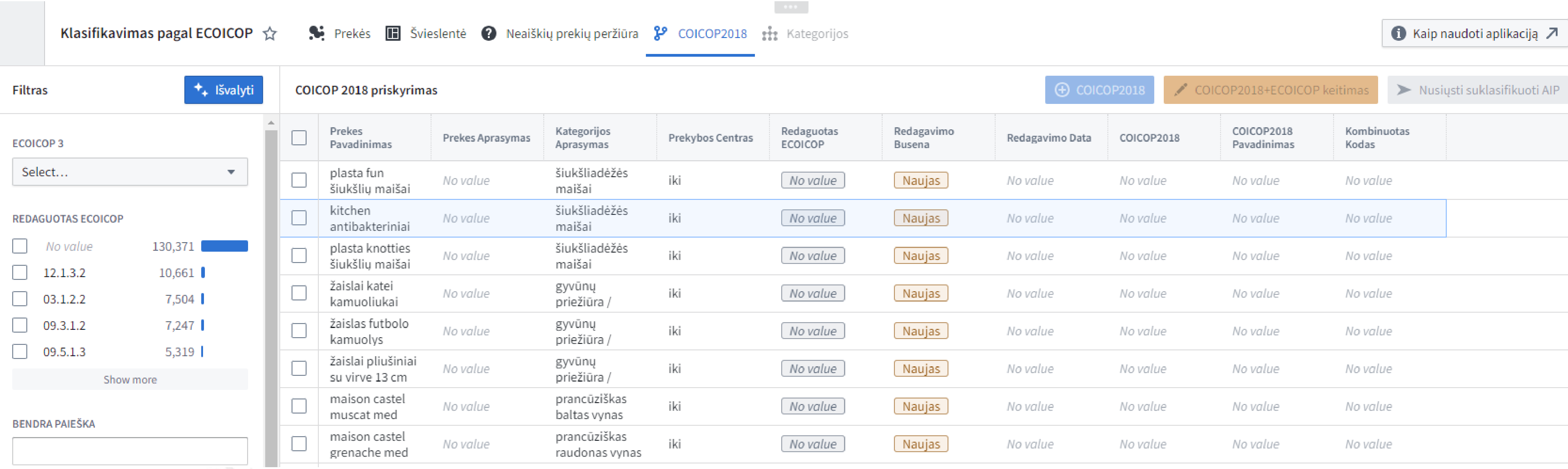
Data classification (ECOICOP)

Data classification pipeline:



Data classification (ECOICOP)

Application for manual data classification (building training data set):



Algorithms for classification

- Currently running: **SVM** (A support vector machine) + **LR** (logistic regression)
- Python (scikit-learn), PySpark
- **Train** data set: **35095**; **Test** data set **8774**

Tested models (more that 13):

NAME	DATE SUBMITTED	SUBMITTER	MODEL OWNER	EVALUATION	REVIEWS
ECOICOP_LR_SVM_combination_with…	Thu, Mar 28, 2024, 1:28 PM				0
ECOICOP_LR_SVM_combination_with_pr…	Thu, Mar 28, 2024, 9:45 AM				0
ECOICOP_LR_SVM_combination_with…	Mon, Mar 4, 2024, 2:14 PM				0
ECOICOP_LR_classifier 1.1	Wed, Feb 28, 2024, 2:15 PM				0
ECOICOP_SVM_classifier_with_probabilit…	Wed, Feb 21, 2024, 4:44 PM				0
ECOICOP_SVM_classifier_with_proba…	Wed, Feb 21, 2024, 3:28 PM				1
ECOICOP_SVM_classifier_with_probabilit…	Wed, Feb 21, 2024, 3:14 PM				0
ECOICOP_SVM_classifier_with_probabilit…	Wed, Feb 21, 2024, 2:37 PM				0
ECOICOP_SVM_classifier_with_probabilit…	Wed, Feb 21, 2024, 1:24 PM				0
ECOICOP_SVM_classifier_1.2	Mon, Feb 12, 2024, 2:25 PM				1
ECOICOP_SVM_classifier_with_proba…	Thu, Jan 4, 2024, 4:56 PM				0
ECOICOP_SVM_classifier_with_proba…	Wed, Jan 3, 2024, 11:26 AM				0
test_MA_model_2023-11-21T13:23:42…	Mon, Nov 27, 2023, 11:41 AM				0

Model input and output:

Model API | Objective API | View as code

Inputs (1)

df_in	Dataset	Required
prekes_id	String	Required
prekybos_centras	String	Required
prekes_apibrezimas	String	Required

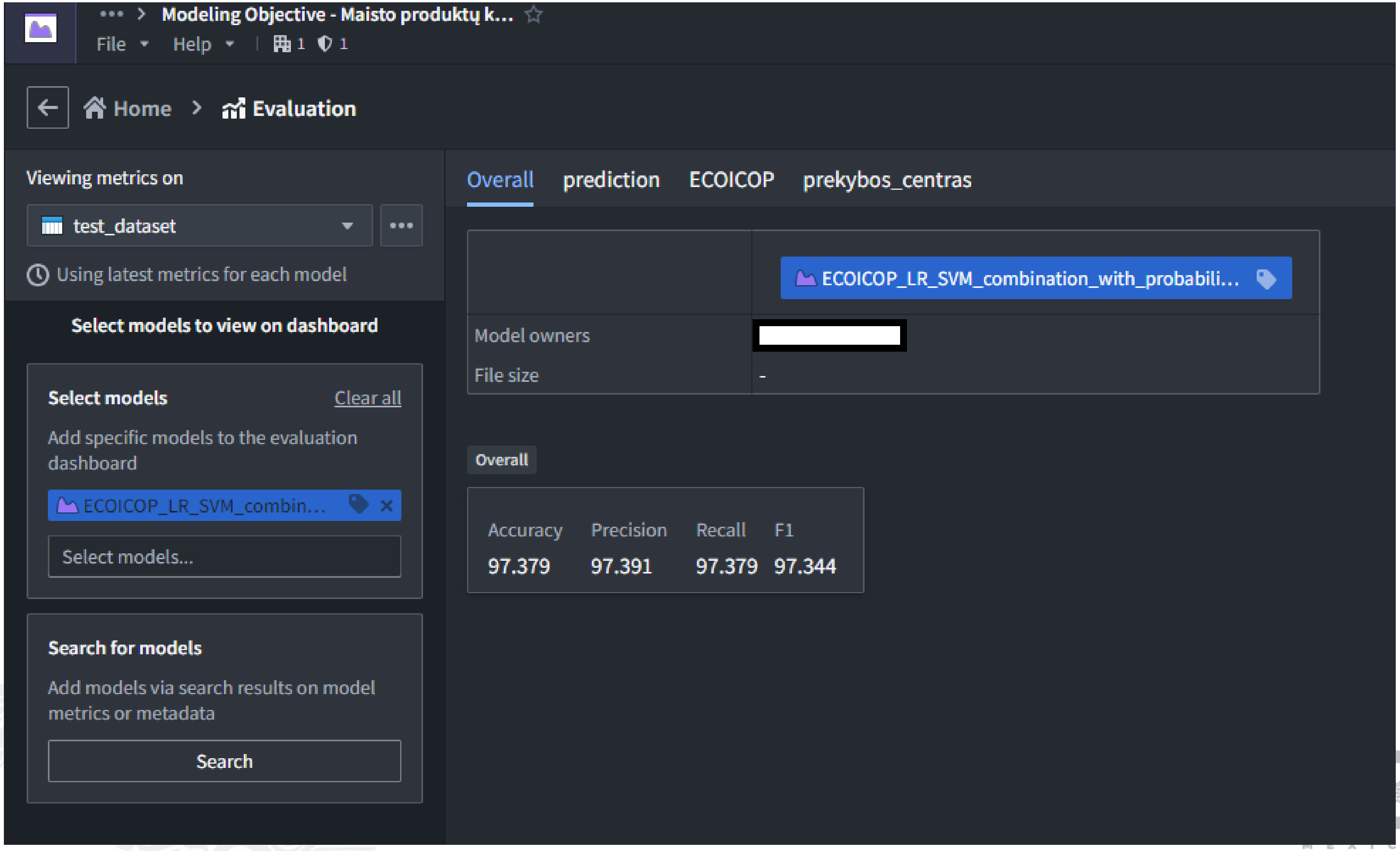
Outputs (1)

df_out	Dataset	Required
prekes_id	String	Required
prekybos_centras	String	Required
prekes_apibrezimas	String	Required
svm_prediction	String	Required
svm_probability_value	String	Required
lr_prediction	String	Required
lr_probability_value	String	Required

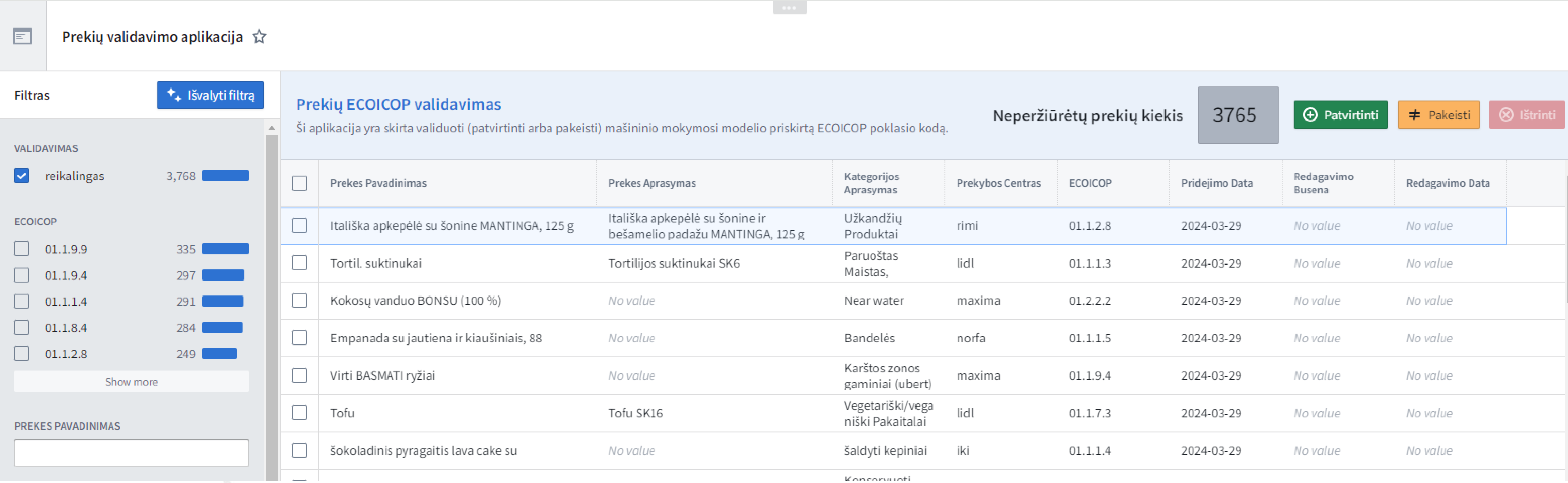
Model:

```
def train_model_combination(training_df):

    X_train = training_df['prekes_apibrezimas']

    y_train = training_df['ECOICOP']

    model = VotingClassifier(
        estimators=[
            ('lr', Pipeline([('features', CountVectorizer()), ('classifier', LogisticRegression())])),
            ('svm', Pipeline([('features', TfidfVectorizer()), ('classifier', SVC(probability=True))]
        ],
        voting='soft'
    )

    model.fit(X_train, y_train)

    return model
```

Classification accuracy



Validation of classification results (manual)


Prekių validavimo aplikacija

Filtras

Išvalyti filtrą

VALIDAVIMAS

reikalingas 3,768

ECOICOP

01.1.9.9 335

01.1.9.4 297

01.1.1.4 291

01.1.8.4 284

01.1.2.8 249

Show more

PREKES PAVADINIMAS

Prekių ECOICOP validavimas

Ši aplikacija yra skirta validuoti (patvirtinti arba pakeisti) mašininio mokymosi modelio priskirtą ECOICOP poklasio kodą.

Neperžiūrėtų prekių kiekis 3765

Patvirtinti | Pakeisti | Ištrinti

Prekes Pavadinimas	Prekes Aprasymas	Kategorijos Aprasymas	Prekybos Centras	ECOICOP	Pridejimo Data	Redagavimo Busena	Redagavimo Data
Itališka apkepėlė su šonine MANTINGA, 125 g	Itališka apkepėlė su šonine ir bešamelio padažu MANTINGA, 125 g	Užkandžių Produktai	rimi	01.1.2.8	2024-03-29	No value	No value
Tortil. suktinukai	Tortilijos suktinukai SK6	Paruoštas Maistas,	lidl	01.1.1.3	2024-03-29	No value	No value
Kokosų vanduo BONSU (100 %)	No value	Near water	maxima	01.2.2.2	2024-03-29	No value	No value
Empanada su jautiena ir kiaušiniais, 88	No value	Bandelės	norfa	01.1.1.5	2024-03-29	No value	No value
Virti BASMATI ryžiai	No value	Karštos zonos gaminiai (ubert)	maxima	01.1.9.4	2024-03-29	No value	No value
Tofu	Tofu SK16	Vegetariški/veganiški Pakaitalai	lidl	01.1.7.3	2024-03-29	No value	No value
šokoladinis pyragaitis lava cake su	No value	šaldyti kepiniai	iki	01.1.1.4	2024-03-29	No value	No value
		Konservuoti					

Application of AI (LLM) for classification

Problem: ECOICOP -> COICOP 2018

Inputs

preke

Use LLM Function output

Collapse

Prompt

Show help

Instructions (System prompt)

Find the appropriate COICOP2018 class (object type "[PTK] Ecoicop Coicop 2018 Combined") for the given product with the same ECOICOP as the product's ECOICOP.
Check the EcoicopCoicop2018Combined classes with the same ECOICOP as the product's ECOICOP.
Choose the most suitable class based on the product title, description, and category description.
Apply an action "Priskirti prekei COICOP2018 (AIP)" using the chosen class. Mind the format of the LtPtkPrekiuLentele argument, it shoud look like this: ["primary_key"].

Tools

Tools enable your LLM to access to your ontology or custom functions

[PTK] Ecoicop Coicop 2018 Combined
Query objects

Priskirti prekei COICOP2018 (AIP)
Apply actions

Posibility to use different models:

- Mistral AI Mixtral 8x7B - (Mistral AI's Mixtral 8x7B Instruct)
- OpenAI GPT 3.5 Turbo (16K) - (OpenAI's GPT 3.5 Turbo (16K) chat model)
- OpenAI GPT4 (Default)- (OpenAI's GPT4 chat model)
- OpenAI GPT4 32K - (OpenAI's GPT4 32K chat model)
- OpenAI GPT4 Turbo - (OpenAI's GPT4 Turbo chat model)

Conclusions:

- Integration of scanner data is complex process
- Requires:
    - Special methodological knowledge
    - Technological capabilities and solutions
    - Staff involvement

Near future plans:

- Index calculation

Thank you

IAOS-ISI 2024
MEXICO CONFERENCE

Nowcasting industrial production index with high-frequency toll data

Peter Knížat
Statistical Office of the Slovak Republic
University of Economics in Bratislava

INEGI International Statistical Institute IAOS IMPROVING OFFICIAL STATISTICS

Agenda

- Toll data analysis, processing and aggregation
- Estimation of index from toll data – comparison with Industrial Production Index
- Empirical Mode Decomposition – identification of trend and cyclicality
- Results and conclusions

Assumption and hypothesis

The fluctuation in the industrial production output in Slovakia can be detected through freight. The freight is estimated using toll data that are daily records of all vehicles (trucks) passing through satellite-monitored sections of roads.

Data processing and aggregation

Original data records

Monthly files of daily in-and-out passages of all vehicles (trucks) through section of roads

Data filtering

Filter out all vehicles that do not carry any goods or commodities for the industrial production

Data aggregation

Monthly aggregation – an observation unit is a vehicle ID, for which all passages though sections of roads are counted

Estimation of index

A theoretical concept of the price index formulae is used

Index formulae – bilateral

- We use two types of indices – (unweighted) bilateral and multilateral.

- **Jevons** index that is based on the geometric average:

$$I_{Jevons}^{0,t} = \prod_{i=1}^{N} \left(\frac{q_i^t}{q_i^0} \right)^{\frac{1}{N}}, \qquad t = 1, \dots, T$$

- where q_i^0 and q_i^t refer to a count of passages in the base period 0 and the current period *t* for each vehicle *i*

- Jevons index can also be expressed for month-on-month changes but these are normally chained from the base period

Index formulae – multilateral

- **Time-Product Dummy** index that is a (fixed-effects) regression based index:

$$ln\, q_i^t = \partial^0 + \sum_{t=1}^{T} \partial^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t, \qquad t = 0, \dots, T$$

- D_i^t is a dummy variable that takes the value 1 if the vehicle is observed in month *t* and 0 otherwise and D_i is a dummy for each observation (fixed-effects are the estimated parameters γ_i) – we can use the Ordinary Least-Squares method for the estimation of parameters

- Time-Product Dummy index is estimated as

$$I_{TPD}^{0,t} = exp(\,\hat{\partial}^t) = \frac{\prod_{i\epsilon S^t} (q_i^t)^{\frac{1}{N^t}}}{\prod_{i\epsilon S^0} (q_i^0)^{\frac{1}{N^0}}} exp[\bar{\hat{\gamma}}_i^0 - \bar{\hat{\gamma}}_i^t], \qquad t = 1, \dots, T$$

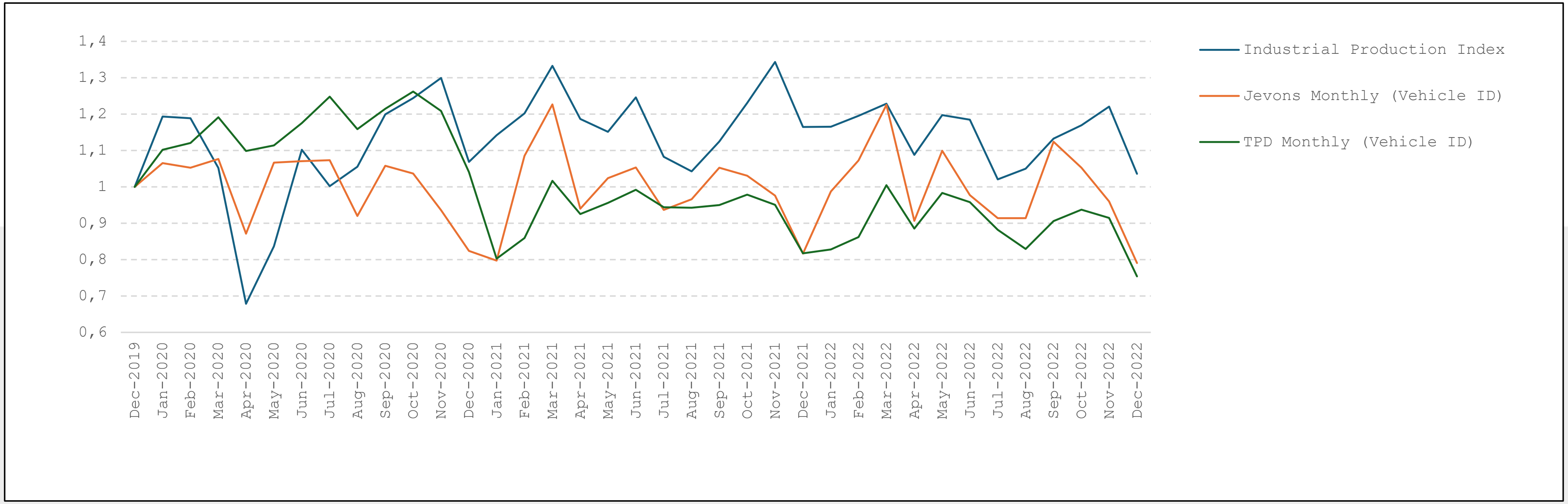
Empirical Mode Decomposition

- **Empirical Mode Decomposition** is suitable for decomposing time series that exhibit a strong nonlinearity and non-stationarity:

$$I^{0,t} = \sum_{j=1}^{n} IMF_j(t) + \varepsilon_n(t), \qquad t = 1, \dots, T$$

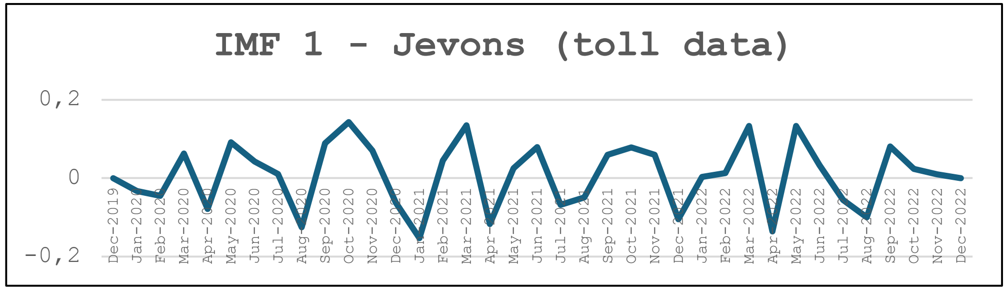
- where $IMF_j(t)$ are intrinsic mode functions, its extraction is obtained through the cubic splines interpolation that are fitted around local maxima and minima of original time series – these are fitted iteratively until the residual term $\varepsilon_n(t)$ is either a monotonic trend or a constant

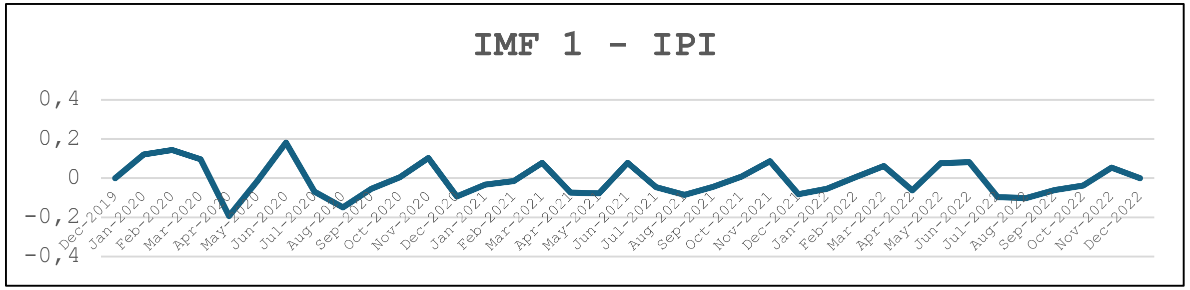
Results – Toll index vs Industrial Production Index

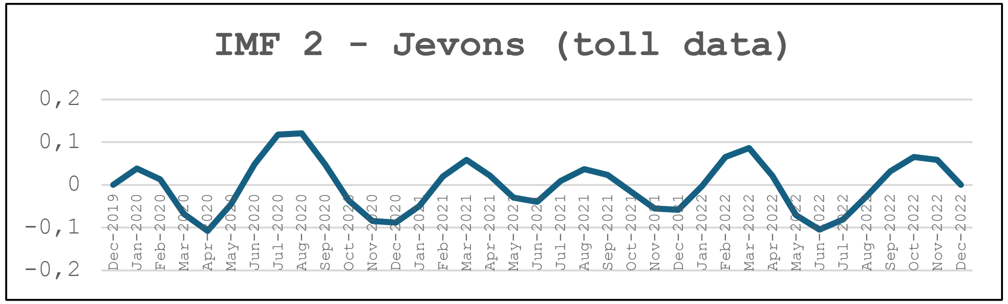


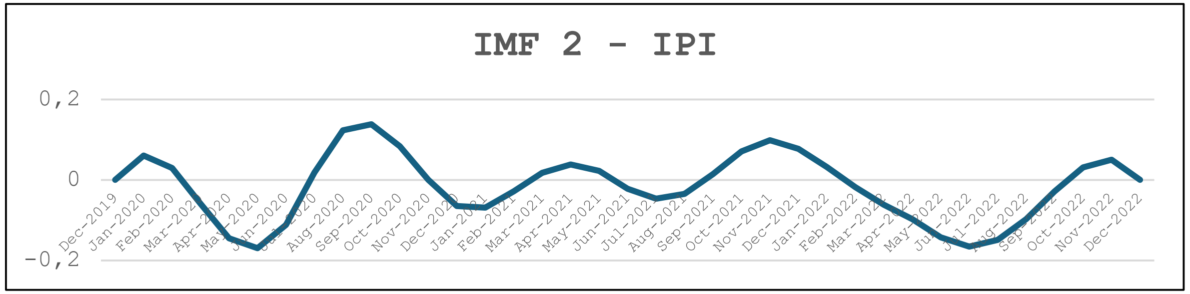
- Both Jevons and TPD index seem to capture cyclicality of IPI index
- Most of time periods, IPI index is above both toll indices
- Jevons index is more volatile than TPD index (except 2020)

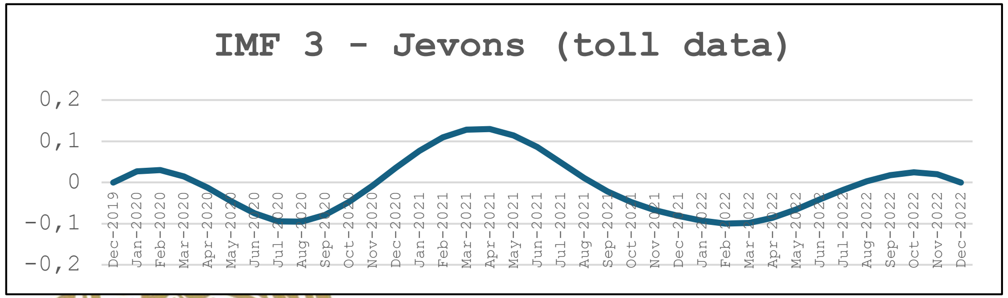
Index time series decomposition – Empirical Mode Decomposition

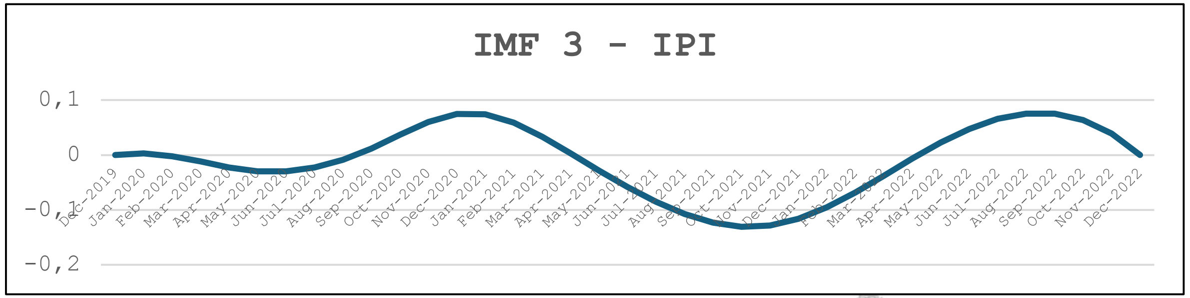












- We observe that IMFs of Jevons index have a similar pattern as IMFs of IPI → economic cyclicality of the industrial output is captured by freight

Conclusions

Nowcasting Industrial Production Index

The estimated toll index has a high potential for nowcasting industrial production or specific industries

01

02

The business and seasonality cycles can be detected and used by public and private sector economists

Identification of economic cycles

Further research

- Transport statistics
- Forecasting environmental variables
- More complex statistical models

03

Thank you Mexico for hosting us