IAOS-ISI 2024
MEXICO CONFERENCE

Application of AI and Data Science techniques

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

Chair: Tomas Rudys

Participants:

Norberts Talers: Computer Assisted Web interviews on mobile phones

Abel Coronado: INEGI's Advancements in Remote Sensing for a Comprehensive View of Mexico's Agricultural Landscape Using Data Science Techniques

Mark Motivans: Using record linkage of administrative records to improve federal justice statistics in the United States

Caiphus Mashaba:* An exploration of statistical models on planned and unplanned survey reporting domains

Alban Manishimwe: Application of AI to bridge the teacher to pupil ratio in Uganda

*Work presentation not available or non-existent

INEGI
IAOS IMPROVING OFFICIAL STATISTICS
isi International Statistical Institute

IAOS-ISI 2024
MEXICO CONFERENCE

Computer Assisted Web interviews on mobile devices

Norberts Talers,
Deputy director general,
Cenral Statistical Bureau of Latvia

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

Contents of the presentation

- historical insight to data collection

- outline of the project

- some statistics

- conclusions

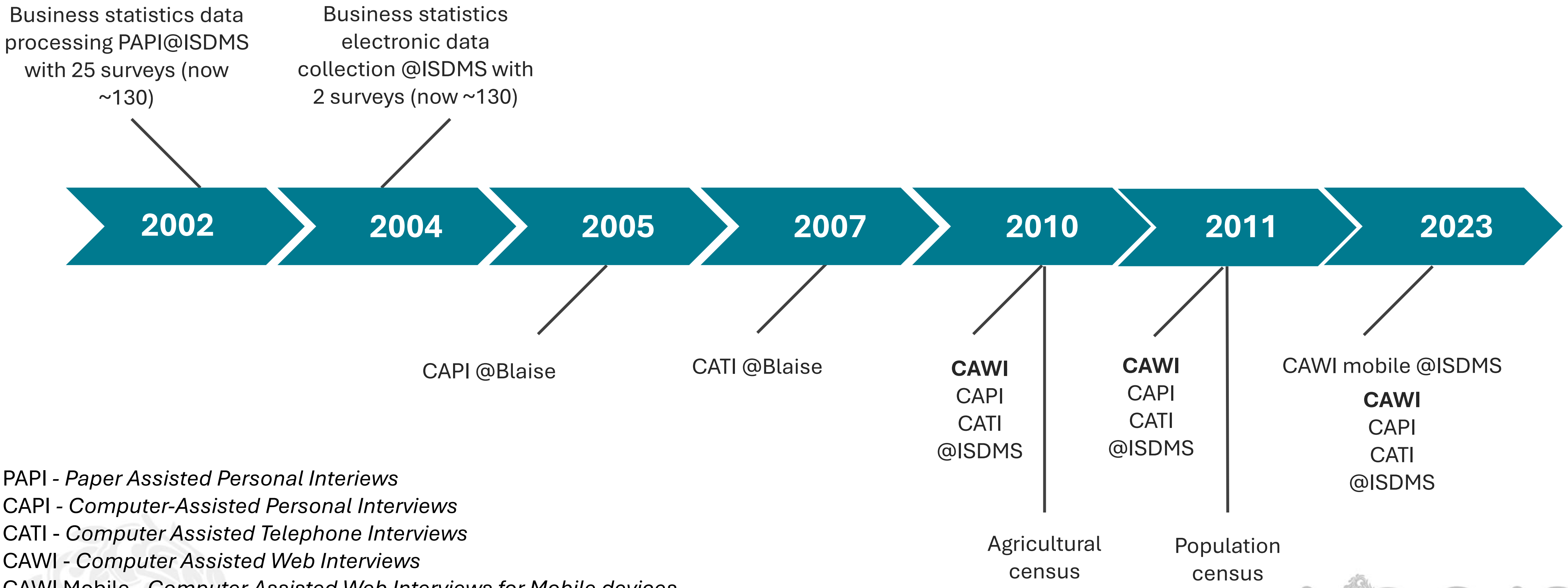
CAWI mobile in CSB of Latvia data collection environment

- first type of surveys to be collected electronically via web forms was business statistics
- social statistics Computer Assisted Data collection was next
- social statistics Computer Assisted Telephone interviews followed
- social statistics Computer Assisted Web interviews was the next step
- some of surveys previously considered as business statistics moved to CAWI
- CAWI mobile for social statistics surveys launched

Different environments

- business statistics data collection:
    - initially manual data entry from paper questionnaires
    - later - electronic data collection via web forms
        - Integrated Statistical Data Management System (ISDMS)
- social statistics (and later agricultural statistics) data collection:
    - initially manual data entry from paper questionnaires
    - later electronic data collection CAPI & CATI – Blaise
    - then CAPI & CATI - Integrated Statistical Data Management System
    - then also CAWI
    - finally, also CAWI mobile

The path to CAWI mobile



2002: Business statistics data processing PAPI@ISDMS with 25 surveys (now ~130)

2004: Business statistics electronic data collection @ISDMS with 2 surveys (now ~130)

2005: CAPI @Blaise

2007: CATI @Blaise

2010: **CAWI** CAPI CATI @ISDMS; Agricultural census

2011: **CAWI** CAPI CATI @ISDMS; Population census

2023: CAWI mobile @ISDMS; **CAWI** CAPI CATI @ISDMS

PAPI - *Paper Assisted Personal Interiews*
CAPI - *Computer-Assisted Personal Interviews*
CATI - *Computer Assisted Telephone Interviews*
CAWI - *Computer Assisted Web Interviews*
CAWI Mobile - *Computer Assisted Web Interviews for Mobile devices*

What is Integrated Statistical Data Management System

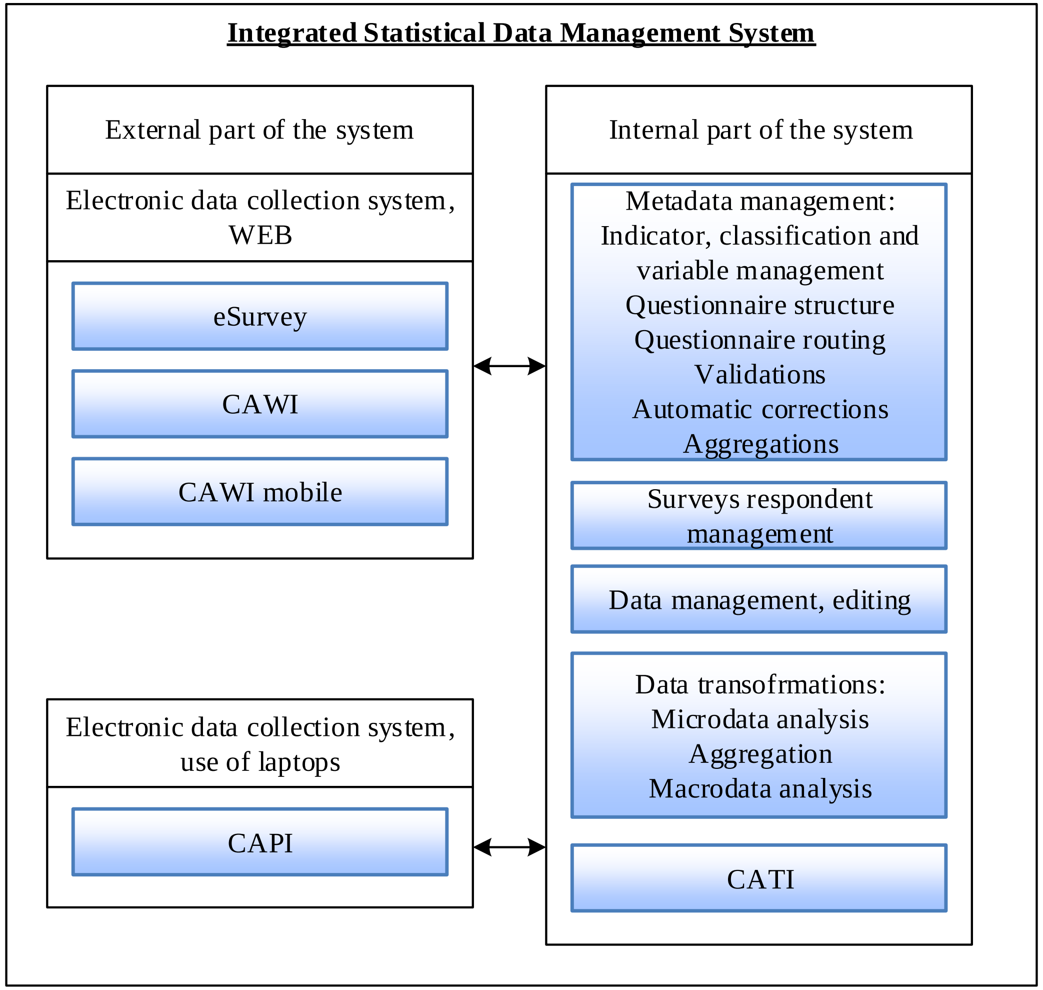
- centralized system

- metadata based system –

no programming for new questionnaires, surveys

- supports full data life cycle

- Yearly approx. 130 surveys, 30K businesses and 30K persons surveyed



Integrated Statistical Data Management System

External part of the system

Electronic data collection system, WEB

eSurvey

CAWI

CAWI mobile

Electronic data collection system, use of laptops

CAPI

Internal part of the system

Metadata management:
Indicator, classification and variable management
Questionnaire structure
Questionnaire routing
Validations
Automatic corrections
Aggregations

Surveys respondent management

Data management, editing

Data transofrmations:
Microdata analysis
Aggregation
Macrodata analysis

CATI

CAPI – CATI – CAWI

- What are the differences in the questionnaire?
- - CAPI – interviewer asks the questions to respondent, capable of explaining
- - CATI – almost the same, but interview is over phone, not face to face
- - CAWI – respondent is on his/her own with the questionnaire via web browser on the computer

CAWI mobile

- Although technically CAWI web pages would open on mobile devices, but:

- - functionality of the pages is not intended to small screens and tapping

- - the questions, answers and hints are designed for a big screen

CAWI mobile

- - We started using electronic means of communication with respondents at some point for reminders

- - People read them in their mobile devices

- - Trigger for CAWI mobile - SMS reminders with survey link

    - Data input is not designed for mobile devices (*respondents in most cases tries open the link from mobile device*). CAWI data entry is not ready for the web browsers used on smartphones and respondents interrupt the data entry
    - Questions and answers are not designed for mobile devices (*too long, to complicated etc.)*

CAWI mobile - start

- - Financial source – European Comission DG Eurostat 2020 Grant project for improvement of various aspects of EU – SILC survey
  - one of the many targets of the project – CAWI mobile module implementation

- - As CSB of Latvia has ISDMS – a common system for data collection the improvements aimed at one survey will serve as a platform for all other surveys where respondent is a person

Reference to EU - SILC

The EU statistics on income and living conditions (EU-SILC) aim to collect timely and comparable cross-sectional and longitudinal data on income, poverty, social exclusion, and living conditions.

CAWI mobile – the project

- - ISDMS is an outsource developed system – we needed a technical specification to implement CAWI mobile subsystem

- - Internally we have to think through the design of the mobile version of the system

- - We want to have CAWI mobile as an additional data collection mode which can be optional

-

- - Metadata descriptions shall be used to prepare CAWI mobile survey in the same way as the other modes

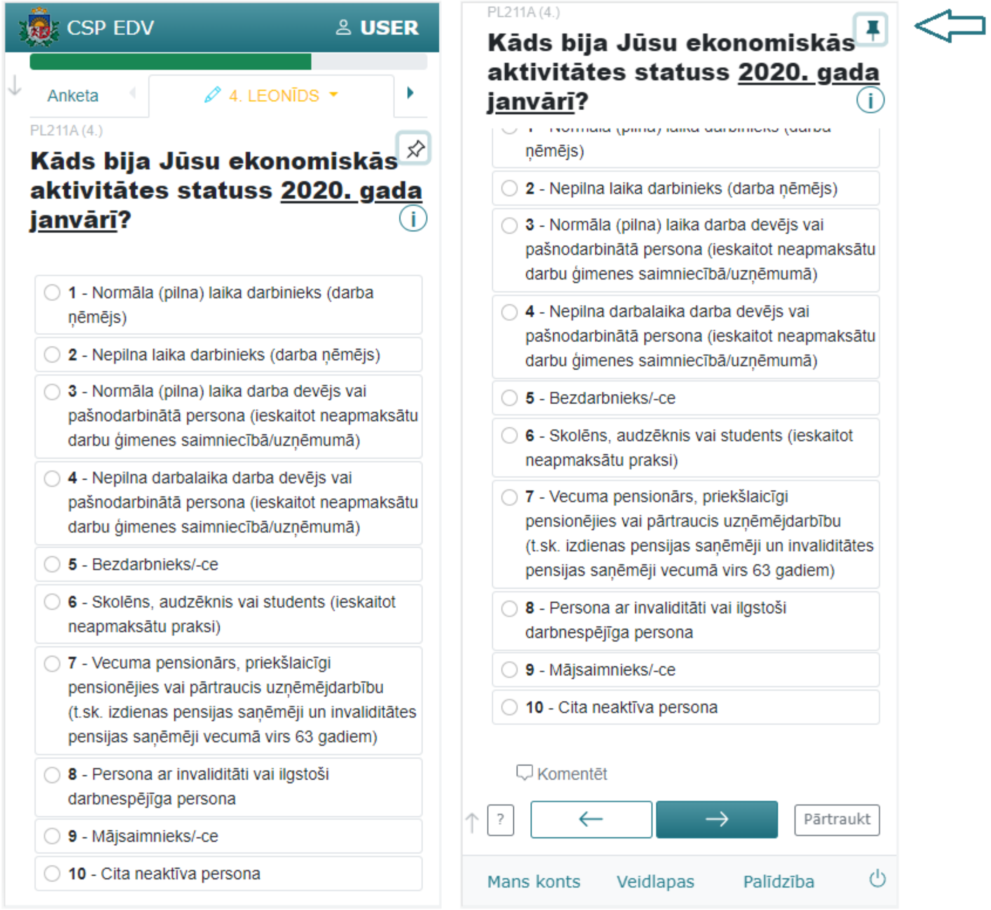
CAWI mobile – the project, cont.

- - Within ISDMS it should be survey – wise to be able to use mobile version or not

- - The functionality should remain the same – be it on laptop or mobile device, but it should be functionally possible to use it

- - It must be possible for the respondent to start survey in CAWI mobile and to continue in CAWI – thus data matrix must remain

Functional changes



CSP EDV USER

Anketa 4. LEONĪDS

PL211A (4.)

Kāds bija Jūsu ekonomiskās aktivitātes statuss 2020. gada janvārī?

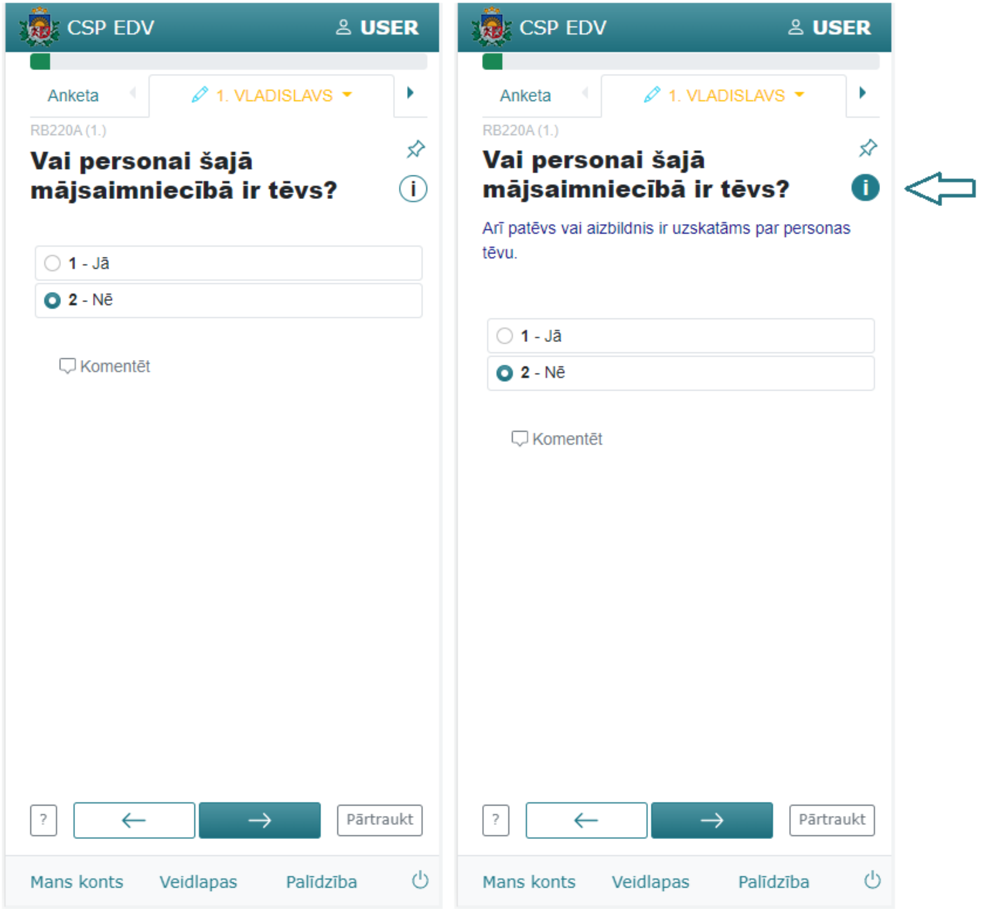
- ○ **1** - Normāla (pilna) laika darbinieks (darba ņēmējs)
- ○ **2** - Nepilna laika darbinieks (darba ņēmējs)
- ○ **3** - Normāla (pilna) laika darba devējs vai pašnodarbinātā persona (ieskaitot neapmaksātu darbu ģimenes saimniecībā/uzņēmumā)
- ○ **4** - Nepilna darbalaika darba devējs vai pašnodarbinātā persona (ieskaitot neapmaksātu darbu ģimenes saimniecībā/uzņēmumā)
- ○ **5** - Bezdarbnieks/-ce
- ○ **6** - Skolēns, audzēknis vai students (ieskaitot neapmaksātu praksi)
- ○ **7** - Vecuma pensionārs, priekšlaicīgi pensionējies vai pārtraucis uzņēmējdarbību (t.sk. izdienas pensijas saņēmēji un invaliditātes pensijas saņēmēji vecumā virs 63 gadiem)
- ○ **8** - Persona ar invaliditāti vai ilgstoši darbnespējīga persona
- ○ **9** - Mājsaimnieks/-ce
- ○ **10** - Cita neaktīva persona

PL211A (4.)

Kāds bija Jūsu ekonomiskās aktivitātes statuss 2020. gada janvārī?

- ņēmējs)
- ○ **2** - Nepilna laika darbinieks (darba ņēmējs)
- ○ **3** - Normāla (pilna) laika darba devējs vai pašnodarbinātā persona (ieskaitot neapmaksātu darbu ģimenes saimniecībā/uzņēmumā)
- ○ **4** - Nepilna darbalaika darba devējs vai pašnodarbinātā persona (ieskaitot neapmaksātu darbu ģimenes saimniecībā/uzņēmumā)
- ○ **5** - Bezdarbnieks/-ce
- ○ **6** - Skolēns, audzēknis vai students (ieskaitot neapmaksātu praksi)
- ○ **7** - Vecuma pensionārs, priekšlaicīgi pensionējies vai pārtraucis uzņēmējdarbību (t.sk. izdienas pensijas saņēmēji un invaliditātes pensijas saņēmēji vecumā virs 63 gadiem)
- ○ **8** - Persona ar invaliditāti vai ilgstoši darbnespējīga persona
- ○ **9** - Mājsaimnieks/-ce
- ○ **10** - Cita neaktīva persona

The question with many answers can be pinned to the top while scrolling



CSP EDV USER

Anketa 1. VLADISLAVS

RB220A (1.)

Vai personai šajā mājsaimniecībā ir tēvs?

- ○ **1** - Jā
- ● **2** - Nē

RB220A (1.)

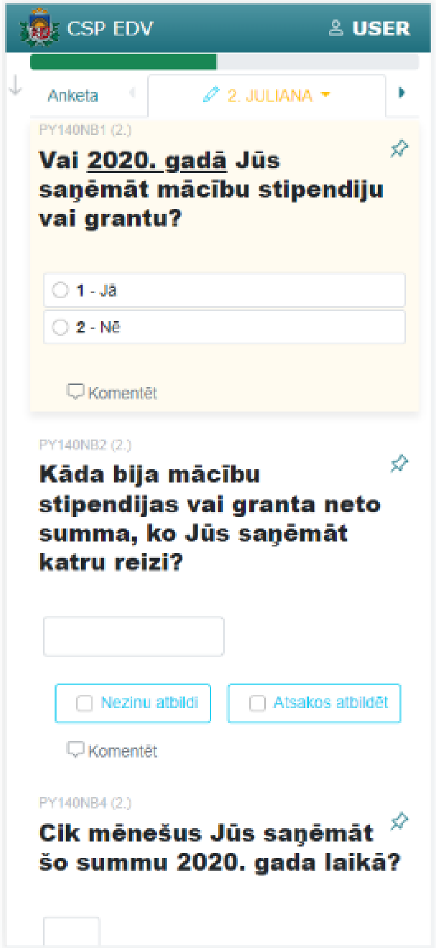
Vai personai šajā mājsaimniecībā ir tēvs?

Arī patēvs vai aizbildnis ir uzskatāms par personas tēvu.

- ○ **1** - Jā
- ● **2** - Nē

Additional description on question is hidden, to leave more space for the main question and answers, but it is easily accessible under the “i” button

Functional changes, cont.



PY140NB1 (2.)

Vai 2020. gadā Jūs saņēmāt mācību stipendiju vai grantu?

○ 1 - Jā

○ 2 - Nē

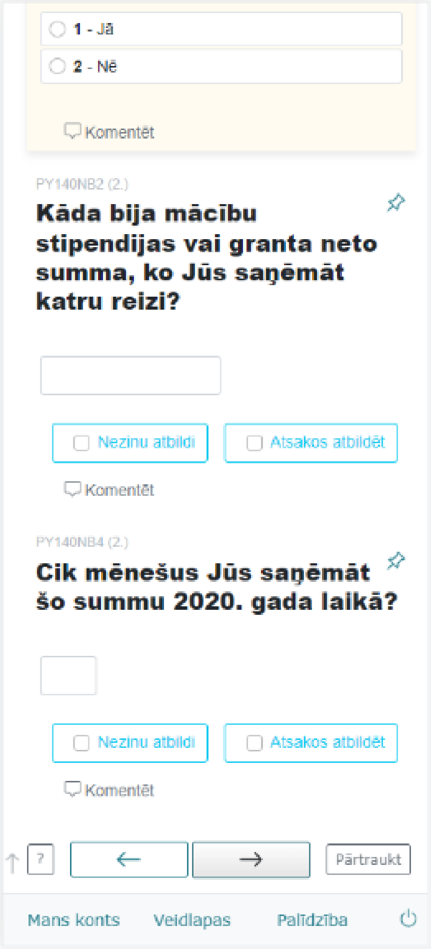
PY140NB2 (2.)

Kāda bija mācību stipendijas vai granta neto summa, ko Jūs saņēmāt katru reizi?

□ Nezinu atbildi □ Atsakos atbildēt

PY140NB4 (2.)

Cik mēnešus Jūs saņēmāt šo summu 2020. gada laikā?



○ 1 - Jā

○ 2 - Nē

PY140NB2 (2.)

Kāda bija mācību stipendijas vai granta neto summa, ko Jūs saņēmāt katru reizi?

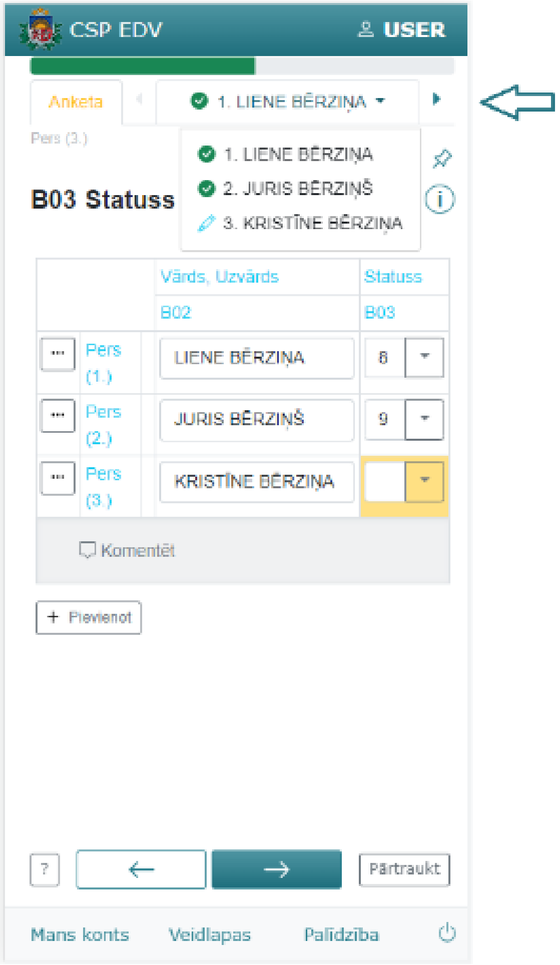
□ Nezinu atbildi □ Atsakos atbildēt

PY140NB4 (2.)

Cik mēnešus Jūs saņēmāt šo summu 2020. gada laikā?

□ Nezinu atbildi □ Atsakos atbildēt

Respondent can scroll through the questionnaire, instead of getting questions on by one



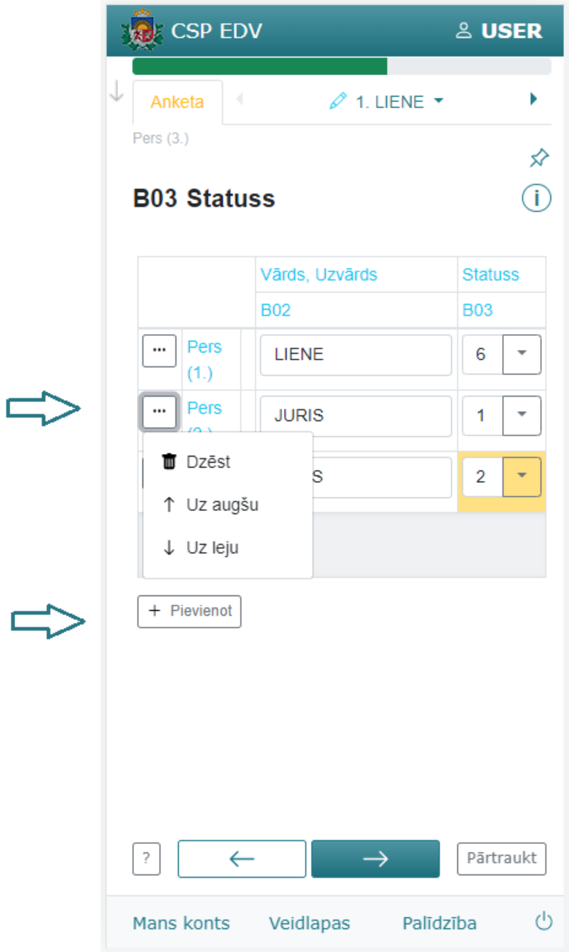
1. LIENE BĒRZIŅA
2. JURIS BĒRZIŅŠ
3. KRISTĪNE BĒRZIŅA

B03 Statuss

	Vārds, Uzvārds	Statuss
	B02	B03
Pers (1.)	LIENE BĒRZIŅA	8
Pers (2.)	JURIS BĒRZIŅŠ	9
Pers (3.)	KRISTĪNE BĒRZIŅA	

The selection of the respondent within the household is adapted for smartphone screen and have the dropdown style (previously tab style)

Functional changes, cont.

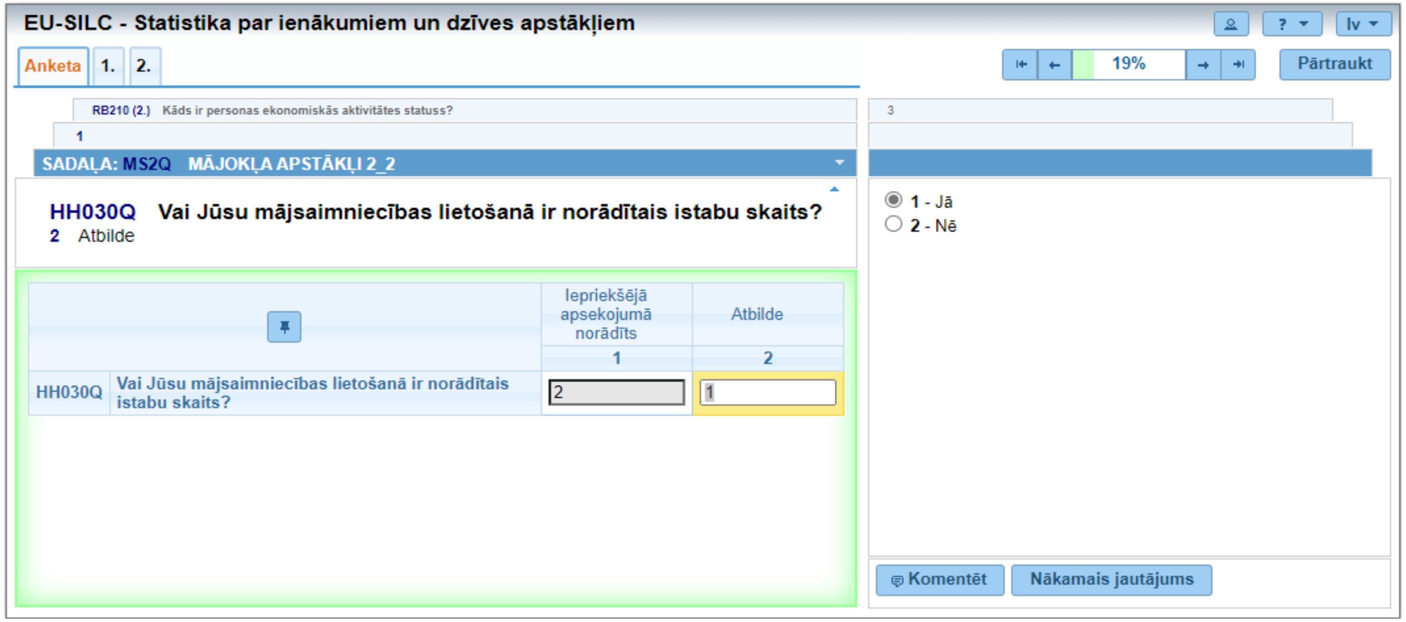


B03 Statuss

	Vārds, Uzvārds	Statuss
	B02	B03
Pers (1.)	LIENE	6
Pers (2.)	JURIS	1
	S	2

Functional buttons of adding/removing text and moving within the table are modified and aligned, hidden under a button

Redesign of questionnaire



EU-SILC - Statistika par ienākumiem un dzīves apstākļiem

RB210 (2.) Kāds ir personas ekonomiskās aktivitātes statuss? 3

1

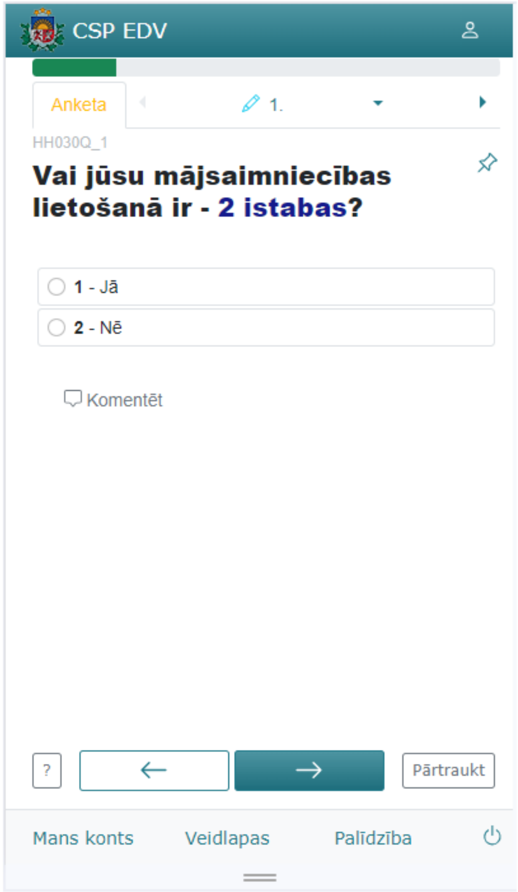
SADAĻA: MS2Q MĀJOKĻA APSTĀKĻI 2_2

HH030Q Vai Jūsu mājsaimniecības lietošanā ir norādītais istabu skaits?
2 Atbilde

		Iepriekšējā apsekojumā norādīts	Atbilde
		1	2
HH030Q	Vai Jūsu mājsaimniecības lietošanā ir norādītais istabu skaits?	2	1

● 1 - Jā
○ 2 - Nē

Tables of information changed to a standard question with answers

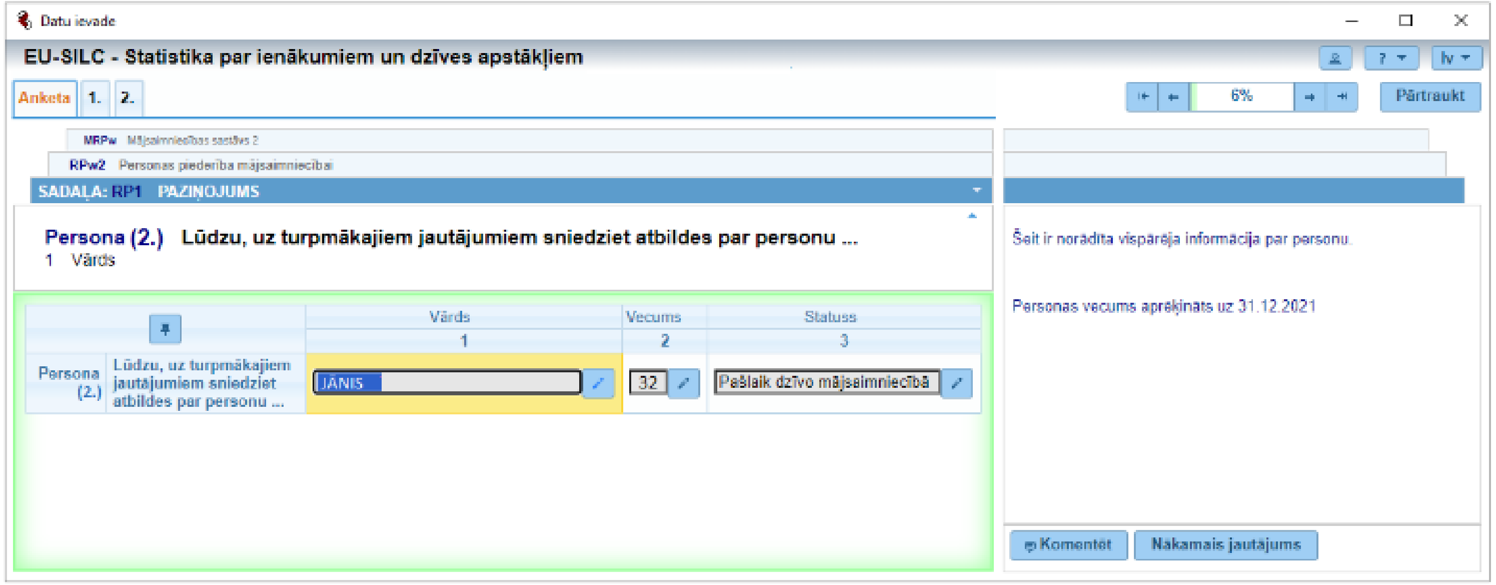


HH030Q_1

Vai jūsu mājsaimniecības lietošanā ir - 2 istabas?

○ 1 - Jā
○ 2 - Nē

Redesign of questionnaire, cont.



Datu ievade

EU-SILC - Statistika par ienākumiem un dzīves apstākļiem

MRPw Mājsaimniecības sastāvs 2

RPw2 Personas piederība mājsaimniecībai

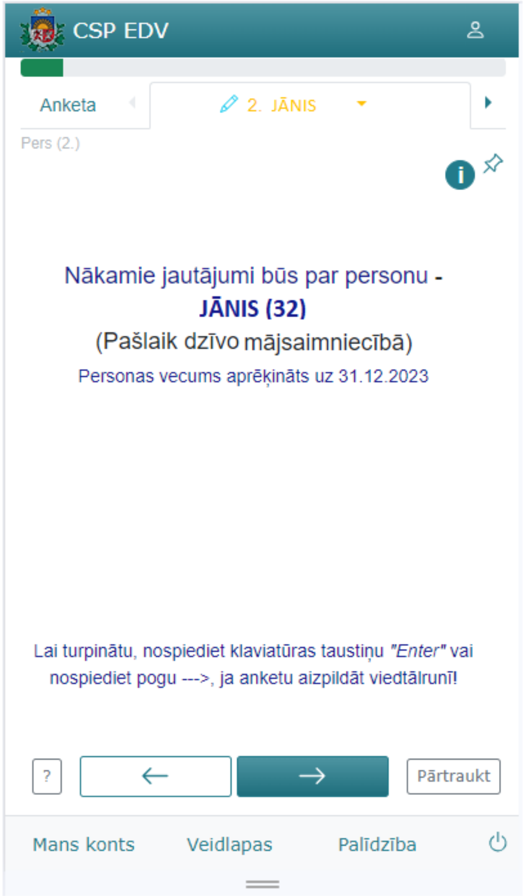
SADAĻA: RP1 PAZIŅOJUMS

Persona (2.) Lūdzu, uz turpmākajiem jautājumiem sniedziet atbildes par personu ...
1 Vārds

		Vārds	Vecums	Statuss
		1	2	3
Persona (2.)	Lūdzu, uz turpmākajiem jautājumiem sniedziet atbildes par personu ...	JĀNIS	32	Pašlaik dzīvo mājsaimniecībā

Šeit ir norādīta vispārēja informācija par personu.

Personas vecums aprēķināts uz 31.12.2021

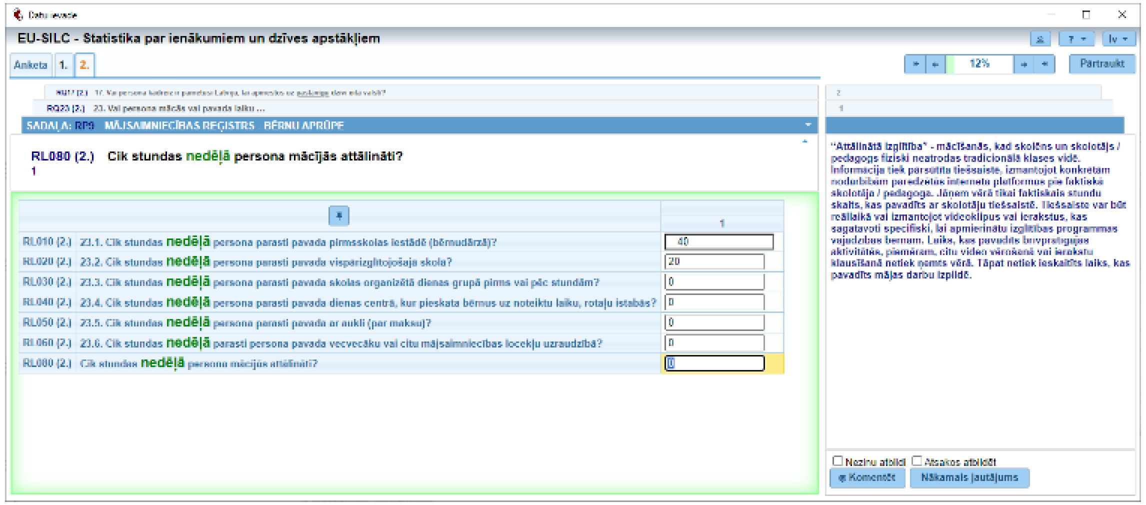


Nākamie jautājumi būs par personu - **JĀNIS (32)**
(Pašlaik dzīvo mājsaimniecībā)
Personas vecums aprēķināts uz 31.12.2023

Lai turpinātu, nospiediet klaviatūras taustiņu *"Enter"* vai nospiediet pogu --->, ja anketu aizpildāt viedtālrunī!

Table changed to informational screen with respective variables shown within informational note

Redesign of questionnaire, cont.



Datu ievade

EU-SILC - Statistika par ienākumiem un dzīves apstākļiem

RQ17 (2.) 17. [illegible]

RQ23 (2.) 23. Vai persona mācās vai pavada laiku ...

SADAĻA: RP9 MĀJSAIMNIECĪBAS REĢISTRS - BĒRNU APRŪPE

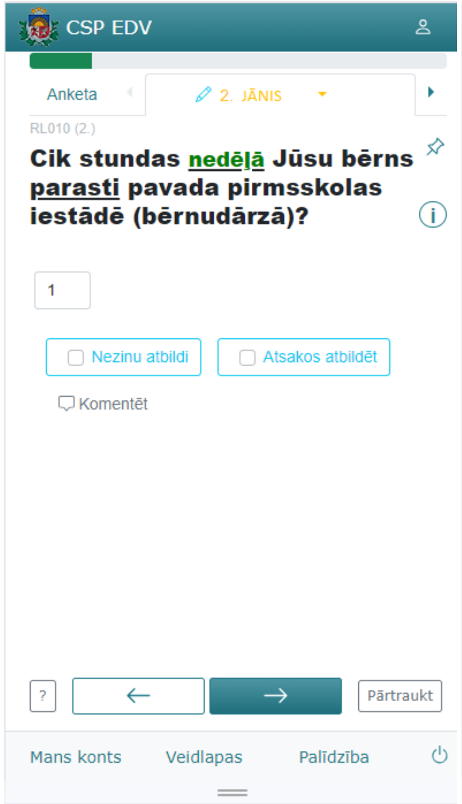
RL080 (2.) Cik stundas nedēļā persona mācījās attālināti?
1

	1
RL010 (2.) 23.1. Cik stundas nedēļā persona parasti pavada pirmsskolas iestādē (bērnudārzā)?	40
RL020 (2.) 23.2. Cik stundas nedēļā persona parasti pavada vispārizglītojošajā skolā?	20
RL030 (2.) 23.3. Cik stundas nedēļā persona parasti pavada skolas organizētā dienas grupā pirms vai pēc stundām?	0
RL040 (2.) 23.4. Cik stundas nedēļā persona parasti pavada dienas centrā, kur pieskata bērnus uz noteiktu laiku, rotaļu istabās?	0
RL050 (2.) 23.5. Cik stundas nedēļā persona parasti pavada ar aukli (par maksu)?	0
RL060 (2.) 23.6. Cik stundas nedēļā parasti persona pavada vecvecāku vai citu mājsaimniecības locekļu uzraudzībā?	0
RL080 (2.) Cik stundas nedēļā persona mācījās attālināti?	0

"Attālinātā izglītība" - mācīšanās, kad skolēns un skolotājs / pedagogs fiziski neatrodas tradicionālā klases vidē. Informācija tiek pārsūtīta tiešsaistē, izmantojot konkrētām nodarbībām paredzētās interneta platformas pie faktiskā skolotāja / pedagoga. Jāņem vērā tikai faktiskais stundu skaits, kas pavadīts ar skolotāju tiešsaistē. Tiešsaiste var būt reāllaikā vai izmantojot videoklipus vai ierakstus, kas sagatavoti specifiski, lai apmierinātu izglītības programmas vajadzības bērnam. Laiks, kas pavadīts brīvprātīgajās aktivitātēs, piemēram, citu video vērošanā vai ierakstu klausīšanā netiek ņemts vērā. Tāpat netiek ieskaitīts laiks, kas pavadīts mājas darbu izpildē.

□ Nezinu atbildi □ Atsakos atbildēt

Tables with multiple rows changed in single questions. One more example of explanations hidden under "I" icon and can be lengthy

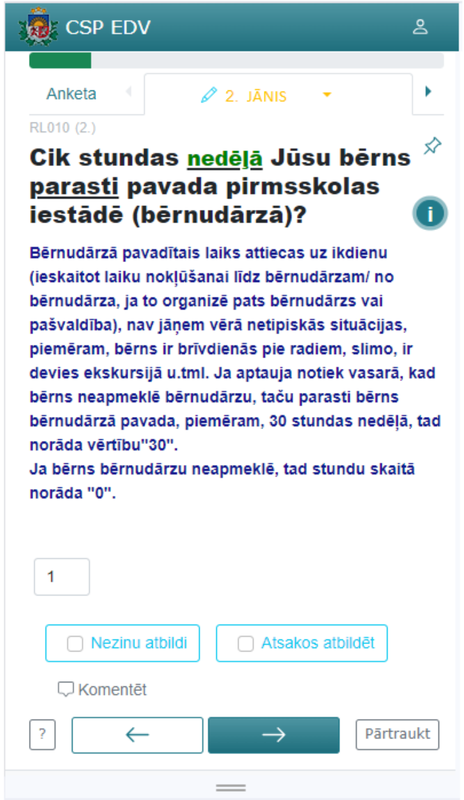


RL010 (2.)

Cik stundas nedēļā Jūsu bērns parasti pavada pirmsskolas iestādē (bērnudārzā)?

1

□ Nezinu atbildi □ Atsakos atbildēt



RL010 (2.)

Cik stundas nedēļā Jūsu bērns parasti pavada pirmsskolas iestādē (bērnudārzā)?

Bērnudārzā pavadītais laiks attiecas uz ikdienu (ieskaitot laiku nokļūšanai līdz bērnudārzam/ no bērnudārza, ja to organizē pats bērnudārzs vai pašvaldība), nav jāņem vērā netipiskās situācijas, piemēram, bērns ir brīvdienās pie radiem, slimo, ir devies ekskursijā u.tml. Ja aptauja notiek vasarā, kad bērns neapmeklē bērnudārzu, taču parasti bērns bērnudārzā pavada, piemēram, 30 stundas nedēļā, tad norāda vērtību"30".
Ja bērns bērnudārzu neapmeklē, tad stundu skaitā norāda "0".

1

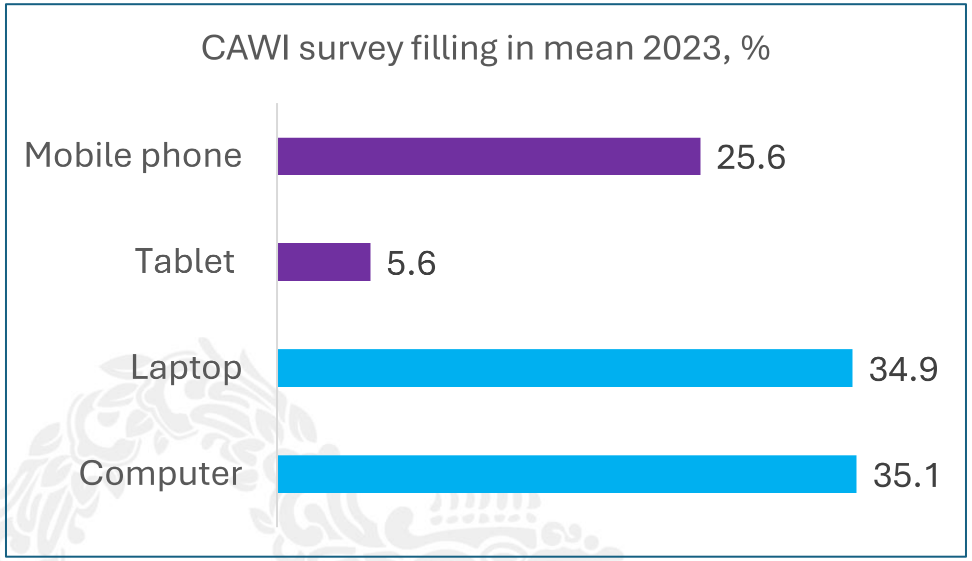
□ Nezinu atbildi □ Atsakos atbildēt

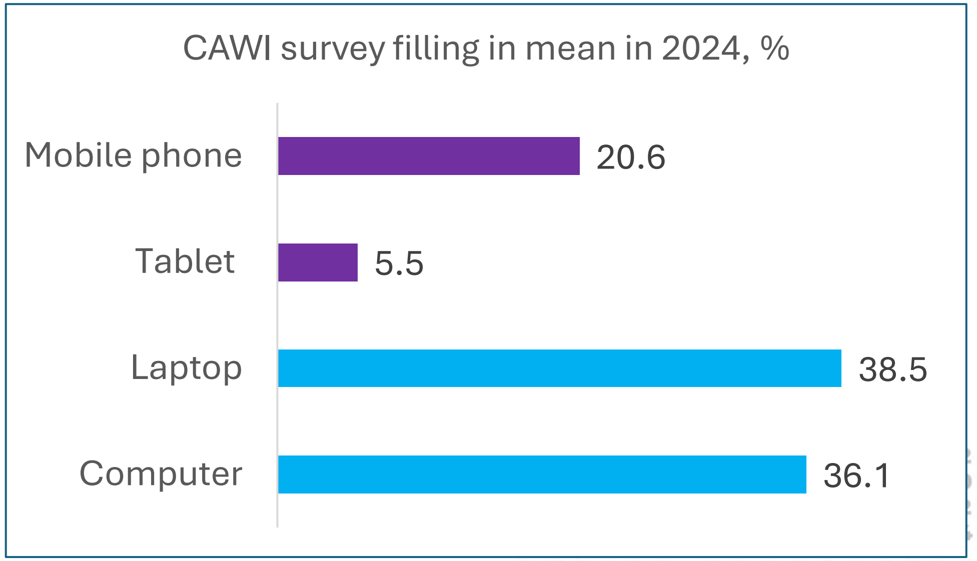
Statistics on response rates, EU - SILC

- EU – SILC survey in CAWI mobile started in 2023 and continues in 2024, respondent feedback is also collected via addition questions

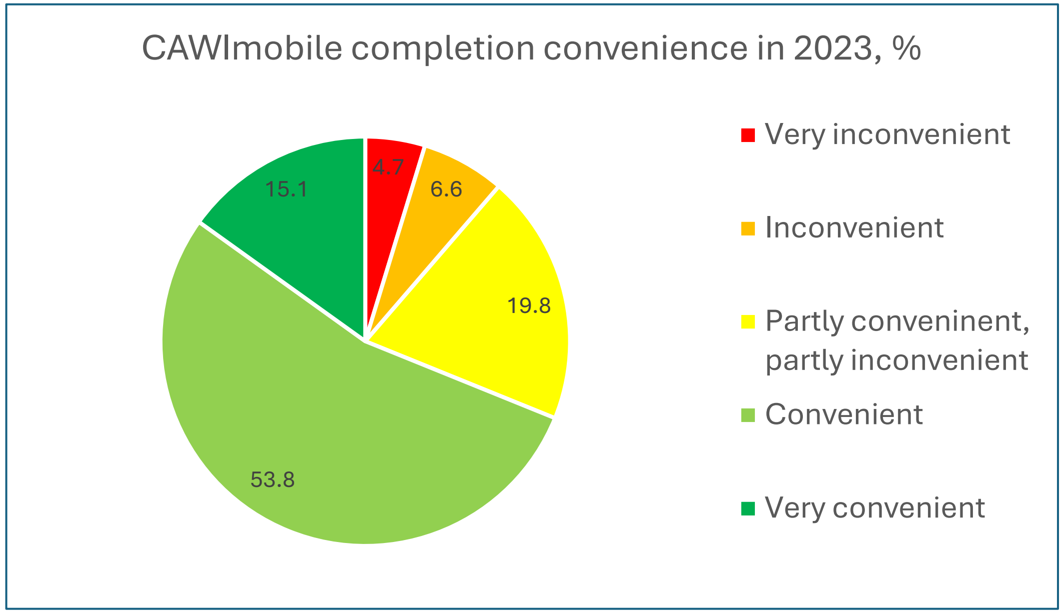
- A total of 390 questionnaires (~6.7% from collected questionnaires) via CAWI were received in 2023

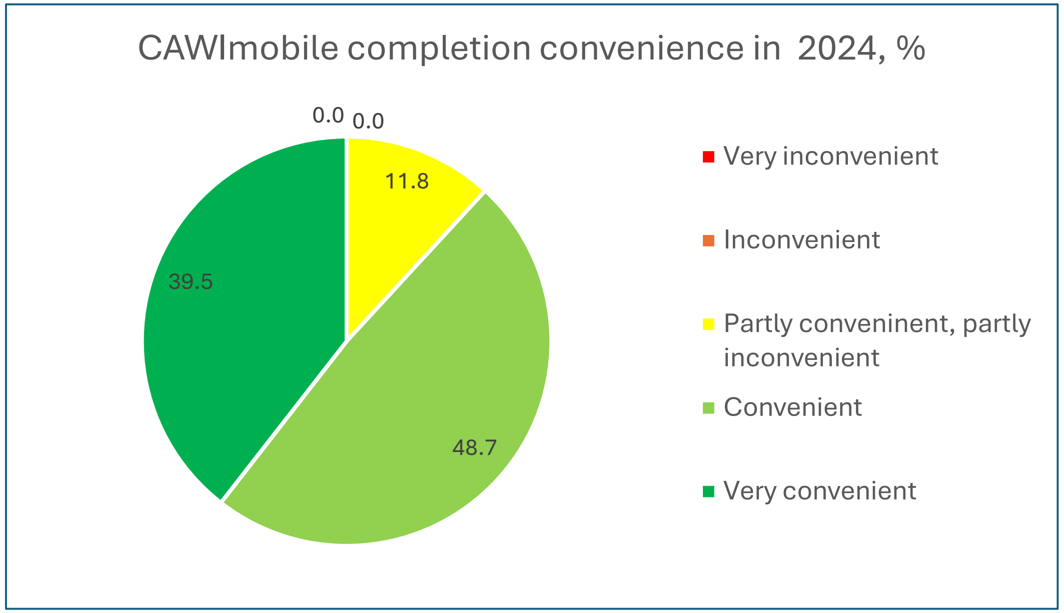
- A total of 291 questionnaires via CAWI were received in 2024 (and counting)





Statistics on response rates, EU – SILC, convenience feedback

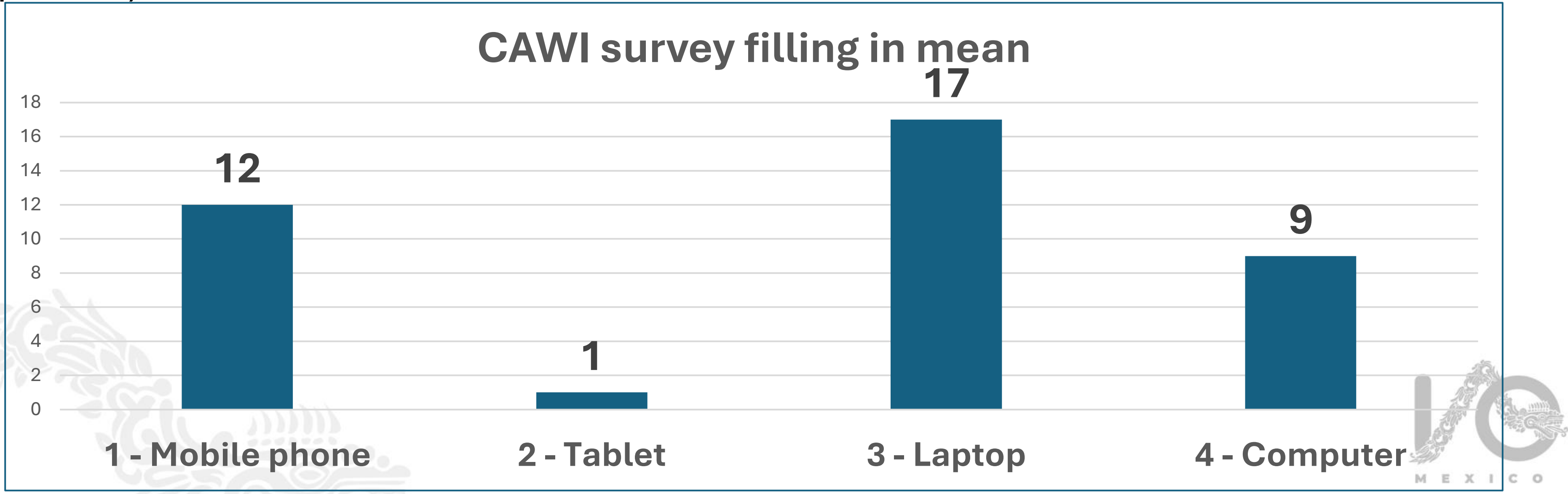




Statistics on response rates, survey “Personal and Professional Trips of Latvian Residents”

- Another survey has started to use CAWI mobile in 2024 - survey “Personal and Professional Trips of Latvian Residents”

- Data has been collected for one week, 45 collected questionnaires (4.2% of total sample size) via CAWI mobile



Survey on the use of information and communication technologies in households and by individuals

- One more survey has started to use CAWI mobile in 2024 - survey on the use of ICT.

- Data collection via CAWI is completed

	Count	%
Sample size	8500	100
CAWI (completed)	513	6.04
out of which CAWI mobile	197	38.40

Conclusions

- Number of respondents in CAWI as such still is quite small

- Number of respondents specifically in CAWI mobile approach is substantial to overall CAWI respondents

- Work in progress:

    - UX/UI further improvements
    - Work on questionnaire forms adjustment to small screen – a huge challenge
    - Questionnaire form length is crucial – contradiction to official statistics survey aiming at collecting as much data as possible
    - Use of administrative data can help a lot, still a lot of methodological issues – pre-print vs shortening the questionnaire vs (not)having required information in administrative data sources at all

Thank you

Norberts Talers,
Deputy director general,
Central Statistical Bureau of Latvia
norberts.talers@csp.gov.lv

IAOS-ISI 2024

MEXICO CONFERENCE

INEGI's Advancements In Remote Sensing For A Comprehensive View Of Mexico's Agricultural Landscape Using Data Science Techniques

Dr. Abel Coronado

INEGI

isi International Statistical Institute

IAOS IMPROVING OFFICIAL STATISTICS

Background

Background in the use of Earth Observations

In 2009, the first exercises to use Earth observations for obtaining agricultural statistical information began. Several projects were developed between 2012 and 2018..

However, the high cost of images and software licenses limited their use.

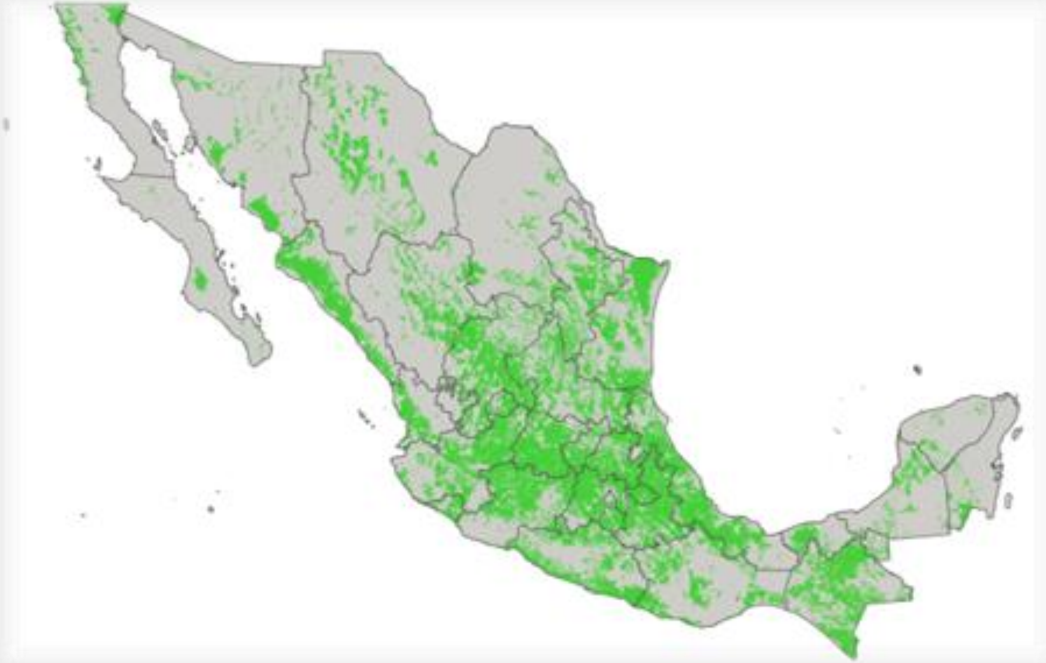
Evolution of Agricultural Census Data

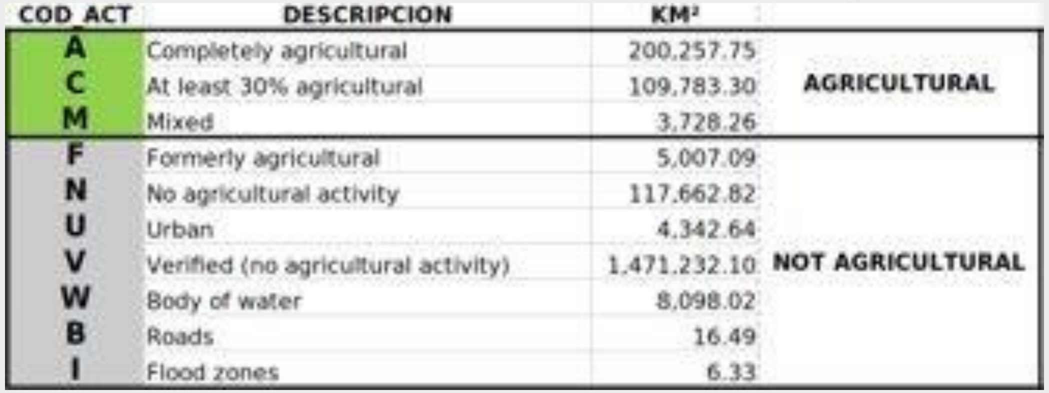
The 2007 Agriculture Census produced a digital archive of all surveyed lands, including their primary activities (agricultural, livestock, or forestry) as attributes. Subsequent update projects began with the 2016 Update of the Agricultural Census Framework (AMCA), which included the following:

- 2016 AMCA, at land level
- 2017 ENA, selected sample only
- 2019 ENA, selected sample only
- 2018-2019 Review of AMCA using satellite imagery
- 2019-2020 Comparison of AMCA with other sources of agricultural boundaries

2022 Agricultural Census (New)

Agricultural Land Use Identification





COD_ACT	DESCRIPCION	KM²	
A	Completely agricultural	200,257.75	AGRICULTURAL
C	At least 30% agricultural	109,783.30	
M	Mixed	3,728.26	
F	Formerly agricultural	5,007.09	NOT AGRICULTURAL
N	No agricultural activity	117,662.82	
U	Urban	4,342.64	
V	Verified (no agricultural activity)	1,471,232.10	
W	Body of water	8,098.02	
B	Roads	16.49	
I	Flood zones	6.33	

Agricultural frontier concept, Territorial distribution of areas in Mexico with agricultural activity, and lands cultivated in the last 5 years.

Problem Statement

Problem Statement

The problem to address is how to produce timely, cost-effective, and reliable estimates of the national agricultural frontier using Earth Observations combined with artificial intelligence algorithms.

Objective

Objective

Calibrate an algorithm using Artificial Intelligence and SENTINEL-2 satellite imagery to estimate the National Agricultural Frontier.

Methodology

Data Sources

Agricultural frontier

2019

Sentinel-2 Geomedian (12)

2019

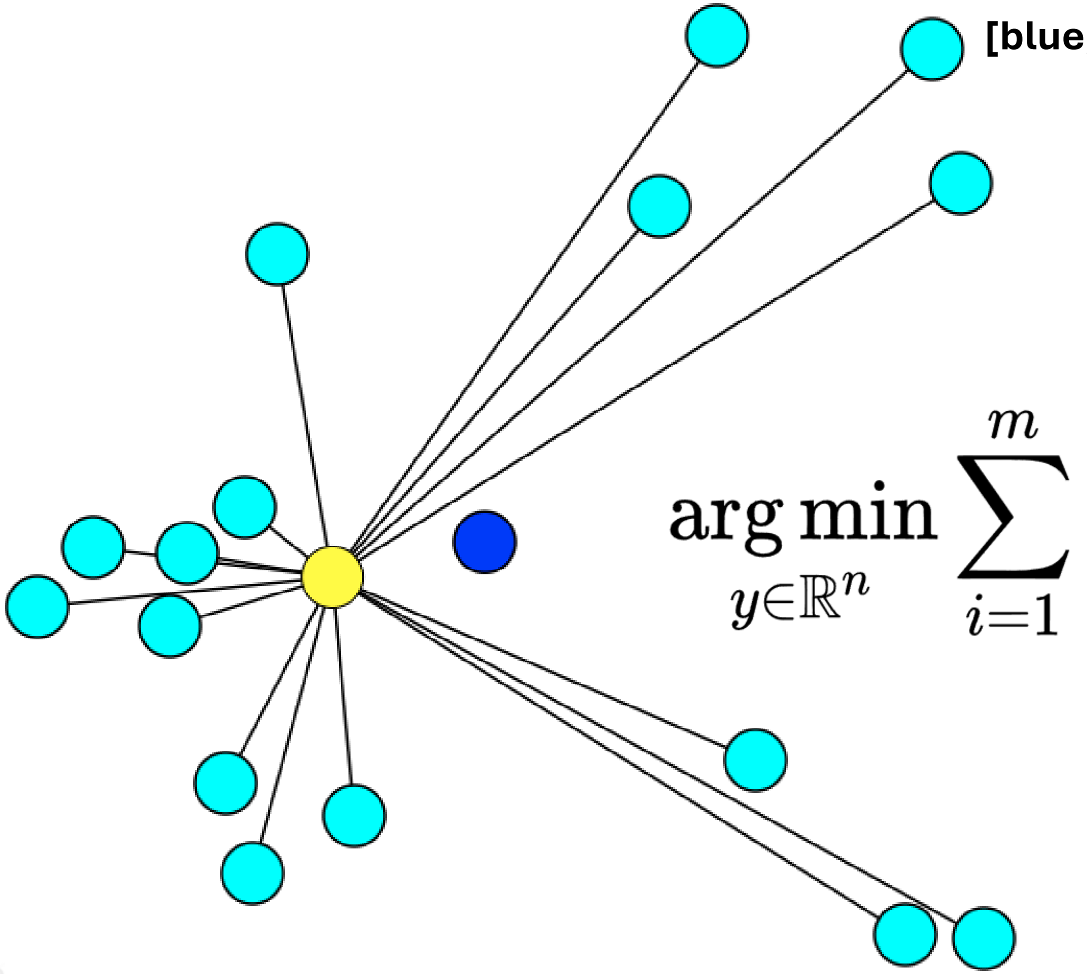
Spectral indexes (20)

2019

Texture filters (48)

2019

Geomedian = Geometric Median

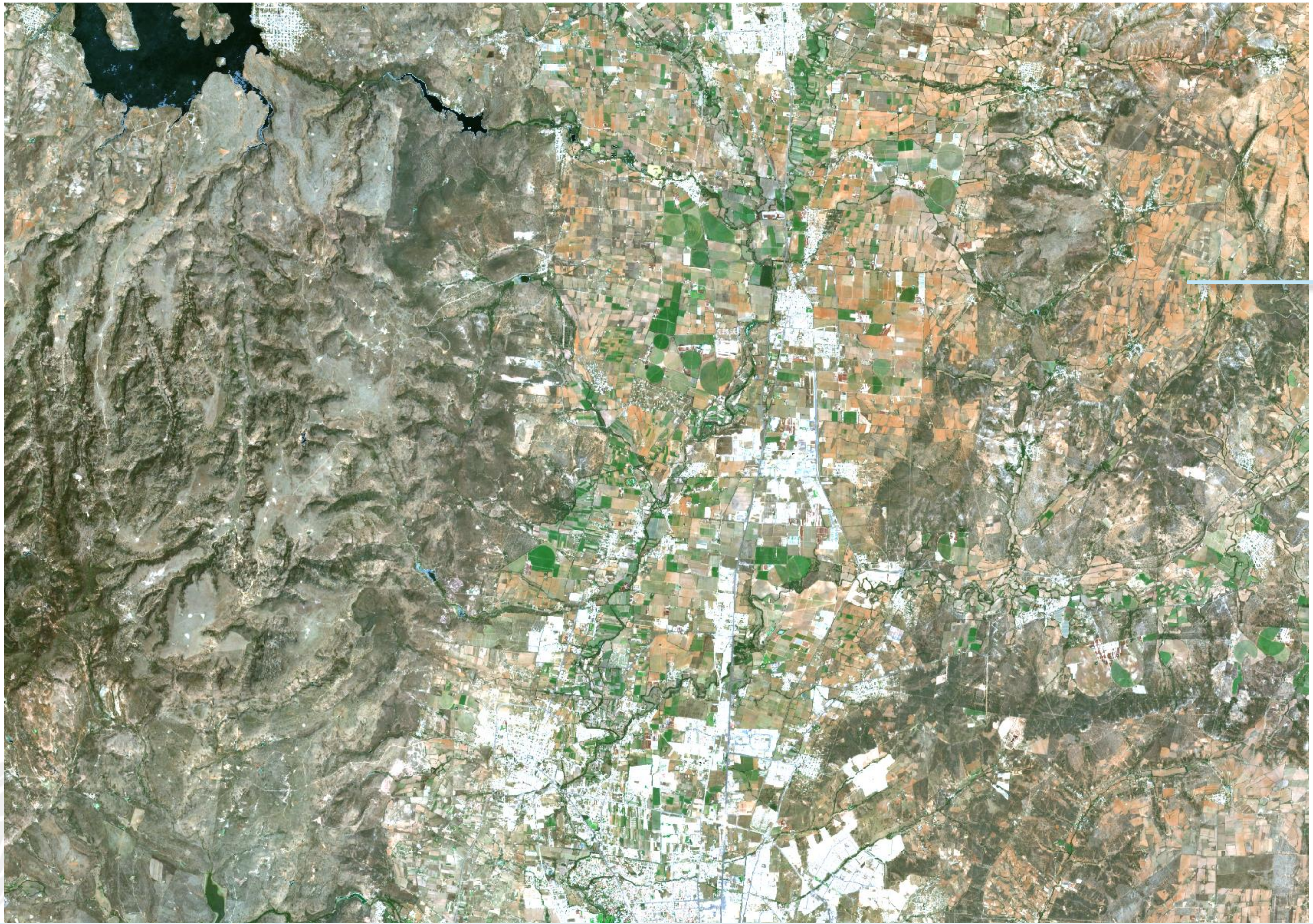


[blue, green, red, nir, swir1, swir2, ...]

$$\underset{y \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{m} \left\| x_i - y \right\|_2$$

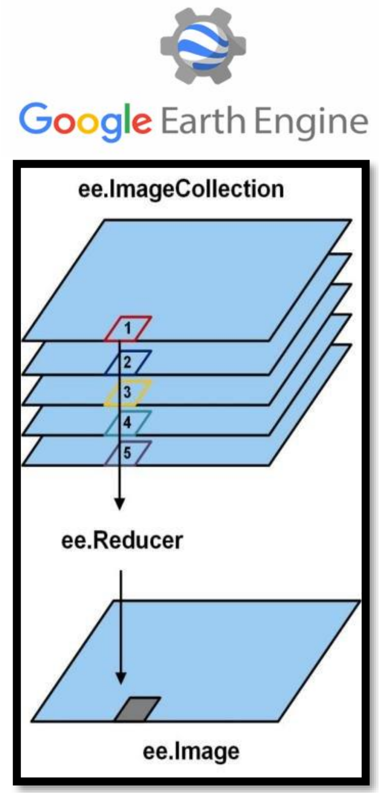
[blue, green, red, nir, swir1, swir2, ...]

Geomedian

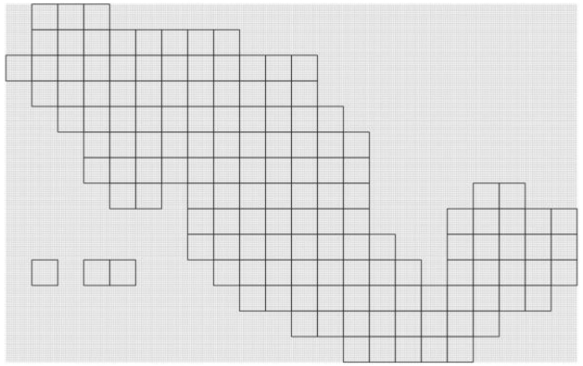


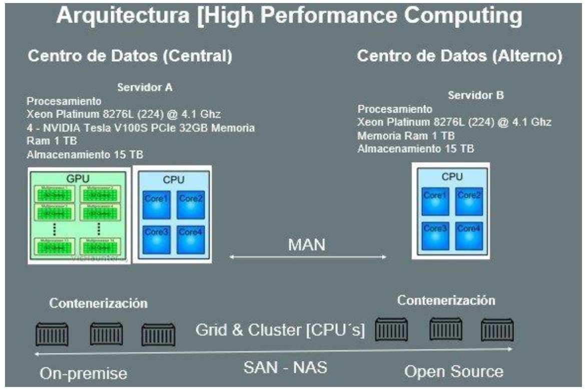
1. Coastal Aerosol
2. Blue
3. Green
4. Red
5. Vegetation 5
6. Vegetation 6
7. Vegetation 7
8. Near-Infrared
9. Vegetation 8
10. Water Vapour
11. Short Wave Infrared 1
12. Short Wave Infrared 2

Geomedian = Geometric Median



ee.Reducer.geometricMedian





Arquitectura [High Performance Computing

Centro de Datos (Central)

Servidor A

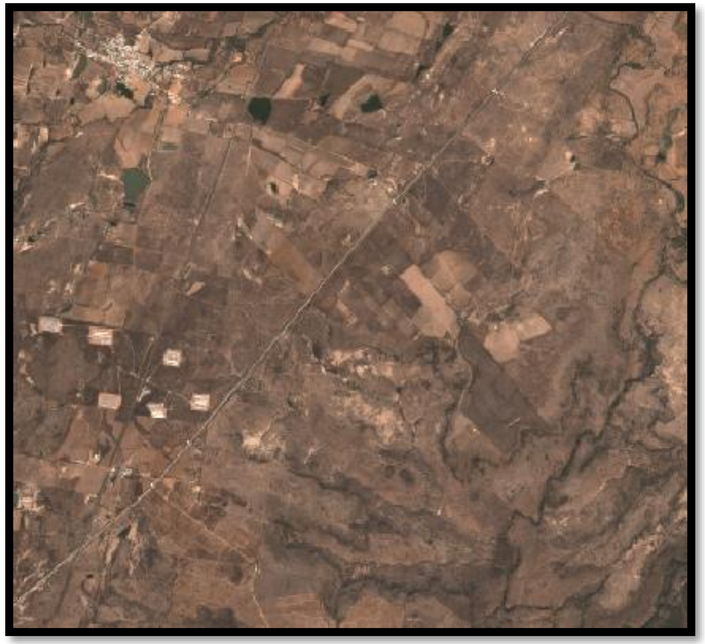
Procesamiento
Xeon Platinum 8276L (224) @ 4.1 Ghz
4 - NVIDIA Tesla V100S PCIe 32GB Memoria
Ram 1 TB
Almacenamiento 15 TB

Centro de Datos (Alterno)

Servidor B

Procesamiento
Xeon Platinum 8276L (224) @ 4.1 Ghz
Memoria Ram 1 TB
Almacenamiento 15 TB

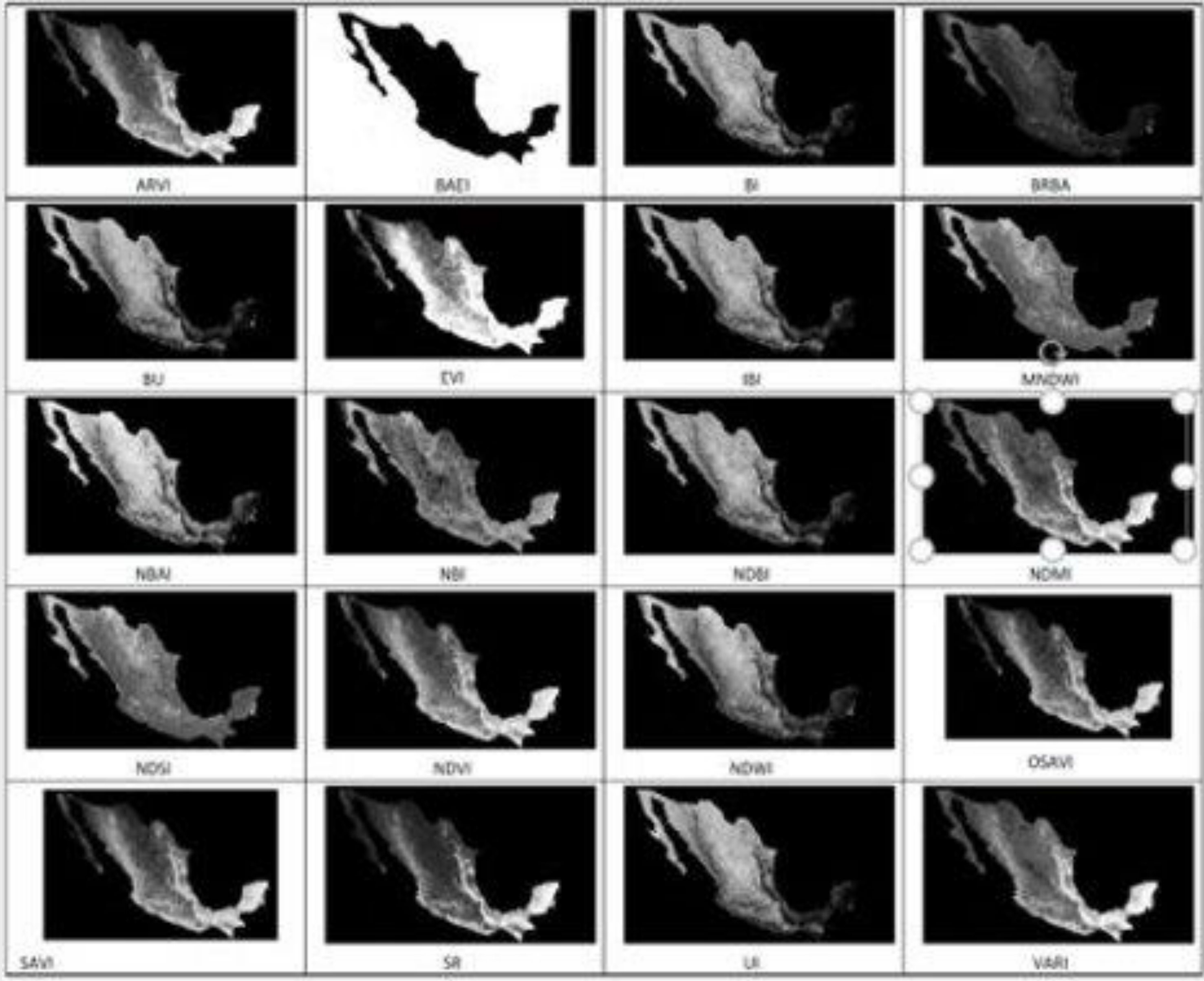
Image Reprojection & Alignment



GeoTIFF Images
12 Bands

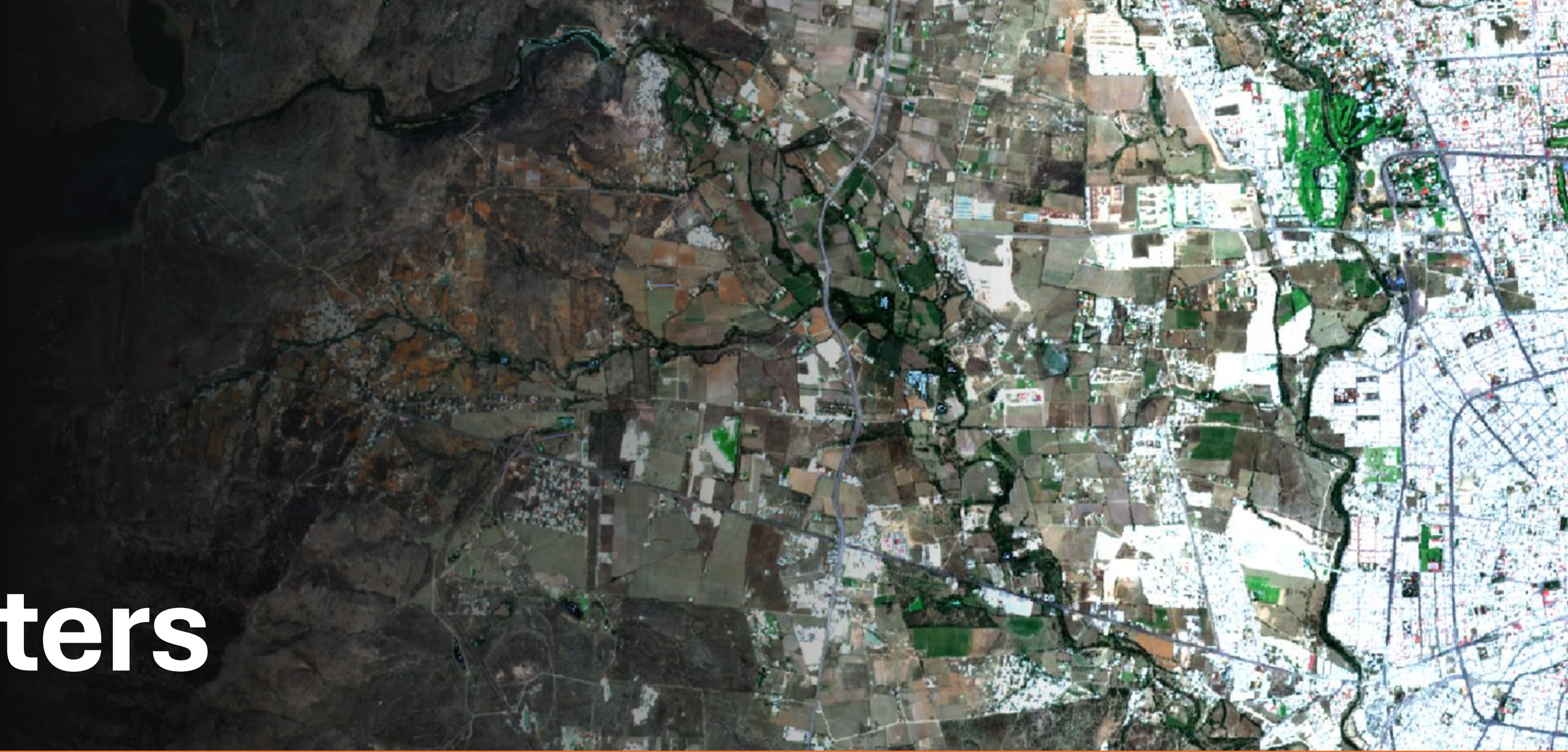
Sources:
https://developers.google.com/earth-engine/apidocs/ee-reducer-geometricmedian
https://www.researchgate.net/figure/The-reducer-operation-provided-by-Google-Earth-Engine-GEE-17_fig3_349430332
https://en.wikipedia.org/wiki/Geometric_median

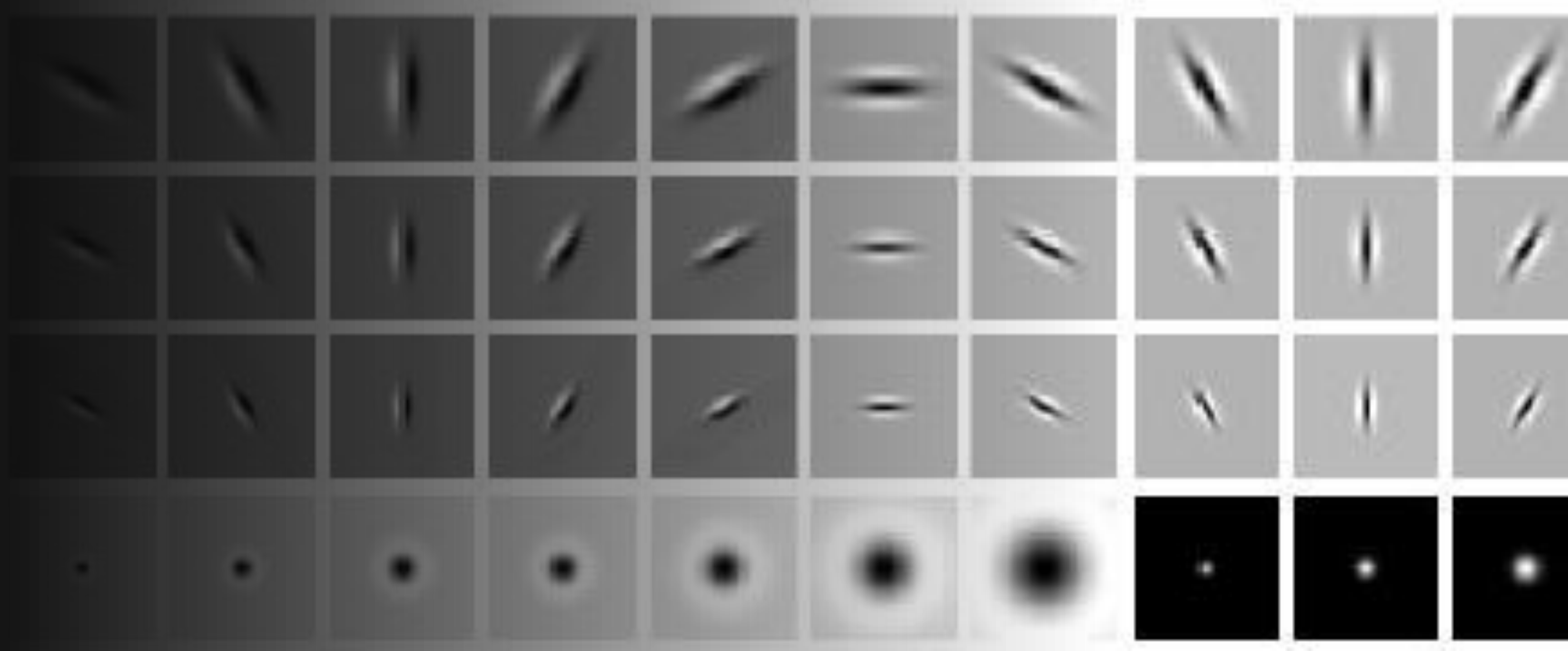
20 Spectral Indexes



Dimensiones de 110,000 pixeles x 70,000 pixeles

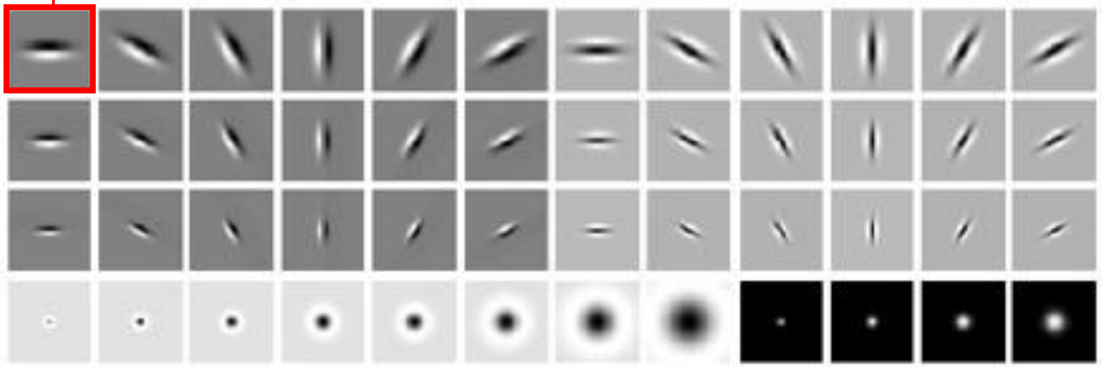
48 Texture Filters





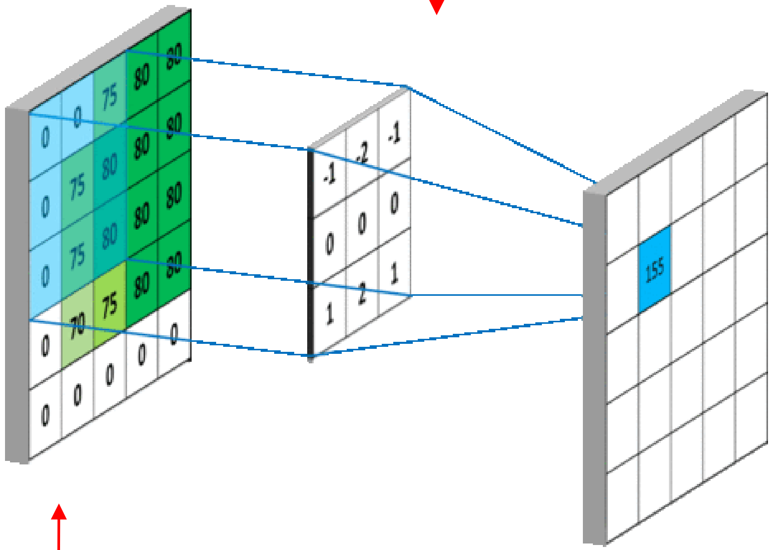
Texture Filters on the Infrared Band

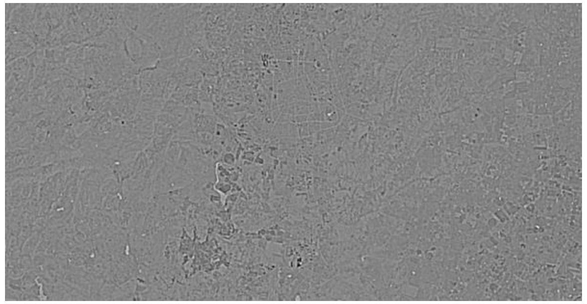
48 Texture Filters



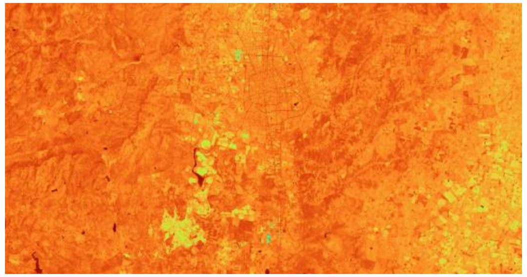
Leung-Malik Texture Filter Bank

Leung, Thomas, and Jitendra Malik. "Representing and recognizing the visual appearance of materials using three-dimensional textons." International journal of computer vision 43, no. 1 (2001): 29-44.

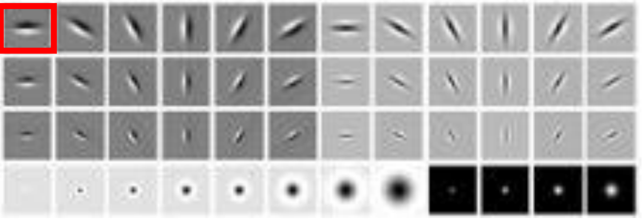




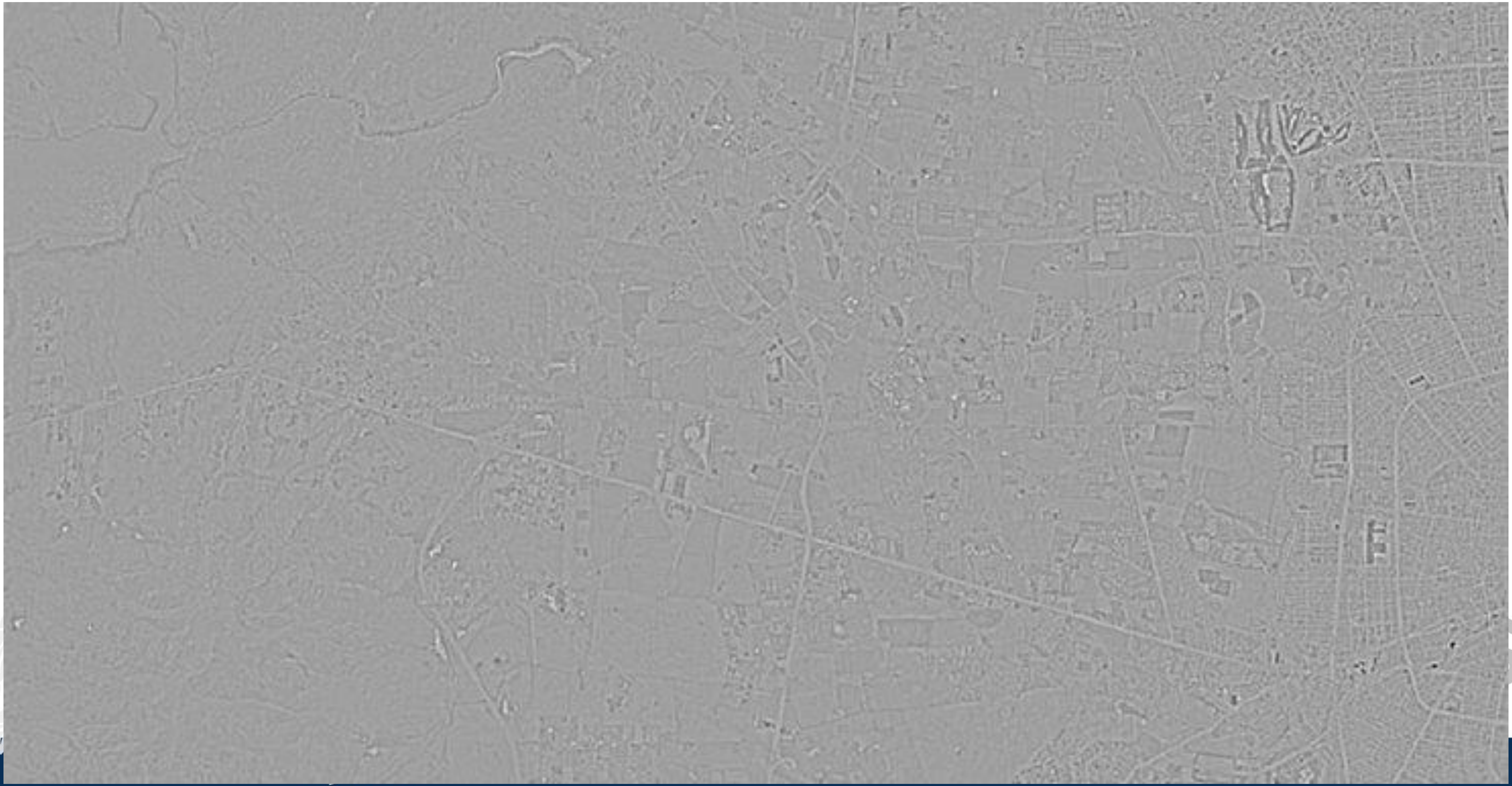
New Filtered Image

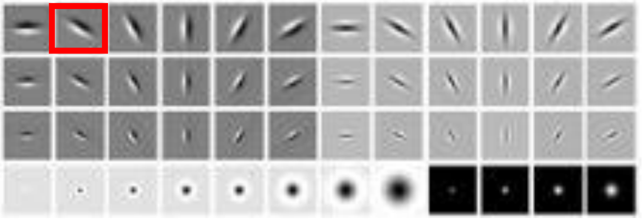


NIR

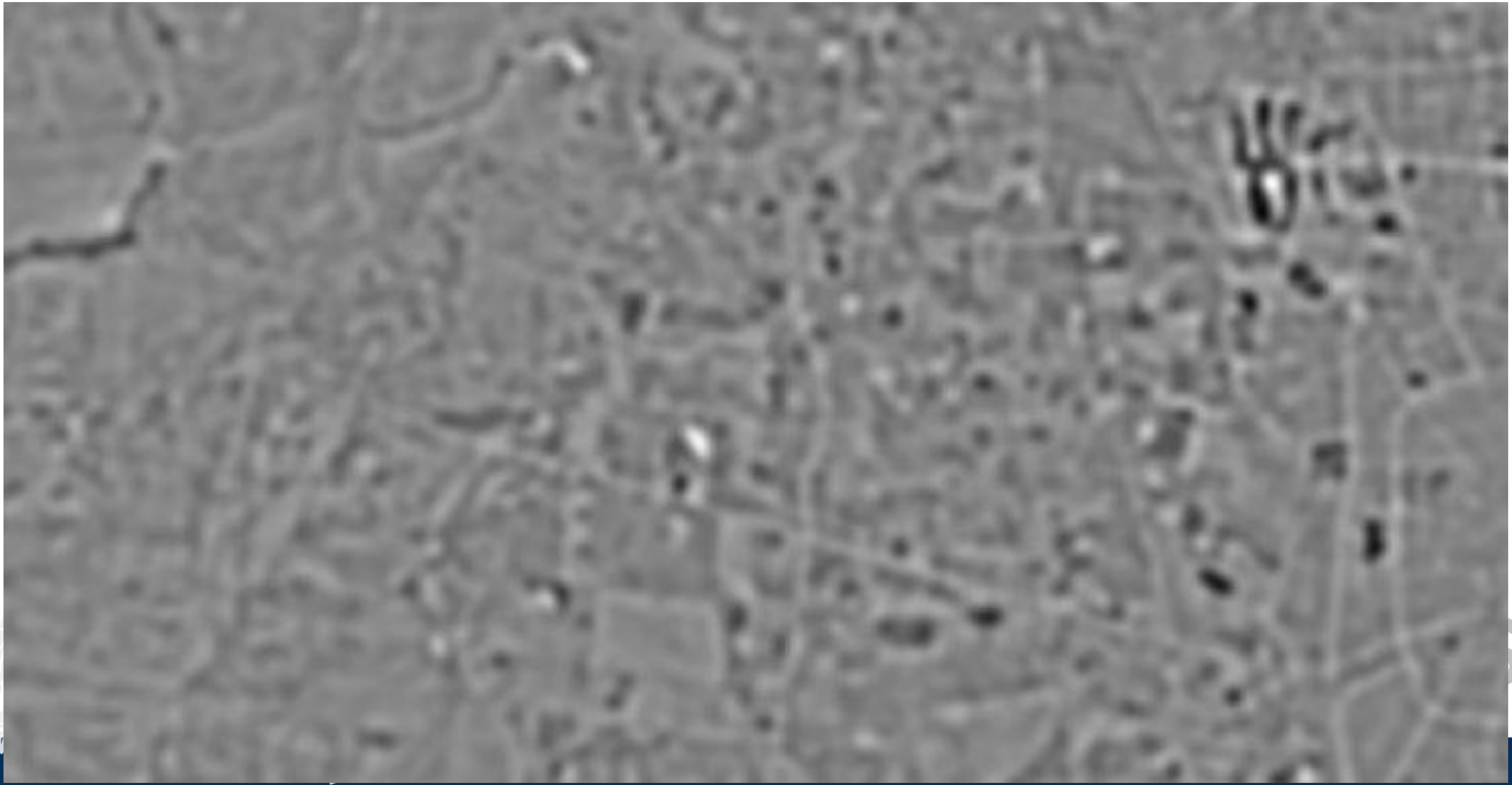


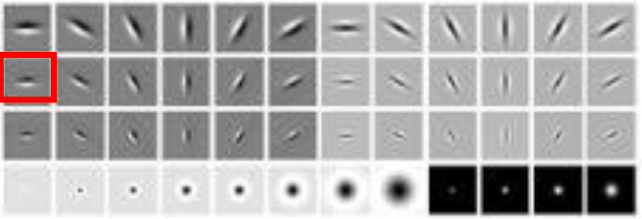
48 Texture Filters



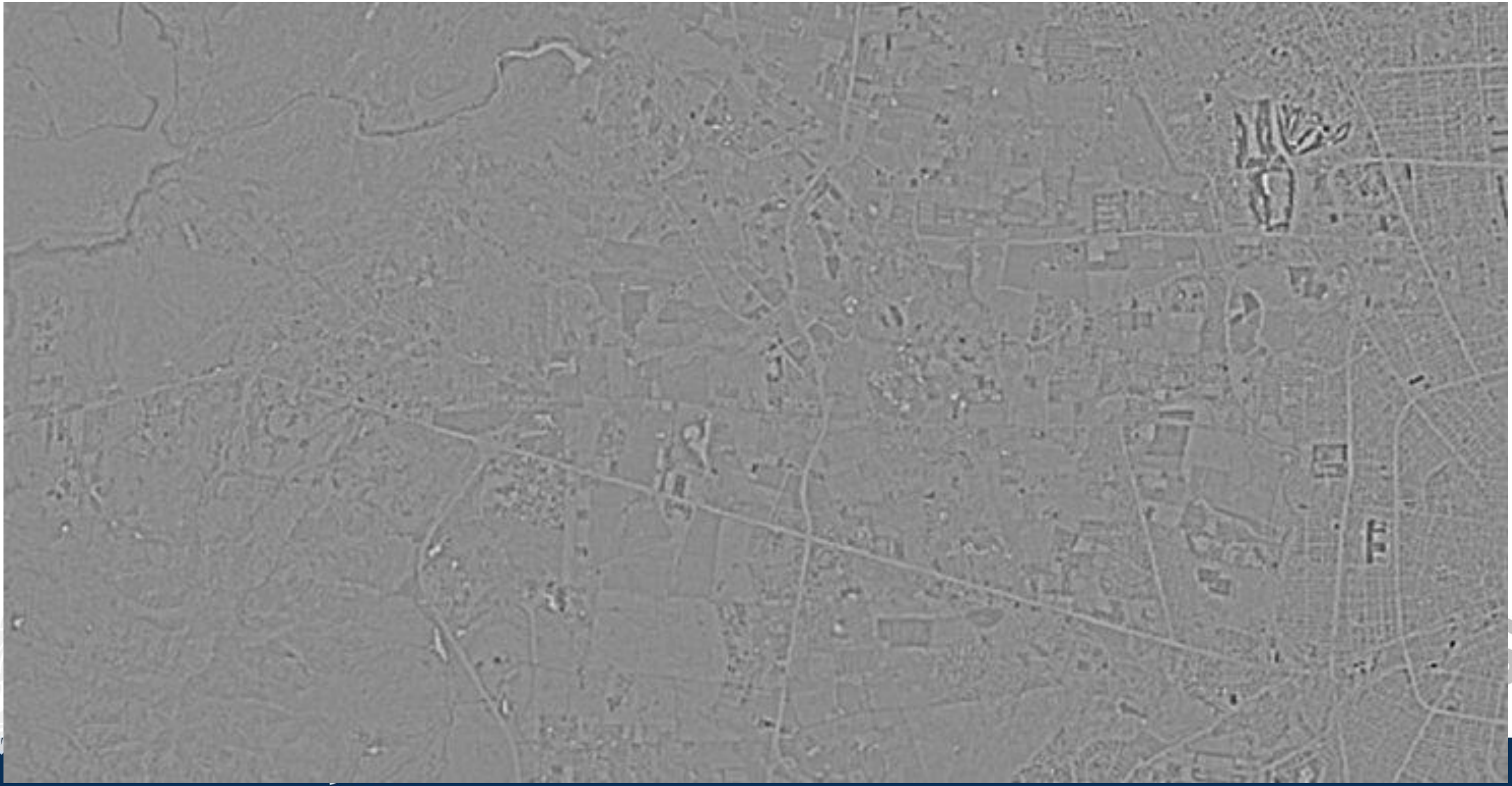


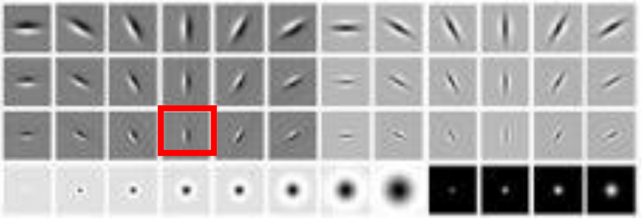
48 Texture Filters



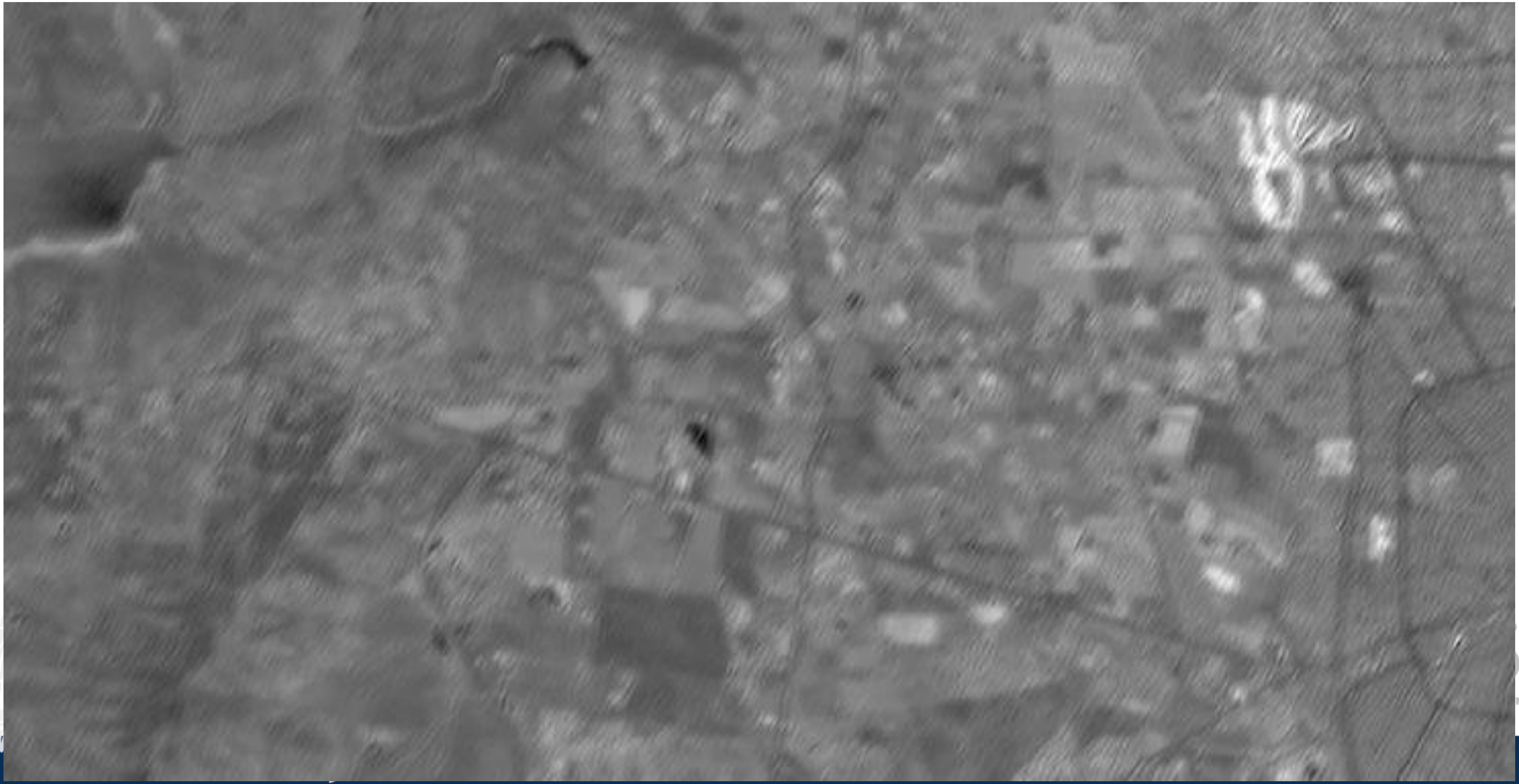


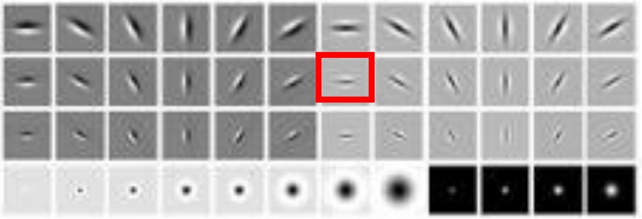
48 Texture Filters



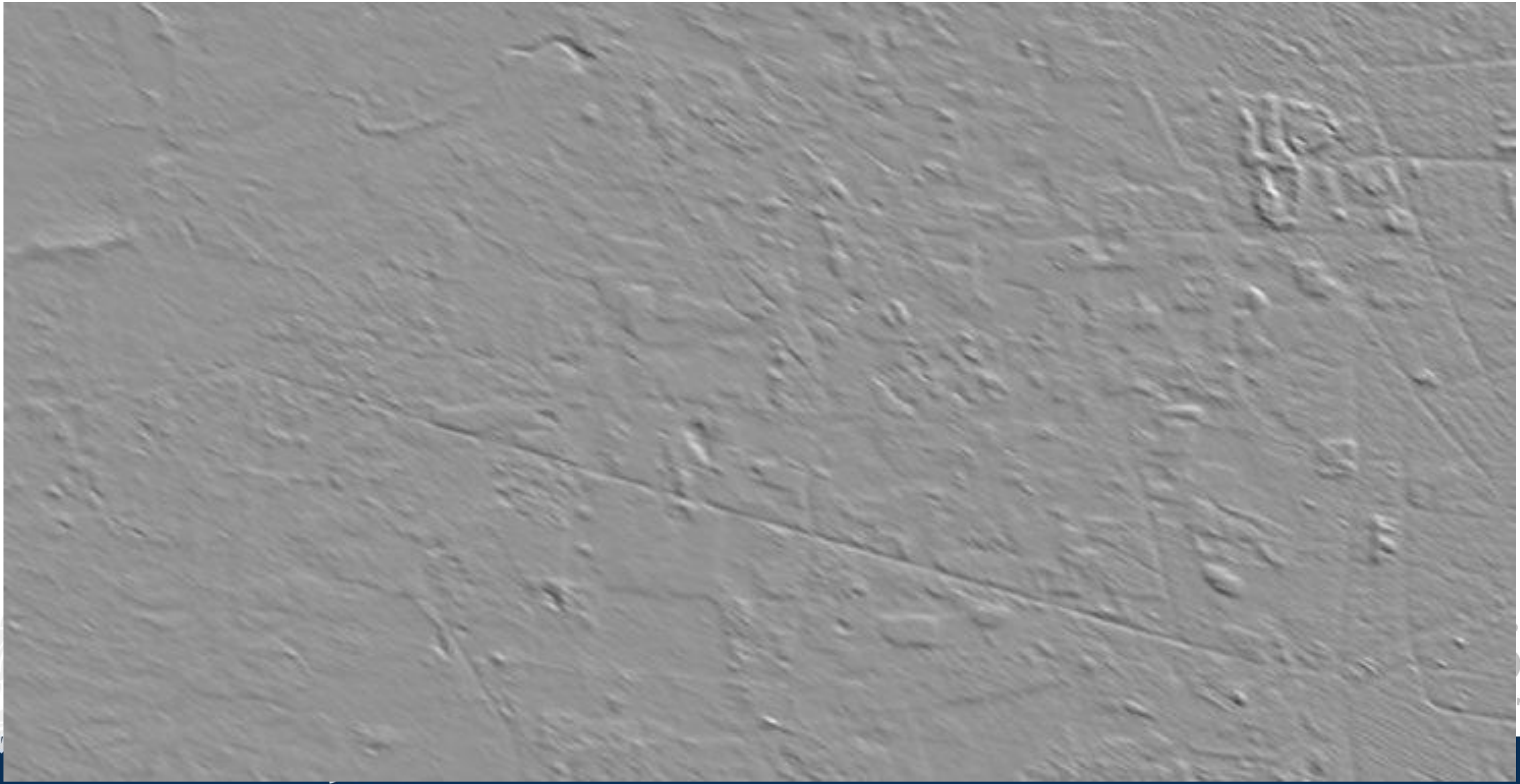


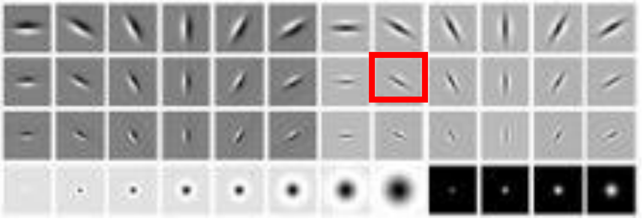
48 Texture Filters



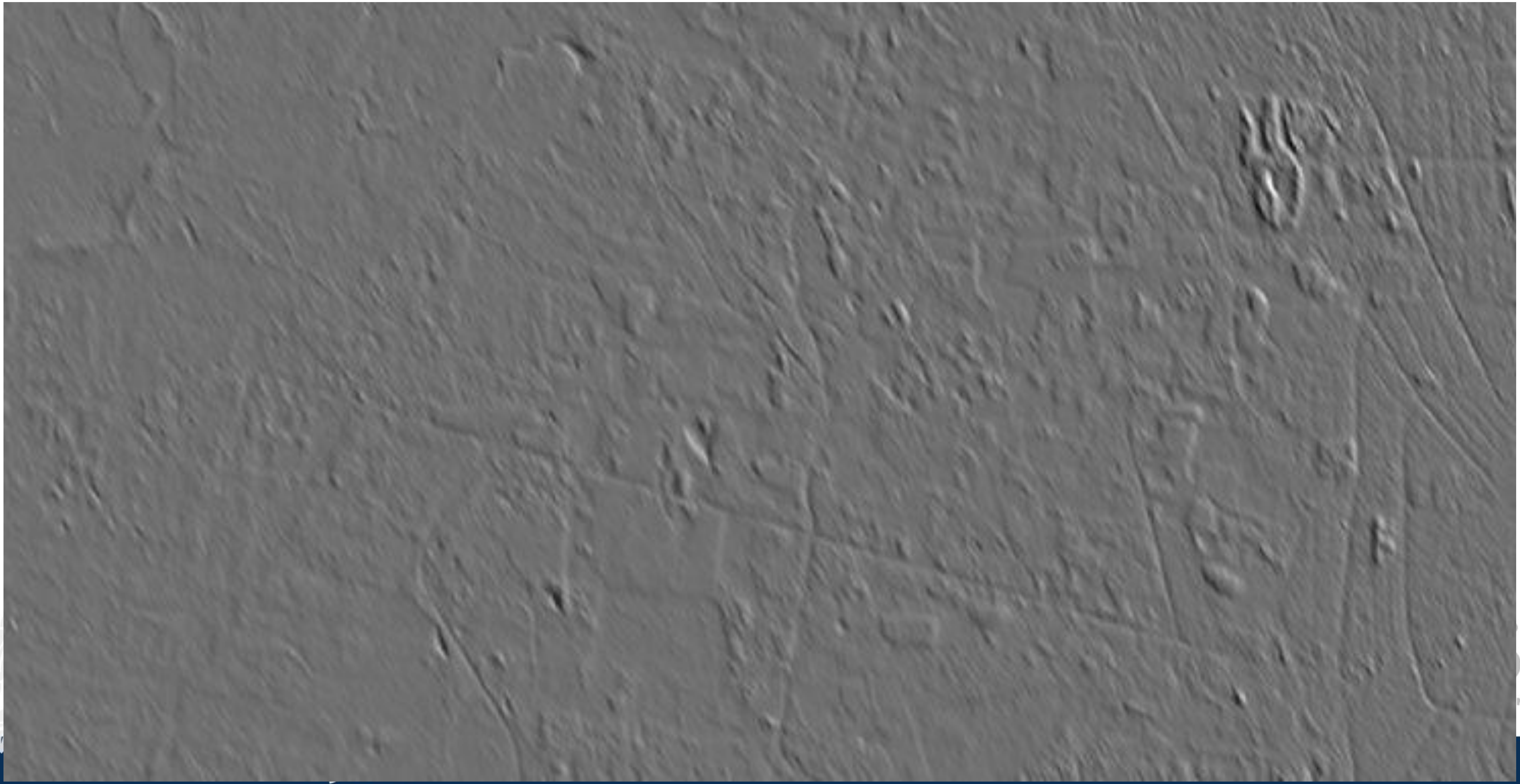


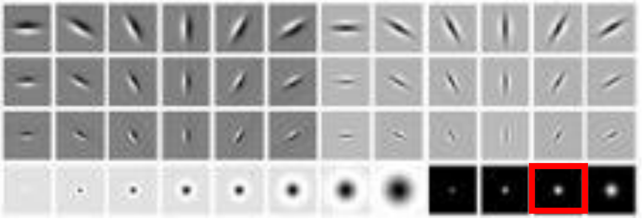
48 Texture Filters



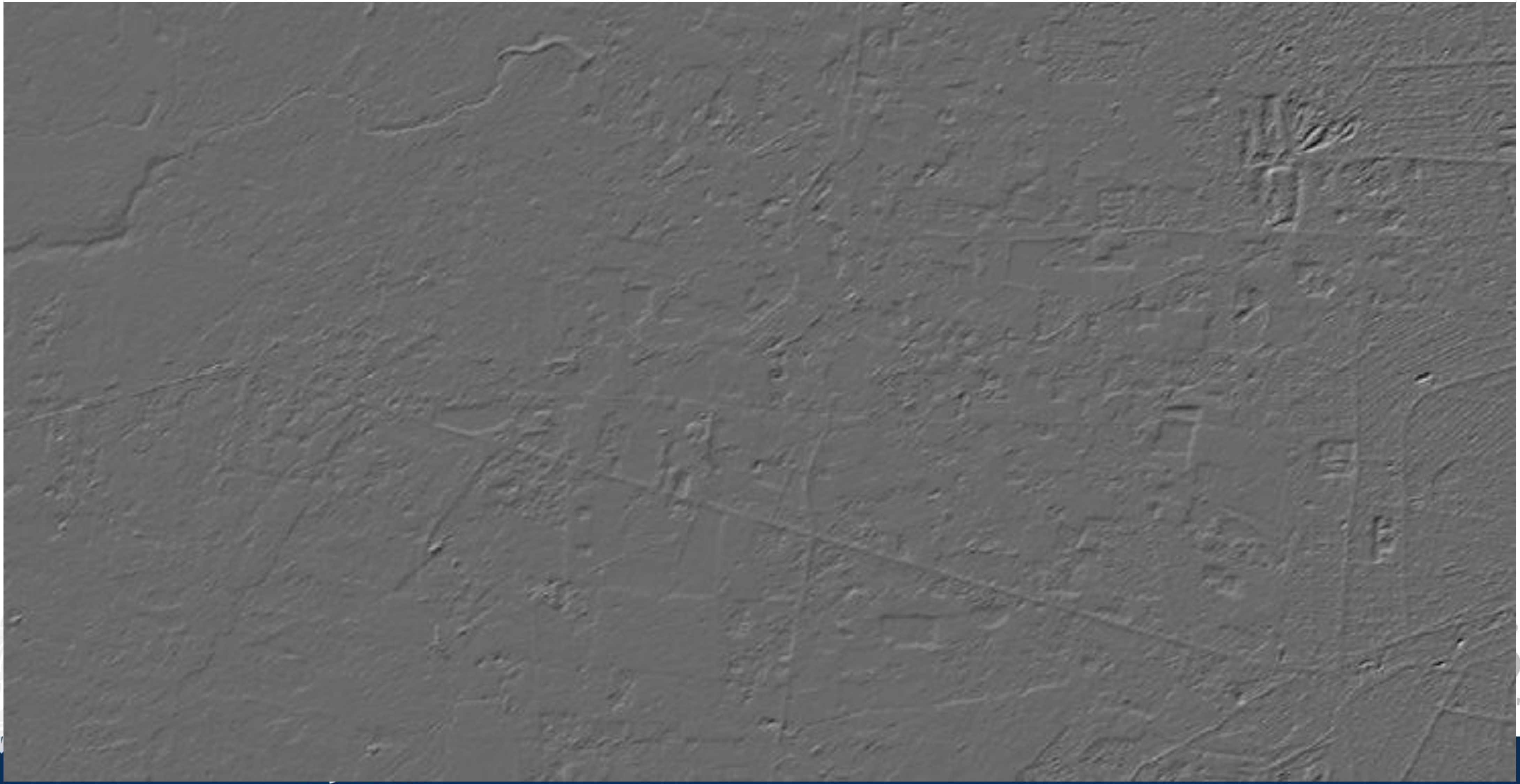


48 Texture Filters



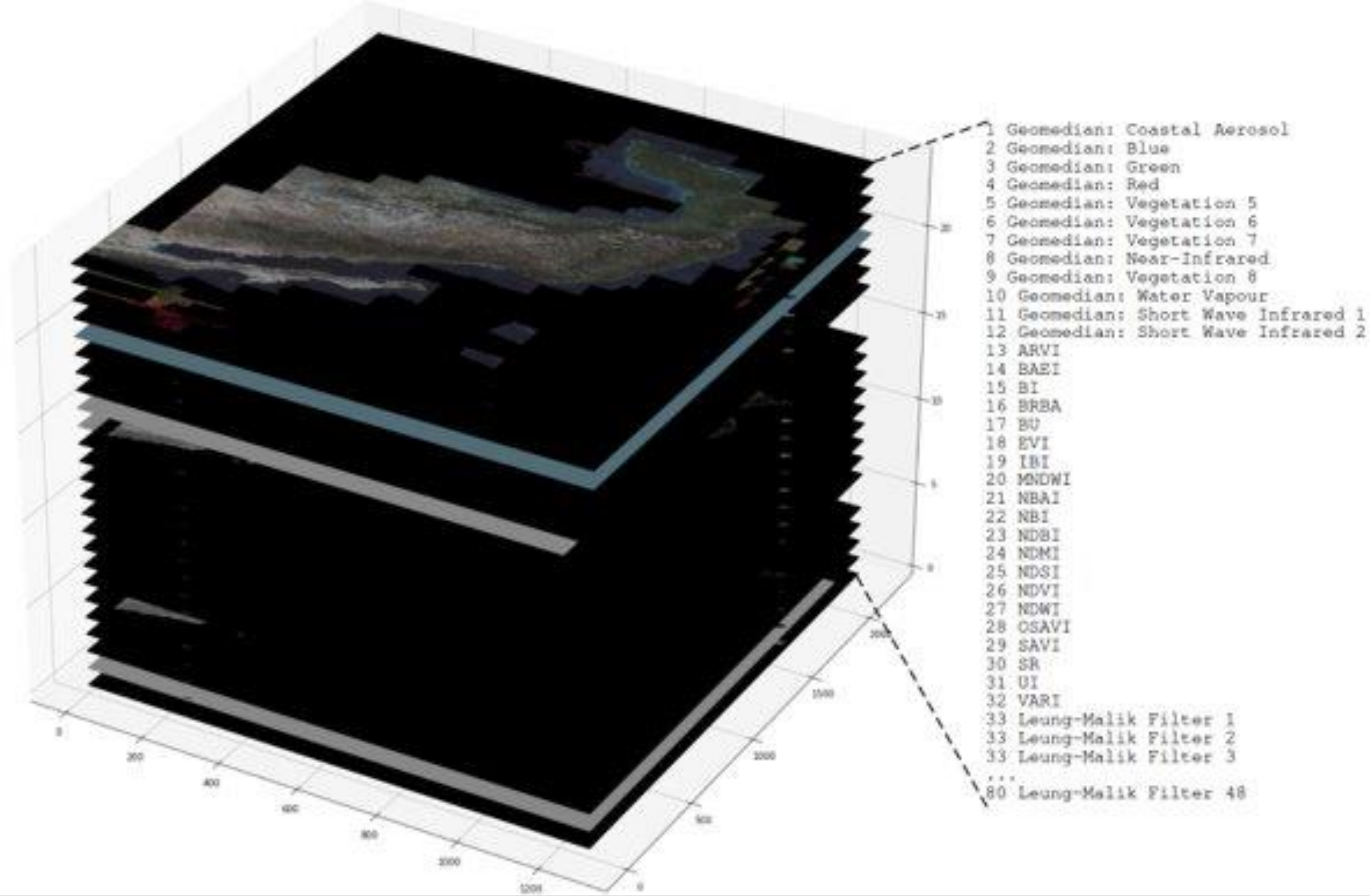


48 Texture Filters



80 Raster Layers

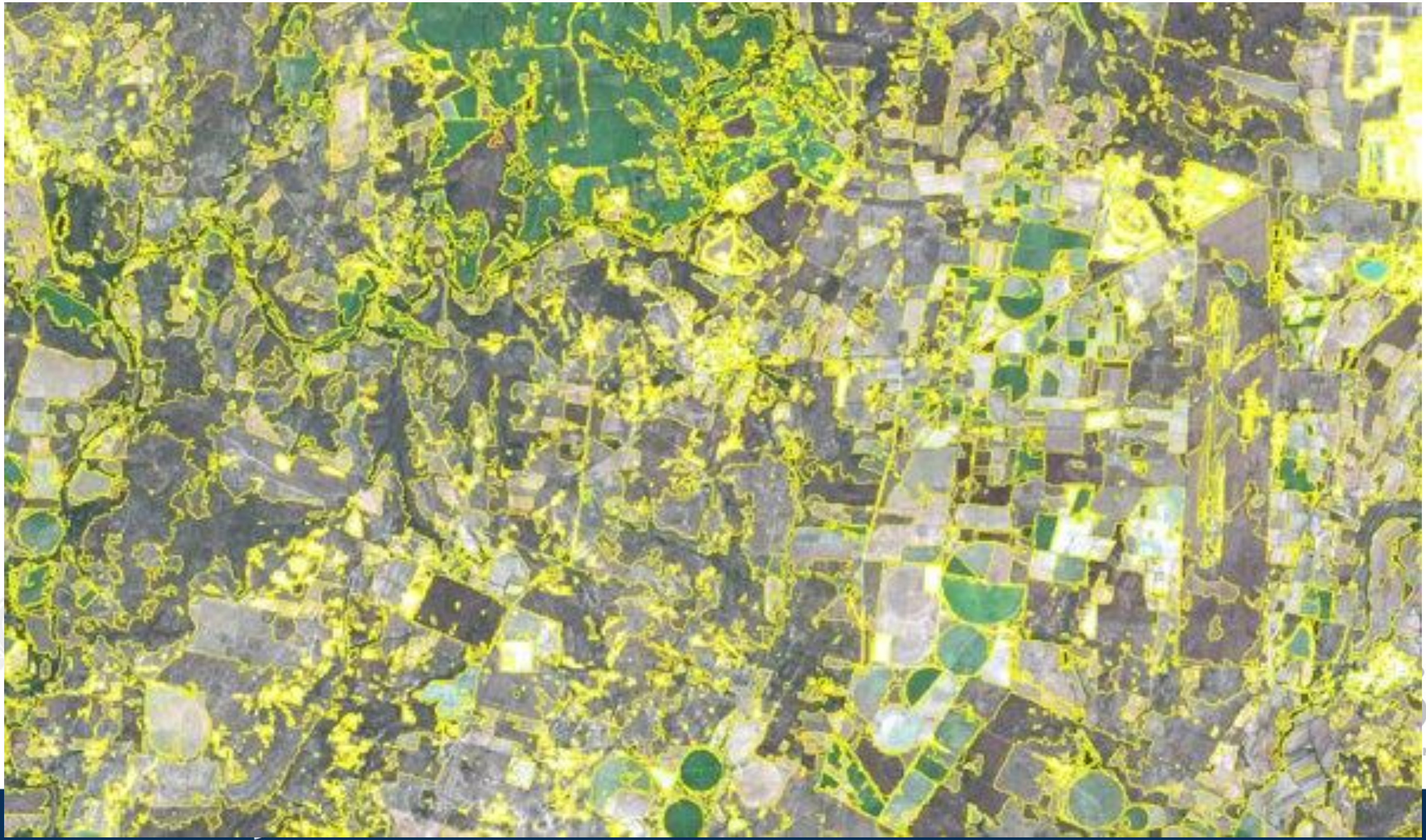
17.2 TB



1 Geomedian: Coastal Aerosol
2 Geomedian: Blue
3 Geomedian: Green
4 Geomedian: Red
5 Geomedian: Vegetation 5
6 Geomedian: Vegetation 6
7 Geomedian: Vegetation 7
8 Geomedian: Near-Infrared
9 Geomedian: Vegetation 8
10 Geomedian: Water Vapour
11 Geomedian: Short Wave Infrared 1
12 Geomedian: Short Wave Infrared 2
13 ARVI
14 BAEI
15 BI
16 BRBA
17 BU
18 EVI
19 IBI
20 MNDWI
21 NBAI
22 NBI
23 NDBI
24 NDMI
25 NDSI
26 NDVI
27 NDWI
28 OSAVI
29 SAVI
30 SR
31 UI
32 VARI
33 Leung-Malik Filter 1
33 Leung-Malik Filter 2
33 Leung-Malik Filter 3
...
80 Leung-Malik Filter 48

Geomedian Segmentation

Geomedian Segmentation



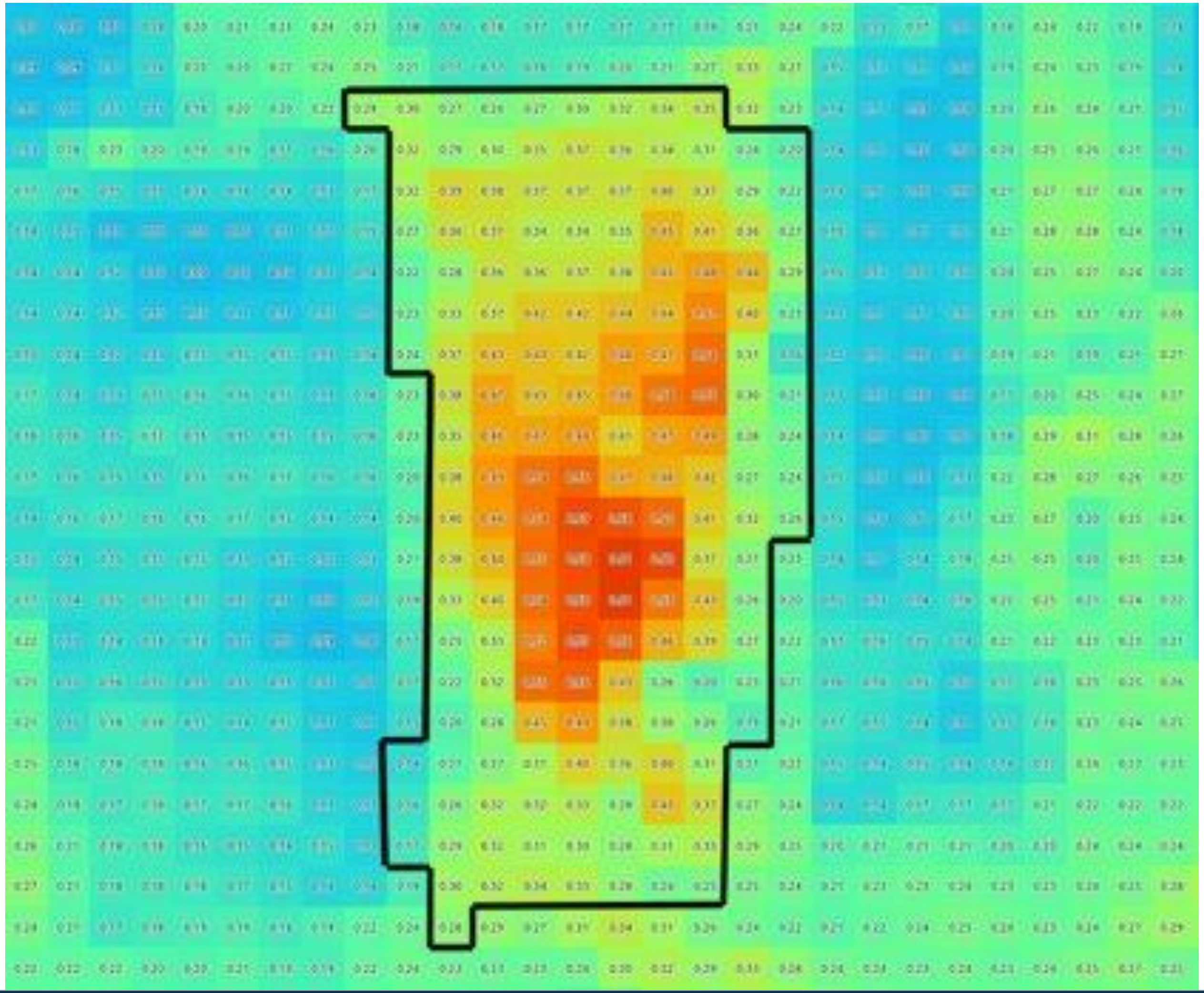
Characterization of Segments

ALL LAYERS

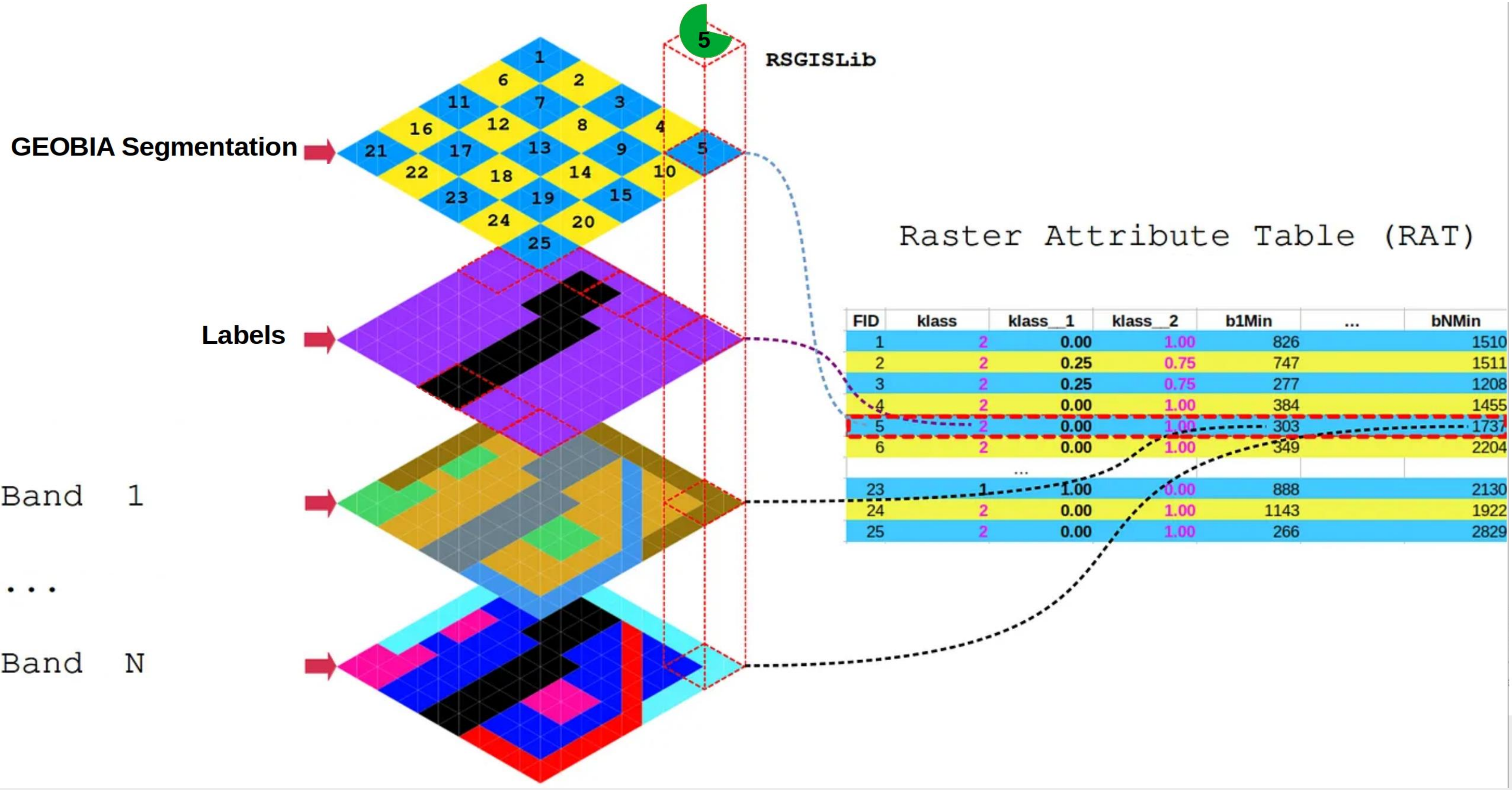
- Minimum
- Maximum
- Average
- Sum
- Standard Deviation

TEXTURE FILTERS

- Percentile 10 - 90

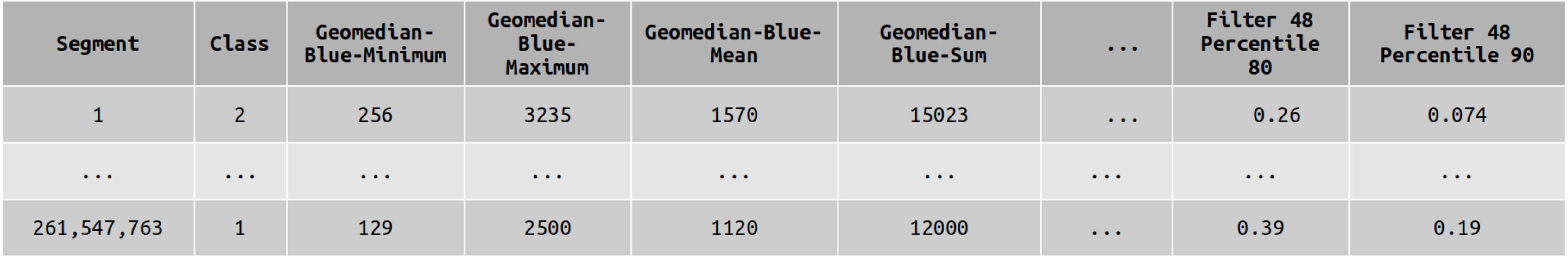


Characterization of Segments



FID	klass	klass__1	klass__2	b1Min	...	bNMin
1	2	0.00	1.00	826		1510
2	2	0.25	0.75	747		1511
3	2	0.25	0.75	277		1208
4	2	0.00	1.00	384		1455
5	2	0.00	1.00	303		1737
6	2	0.00	1.00	349		2204
		...				
23	1	1.00	0.00	888		2130
24	2	0.00	1.00	1143		1922
25	2	0.00	1.00	266		2829

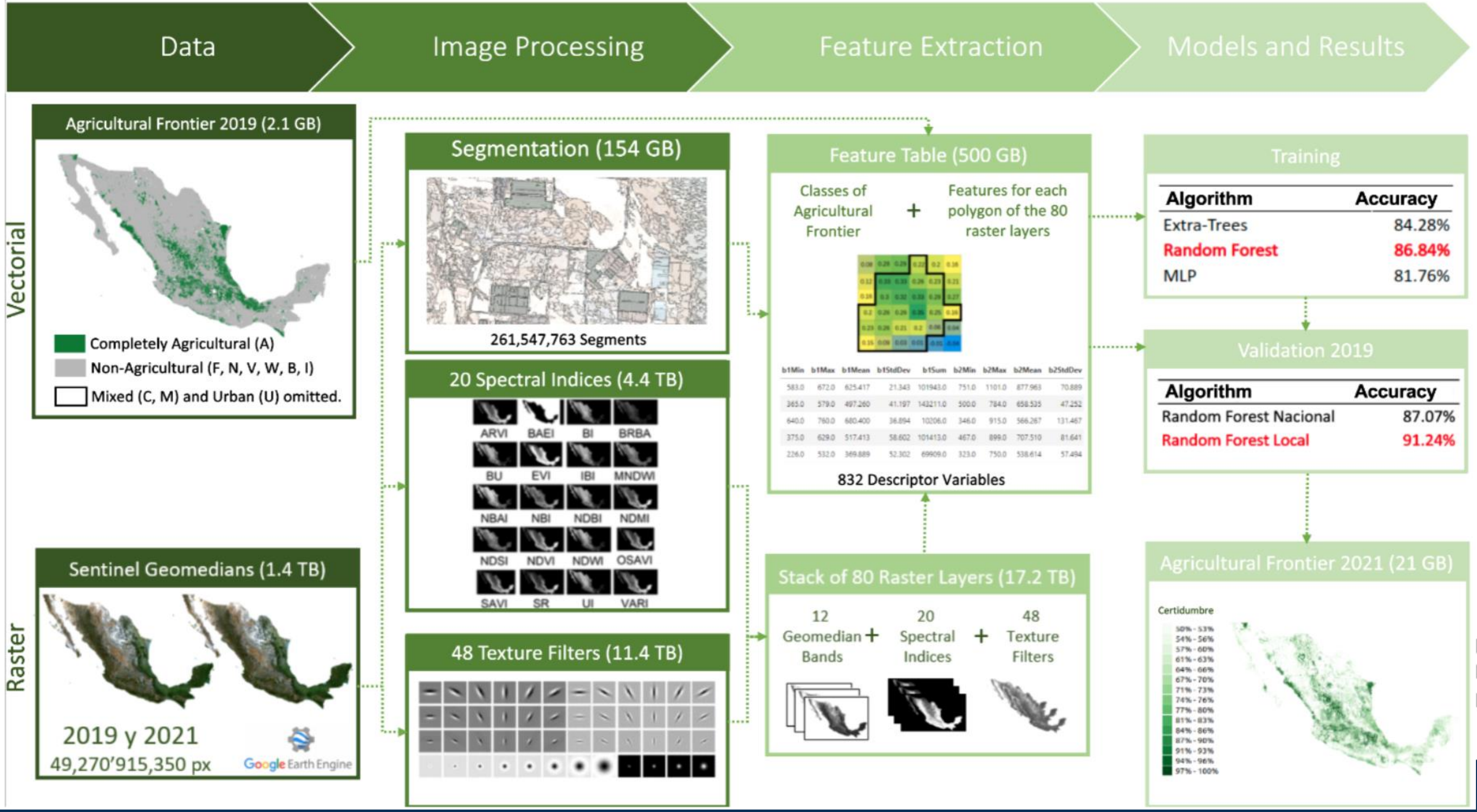
Segment characterization



Segment	Class	Geomedian-Blue-Minimum	Geomedian-Blue-Maximum	Geomedian-Blue-Mean	Geomedian-Blue-Sum	...	Filter 48 Percentile 80	Filter 48 Percentile 90
1	2	256	3235	1570	15023	...	0.26	0.074
...	...	...	...	...	...	...	...	...
261,547,763	1	129	2500	1120	12000	...	0.39	0.19

Data table with 834 columns = Segment + Class + eight hundred thirty-two Variables

Summary Methodology



Data

Image Processing

Feature Extraction

Models and Results

Vectorial

Agricultural Frontier 2019 (2.1 GB)

- 🟩 Completely Agricultural (A)
- ⬜ Non-Agricultural (F, N, V, W, B, I)
- □ Mixed (C, M) and Urban (U) omitted.

Raster

Sentinel Geomedians (1.4 TB)

2019 y 2021

49,270'915,350 px

Google Earth Engine

Segmentation (154 GB)

261,547,763 Segments

20 Spectral Indices (4.4 TB)

48 Texture Filters (11.4 TB)

Feature Table (500 GB)

Classes of Agricultural Frontier + Features for each polygon of the 80 raster layers

b1Min	b1Max	b1Mean	b1StdDev	b1Sum	b2Min	b2Max	b2Mean	b2StdDev
583.0	672.0	625.417	21.343	101943.0	751.0	1101.0	877.963	70.889
365.0	579.0	497.260	41.197	143211.0	500.0	784.0	658.535	47.252
640.0	760.0	680.400	36.894	10206.0	346.0	915.0	566.267	131.467
375.0	629.0	517.413	58.602	101413.0	467.0	899.0	707.510	81.641
226.0	532.0	369.889	52.302	69909.0	323.0	750.0	538.614	57.494

832 Descriptor Variables

Stack of 80 Raster Layers (17.2 TB)

12 Geomedian Bands + 20 Spectral Indices + 48 Texture Filters

Training

Algorithm	Accuracy
Extra-Trees	84.28%
Random Forest	86.84%
MLP	81.76%

Validation 2019

Algorithm	Accuracy
Random Forest Nacional	87.07%
Random Forest Local	91.24%

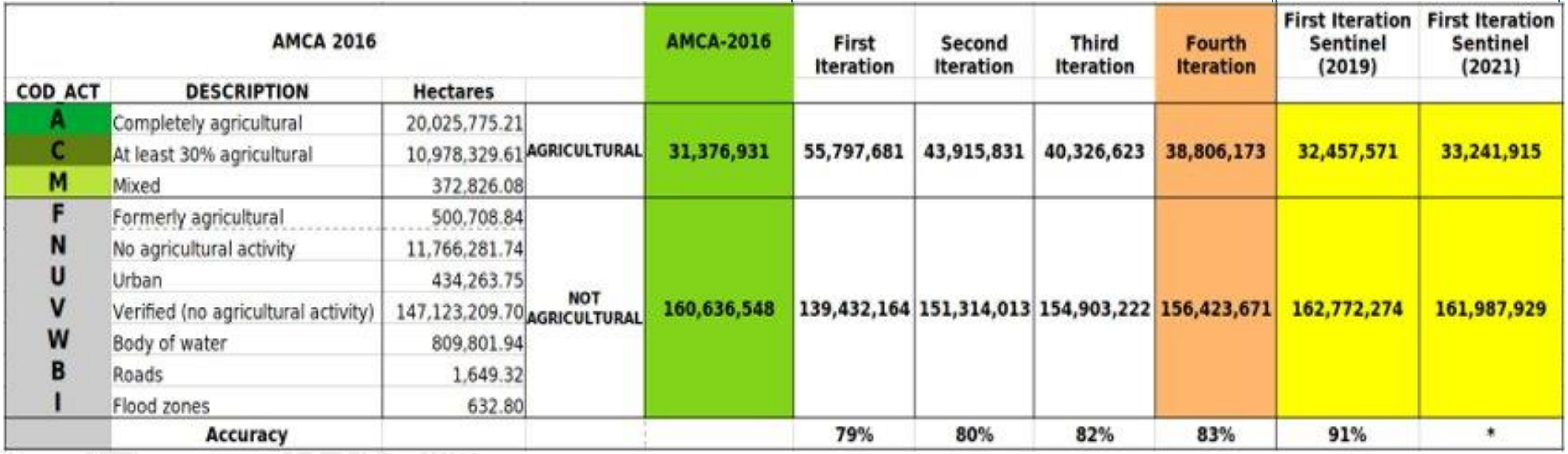
Agricultural Frontier 2021 (21 GB)

Results

Results

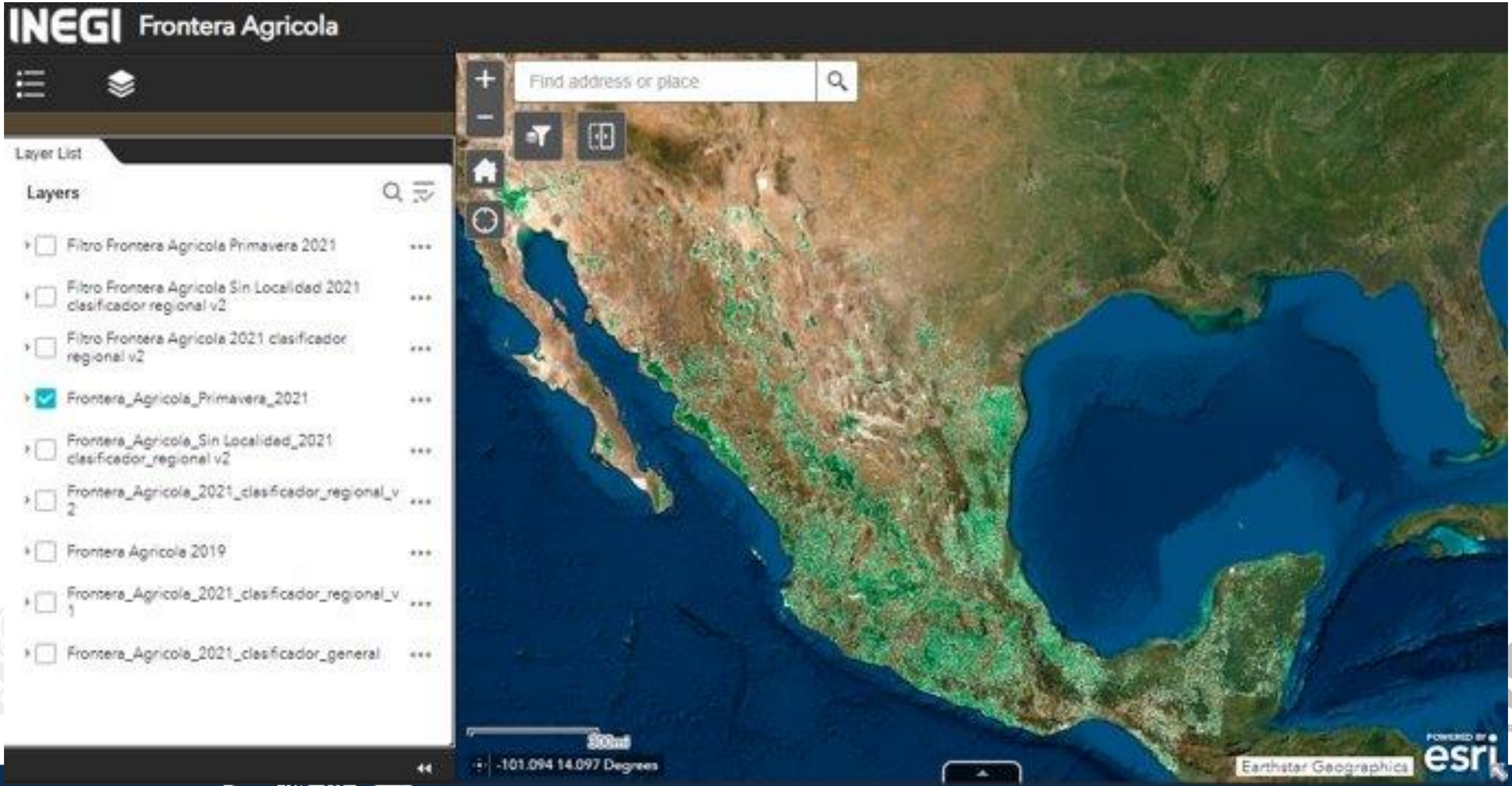
Landsat

Sentinel-2



AMCA 2016				AMCA-2016	First Iteration	Second Iteration	Third Iteration	Fourth Iteration	First Iteration Sentinel (2019)	First Iteration Sentinel (2021)
COD_ACT	DESCRIPTION	Hectares								
A	Completely agricultural	20,025,775.21	AGRICULTURAL	31,376,931	55,797,681	43,915,831	40,326,623	38,806,173	32,457,571	33,241,915
C	At least 30% agricultural	10,978,329.61								
M	Mixed	372,826.08								
F	Formerly agricultural	500,708.84	NOT AGRICULTURAL	160,636,548	139,432,164	151,314,013	154,903,222	156,423,671	162,772,274	161,987,929
N	No agricultural activity	11,766,281.74								
U	Urban	434,263.75								
V	Verified (no agricultural activity)	147,123,209.70								
W	Body of water	809,801.94								
B	Roads	1,649.32								
I	Flood zones	632.80								
	Accuracy				79%	80%	82%	83%	91%	*

Web App



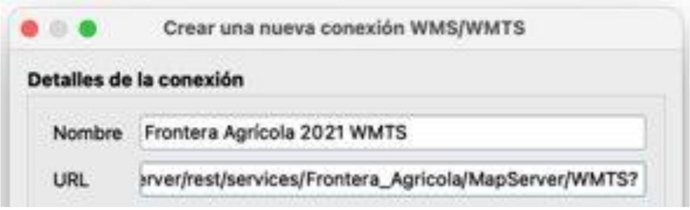
INEGI Frontera Agricola

Layer List

Layers

- ☐ Filtro Frontera Agricola Primavera 2021
- ☐ Filtro Frontera Agricola Sin Localidad 2021 clasificador regional v2
- ☐ Filtro Frontera Agricola 2021 clasificador regional v2
- ☑ Frontera_Agricola_Primavera_2021
- ☐ Frontera_Agricola_Sin Localidad_2021 clasificador_regional v2
- ☐ Frontera_Agricola_2021_clasificador_regional_v2
- ☐ Frontera Agricola 2019
- ☐ Frontera_Agricola_2021_clasificador_regional_v1
- ☐ Frontera_Agricola_2021_clasificador_general

WMTS

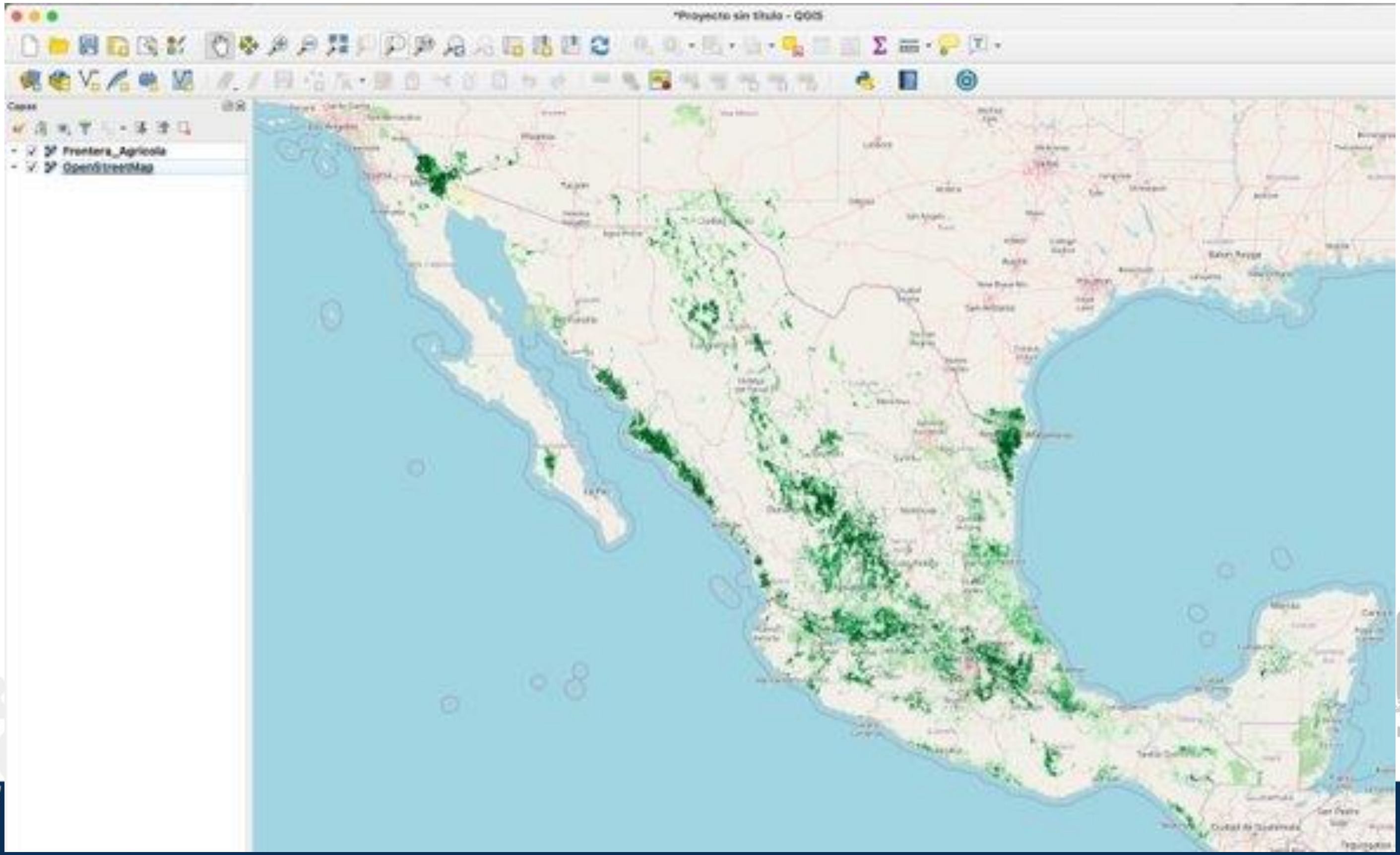


Crear una nueva conexión WMS/WMTS

Detalles de la conexión

Nombre: Frontera Agrícola 2021 WMTS

URL: erver/rest/services/Frontera_Agricola/MapServer/WMTS?

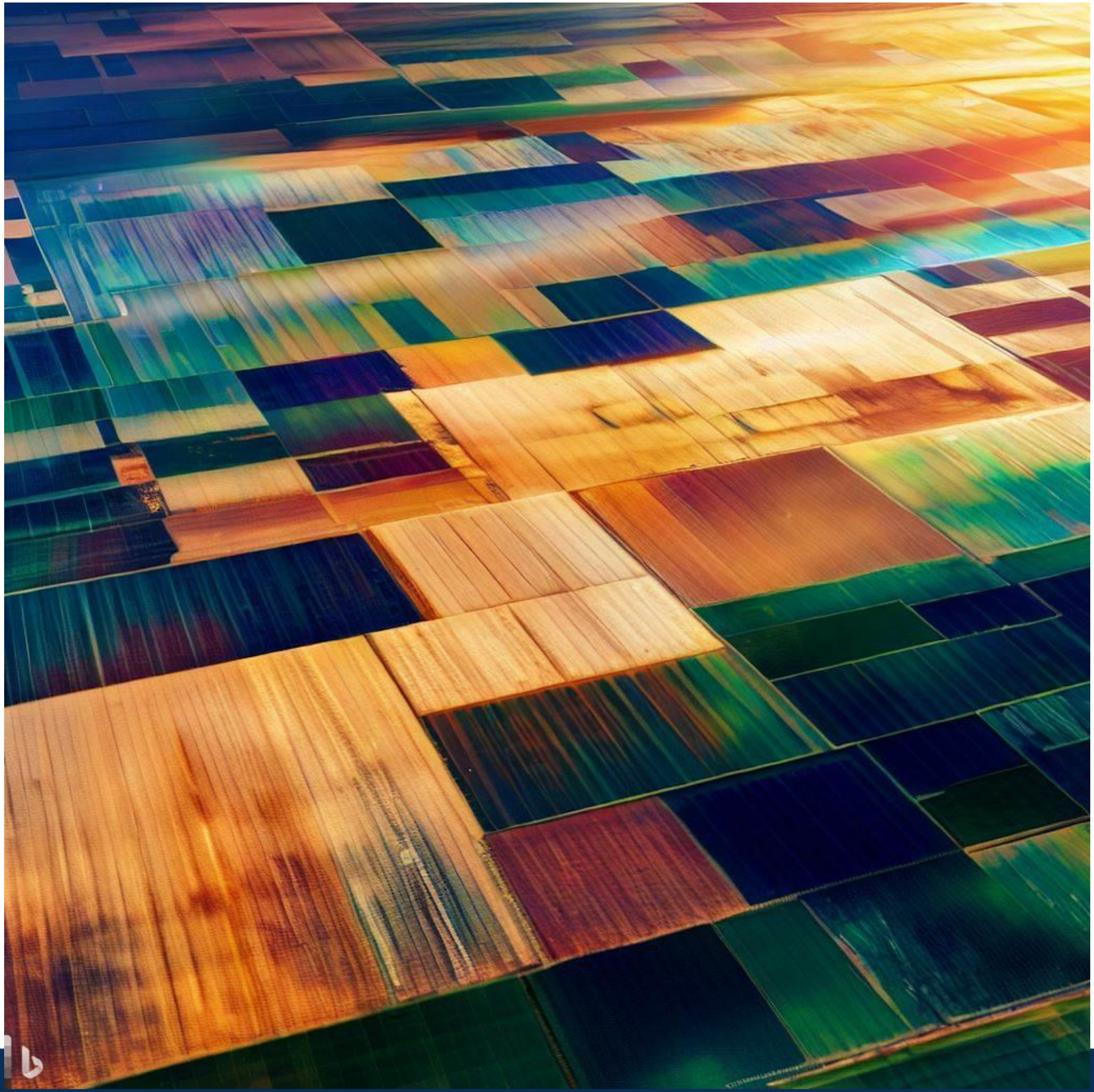


Next Steps

- Compare the result of the new run with the results of the 2022 Census.

- Identify areas with good and poor algorithm performance

- Algorithm refinement from review results



Conclusions

Conclusions

Open Data

Big Data

Successful Collaboration

Replicable Methodology

Relevant Results

IAOS-ISI 2024
MEXICO CONFERENCE

THANK YOU

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

IAOS-ISI 2024

MEXICO CONFERENCE

Using Record Linkage of Administrative Records to Improve Federal Justice Statistics in the United States

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

Mark Motivans, Ph.D.
Statistician
U.S. Bureau of Justice Statistics

BJS | Bureau of Justice Statistics

The views, opinions, and content expressed in this presentation do not necessarily reflect the views, opinions, or policies of the Bureau of Justice Statistics (BJS), the Office of Justice Programs (OJP), or the U.S. Department of Justice (DOJ).

IAOS-ISI 2024
MEXICO CONFERENCE

Outline

- Overview of the Federal Justice Statistics Program (FJSP)
- Using Record Linkage to Improve Statistics
- Conclusion

BJS is the statistical agency of the U.S. Department of Justice

- The Bureau Justice of Statistics (BJS) is one of thirteen federal statistical agencies in the Executive Branch of the U.S. Government

- The mission of BJS is to collect, analyze, publish, and disseminate information on crime, criminal offenders, victims of crime, and the operation of justice systems at all levels of government

The Federal Justice Statistics Program (FJSP) is made up of nearly **30 years** of standardized administrative **data** from **six agencies**

- The FJSP was created by BJS in 1982 to serve as a national clearinghouse of federal criminal case processing data (1994-2022)

- The primary goal is to enhance uniformity in statistics across the federal criminal case process

- Methods standardize differences in administrative data files to improve uniformity, including—
  - Standard unit of count (person-case)
  - Common reporting period
  - Common offense classification
  - Standard disposition outcomes

6 federal justice agencies provide data annually

STAGE/DATA SOURCE

Arrest and booking

- **U.S. Marshals Service**: Persons arrested for federal offenses and booked by the U.S. Marshals Service
- **Drug Enforcement Administration**: Persons arrested by DEA agents

Prosecution

- **Executive Office for U.S. Attorneys**: Persons investigated by U.S. attorneys' offices

Pretrial release

- **Administrative Office of the U.S. Courts**: Persons supervised by pretrial services officers

Adjudication/Sentencing

- **Administrative Office of the U.S. Courts**: Persons in cases adjudicated in U.S. district courts
- **U.S. Sentencing Commission**: Defendants sentenced in U.S. district courts

Appeals

- **Administrative Office of the U.S. Courts**: Criminal appeals heard in U.S. Courts of Appeals

Corrections

- **Administrative Office of the U.S. Courts**: Persons under federal supervision in the community
- **Federal Bureau of Prisons**: Persons admitted, released and present in federal prison system

BJS has developed methods to **link records across stages** of the criminal justice system

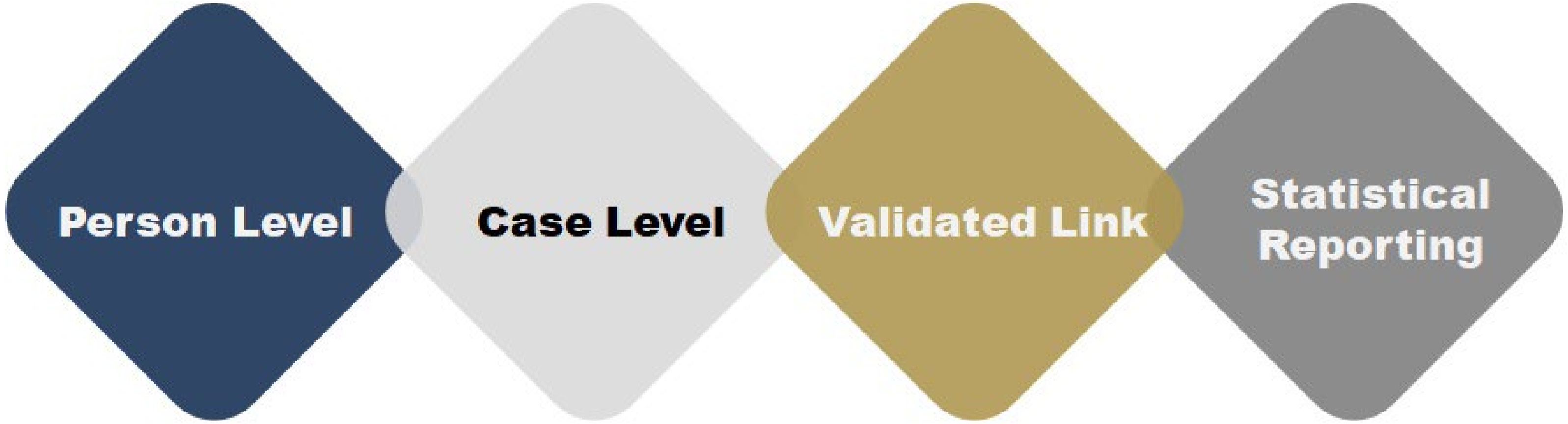
- There isn't a single identifier collected by each agency that permits linking across agencies

- The Dyad Link File (DLF) approach uses algorithms that employ exact and fuzzy matching of person and case identifiers that are available in each agency's data

- This strategy focuses on establishing links between pairs of agency files (or "dyads") from adjacent stages of the case process

- Information is at varying levels of quality and completeness, affecting success in linking records

Linking Strategy



Person Level

Person-level identifiers (include name, FBI number, date of birth, and more)

Case Level

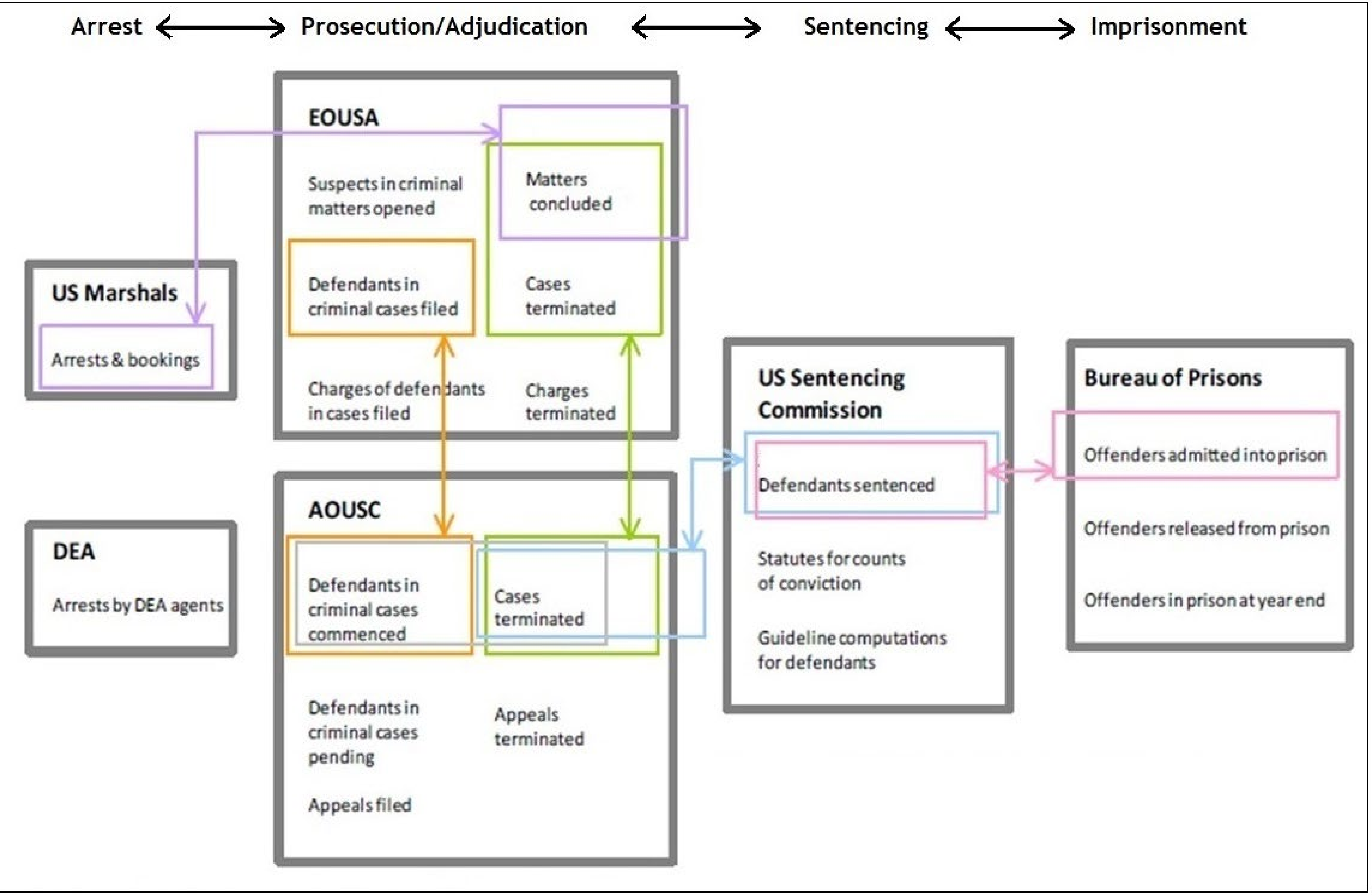
Case-level identifiers (include court docket number, federal district, judge, and more)

Validated Link

Algorithms provide summary measure of match quality using Levenshtein Distance

Statistical Reporting

Linked data are used in statistical publications and development is underway for inclusion in data tool



Arrest ⟷ Prosecution/Adjudication ⟷ Sentencing ⟷ Imprisonment

US Marshals

Arrests & bookings

DEA

Arrests by DEA agents

EOUSA

Suspects in criminal matters opened

Matters concluded

Defendants in criminal cases filed

Cases terminated

Charges of defendants in cases filed

Charges terminated

AOUSC

Defendants in criminal cases commenced

Cases terminated

Defendants in criminal cases pending

Appeals terminated

Appeals filed

US Sentencing Commission

Defendants sentenced

Statutes for counts of conviction

Guideline computations for defendants

Bureau of Prisons

Offenders admitted into prison

Offenders released from prison

Offenders in prison at year end

Diagram of available data and links.

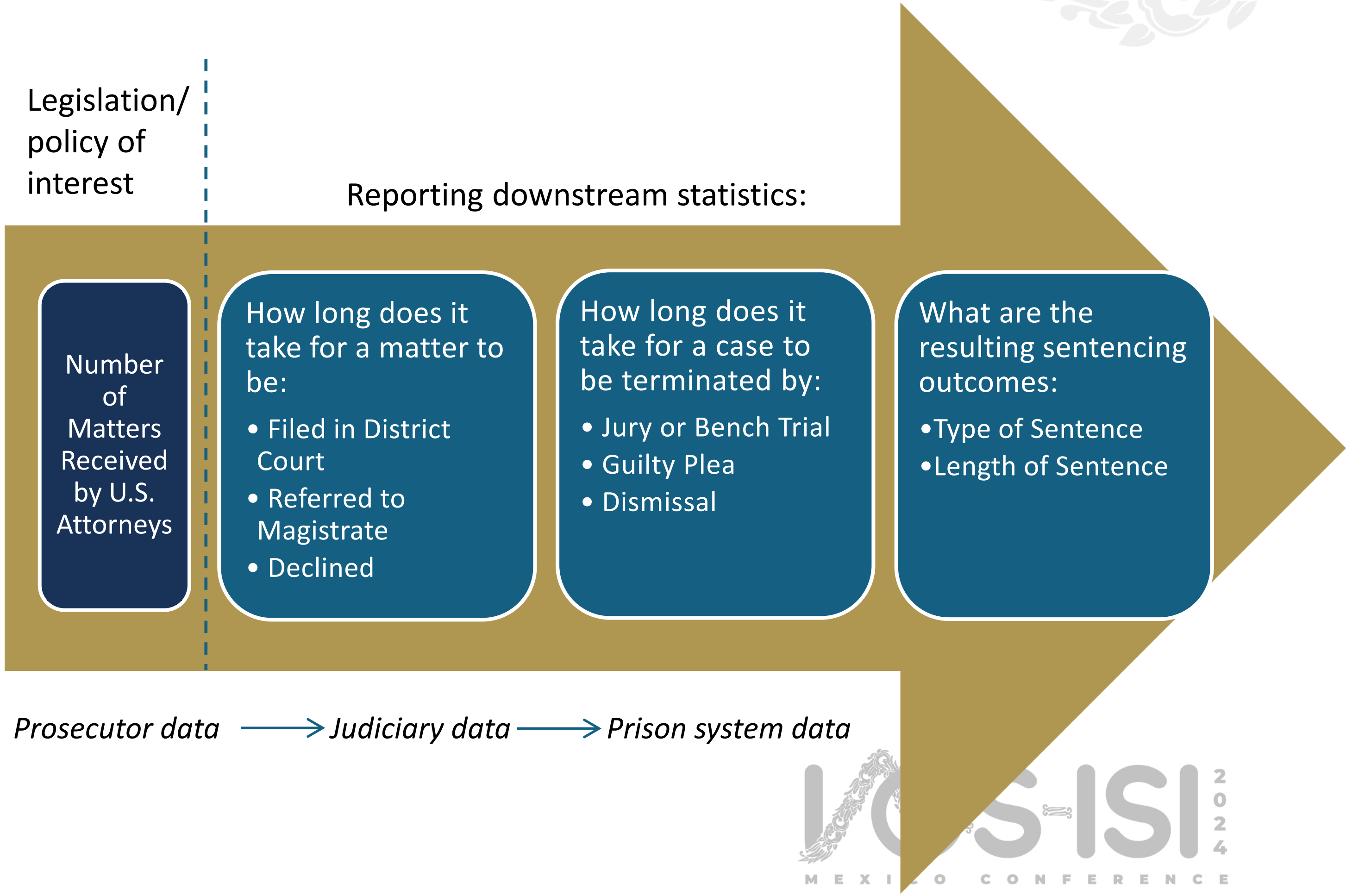
- Once the link is made, personal identifiers are removed and replaced with a sanitized identification number
- This allows users to link case records without needing confidential identifying information

Using Record Linkage to Improve Statistics

- Statistics are typically reported using cross-sectional data
- Ignores the element of time, which is especially important in a mostly linear system like the federal criminal justice system
- Examining the time it takes to move through the system can help to identify resource discrepancies and measure the value of policy changes
- The linked data files allows measure of the lag for different offense types and changes over time

Linked data can be used to investigate the cascading effects of policy decisions made at earlier stages

Changes in policy and laws have "downstream" impacts on later stages and linked data can be used to generate statistics for tracked cohorts



Legislation/ policy of interest

Reporting downstream statistics:

Number of Matters Received by U.S. Attorneys

How long does it take for a matter to be:

- Filed in District Court
- Referred to Magistrate
- Declined

How long does it take for a case to be terminated by:

- Jury or Bench Trial
- Guilty Plea
- Dismissal

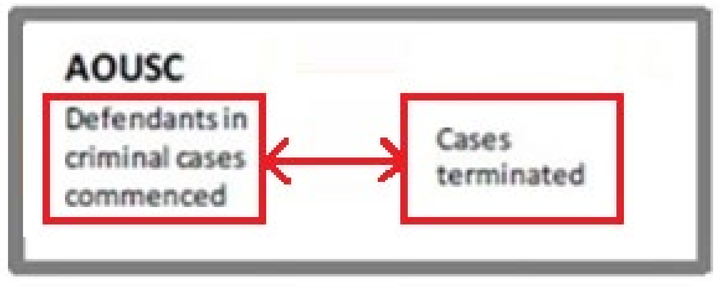
What are the resulting sentencing outcomes:

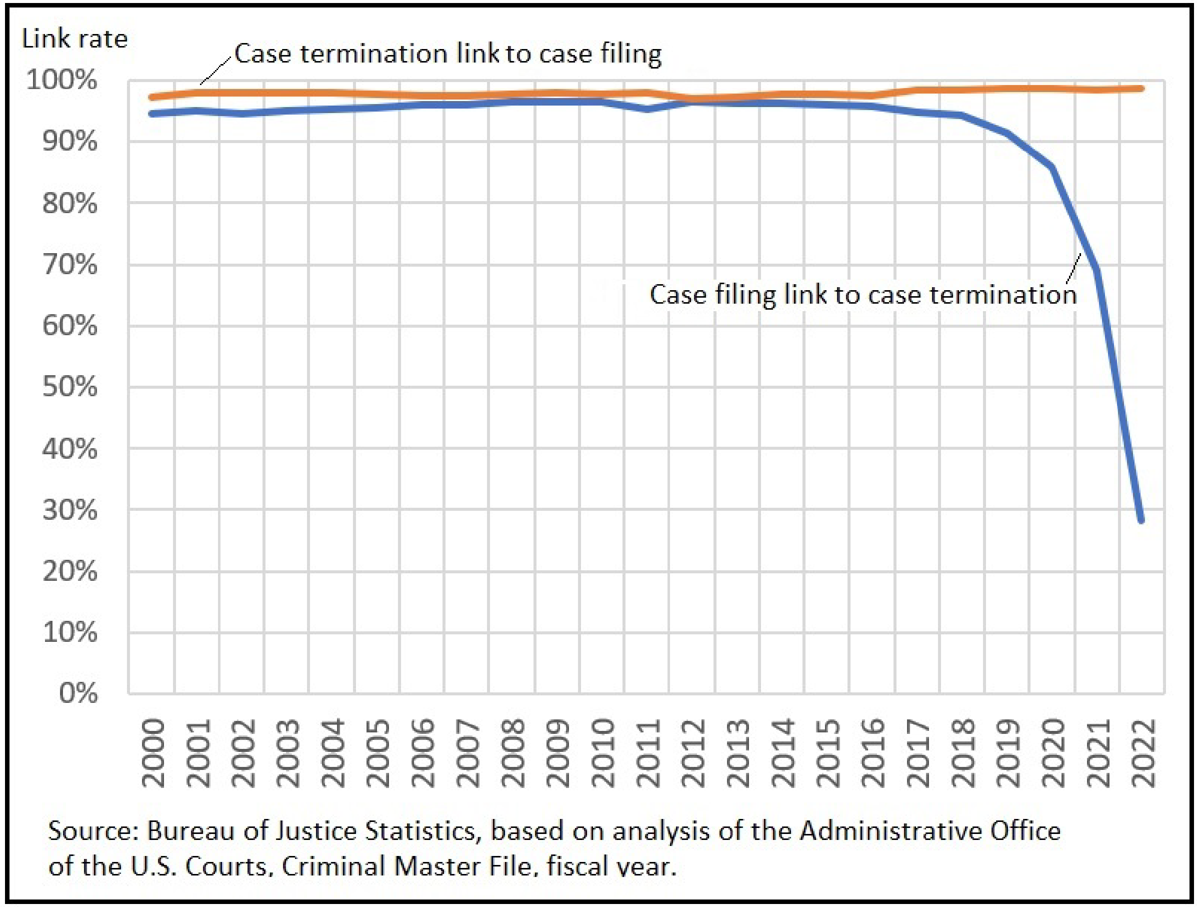
- Type of Sentence
- Length of Sentence

Prosecutor data → *Judiciary data* → *Prison system data*

Example 1: Intra-agency links join records within the same agency

- Permits tracking a cohort within an agency over time
- Censoring occurs when the time required for the case processing event to occur exceeds the available observation period



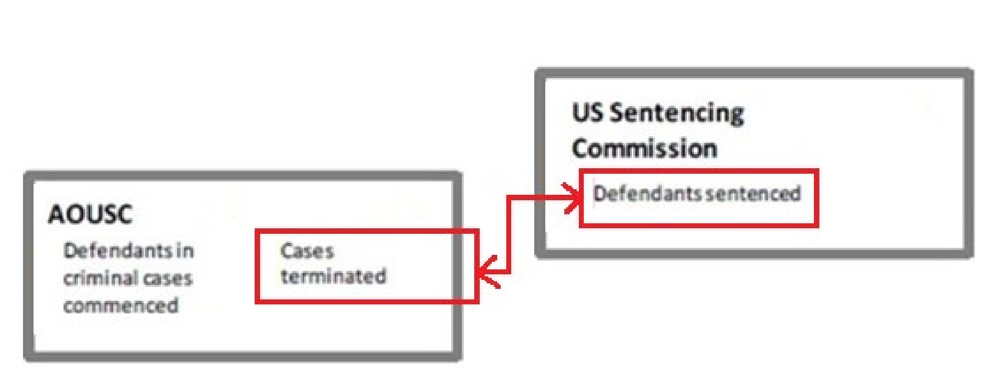


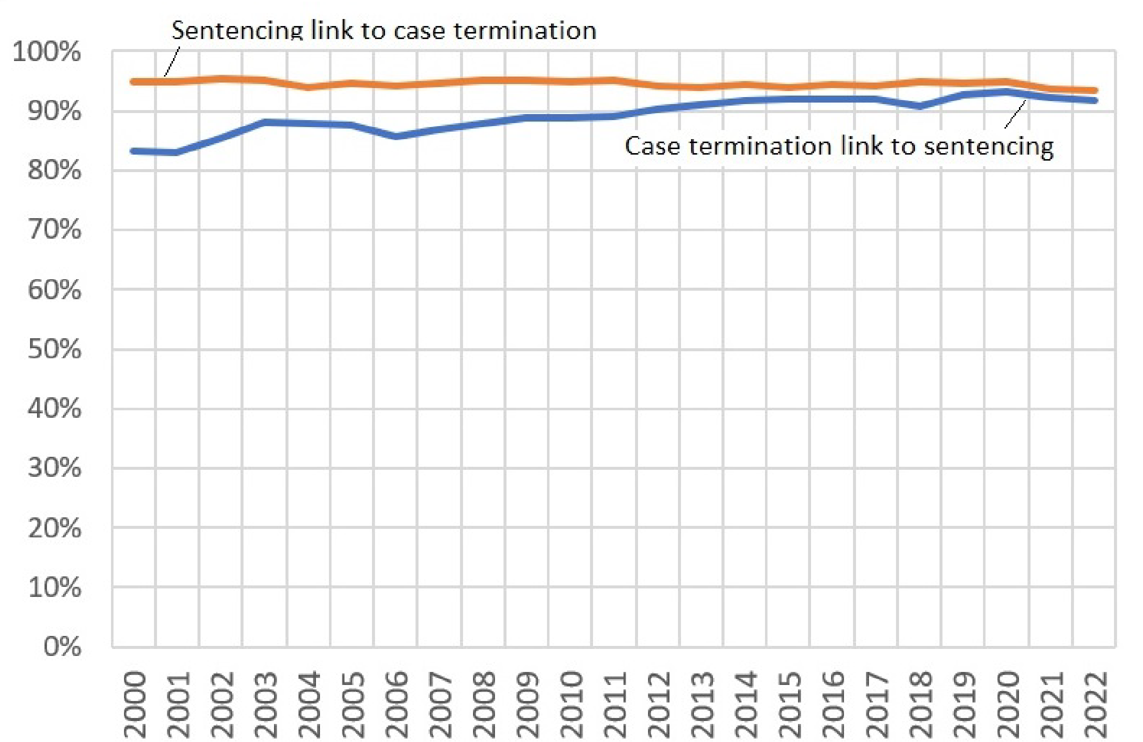
Source: Bureau of Justice Statistics, based on analysis of the Administrative Office of the U.S. Courts, Criminal Master File, fiscal year.

Intra-agency linked example: *Percent of records that link between cases filed (AOUSC) and cases terminated (AOUSC), FY 2000-2022.*

Example 2: Inter-agency links join records between two agencies

- Permits tracking a cohort across two agencies over time
- Maximizes use of shared identifiers between agencies





Source: Bureau of Justice Statistics, based on analysis of the Administrative Office of the U.S. Courts, Criminal Master File and U.S. Sentencing Commission, Monitoring File, fiscal year.

Inter-agency link example*: Percent of records that link between cases terminated (AOUSC) and cases sentenced (USSC), FY 2000-2022.*

Next Steps

- Work with agencies to standardize a “best practice” for creating and testing linked datasets

- Redesign documentation that demonstrate linking rates

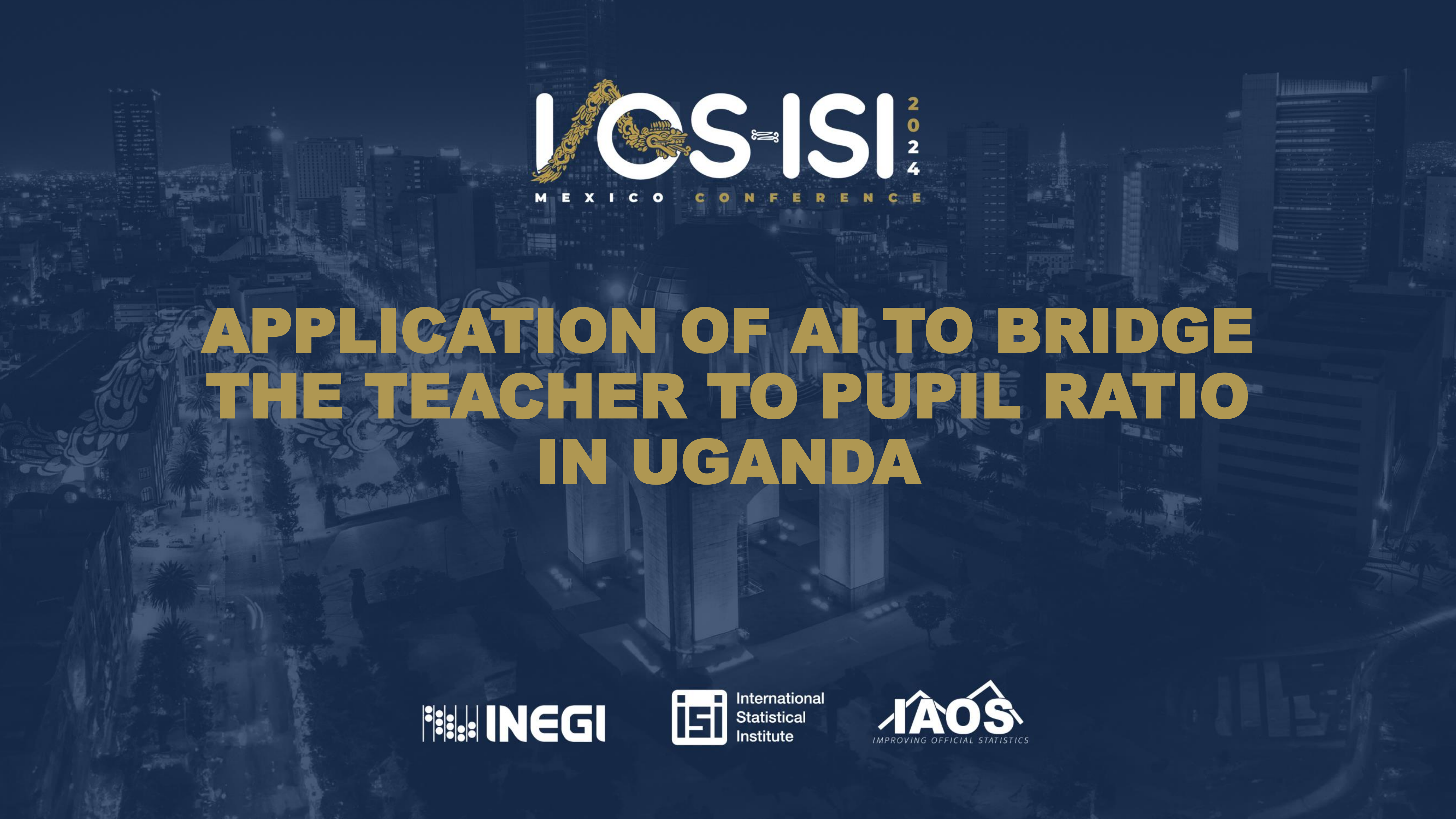
- Incorporate linked statistics in data tool

- Assess and create additional link files as necessary

Thank you

Email: Mark.Motivans@USDOJ.GOV

Acknowledgements: These data are part of a series designed by the Urban Institute (Washington, DC) and the Bureau of Justice Statistics. Data for 2013 and on were prepared by Abt Global (Rockville, MD).



IAOS-ISI 2024

MEXICO CONFERENCE

APPLICATION OF AI TO BRIDGE THE TEACHER TO PUPIL RATIO IN UGANDA

INEGI

International Statistical Institute

IAOS IMPROVING OFFICIAL STATISTICS

AUTHORS:

Alban Manishimwe, Chris Ndatira Mukiza, Edgar Mfite Niyimpa

Uganda Bureau of Statistics

CONTENTS

- □ INTRODUCTION
- □ METHODOLOGY
- □ RESULTS
- □ DISCUSSION
- □ CONCLUSION

INTRODUCTION

- The pupil-to-teacher ratio is defined as the average number of pupils per teacher(UNESCO).

- According to the United Nations, Education, Scientific and Cultural Organization(UNESCO), there is an urgent call to massively recruit about 69 million teachers globally to achieve Sustainable Development Goal 4(SDG-4) (UNESCO, n.d.). This massive recruitment is imperative to fill the existing shortage of teachers globally.

DEFINITION

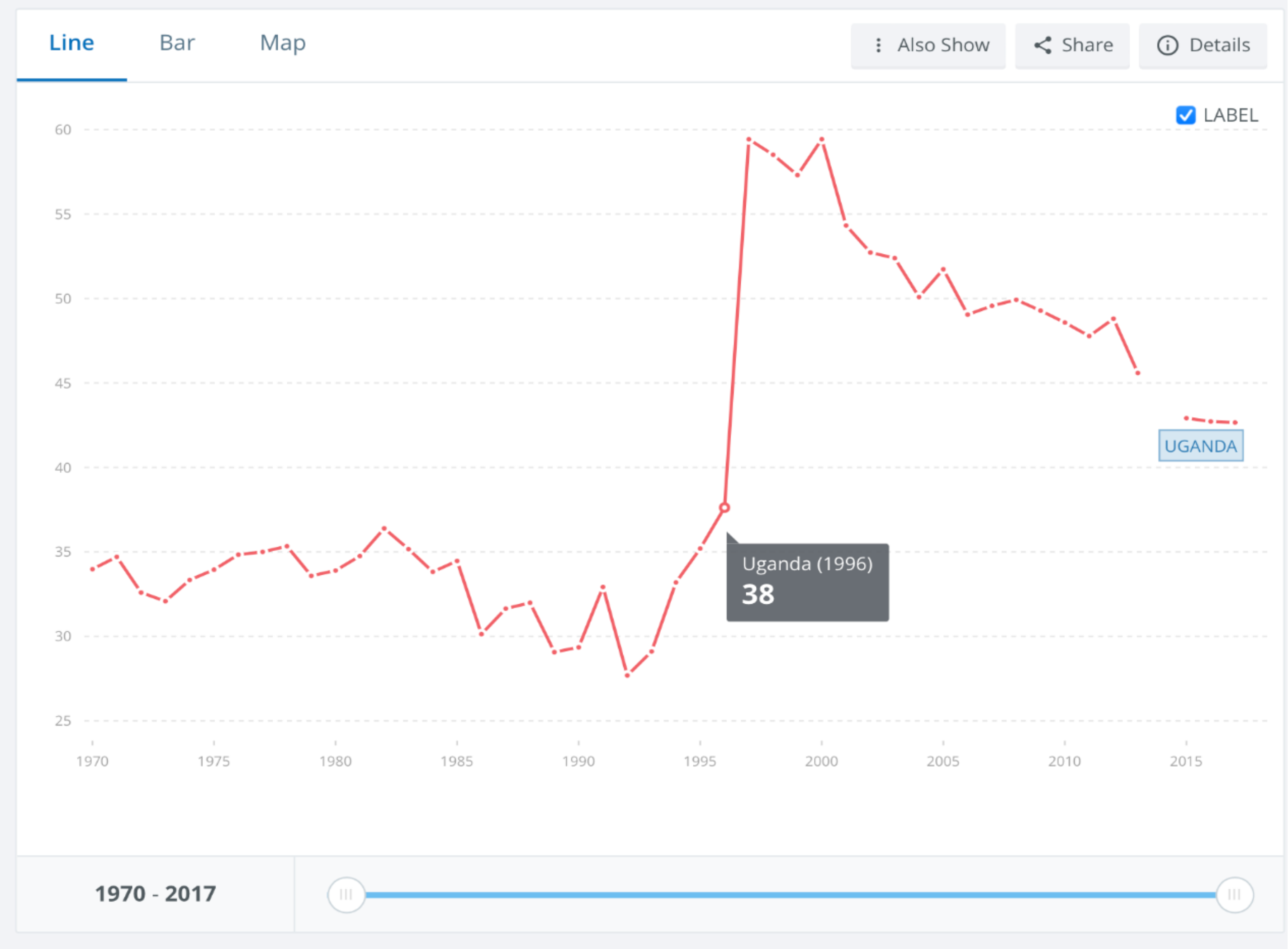
- Human capital development is crucial for any economy to achieve sustainable development; The key input and determinant of human capital development, particularly in economies looking to move toward upper middle-income status – is access to quality education. Education can equip a national workforce with skills, knowledge, and creativity to compete in the knowledge-based global economy(Runde et al., 2017).
- A study done (by Solheim & Opheim, 2019) reducing teacher–pupil ratio boosts academic excellence.
- Moreover, this study elucidates that students excel; when teachers differentiate material for each student's zone of proximal development, provide frequent formative feedback, and build close relationships; this is possible when the gap between teacher-pupil ratio is narrowed.

EDUCATION SECTOR IN UGANDA

- The government of Uganda has implemented several policies and programs to develop the education sector, which is responsible for human capital transformation. One such policy is the Universal Primary Education Policy in 1997, which seeks to improve the literacy and enrolment rates for elementary education, and accordingly, the gross enrolment in primary schools increased from a total of 3.1 million in 1996 to 5.3 million in 1997, an increase of 73 percent in one year(Wabwire, 2022).

- Uganda's education system is comprised of an early childhood program that caters for children aged 3-5 years (pre-primary education), followed by seven (7) years of primary education, followed by four (4) years of Ordinary (O) Level secondary education, two (2) years of Advanced (A) Level secondary education and the final tier is three (3) to five (5) years of Tertiary education(UBOS,2022)

TEACHER TO PUPIL RATIO IN UGANDA



The Pupil Teacher Ratio (PTR) has remained constant, though, at 43 pupils per teacher since 2015(Uganda Bureau of Statistics, 2022)

EDUCATION SECTOR IN UGANDA

- The government of Uganda has implemented several policies and programs to develop the education sector, which is responsible for human capital transformation. One such policy is the Universal Primary Education Policy in 1997, which seeks to improve the literacy and enrolment rates for elementary education, and accordingly, the gross enrolment in primary schools increased from a total of 3.1 million in 1996 to 5.3 million in 1997, an increase of 73 percent in one year(Wabwire, 2022).
- Uganda's education system is comprised of an early childhood program that caters for children aged 3-5 years (pre-primary education), followed by seven (7) years of primary education, followed by four (4) years of Ordinary (O) Level secondary education, two (2) years of Advanced (A) Level secondary education and the final tier is three (3) to five (5) years of Tertiary education(UBOS,2022)

EDUCATION AND ARTIFICIAL INTELLIGENCE

- Artificial Intelligence in Education has been mostly used for the last 40 years(Vincent A.W.M.M. Aleven & Kenneth R. Koedinger, 2002).

The main three applications of AI in Education are:

I. Intelligent tutoring systems that track student progress, difficulties, and errors, going through structured subject content to provide feedback and adjust the level of difficulty to create an optimal learning path;

II. Support writing assignments and, conversely, automate the assessments of writing assignments, including identifying plagiarism and other forms of cheating

III. Immersive learning experiences and games.

For this research, Artificial Intelligence was applied to work with the policymakers to distribute and allocate teachers based on demography and expertise factors

METHODOLOGY

- The development of this solution followed agile method of software engineering. Agile is a methodology where continuous iterations and testing take place during the entire Software Development Life Cycle (SDLC) of a product(Srivastava et al., 2017), this choice was based on the fact that Scrum which is an approach for Agile methodology is the mostly used method of software development(Cobb, 2015).

- Scrum was designed to increase the speed of development, align individual and organizations' mottos, define a culture focusing on performance, support shareholder value creation, have good communication of performance at all levels, and improve individual development and quality of life(Srivastava et al., 2017)

OUR IMPLEMENTATION FOR SCRUM IMPLEMENTATION

- The workflow of scrum consists of the Scrum Master, Product Owner, and the scrum team who work together to continuously iterate and evolve the product.
- Our AI solution underwent 2 iterations for 6 weeks, with each scrum cycle lasting 3 weeks, which is the acceptable scrum cycle(Srivastava et al., 2017).

FUNCTIONAL AND NON FUNCTIONAL REQUIREMENTS

- The functional requirements for this study are:
  - Generating a map of Schools in Uganda with different pupil-to-teacher ratios.
  - The features contributing to the high pupil-to-teacher ratio in Uganda.
  - Developing an AI classifier to distribute or predict the total number of teachers given required for a school, given the number of pupils and different demographic indicators as well.
- The non-functional requirements for this study are:
  - Scalability: -The ability of a software or system to perform well given the expanding environment or workload.
  - Usability: - The ability of a software or system to satisfy user needs with effectiveness and efficiency.
  - Accessibility: -The ability of software or system to reach as many people as possible.

DATASETS DESCRIPTION

We used two datasets namely:

- Dataset on pupil enrolment with data from 2010-2015, covering 117 districts.
- Dataset on teachers with data from 2010 – 2015 teaching in the districts as described in the pupil enrolment dataset.

The dataset on pupil enrolment consists of 11,703 data examples of pupils from primary one (P.1) to primary seven (P.7).

The dataset on teachers consists of 11,703 data examples of teachers from the same district as the pupils

DATA PRE-PROCESSING

- This research utilized normalization which involves scaling data to a common range using min-max normalization.
- For the categorical data (the type of school feature, regions, and districts) the researchers applied one hot encoding to transform the values into a format that the Machine learning algorithm can understand.
- The pandas.get_dummies function which is from the pandas library API (Application Programming Interface), which was used in this case, is used to convert categorical variables into dummy/indicator variables.

MODEL DEVELOPMENT

- The model was developed after a comprehensive analysis of the most used models which are Random Forest Regression, Linear regression, Decision trees, and K-Nearest Neighbors. The dataset was split into training (80%), and testing (20%). After the splitting, the features were scaled using a standard Scaler.
- After training the models, we evaluated their performances using Mean Squared Error, Root Mean Squared Error and R2 Score.

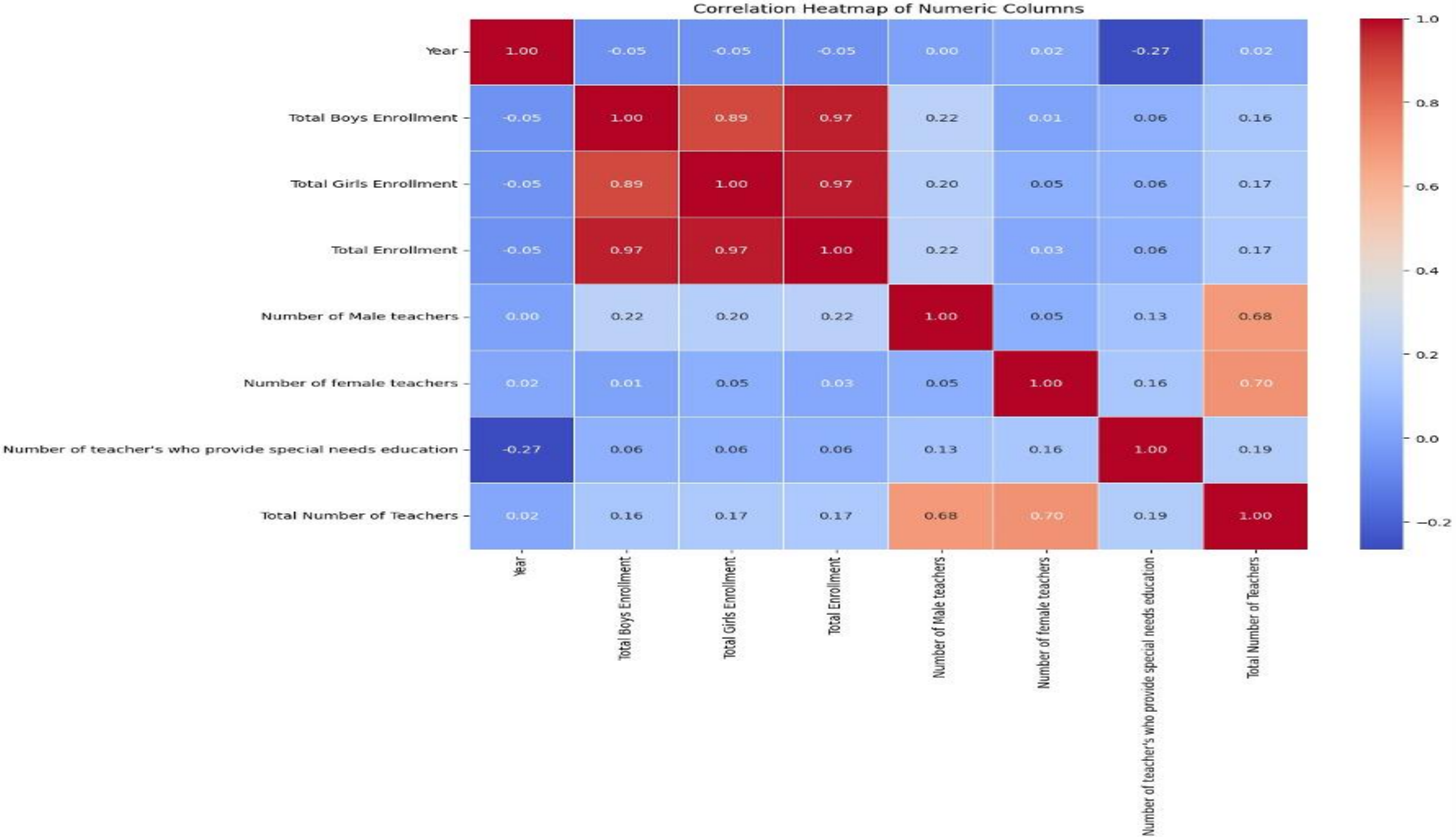
FLASK APPLICATION

- A flask application was developed as a server interface for Application Programming Interface (API) linkage, the selection of flask was based on its popularity and ability to make core functionality simple but extensive in terms of development(Neema Mduma et al., 2019)

RESULTS

- FEATURE ENGINEERING:

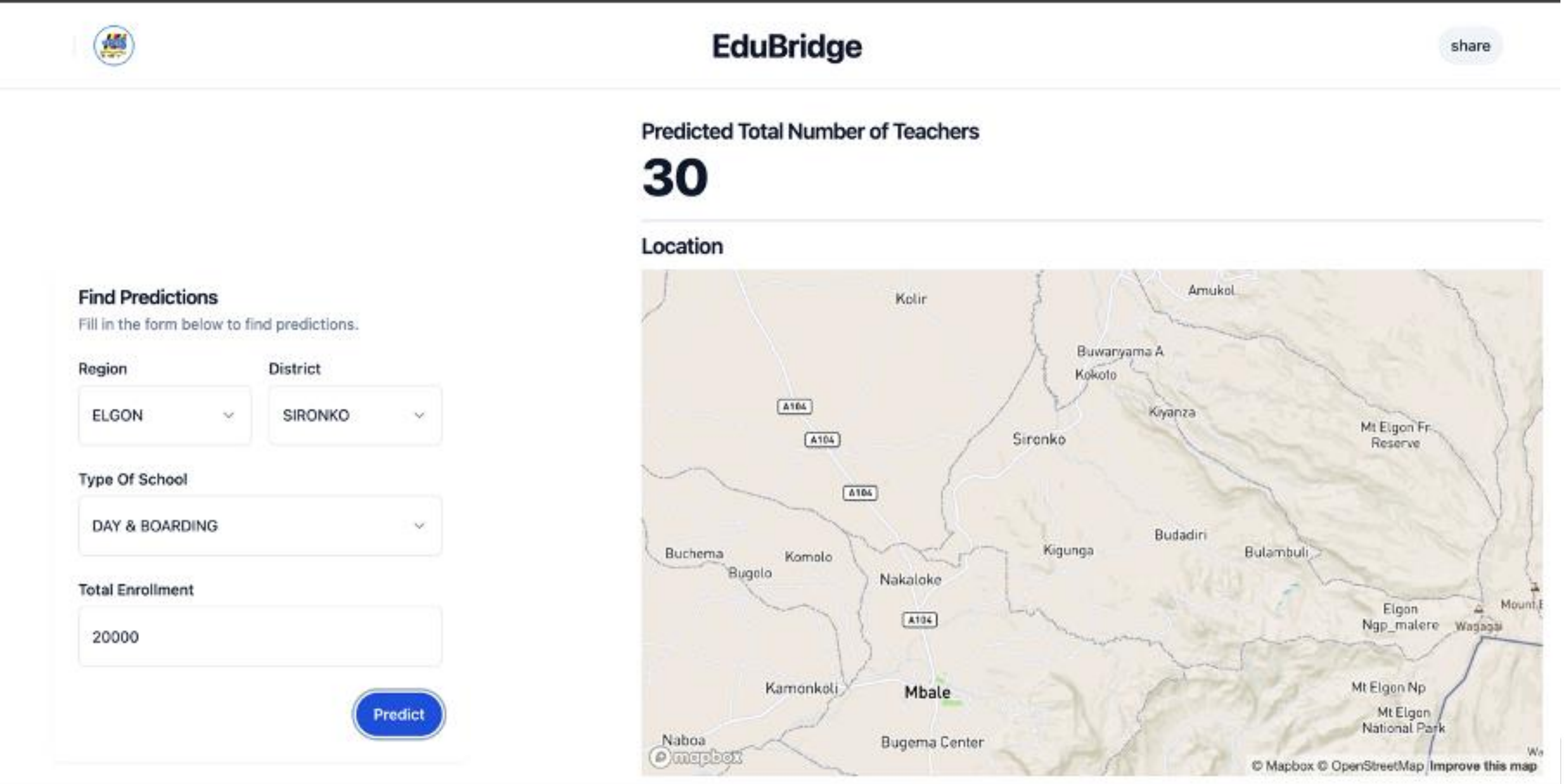
The results demonstrated in the figure below indicate that there is a positive correlation between the total number of teachers and other features in the dataset: the number of teachers who provide special needs education (0.19); the number of female teachers (0.70); the number of male teachers (0.68); total enrolment (0.17); total girls enrolment (0.17); total boys enrolment (0.16); year (0.02)



	Year	Total Boys Enrollment	Total Girls Enrollment	Total Enrollment	Number of Male teachers	Number of female teachers	Number of teacher's who provide special needs education	Total Number of Teachers
Year	1.00	-0.05	-0.05	-0.05	0.00	0.02	-0.27	0.02
Total Boys Enrollment	-0.05	1.00	0.89	0.97	0.22	0.01	0.06	0.16
Total Girls Enrollment	-0.05	0.89	1.00	0.97	0.20	0.05	0.06	0.17
Total Enrollment	-0.05	0.97	0.97	1.00	0.22	0.03	0.06	0.17
Number of Male teachers	0.00	0.22	0.20	0.22	1.00	0.05	0.13	0.68
Number of female teachers	0.02	0.01	0.05	0.03	0.05	1.00	0.16	0.70
Number of teacher's who provide special needs education	-0.27	0.06	0.06	0.06	0.13	0.16	1.00	0.19
Total Number of Teachers	0.02	0.16	0.17	0.17	0.68	0.70	0.19	1.00

TEACHER DISTRIBUTION INTERFACE

- The deployed AI interface allows a policymaker to automatically allocate and re-distribute new and existing labor while policymakers maintain their responsibility for planning, communication, and coordination. From the interface, the policymaker selects: the region; district; type of school; and total enrolment for the school, and then the Artificial Intelligence model can select the exact location for the school and the number of teachers that are required for that particular school



EduBridge

Predicted Total Number of Teachers

30

Location

Find Predictions

Fill in the form below to find predictions.

Region: ELGON

District: SIRONKO

Type Of School: DAY & BOARDING

Total Enrollment: 20000

MODEL EVALUATION

Model	Train MSE	Train RMSE	Train R2	Test MSE	Test RMSE	Test R2
Random Forest	0.06345315	0.25189909559484003	0.9974348207464513	0.17626439395831173	0.4198385331985521	0.9929137824521647
Linear Regression	1.9538743296140562	1.3978105485415597	0.9210119978305032	2.1001265160725113	1.4491813261536706	0.9155702803233934
KNN Regressor	1.9562433249867621	1.3986576868507756	0.92091623	4.606038848950994	2.1461684111343624	0.8148270754831681
Decision Tree	0.06096345	0.24690778139191175	0.99753547	0.2516167797465886	0.5016141741882785	0.9898844502855648

DISCUSSION AND CONCULSION

- An Artificial Intelligence Model-based prototype has been developed and deployed for this study to establish the location of the required school, and the total number of teachers that would be required for a particular school to bridge the pupil-to-teacher ratio in Uganda.
- The developed system whose requirements have been established in this paper narrates the futurist approach to policy-making in the education sector in Uganda. By taking advantage of the digital Census that the Uganda Bureau of Statistics is to undertake in April 2024, the findings of the census should be incorporated for the next iteration of the modelling, this will ensure wide coverage of schools in the Country

ACKNOWLEDGMENT

The authors would like to thank the Uganda Bureau of Statistics for supporting this work in every way possible.

REFERENCES

1. Cedrick Joseph Wabwire. (2022). What's with overwhelming pupil numbers in schools. *The Daily Monitor*.
2. Cobb, C. G. ,. (2015). The project manager's guide to mastering Agile: Principles and practices for an adaptive approach. *John Wiley & Sons*.
3. Glinz, M. (2007). On Non-Functional Requirements. *15th IEEE International Requirements Engineering Conference (RE 2007)*, 21–26. https://doi.org/10.1109/RE.2007.45
4. Neema Mduma, Khamisi Kalegele, & Dina Machuve. (2019). An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools. *Journal of Information Systems Engineering & Management*.
5. Obaid, H. S., Dheyab, S. A., & Sabry, S. S. (2019). The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning. *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, 279–283. https://doi.org/10.1109/IEMECONX.2019.8877011
6. Pandas. (2024). *Pandas API documentation*.
7. Runde, D. F., Rice, C., & Yayboke, E. (2017). Education and Human Capital Development. In *Innovation-Led Economic Growth* (Transforming Tomorrow's Developing Economies through Technology and Innovation). Center for Strategic and International Studies (CSIS). http://www.jstor.org/stable/resrep23182.6
8. Solheim, O. J., & Opheim, V. (2019). Beyond class size reduction: Towards more flexible ways of implementing a reduced pupil–teacher ratio. *International Journal of Educational Research*, *96*, 146–153. https://doi.org/https://doi.org/10.1016/j.ijer.2018.10.008
9. Srivastava, A., Bhardwaj, S., & Saraswat, S. (2017). SCRUM model for agile methodology. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 864–869. https://doi.org/10.1109/CCAA.2017.8229928
10. Uganda Bureau of Statistics. (2022). *Statistical Abstract*.
11. UNESCO. (n.d.). UIS FACT SHEET OCTOBER 2016, No. 39. In *THE WORLD NEEDS ALMOST 69 MILLION NEW TEACHERS TO REACH THE 2030 EDUCATION GOALS*.
12. UNESCO. (2023). *Global Education Monitoring Report*.
13. Vincent A.W.M.M. Aleven, & Kenneth R. Koedinger. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science: A Multi Disciplinary Journal*.

IAOS-ISI 2024
MEXICO CONFERENCE

Thank you

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS