Scaling Science Based Solutions for Sustainable Development-Regional Innovations in Technology and Data for Achieving SDGs

Big Data, Artificial Intelligence and Official Statistics

Side Event 28.02.2025(Meeting Room G)

Shailja Sharma –Director(SIAP)

UNITED NATIONS
SIAP
Statistical Institute for Asia and the Pacific

What is Official Statistics

- Data and information collected, analyzed, and published by government agencies (National Statistical Offices, Ministries Departments)
- other authorized organizations
- provides a reliable picture of a country's economy, demographics, society, and environment,
- used for policymaking and research purposes;

Big Data and Artificial intelligence for Official Statistics

Big data

Large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

Artificial Intellligence

A machine based system that uses Input to deduce output. Basically intelligence demonstrated by machines after they learn.

Applications of AI in Official Statistics

Data Collection

- AI is generating Big Data using technologies as Web Technologies, Remote sensing etc.

- AI is Assisting statistical organizations in leveraging new data sources as social media data to gather additional insights supplementing Official Statistics.

➢ quality of OS is improving by increased access to Data, improved Data collection and more accurate insights.

Applications of AI in Official Statistics

Data Analysis

- AI can significantly assist in data analysis by automating repetitive tasks, identifying patterns and trends within large datasets, providing insights through predictive analytics,

- detecting anomalies,

➢ enabling better decision-making by processing complex information faster than humans, all through techniques like machine learning, automation of tasks and enhanced data visualization.

Artificial intelligence in support of SDG13

Environment statistics:

Semantics (shared vocabulary among different data sets) and machine reasoning is used to facilitate the processing of satellite imagery and other additional layers of data for example:

- Track changes in land cover
- Measuring extent of ecosystems and changes over time
- Measuring extent of wetlands and other critical water-related ecosystem

Artificial intelligence in support of SDG13

Climate change:

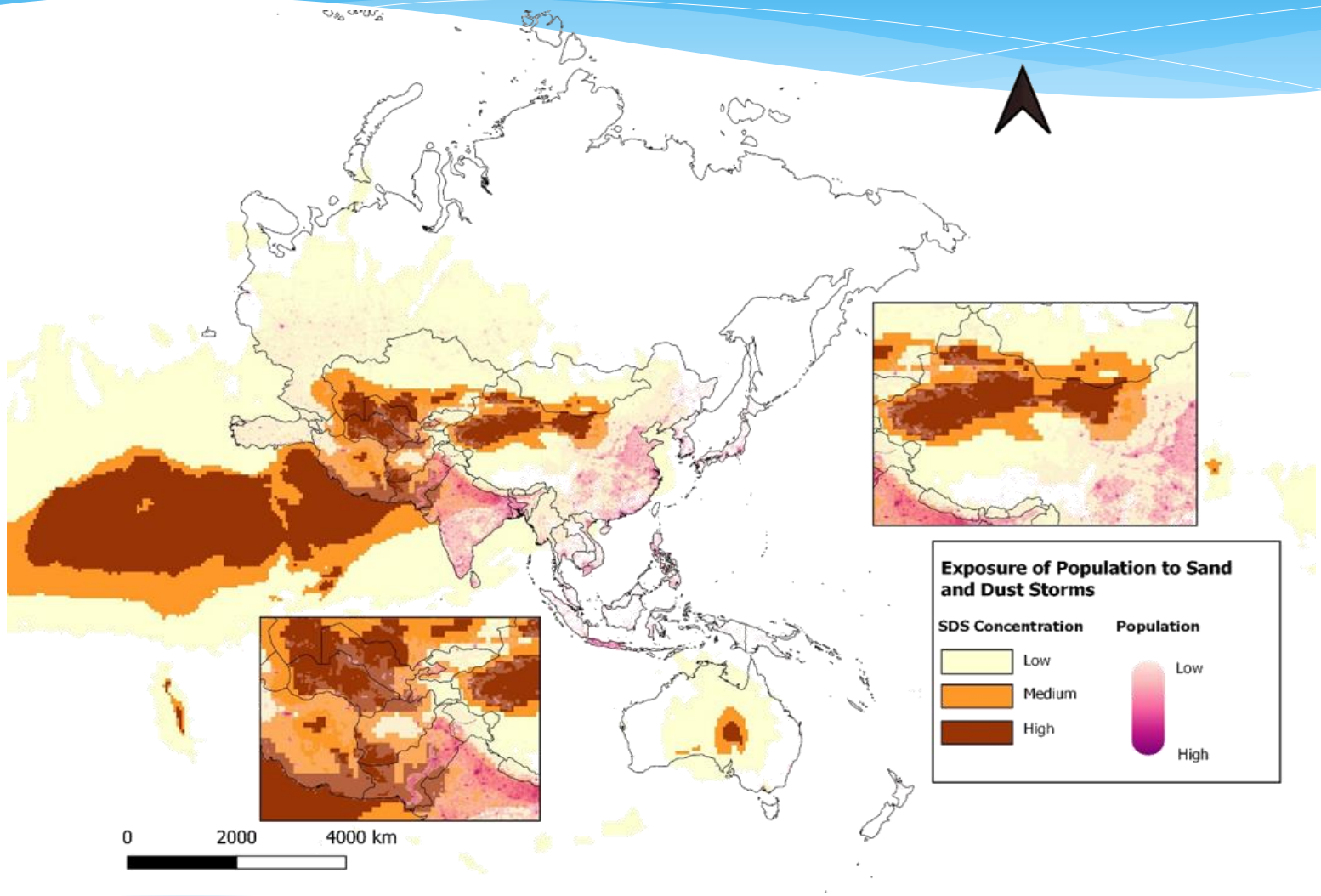
- Estimation of carbon stocks and carbon cycle
- Estimation of carbon sequestration (process of capturing and storing Atmospheric carbon dioxide)
- Estimation of areas that are vulnerable to impacts from climate change (related also to disaster risk reduction)
- Forest related information (key ecosystem to carbon sequestration)
- Changes in the services provided by ecosystems as a result of climate change (e.g. reduced water supply)

Artificial intelligence in support of SDG13

Disaster risk reduction

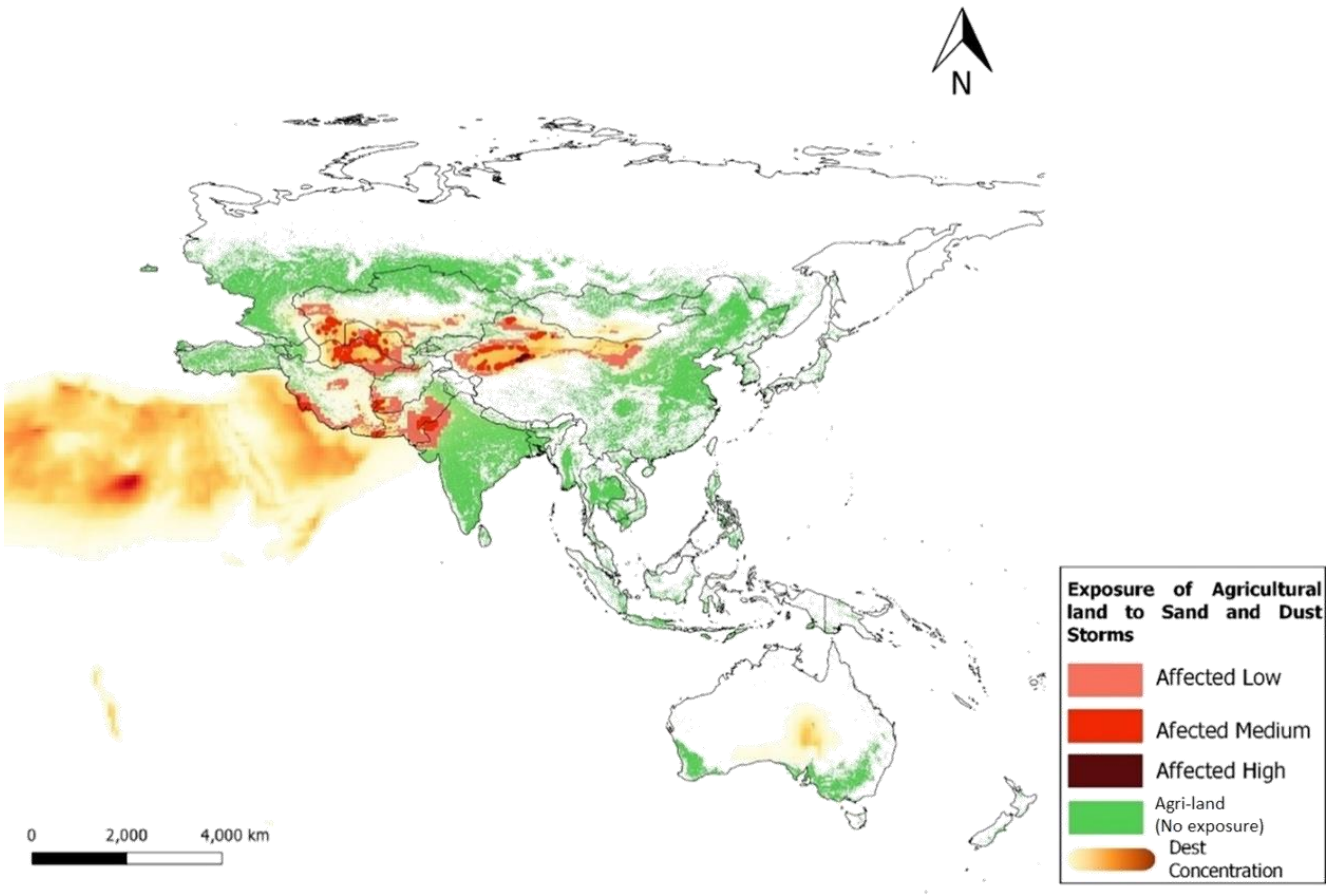
- Identification of vulnerable areas (e.g. vulnerability maps for certain types of risks such as floods, sandstorms, etc.)
- Geospatial data analysis in support of post disaster recovery (e.g. identifying impacted areas)

Exposure of Population to Sand and Dust Storms



Source: APDIM

Exposure of Agricultural land to Sand and Dust Storms



Source: APDIM

Is child marriage linked to environmental factors? SDG5

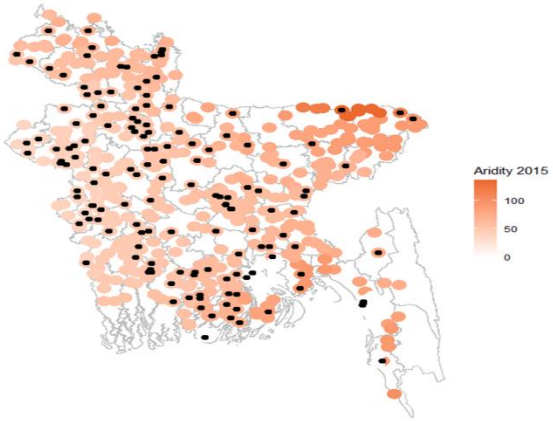
Child marriage is linked to several known factors (education, wealth) but also to (exogenous) environmental factors.

How:

- From household survey, one can identify geospatial attributes (location) ← **DHS**
- From satellite images, one can integrate local environmental factors ← **Big Data**

→Matching data sets at a very small spatial unit
→Use logistic and ML models to **assess the role of environmental factors**.

Results
High Aridity Index
Linked to child marriage



Can web search and AI be used for collecting and analyzing femicide? SDG5

Natural Language Processing (NLP) can help monitor trends and circumstances that may lead to femicide

How:

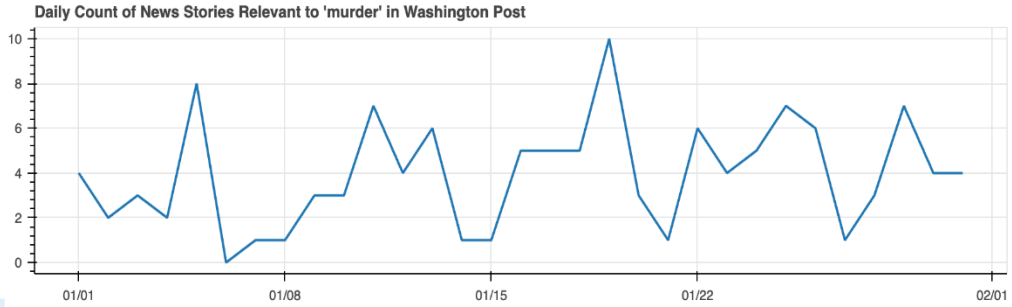
- Collecting feminicides data from news media ← **Big Data** (using API)
- Analyze huge amount of text and news using pre-trained NLP models ← **AI**

→ Use Machine Learning (AI) models to accurately **identify femicides**.

Results:

- Multi dimensional collection of feminicides over time (circumstances, types, locations, ..)
- Collection of non- or mis-reported cases.
- Better analysis of trends and factors

Source:UNWomen



Does online VAW increase in times of crisis? SDG5

Use of big data from online searches and social media posts provide important insights on factors (crisis, disasters) linked with online VAW-related posts

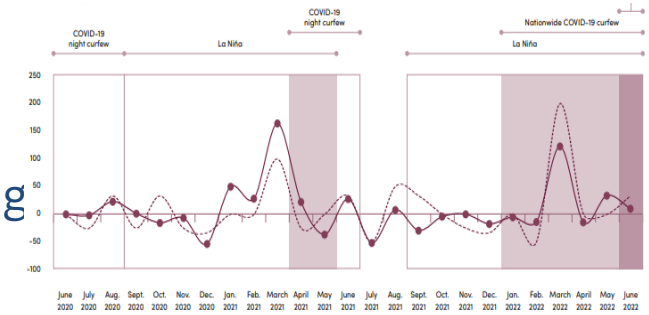
How:

- Keyword search from Google searches, social media posts ← **Big Data**
- Identification with related events over time (crisis, disasters) ← **Analytics**

→Clustering of keywords and typology of VAW
→**Identify VAW** on social medias – several countries

Results:

- An increase in VAW-related terms was noticed following crises (curfews, floods, landslides) across all countries.



Note: Darker shades of maroon refer to overlapping crises. The darker the shade, the higher the number of overlapping crises.

Source: UNWomen

Are the Level of infrastructure development and adolescent birth rates related ? (ongoing work) SDG5

In Cambodia, inadequate access to education and infrastructures may be linked to high adolescent birth rates

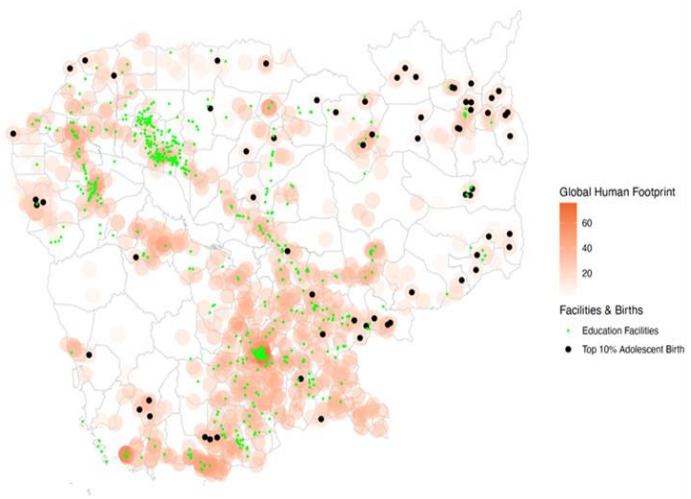
How:

- From household survey, one can identify geospatial attributes (location) ← **DHS**
- From satellite images, one can integrate urbanization levels ← **Big Data**

→ Spatial identifiers allows spatial linkages
→ Use spatial models to **assess the impact of levels of urbanization**.

Results:

- Adolescent birth rate and degree of urbanization are (spatially) **negatively correlated**



- Source: UNWomen

Sustained and inclusive Economic Growth SDG8

Use of mobile positioning data (MPD) for tourism statistics – INDONESIA

Passive MPD

- Mobile network operators have information on the call detail records and location-based advertising/signaling which are then converted to coordinate points and location administrative unit. An algorithm is developed to transform the transaction data between the Base Tranceiver Station (BTS) and customer fixed point location (staypoint) of an anonymized customer in order to track the customer movement.

Determination of a Trip

The movement is then classified as domestic tourism, commuter, circular, or other, consistent with internationally accepted definition of usual environment and travel.

Sources: The use of mobile positioning data to capture tourism data in Indonesia, BPS Domestic Tourism Statistics 2023)

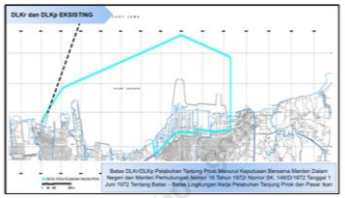
Sustained and inclusive Economic Growth SDG8

Utilization of AIS data for Transportation Statistics

BPS Statistics Indonesia

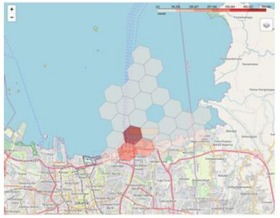
Automatic Identification System

AIS is an innovative source of data on maritime traffic. BPS Statistics Indonesia has explored how it can use the AIS to estimate and analyze traffic statistics and the time spent of ships at Indonesian ports.



Source: Ministerial Decree No. 16 of 1972

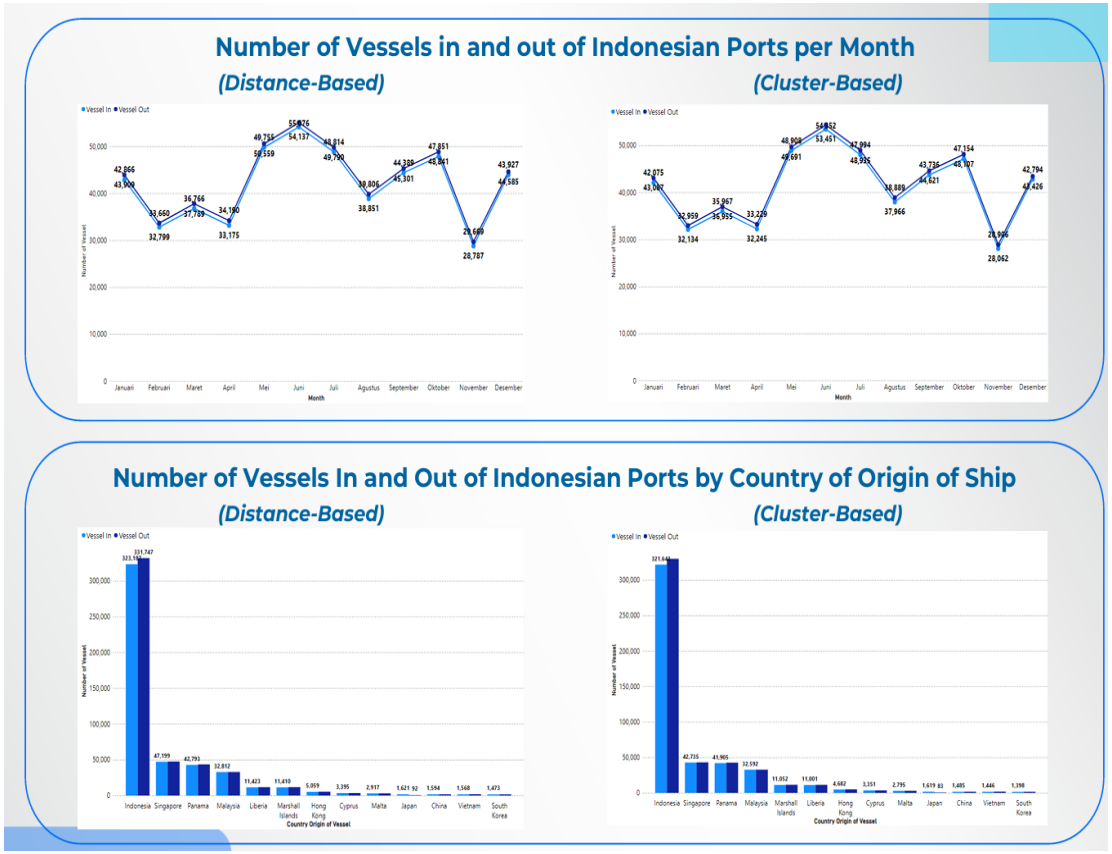
Source: SIRI, Samudera Indonesia



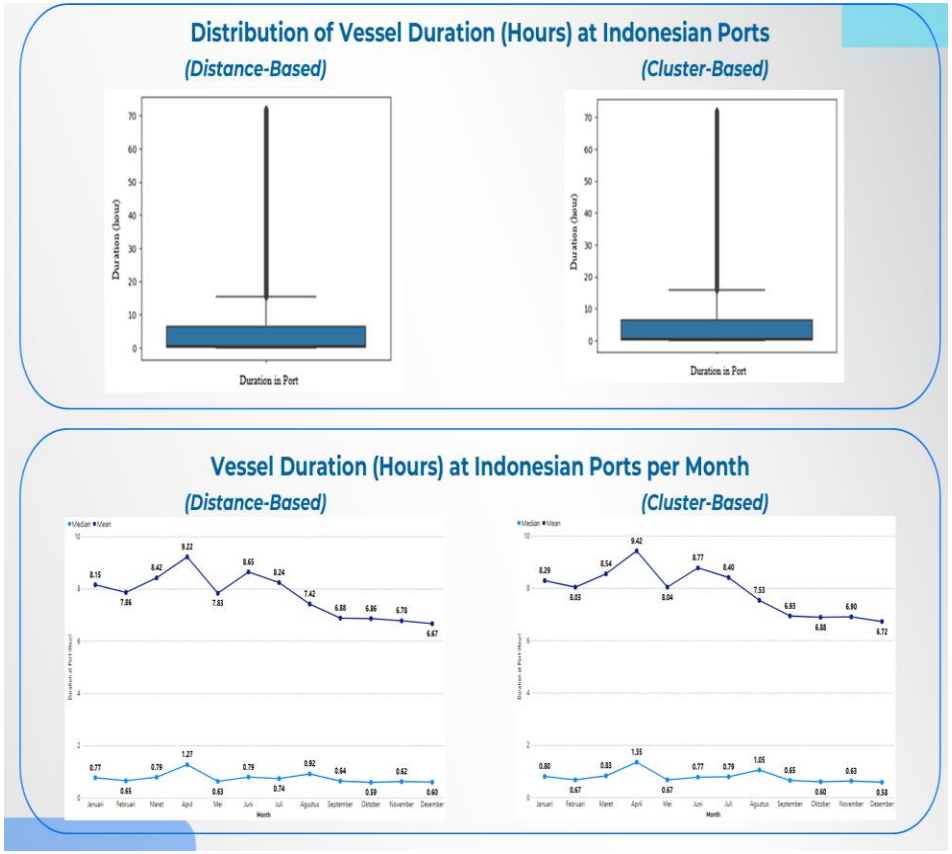
Source: BPS Statistics Indonesia, 2024

Source: BPS Statistics Indonesia, 2024

Graphics


Number of Vessels in and out of Indonesian Ports per Month
(Distance-Based)
(Cluster-Based)
Vessel In
Vessel Out
Number of Vessel
Month
Januari
Februari
Maret
April
Mei
Juni
Juli
Agustus
September
Oktober
November
Desember
Number of Vessels In and Out of Indonesian Ports by Country of Origin of Ship
(Distance-Based)
(Cluster-Based)
Indonesia
Singapore
Panama
Malaysia
Liberia
Marshall Islands
Hong Kong
Cyprus
Malta
Japan
China
Vietnam
South Korea
Country Origin of Vessel

Graphics


Distribution of Vessel Duration (Hours) at Indonesian Ports
(Distance-Based)
(Cluster-Based)
Duration (hour)
Duration in Port
Vessel Duration (Hours) at Indonesian Ports per Month
(Distance-Based)
(Cluster-Based)
Median
Mean
Januari
Februari
Maret
April
Mei
Juni
Juli
Agustus
September
Oktober
November
Desember
Month

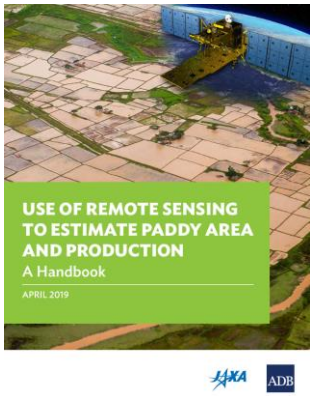
AI in support of Area Production Crop estimates

Remote sensing for agricultural statistics

Savannakhet, Lao People's Democratic Republic;
Nueva Ecija, Philippines;
Ang Thong, Thailand; and
Thai Binh, Viet Nam

From synthetic aperture radar data to production estimates

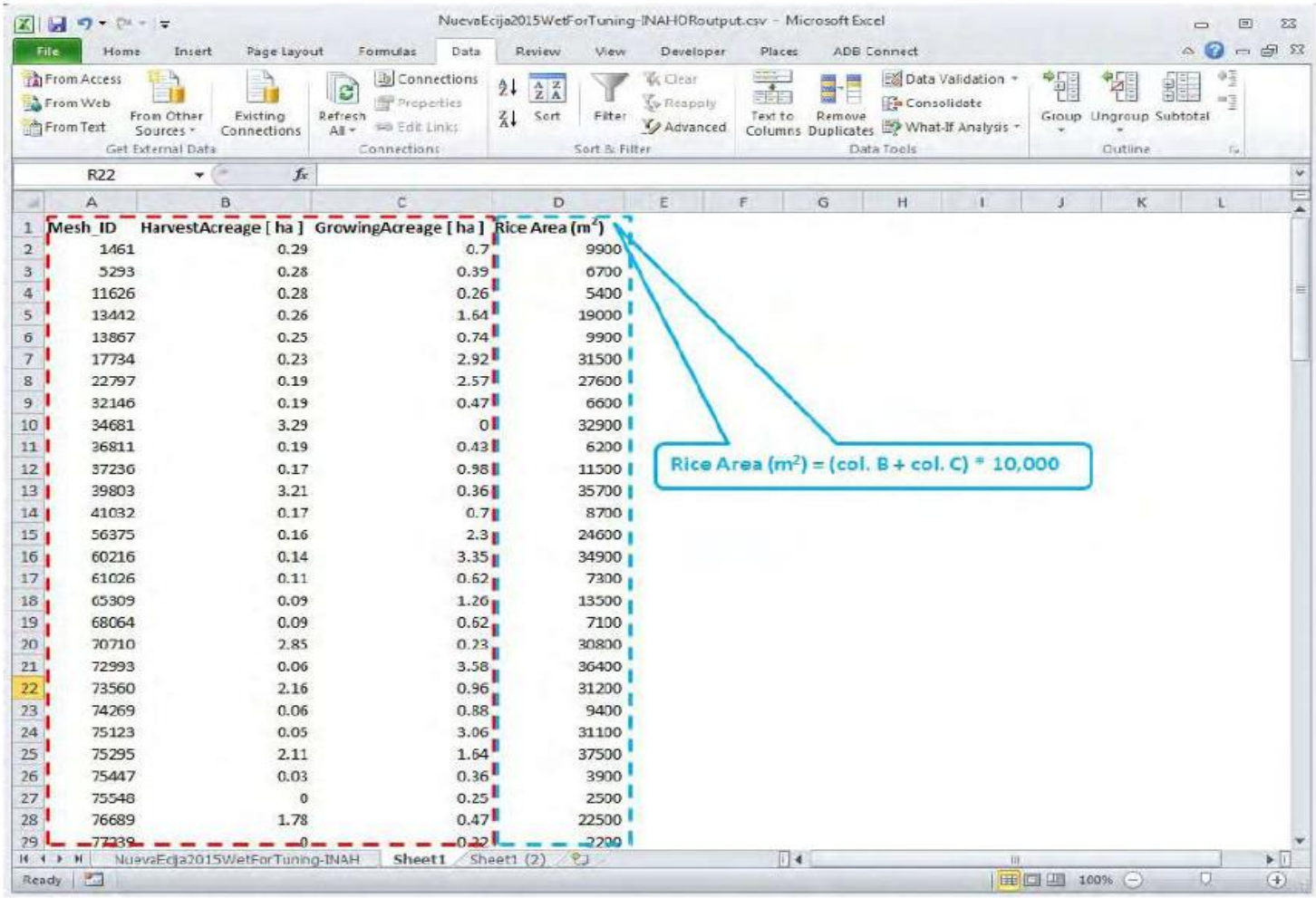
ADB implemented a project that used satellite images are used to identify and estimate paddy area using International Asian Harvest Monitoring system for Rice (INAHOR-AD) software. The Rice Crop Mapper module and Rice Production Calculator of the INAHOR-AD were used to detect rice area and to calculate rice production estimates. Ground truthing activities were conducted to validate the estimates



Source: ADB's Use of Remote Sensing to estimate paddy area and production

Table on Total Area under the crop

Figure 6.21: Calculating Total Paddy Area from INAHOR-AD Estimates



Mesh_ID	HarvestAcreage [ha]	GrowingAcreage [ha]	Rice Area (m²)
1461	0.29	0.7	9900
5293	0.28	0.39	6700
11626	0.28	0.26	5400
13442	0.26	1.64	19000
13867	0.25	0.74	9900
17734	0.23	2.92	31500
22797	0.19	2.57	27600
32146	0.19	0.47	6600
34681	3.29	0	32900
36811	0.19	0.43	6200
37236	0.17	0.98	11500
39803	3.21	0.36	35700
41032	0.17	0.7	8700
56375	0.16	2.3	24600
60216	0.14	3.35	34900
61026	0.11	0.62	7300
65309	0.09	1.26	13500
68064	0.09	0.62	7100
70710	2.85	0.23	30800
72993	0.06	3.58	36400
73560	2.16	0.96	31200
74269	0.06	0.88	9400
75123	0.05	3.06	31100
75295	2.11	1.64	37500
75447	0.03	0.36	3900
75548	0	0.25	2500
76689	1.78	0.47	22500
77339	0	0.22	2200

Source: Screen shot generated by Remote Sensing Technology Center of Japan staff. 2015. Tokyo.

Artificial intelligence

Europe's one-stop-shop for AI-ML for official statistics

- AI/ML on earth observation data, satellite imagery
- Editing focus – statistically valid and efficient editing in official statistics
- Imputation focus – statistically valid and efficient imputation in official statistics
- From text to code – Experiences and potential of the use of AI/ML for classifying and coding
- Applying ML for estimating firm-level supply chain networks
- Use of generative large language models in statistics
- Generation of synthetic data in official statistics: techniques and applications

(Source: https://unstats.un.org/bigdata/events/2025/ai-data-science/webinar1/index.cshtml#anchor1)

Grateful thanks are due to my SIAP Colleagues Sokol, Christopher and Chesca (Statistician Lecturers) for providing me the images for the presentation

UNITED NATIONS
siap
Statistical Institute for Asia and the Pacific