Food and Agriculture Organization
of the United Nations

>> FAO Statistics Division

DataLab@FAO: a new approach to fill data gaps

Carola Fabi, Senior Statistician, 27° AFCAS November 2021

The context

Crisis of traditional data collection systems

- Reduced **budgets**
- Necessity to have **timeliness** information to support decisions
- **Cultural changes**
- New **competitors** on the market
- New methods and new technologies to extract data from **unstructured sources**

FAO, end of 2019 → "Launch" of the Data Lab

Develop **geospatial tools** and tagging systems at sub-national level

Support the Hand-in-Hand Initiative (HIHI)
through detailed and recent data in 43 priority countries

Develop **text-mining tools** to extract and analyse policy documents

Promote the use of **non-official and unstructured data** in domains and geographical areas where little official data is available

Covid-19: the first challenge for the DataLab

COVID-19, February 2020

FAO Big Data tool on Covid-19 impact on food value chains

Impact analyses: collection and organization of daily information on food value chain disruptions and trade-related measures

Tweets Semantic Search: collection of tweets related to the impact of the coronavirus pandemic on the food value chains, from more than 270 newspapers accounts worldwide

News search: collection of news related to the impact of the coronavirus pandemic on the food value chains;

Daily food prices: monitor of the consumer prices of 14 food products in all countries, with updated percentage prices changes;
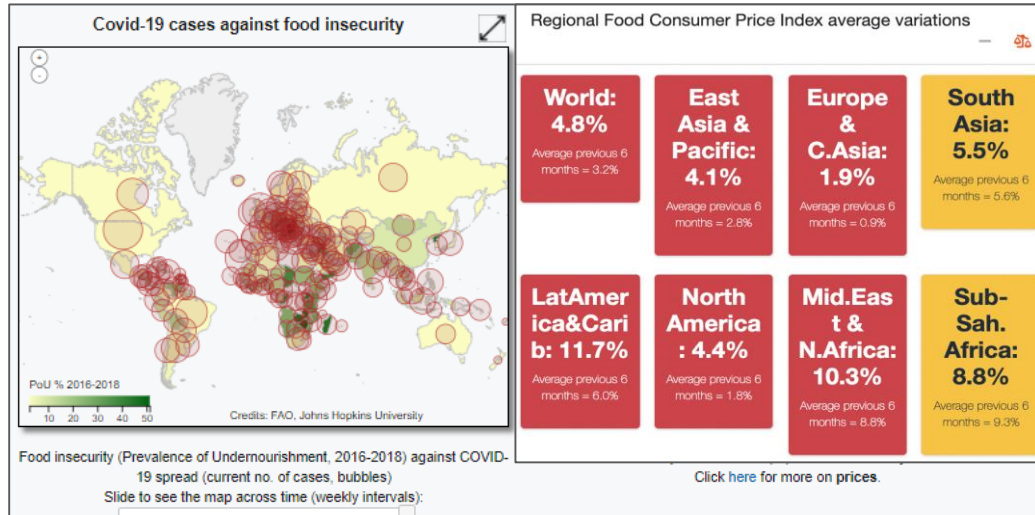
Weekly news digest: collection and summary of the most relevant news on the impact of COVID-19 on the food value chains, complemented with the results of the COVID-19 impact, food prices and tweets analyses
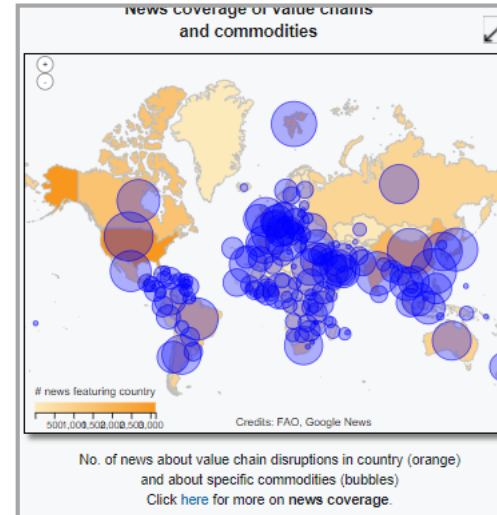
Covid-19: the first challenge for the DataLab

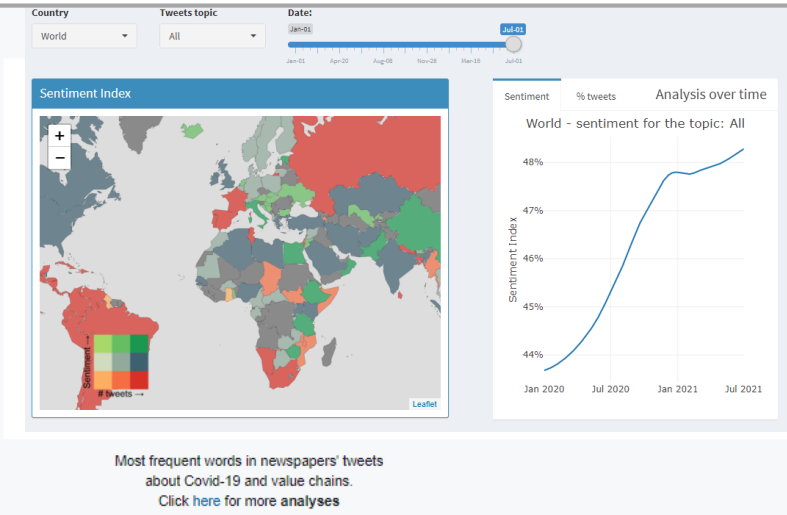FAO Big Data tool on Covid-19 impact on food value chains

COVID cases
and PoU

Nowcasting CPI's and
monitoring
Accelerations in daily
prices

Food Chain
disruptions
News distribution

Text mining tweets and
sentiment analysis

Covid-19: lesson learned from the first output

- Data sources: tested the reliability of web-scraped information and non-conventional sources
- Methods : text mining methods and models that have been generalized in other contexts
- Dissemination: interactive charts and maps, daily information in real time
- Approach: first attempt to combine different information for a more comprehensive view of a phenomenon
- Modalities: new collaborations inside the Organization or with external Entities

Relevance: thousands of unique visitors of the web site (for instance: 1,000 daily unique visitors in the first days of October 2020)

Core activities in 2021: promote the use of Big Data for official statistics and more

Non-official data, unstructured data and big data are the main sources of information of the Data Lab. They serve a purpose:

- Support the **Hand-in-Hand** Initiative
  - Scraping **sub-national** agriculture statistics (area and production) for economic analysis
  - Estimating **poverty maps** in the absence of household data
- Contribute to fill **FAOSTAT** data gaps and replace estimates with data
- Inform **Early Warning** systems
- Inform **SDG** indicator 12.3.1.a Food Loss Index
- Enable a quantitative analysis of **policy documents**

HIHI support – disaggregated agriculture statistics

Agricultural production data at national and sub-national level

To reduce respondent burden, the DataLab scraped agricultural production and area data. In the first 13 countries, these data came from national official websites or publications. The data is validated against FAOSTAT data and in some cases replaces FAOSTAT estimates.

The sub-national data is published in the HiH-GIS platform and on the Data Lab Website.

Sub-national validateddata

From the table below it is possible to download the report with the main validation findings and the related data (in comma separated format).

| COUNTRY NAME	SUB-NATIONAL LEVELS	DATA AVAILABILITY PERIOD	NUMBER OF COMMODITIES	VALIDATION STATUS	FINDINGS REPORT	CSV DATA
Burkina Faso	2	1995-2018	16	Finished	Data Validation for Burkina Faso.pdf	Burkina Faso - Sub-National Level 1.csv Burkina Faso - Sub-National Level 2.csv
El Salvador	2	2010-2017	4	Finished	Data Validation for El Salvador.pdf	El Salvador - Sub-National Level 1.csv El Salvador - Sub-National Level 2.csv

HIHI support: Use of geospatial data to obtain indicators for welfare

A Neural Network model was developed to determine land cover classes and changes in land cover without ground-truthing data

The land cover section

The image for each Governorate/District is divided into "tiles", i.e. in a set of 64*64 pixels (640*640 meters) images
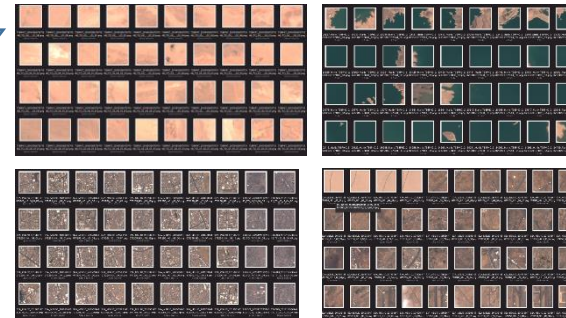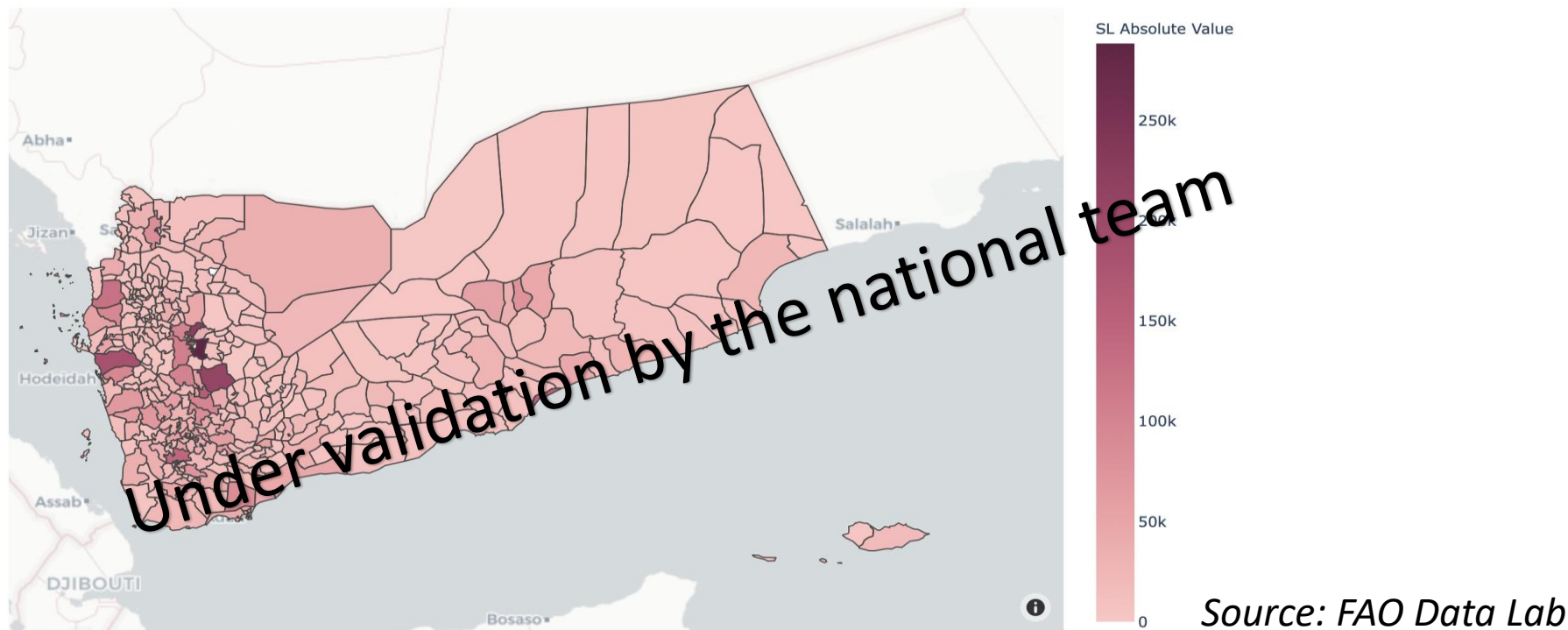
Each tile is assigned to a specific Land Cover

Governorate	Percentage of Land cover for 2015					Percentage of Land cover for 2020				
	NUA	R	UA	C	W	NUA	R	UA	C	W
Ad_Dali	0.96174	0.00374	0.00663	0.02758	0.00032	0.92807	0.00214	0.04436	0.02469	0.00075
Al_Hodeidah	0.71777	0.02001	0.00865	0.25083	0.00274	0.60292	0.01730	0.01373	0.36398	0.00208
Raymah	0.97959	0.00311	0.00710	0.00998	0.00022	0.95651	0.00133	0.01908	0.02285	0.00022
Aden	0.68358	0.20295	0.09111	0.01037	0.01200	0.54888	0.21245	0.14910	0.07701	0.01256
Al_Bayda	0.95093	0.01021	0.00794	0.02977	0.00116	0.86739	0.00492	0.01429	0.11109	0.00230
Sanaa_City	0.40997	0.11727	0.33980	0.13112	0.00185	0.14351	0.03291	0.47532	0.34552	0.00274
Dhamar	0.93563	0.01670	0.00548	0.04209	0.00010	0.87221	0.00562	0.01256	0.10951	0.00010
Amran	0.91250	0.02330	0.00212	0.06194	0.00014	0.90706	0.01585	0.00647	0.07033	0.00030
Marib	0.98015	0.01197	0.00044	0.00675	0.00069	0.96760	0.01129	0.00175	0.01462	0.00474
Ibb	0.97612	0.00049	0.01250	0.01082	0.00007	0.95685	0.00080	0.01418	0.02796	0.00020
Lahj	0.97774	0.01290	0.00425	0.00494	0.00017	0.96926	0.01292	0.00758	0.00996	0.00028
Shabwah	0.99150	0.00557	0.00093	0.00185	0.00014	0.98512	0.00860	0.00185	0.00416	0.00027
Hadramawt	0.99153	0.00322	0.00105	0.0 96	0.00025	0.95830	0.00605	0.00492	0.03061	0.00011
Sanaa	0.92761	0.02525	0.01827		0.00018	0.87973	0.01592	0.04266	0.06066	0.00102
Sadah	0.95837	0.01079	0.00114		014	0.95230	0.00843	0.00220	0.03647	0.00060
Al_Mahwit	0.98610	0.00377	0.00362			0.97654	0.00159	0.00594	0.01564	0.00029
Hajjah	0.94538	0.00535	0.00148	0.04 43		0.95458	0.00473	0.00247	0.03804	0.00017
Al_Jawf	0.99120	0.00347	0.00006	0.00521	0.000	0.98938	0.00412	0.00008	0.00627	0.00014
Al_Maharah	0.99214	0.00421	0.00046	0.00054	0.00266	0.99486	0.00346	0.00069	0.00004	0.00096
Abyan	0.98107	0.01002	0.00079	0.00806	0.0000	0.97682	0.00724	0.00290	0.01297	0.00007
Taiz	0.89134	0.04342	0.02325	0.03948	0.00251	0. 454	0.00902	0.02164	0.02413	0.00067
Socotra	0.94308	0.01324	0.00225	0.00315	0.03828	0.5 799	0.00832	0.00404	0.00333	0.04633

Source: FAO Data Lab

HIHI support: Use of geospatial data to estimate welfare indicators

The Data Lab developed a vulnerability map for Yemen in the absence of the usual data. Satellite images, scraped data (documents and databases), and prices were used to estimate the vulnerability at district level.

Source: FAO Data Lab

The boundaries and names shown and the designations used on these map(s) do not imply the expression of any opinion whatsoever on the part of FAO concerning the legal status of any country, territory, city or area of its authorities, or concerning the delimitation of its frontier and boundaries. Dashed lines on maps represent approximate border lines for which there may not yet be full agreement

This map displays the number of vulnerable people in each district

The vulnerability indicator is based on a dataset of more than 100 variables
-   socio-economic variables from various sources
-   Land Cover classes through a Neural Network model

Support SDG: Inform SDG indicator 12.3.1.a Food Loss Index

An online Food Loss and Waste database give access to structured data from literature and reports

To overcome data scarcity in the area of food losses, FAO developed a whole data flow and set of tools to structure data from unstructured data sources

dissemination

Data sources:

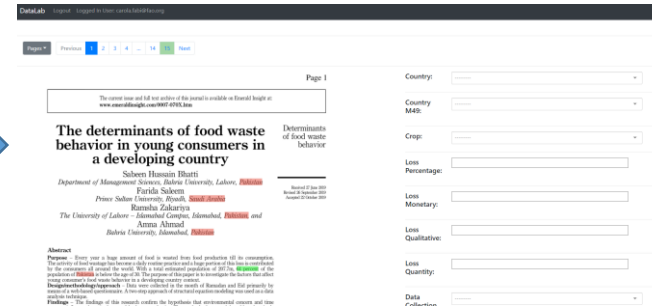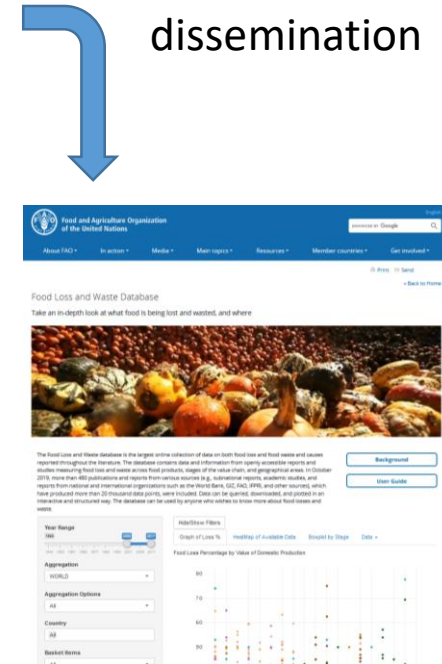queries to search engines, access to a repository, manual addition of files

Retrieved objects:

a set of files (FLW literature and reports)

SolR Database:

Conversion and relevance

Postgres Database:

Guided extraction tool