UN Big Data Hackathon

WEBINAR

UN Big Data Hackathon

Big Data Sources & Analysis Webinar

The 2022 UN Big Data Hackathon in numbers…

4 days

60 countries

450 teams

1000+ participants

Available Data Sources

- All public data sets can be used in the UN Big Data Hackathon.

- The use of private and/or copyrighted datasets is not allowed for any team.

Data Sources Summary

	Youth Track	Big Data Experts Track
AIS data (UNGP x IMO)	✗	✅
Open data on AWS registry	✅	✅
Lloyd's Register Foundation - World Risk Poll (Gallup)	✅	✅
UNICEF data portal	✅	✅
The Humanitarian Data Exchange (HDX) data portal	✅	✅
World Bank open data	✅	✅

Other data sources (ex. IMF, UN, WHO...) will be provided

Data and Platform

- **For Youth track: AWS**

  All public open data sets can be used in the hackathon

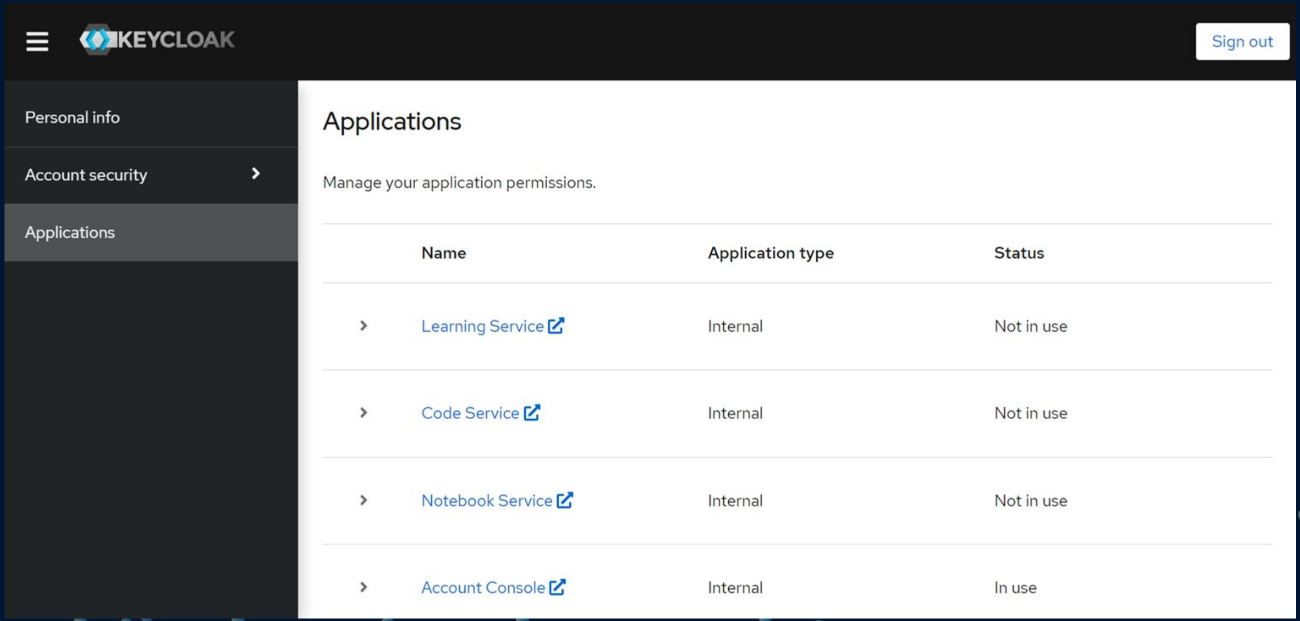
  Eg : **Registry of Open Data on AWS**

  Relevant data sets within the registry of open data on AWS and many other open data sources will be made available directly on the AWS platform.

  More details on the AWS platform and the data sets available will be discussed during the webinar on the 31st of October

Data and Platform: Big Data Experts track

- **Data Source:** AIS (Automatic Identification System)
- **Platform:** UN Global Platform
- **Link:** https://id.officialstatistics.org/



A brief definition of AIS data

The Automatic Identification System (AIS) is an automated, autonomous tracking system which is extensively used in the maritime world for the exchange of navigational information between AIS-equipped terminals, originally developed for collision avoidance

A bit of background

- Developed by the International Maritime Organisation (IMO) in 2004, solely for collision avoidance among large vessels at sea that are not within range of shore-based systems
- Fully automatic transceiver system
- Global coverage
- Real-time data tracked by several data providers, and is made available to the AIS community online

AIS data example

Vessel ID, vessel name, vessel type, vessel size, and the nationality of the ship

FID	mmsi	imo	vessel_name	callsign	vessel_type	vessel_type_code	vessel_type_cargo	vessel_class	length	width	flag_country	flag_code
null	440503000	8815724	55 SHIN YUNG	6MWP	Fishing	30	null	A	55	9	South Korea	440
null	366557000	8419142	MATSON ANCHORAGE	KGTX	Cargo	70	null	A	216	24	USA	366
null	440055000	9019509	ORYONG 325	6MNZ	Fishing	30	null	A	56	10	South Korea	440
null	367542320	null	WALTER L GIBBS	WDG5004	Towing	31	null	A	27	10	USA	367
null	538008215	9844277	OLYMPIC LIFE	V7A2092	Tanker	80	null	A	333	60	Marshall Islands	538
null	345070040	9242106	DONA BLANCA	XCDC	Passenger	60	null	A	22	5	Mexico	345
null	735057514	null	DARWIN	HC2113	Passenger	60	null	A	20	5	Ecuador	735
null	367651380	440	ELK	WDH7758	Cargo	70	null	A	58	15	USA	367
null	366998130	null	TAYLOR MARIE	WDC2822	Tug	52	null	A	22	8	USA	366
null	218791000	9612997	ANTWERPEN EXPRESS	DJCE2	Cargo	79	No Additional Inf...	A	366	48	Germany	218
null	735059299	null	JOLINDA	HC5601	Fishing	30	null	A	45	5	Ecuador	735
null	636016940	9238789	MSC MANU	A8CF3	Cargo	70	null	A	260	32	Liberia	636
null	338392816	null	COOL BREEZE	null	Pleasure Craft	37	null	B	13	5	USA	338
null	636018346	9797187	POLAR CHILE	D5PH8	Cargo	72	Carrying DG,HS or...	A	230	37	Liberia	636
null	563063700	9833541	STI MAGISTER	9V8891	Tanker	80	null	A	183	32	Singapore	563
null	636010032	9018658	SOL DO BRASIL	ELQQ4	Cargo	70	null	A	172	26	Liberia	636
null	338125000	9670339	RUSSELL ADAMS	WDG9047	WIG	20	null	A	81	18	USA	338
null	224559000	8802363	PLAYA DE RODAS	EHQQ	Fishing	30	null	A	55	10	Spain	224
null	316266000	9175298	PLACENTIA PRIDE	VCWB	Tug	52	null	A	38	13	Canada	316
null	710003110	null	PELAGIUS	PR 6983	Tug	52	null	A	30	10	Brazil	710

AIS data example

Destination, geospatial location, speed, and navigational status on the ship

destination	eta	draught	position	longitude	latitude	sog	cog	rot	heading	nav_status	nav_status_code
null	null	0.0	POINT (13.1726333...	-164.43488333	13.17263333	3.7	116.8	0.0	0	Under Way Using E...	0
TACOMA WA	null	9.0	POINT (53.9401883...	-164.57464667	53.94018833	19.3	86.8	16.11514409	86	Under Way Using E...	0
null	null	3.7	POINT (1.6708 -15...	-153.56116667	1.6708	4.0	152.6	0.0	0	Under Way Using E...	0
HOUSTON	null	2.9	POINT (29.7433333...	-94.08	29.74333333	5.0	230.0	0.0	0	Unknown	16
GALVESTON	null	11.0	POINT (28.3352133...	-93.05576667	28.33521333	11.5	303.8	0.0	302	Under Way Using E...	0
null	null	0.0	POINT (18.6533333...	-91.84166667	18.65333333	0.0	212.0	0.0	0	Not Defined	15
CRUCEROS INTERISLAS	null	0.0	POINT (-0.75 -90.31)	-90.31	-0.75	0.0	276.0	0.0	0	At Anchor	1
FOURCHON	null	4.0	POINT (28.35 -90....	-90.66666667	28.35	0.0	26.0	0.0	0	Under Way Using E...	0
US^0EW8>0E70	null	2.8	POINT (30.0466666...	-90.6	30.04666667	0.0	173.0	0.0	0	Unknown	16
KRPUS	null	12.9	POINT (8.24166666...	-86.84666667	8.24166667	19.0	284.0	0.0	0	Under Way Using E...	0
FAENA D PESCA	null	0.0	POINT (-11.474056...	-84.07834	-11.47405667	0.0	0.0	0.0	129	Engaged In Fishing	7
PAROD	null	8.8	POINT (-0.11949 -...	-81.113605	-0.11949	17.9	13.5	0.0	13	Under Way Using E...	0
null	null	0.0	POINT (26.16917 -...	-80.10563	26.16917	0.0	0.0	0.0	0	Unknown	16
BALBOA	null	10.2	POINT (-33.592733...	-71.61748333	-33.59273333	0.0	222.2	0.0	181	Moored	5
BR SLZ	null	12.2	POINT (14.603895 ...	-68.09905	14.603895	11.3	114.2	0.0	116	Under Way Using E...	0
US ILG	null	9.4	POINT (26.2783333...	-64.30666667	26.27833333	17.0	312.0	0.0	0	Under Way Using E...	0
GT GUY	null	4.2	POINT (6.78647833...	-58.17381333	6.78647833	0.0	46.0	0.0	13	Moored	5
FISHING GROUND	null	7.2	POINT (-35.757288...	-55.027085	-35.75728833	10.6	131.1	0.0	131	Moored	5
null	null	0.0	POINT (47.7732133...	-54.01134167	47.77321333	0.0	49.0	0.0	8	Not Defined	15
SAO LUIS	null	4.0	POINT (-2.59382 -...	-44.36726833	-2.59382	0.1	208.6	0.0	0	Underway Sailing	8

AIS data example

Source of the transmission, the date and time of the transmission

source	ts_pos_utc	ts_static_utc	ts_insert_utc	dt_pos_utc	dt_static_utc	dt_insert_utc	vessel_type_main	vessel_type_sub	message_type	eeid	dayIndex
S-AIS	null	null	null	2021-05-08 05:43:34	2021-05-08 05:36:10	2021-05-08 05:43:52	Fishing Vessel	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:43:20	2021-05-08 05:31:08	2021-05-08 05:43:30	Container Ship	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:43:11	2021-05-08 05:36:02	2021-05-08 05:43:30	Fishing Vessel	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:42:59	2021-05-08 05:39:05	2021-05-08 05:43:11	null	null	27	null	739814
S-AIS	null	null	null	2021-05-08 05:43:40	2021-05-08 05:31:50	2021-05-08 05:43:53	null	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:43:28	2021-05-08 05:03:28	2021-05-08 05:43:43	Offshore Vessel	Offshore Tug Supp...	27	null	739814
S-AIS	null	null	null	2021-05-08 05:42:52	2021-05-07 18:08:02	2021-05-08 05:43:11	null	null	27	null	739814
S-AIS	null	null	null	2021-05-08 05:43:13	2021-04-30 02:47:14	2021-05-08 05:43:28	Offshore Vessel	Offshore Support ...	27	null	739814
S-AIS	null	null	null	2021-05-08 05:43:02	2021-05-07 13:09:21	2021-05-08 05:43:21	Service Ship	null	27	null	739814
S-AIS	null	null	null	2021-05-08 05:43:19	2021-05-08 00:27:04	2021-05-08 05:43:43	Container Ship	null	27	null	739814
S-AIS	null	null	null	2021-05-08 05:43:02	2021-05-08 05:33:22	2021-05-08 05:43:20	null	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:43:02	2021-05-08 04:47:33	2021-05-08 05:43:20	Container Ship	null	1	null	739814
T-AIS	null	null	null	2021-05-08 05:43:44	2021-05-08 05:41:45	2021-05-08 05:43:55	null	null	18	null	739814
S-AIS	null	null	null	2021-05-08 05:42:38	2021-05-08 05:32:08	2021-05-08 05:43:08	null	null	3	null	739814
S-AIS	null	null	null	2021-05-08 05:43:00	2021-05-08 04:24:31	2021-05-08 05:43:12	null	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:43:34	2021-05-07 23:01:02	2021-05-08 05:43:53	Other Tanker	Fruit Juice Tanker	27	null	739814
S-AIS	null	null	null	2021-05-08 05:43:16	2021-05-08 05:10:17	2021-05-08 05:43:42	Offshore Vessel	Offshore Tug Supp...	3	null	739814
S-AIS	null	null	null	2021-05-08 05:42:59	2021-05-08 04:40:45	2021-05-08 05:43:12	Fishing Vessel	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:43:34	2021-05-08 05:40:51	2021-05-08 05:43:53	Tug	null	1	null	739814
S-AIS	null	null	null	2021-05-08 05:42:50	2021-05-08 05:35:54	2021-05-08 05:43:08	null	null	1	null	739814

The UN Global Platform:

- Is a cloud based, collaborative environment
- Developed for use with big data – has the functionality to manipulate and work with big data
- Holds big data, methods, algorithms, code and use cases
- Is maintained by the UN Committee of Experts on Big Data and Data Science for Official Statistics
- E-learning course: https://learning.officialstatistics.org/course/view.php?id=84

Data sources - Panel of speakers

Thierry Schlaudecker

Data Management & Visualization Engineer
United Nations International Children's Emergency Fund

Dr. Aaron Ions Gardner

Data and Insight Scientist at Lloyd’s Register Foundation

Faizal Thamrin

Data Manager
OCHA Centre for Humanitarian Data

UNICEF Data Portal

Thierry Schlaudecker

Data Management & Visualization Engineer
United Nations International Children's Emergency Fund

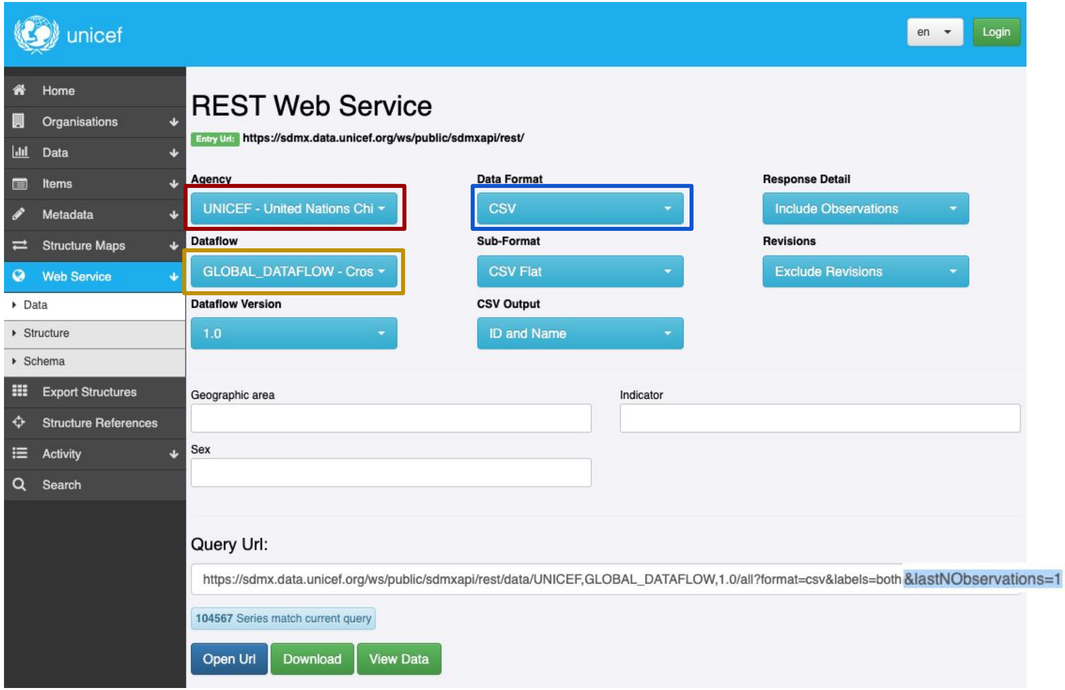
UN Big Data Hackathon

Yves Jaques
Thierry Schlaudecker

October 2022

SDMX Web Services

Available at https://sdmx.data.unicef.org/webservice/data.html



The warehouse supports SDMX, the UN-preferred standard for the exchange of statistical data and metadata. The standard covers not only data structuring, but also the APIs.

The web service builder makes it easy to interactively build the URL to deliver a custom CSV file. To generate a CSV, make sure CSV is selected as the data format.

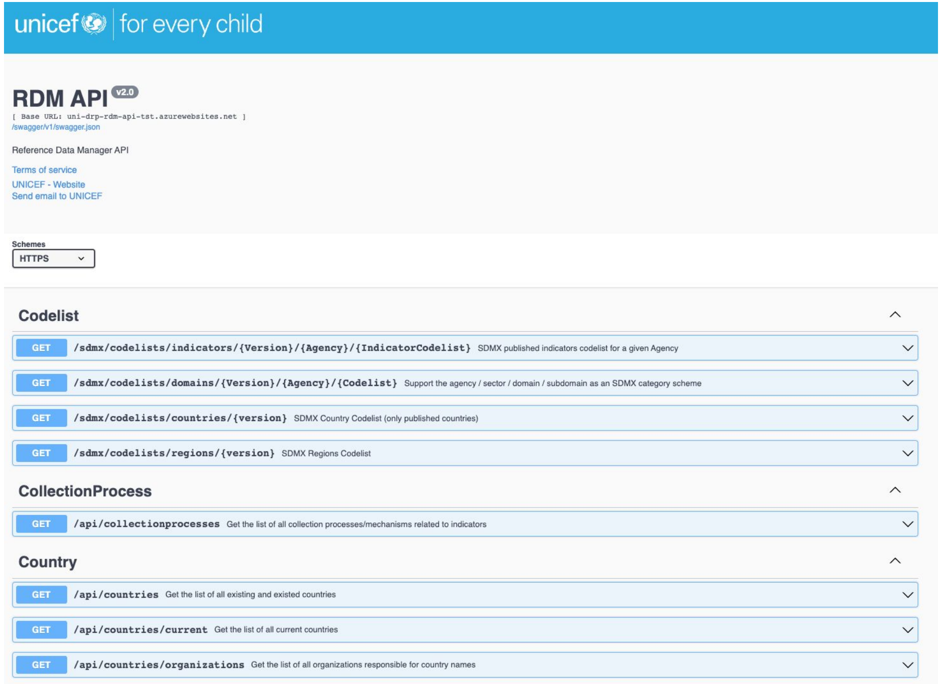
ALL the official data is under the UNICEF agency setting.

Under that Agency there is the GLOBAL dataflow, that has cross-sectoral data that is disaggregated only along a few common dimensions (country/indicator/sex). There are also topic specific dataflows, with many more dimensions.

There is also one “secret” option: add &lastNObservations=1 to the query string to get just the last observation for any particular intersection of dimensions (it’s all modelled on a hypercube). Or &lastNObservations=n with n being the desired number of observations.

Reference Data Manager API

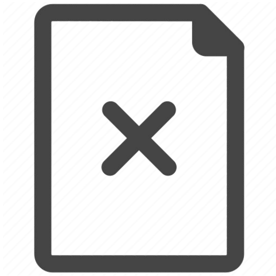
Documentation framework available at https://uni-drp-rdm-api-tst.azurewebsites.net/api/doc/index.html



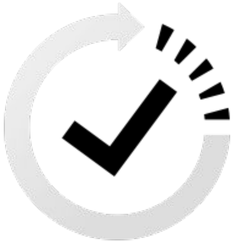
The Reference Data Manager (RDM) is a single source of truth for the most crucial UNICEF Reference Data and Reference Metadata: indicators, and regional aggregations. It holds all of the information about how indicators are calculated, including definitions, computation methods, survey populations, and more.

In the spirit of Open Data, all RDM data are available using publicly available, extensively documented communications interfaces (APIs) using the industry standard, best-practices API documentation framework known as “Swagger”.

Challenges

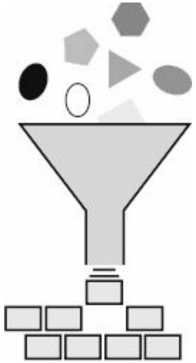


Data availability



Intermittent reporting frequency

Lack of disaggregations



Standardization

HDX Data Portal

Faizal Thamrin

Data Manager
OCHA Centre for Humanitarian Data

centre for humdata

OCHA

HDX

OCHA Centre for Humanitarian Data

Faizal

Partnerships Team

@humdata

centre for humdata

Centre for Humanitarian Data

The Hague, the Netherlands

managed by

OCHA
United Nations
Office for the Coordination
of Humanitarian Affairs

map >

The mission of the Centre is to increase the **use** and **impact** of data in humanitarian response.

What is humanitarian data?

1. Data about the context of the crisis

2. Data about the people affected and their needs

3. Data about the humanitarian response

OUR AIMS

➔ **Speed** of data

We want to speed up the flow of data from collection to use so that humanitarian responders can find and share data that reflects a current day, real-time understanding of a crisis.

➔ **Connections** in the network

We want to increase the number of organisations partnering with the Centre and each other through a shared data infrastructure and shared data goals.

➔ **Increase** use

We want to ensure data is used better and more often by people making critical decisions in a humanitarian response, as well as make data and its related insights more accessible to all.

NEW YORK
(USA)

THE HAGUE
(NETHERLANDS)

GENEVA
(SWITZERLAND)

BUCHAREST
(ROMANIA)

NAIROBI
(KENYA)

DAKAR
(SENEGAL)

BANGKOK
(THAILAND)

JAKARTA
(INDONESIA)

We are a global team

Focus Areas for the Centre

DATA
SERVICES

DATA
RESPONSIBILITY

DATA
LITERACY

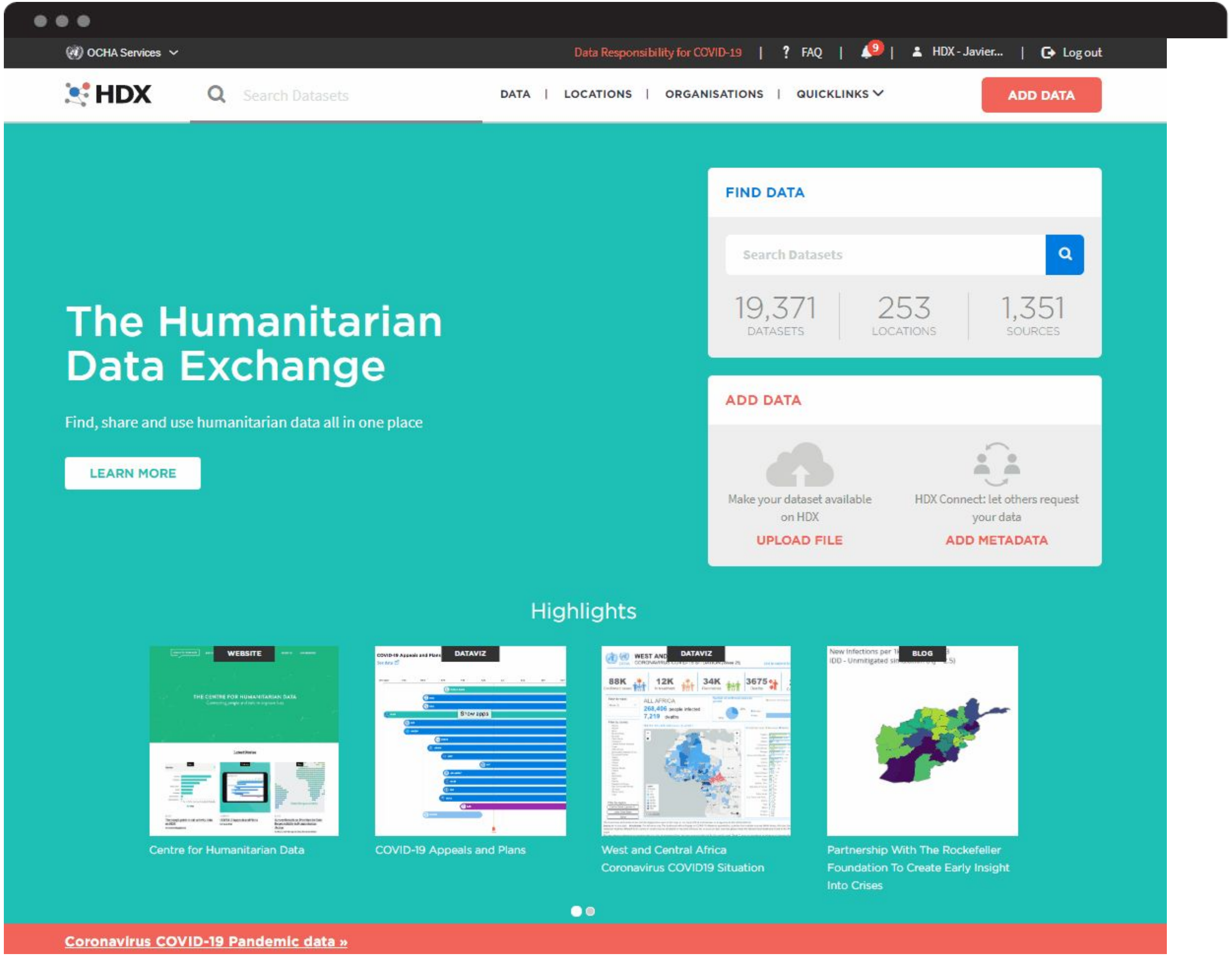
PREDICTIVE
ANALYTICS

OCHA's open platform for sharing data.

The goal of HDX is to make humanitarian data easy to find and use for analysis.

It was launched in 2014 and has become the go-to place for humanitarian data.

http://data.humdata.org



HDX at a Glance (2021)

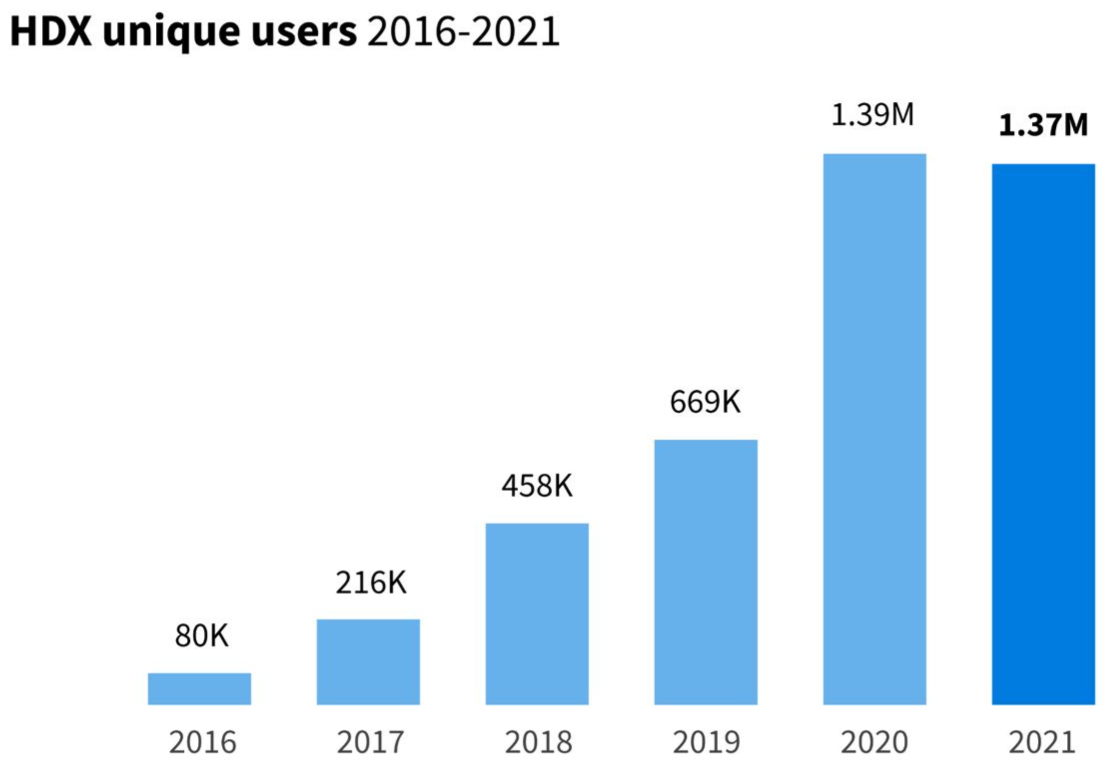
1.4m
UNIQUE USERS IN 2021

1.8m
DOWNLOADS IN 2021

300+
ACTIVE ORGANIZATIONS

19,000+
DATASETS

250+
LOCATIONS



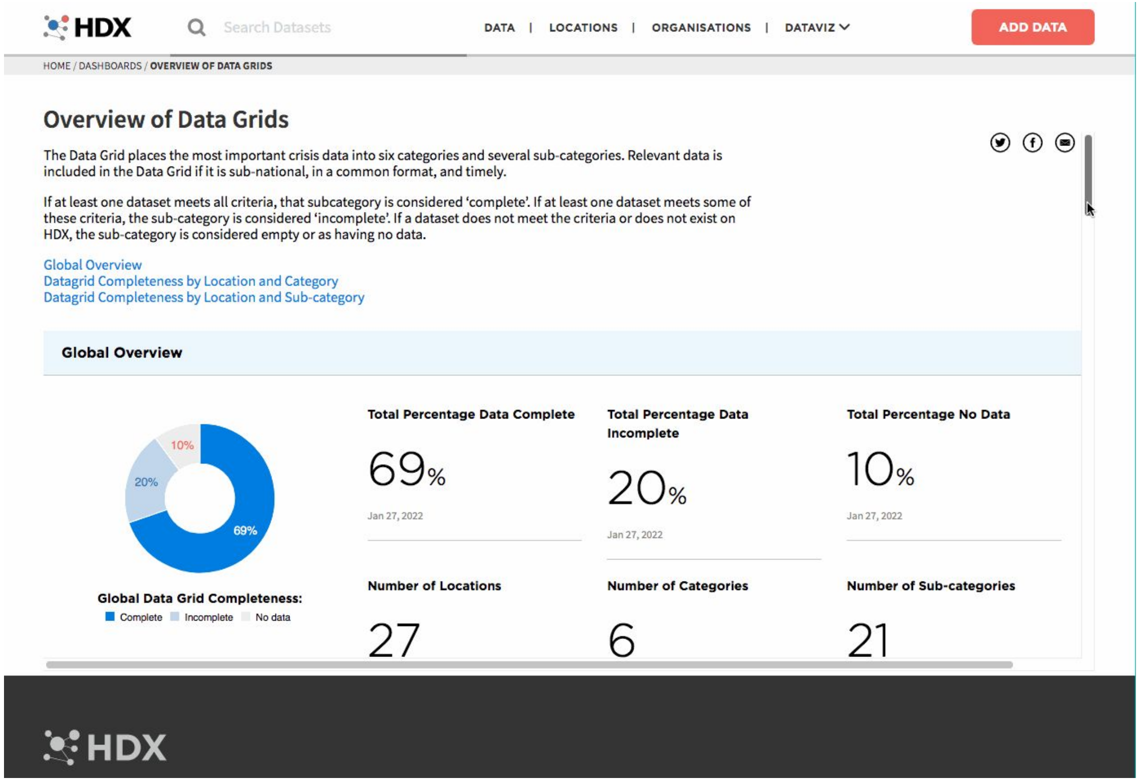
https://centre.humdata.org/hdx-year-in-review-2021/

Featured HDX data grid

Data Grid

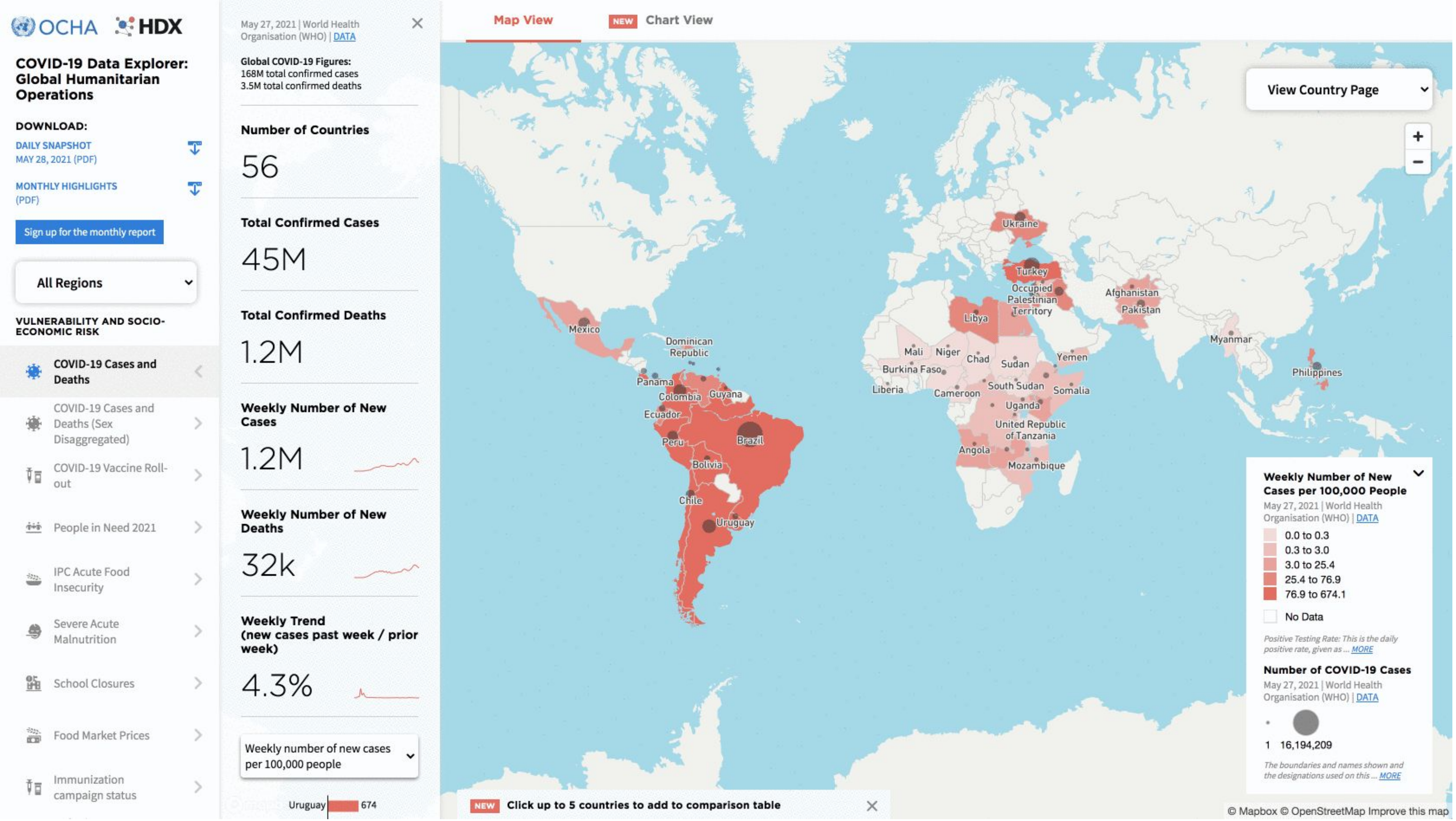
‘Data Grid’ helps users in their quest for good and relevant data. Based on interviews with our users, the **Data Grid** places the most important crisis data into six categories and 27 sub-categories.

Data Grid: The Data Completeness Grid defines six categories and 21 sub-categories and indicates if they are complete, incomplete or missing.

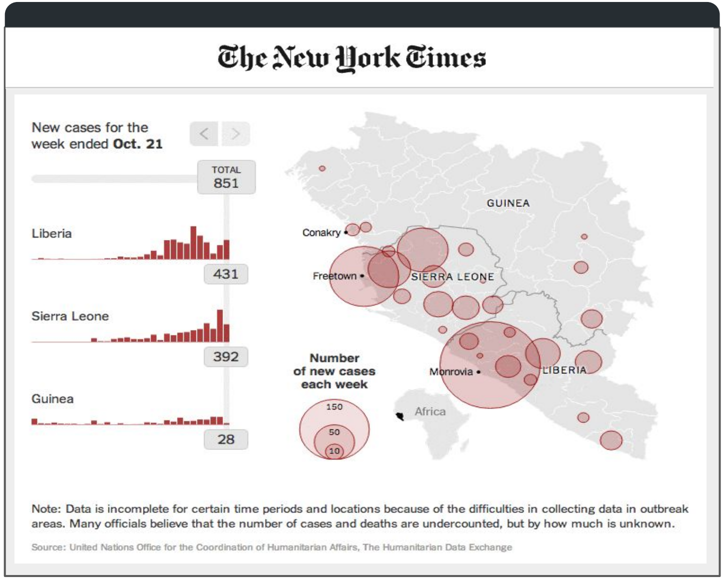


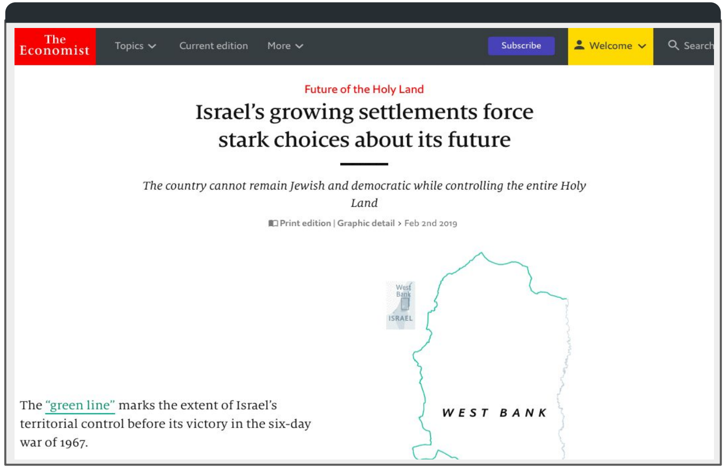
https://data.humdata.org/dashboards/overview-of-data-grids

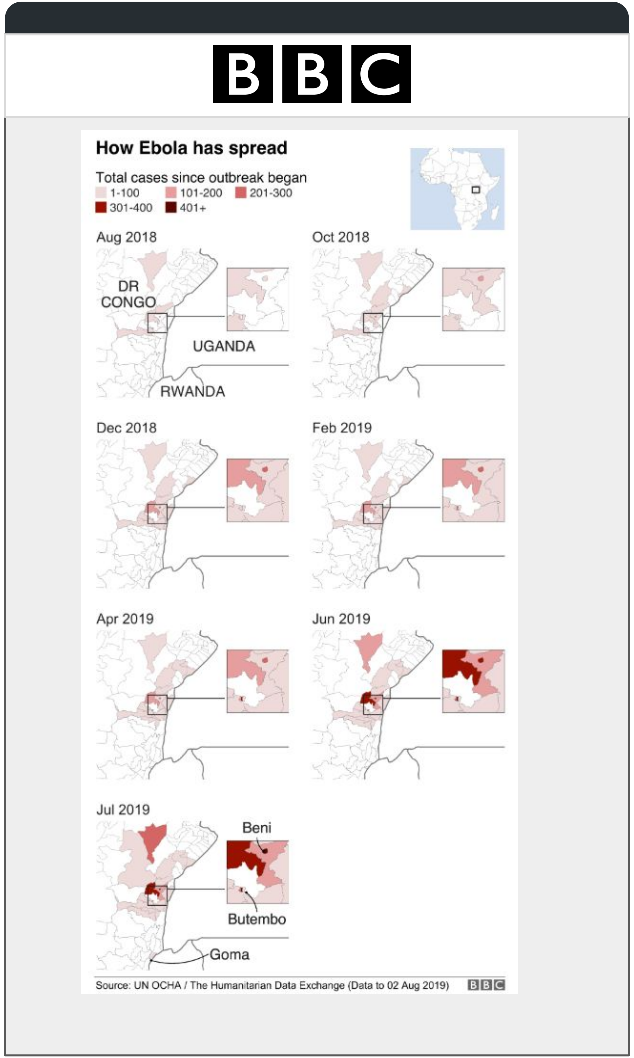
COVID-19 data explorer



What do these three visual journalism pieces have in common?







Live HDX Demo by Faizal

- The HDX homepage
- Data Explorers
- Searching for data
- Checking the metadata
- Downloading data

Thank you and any questions?

centre.humdata.org

@humdata | centrehumdata@un.org
thamrinf@un.org

centre for humdata

OCHA

World risk poll 2021

Dr. Aaron Ions Gardner

Data and Insight Scientist at Lloyd's Register Foundation

World Risk Poll 2021

lrfworldriskpoll.com

@LR_Foundation // #WorldRiskPoll

Dr. Aaron Ions Gardner

Data and Insight Scientist

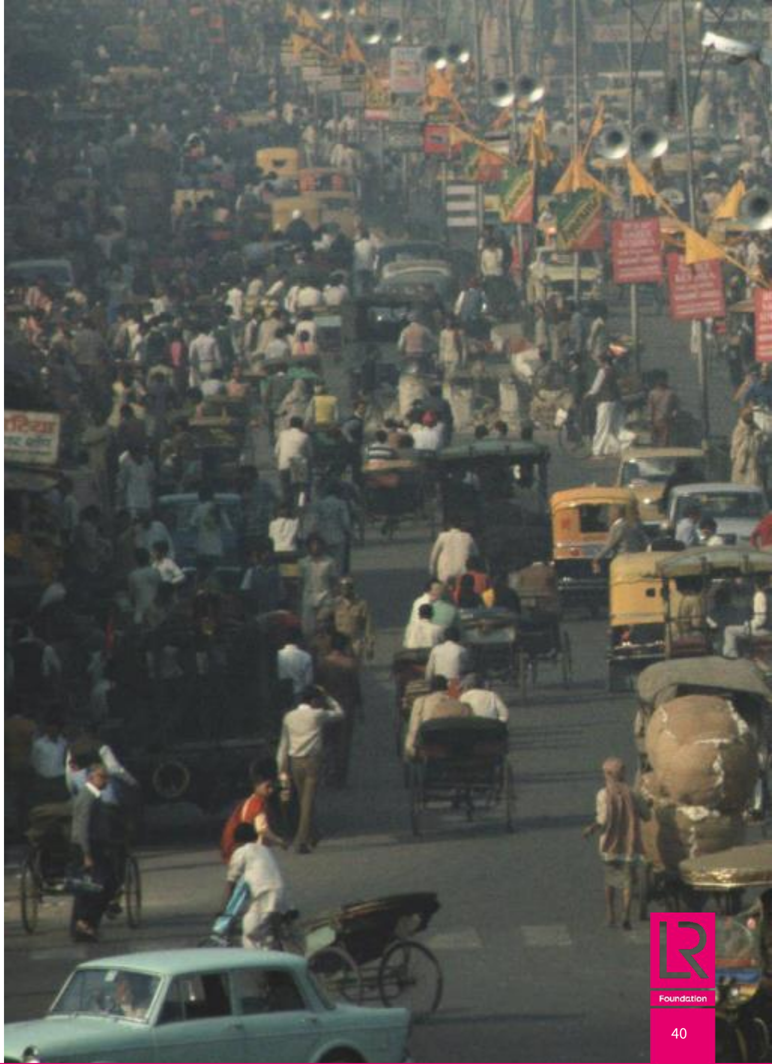
Lloyd's Register Foundation World Risk Poll

- 121 countries, 125,000 interviews
  - Assessing perception and experience of risk
  - In places where little or no official data on safety exists
  - Disaster resilience, violence & harassment at work, data privacy & artificial intelligence
- 2019 > 2021
  - Build on existing data
  - What changed, and what didn't?
  - Impact of Covid-19 on people's sense of safety

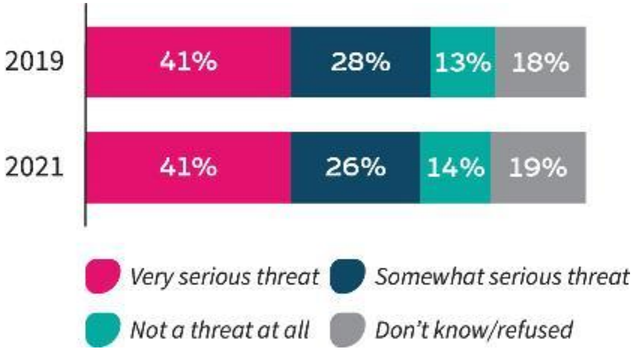
We face many different risks in our daily lives

Road-related accidents/injuries	13%	-5
Crime/violence	12%	-3
Personal health (non-Covid-19)	10%	+1
Covid-19	7%	

Greatest risk to safety in your own words

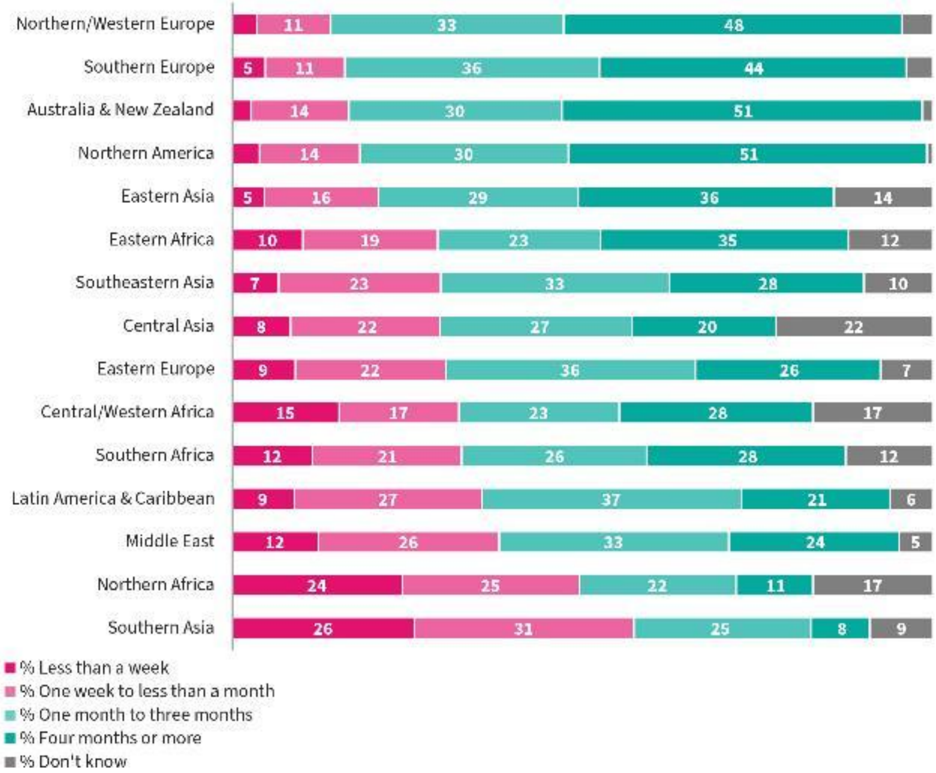


Climate change perceptions unchanged – in spite of competing risks



Do you think that climate change is a threat?

New measure reveals global financial vulnerability



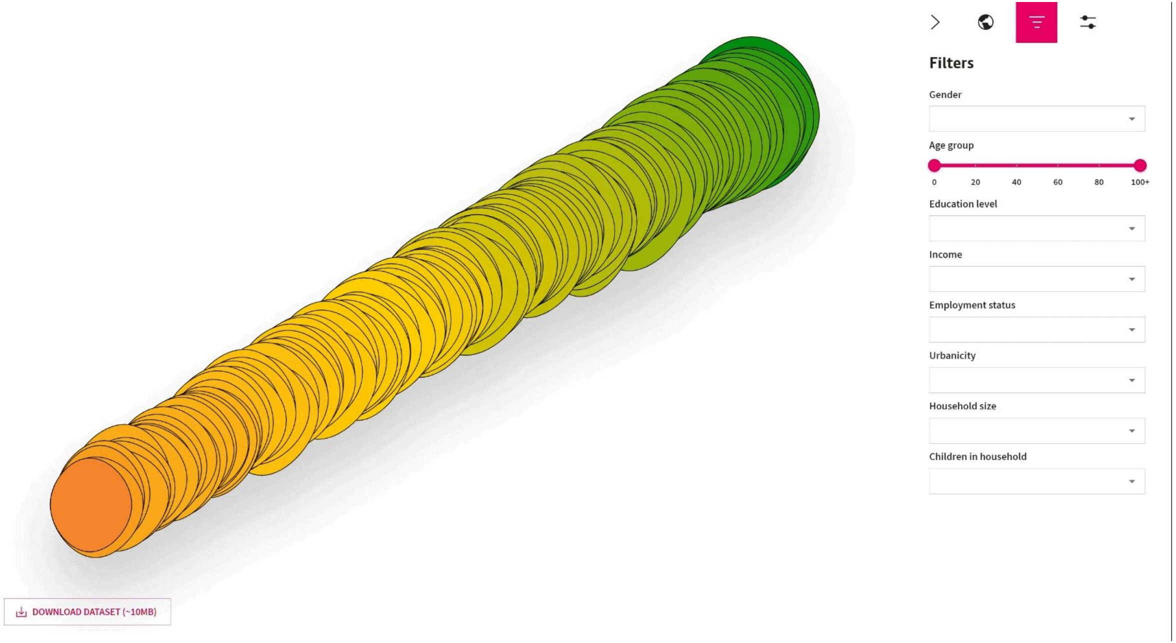
Lloyd's Register Foundation Resilience Index

Individual
Is there anything you could do to protect yourself/family in the event of disaster?

Household
How long could you cover basic needs if you lost all income?

Community
How much do neighbours care about you/your wellbeing?

Society
Have you personally experienced discrimination?



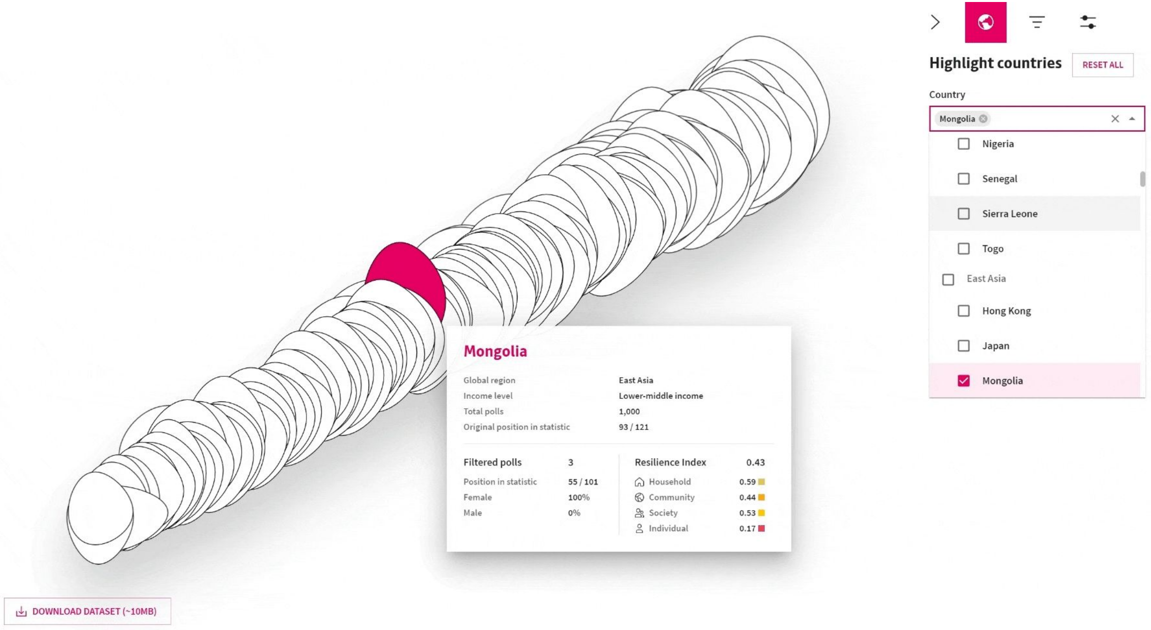
Lloyd's Register Foundation Resilience Index

Individual
Is there anything you could do to protect yourself/family in the event of disaster?

Household
How long could you cover basic needs if you lost all income?

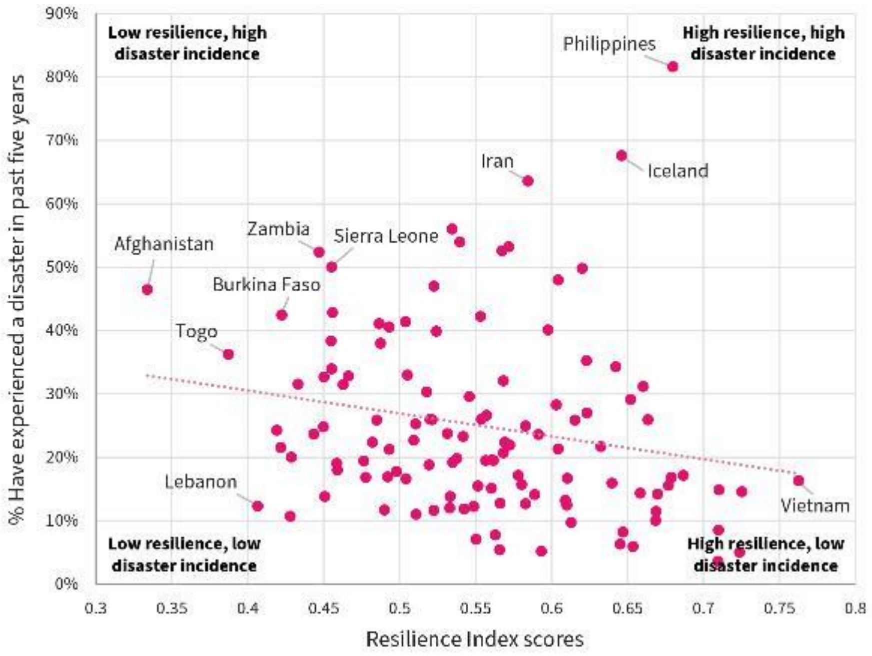
Community
How much do neighbours care about you/your wellbeing?

Society
Have you personally experienced discrimination?



Resilience varies significantly at a global level

Countries experiencing most disasters have low resilience index scores



Disaster follows in the absence of resilience

'We are drowning': Pakistan floods push toxic lake over edge

Heavy rain compounds decades-long environmental catastrophe at country's largest freshwater lake

Rahmat Tunio

Tue 13 Sep 2022 16.45 BST

Pakistan floods: 'The water came and now everything is gone'

31 August

Climate change

'Very Dire': Devastated by Floods, Pakistan Faces Looming Food Crisis

The flooding has crippled Pakistan's agricultural sector, battering the country as it reels from an economic crisis and double-digit inflation that has sent the price of basics soaring.

By Christina Goldbaum and Zia ur-Rehman

Sept. 11, 2022

NEWS | 02 September 2022 | Correction 02 September 2022

Why are Pakistan's floods so extreme this year?

One-third of the country is under water, following an intense heatwave and a long monsoon that has dumped a record amount of rain.

Smriti Mallapaty

'A Monsoon on Steroids.' What To Know About Pakistan's Catastrophic Floods

BY SANYA MANSOOR

AUGUST 31, 2022 12:58 PM EDT

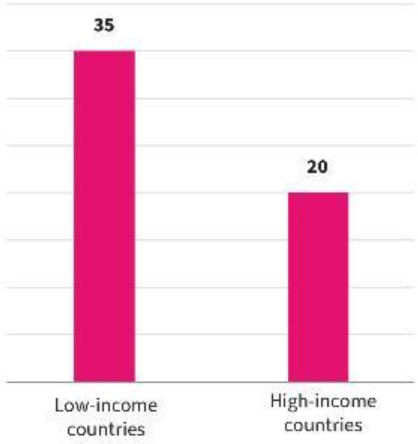
Climate graphic of the week: One third of Pakistan submerged by flooding, satellite data shows

Record rainfall combined with glacial melt devastates estimated 30mn people

Aime Williams in Washington and **Steven Bernard** in London

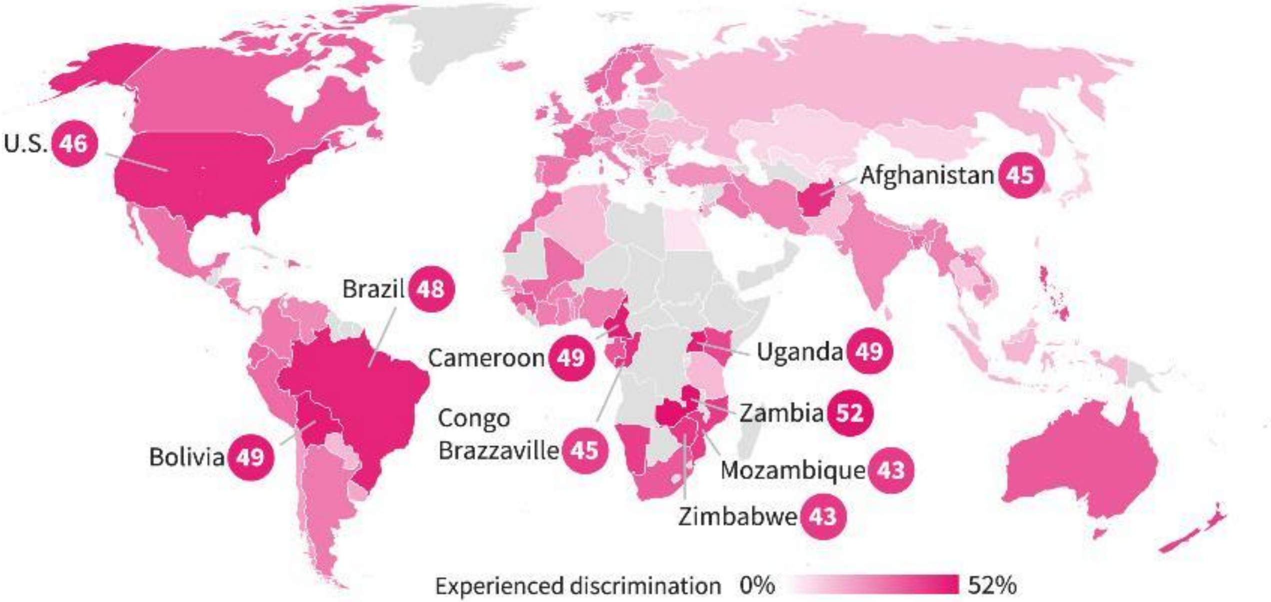
SEPTEMBER 5 2022

Community support is higher in low income countries



Percentage who believe their neighbours care about them 'a lot,' by World Bank country income group

One in five globally has experienced discrimination



Percentage who had experienced discrimination based on one or more of five characteristics: skin colour, nationality/race/ethnicity, sex, religion, disability status

lrfworldriskpoll.com

Explore the poll – stories and visual snapshots

Download the full dataset

Apply for funding to turn the World Risk Poll into action

lrfworldriskpoll.com

@LR_Foundation // #WorldRiskPoll

Dr. Aaron Ions Gardner

Data and Insight Scientist

Experience from the 2021 UN Youth Hackathon's Winning Team

Team Sustainability

from France

Who are we ?

Jean-Philippe Kouadio: Data Scientist, based in Abidjan, Côte d'Ivoire

Marine Jouvin: PhD in Development Economics, based in Bordeaux, France

Oumaïma Boukamel: M&E Manager, based in Bordeaux, France

Our Scope

Analysis focusing on Uganda households.

Analysis based on a sampe of 2225 households surveyed by the *World Bank* and the *Ugandan Office of Statistics.*

Uganda is located in East Africa and has known pretty severe lockdown measures during COVID-19.

Area	
• Total	241,038 km^2 (93,065 sq mi) (79th)
• Water (%)	15.39
Population	
• 2018 estimate	▲ 42,729,036[5][6] (35th)
• 2014 census	▲ 34,634,650[7]
• Density	157.1/km^2 (406.9/sq mi)
GDP (PPP)	2019 estimate
• Total	$102.659 billion[8]
• Per capita	$2,566[8]
GDP (nominal)	2019 estimate
• Total	▲ $30.765 billion[8]
• Per capita	▲ $956[8]

Source: Wikipédia

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda

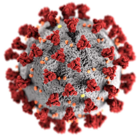
What is vulnerability ?

"Vulnerability is the inability to resist a hazard or to respond when a disaster has occurred. For instance, people who live on plains are more vulnerable to floods than people who live higher up."

unisdr.org

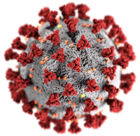
Our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Our objective

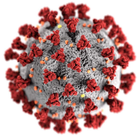
Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda

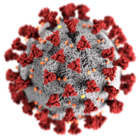


Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security: **What are the household profiles that are most likely to face food insecurity due to COVID ?**

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security: **What are the household profiles that are most likely to face food insecurity due to COVID ?**

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

The data

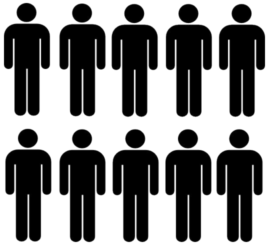
World Bank Microdata Library: contains 3626 studies

THE WORLD BANK
IBRD • IDA

The data

World Bank Microdata Library: contains 3626 studies

What we selected:



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

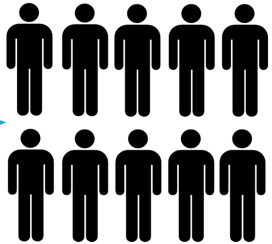
The data

World Bank Microdata Library: contains 3626 studies

What we selected:

THE WORLD BANK
IBRD • IDA

LSMS Survey 19-20 containing data on the socio economic characteristics of households



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

The data

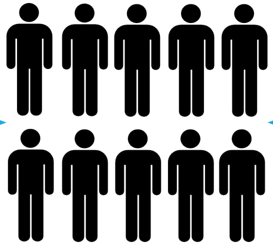
World Bank Microdata Library: contains 3626 studies

What we selected:

THE WORLD BANK
IBRD • IDA

LSMS Survey 19-20 containing data on the socio economic characteristics of households

High Frequency Phone survey on COVID 2020-2021 containing data on the impact and coping of COVID on households



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

The data

Combining both datasets enabled us to have a set of variables that we could use as « predictors » (LSMS variables) and a set of variables that we could use as « predictions » (COVID data).

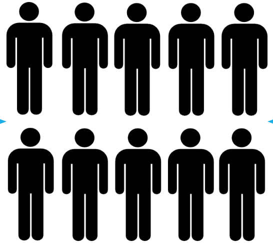
World Bank Microdata Library: contains 3626 studies

What we selected:

THE WORLD BANK
IBRD • IDA

LSMS Survey 19-20 containing data on the socio economic characteristics of households

High Frequency Phone survey on COVID 2020-2021 containing data on the impact and coping of COVID on households



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

Data description

- The LSMS contains two datasets:
  - One dataset at the household level
  - One dataset at the household member level

Data description

- The LSMS contains two datasets:
  - One dataset at the household level
  - One dataset at the household member level
- The high frequency phone survey on COVID contains overall 16 datasets, but we used 8 of them:
  - The cover containing identification information
  - The household roster containing information on the household members
  - A dataset on the level of knowledge of respondents on COVID-19
  - A dataset on the behavior adopted by the respondent to cope with the pandemic
  - A dataset showing the level of access to COVID protection
  - A dataset on the impact of COVID on the crops
  - A dataset on the impact of COVID on income (it is an income level dataset meaning that there is one observation per income source)
  - A dataset on the impact of COVID on food security

Data description

Merging the LSMS datasets:

- Both datasets contained a unique household ID (baselinehhid) that was used to merge both datasets

Data description

Merging the LSMS datasets:

- Both datasets contained a unique household ID (baselinehhid) that was used to merge both datasets

Merging the High Frequency Phone COVID Survey datasets:

- All datasets contained a unique household ID (HHID) that was used to merge all datasets

Data description

Merging the LSMS datasets:

- Both datasets contained a unique household ID (baselinehhid) that was used to merge both datasets

Merging the High Frequency Phone COVID Survey datasets:

- All datasets contained a unique household ID (HHID) that was used to merge all datasets

Merging the High Frequency Phone COVID Survey datasets:

- The dataset containing identification information on the survey also contained the LSMS household ID (baselinehhid) that unabled us to link the datasets.

Data processing and cleaning

STEP 1: Cleaning the two surveys separately

- Check duplicates
- Fix structural errors
- Outliers identification
- Rename columns to make the variables names more transparent and to avoir duplicated of variable names among the different datasets
- Validation and cross-checking

Data processing and cleaning

STEP 2: Synthetizing rosters to get one comprehensive datasets with 1 observation per household

- LSMS: Synthesis of the household member roster (total household size, indicators on education level, education level of the household head, proportion of litterate household members, number of household member per age range and gender etc…)

```r
aggregated_roster<-clean_hh_roster_lsms%>%
  group_by(hhid)%>%
  summarise(hh_size=sum(count),
            f0_14=mean(f0_14),
            f15_64=mean(f15_64),
            f65=mean(f65),
            m0_14=mean(m0_14),
            m15_64=mean(m15_64),
            m65=mean(m65),
            married=mean(married,na.rm=TRUE),
            form_married=mean(form_married,na.rm=TRUE),
            nev_married=mean(nev_married,na.rm=TRUE),
            literacy=mean(literacy,na.rm=TRUE),
            work=mean(work,na.rm=TRUE),
            primary_age=sum(primary*age_5_11),
            secondary_age=sum(secondary*age_12_17),
            tertiary_age=sum(tertiary*age_18),
            primary=sum(primary),
            secondary=sum(secondary),
            tertiary=sum(tertiary),
            tot_5_11=sum(age_5_11),
            tot_12_17=sum(age_12_17),
            tot_18=sum(age_18))


colnames(aggregated_roster)[1]<-"baseline_hhid"
aggregated_roster$prop_primary<-aggregated_roster$primary/aggregated_roster$hh_size
aggregated_roster$prop_secondary<-aggregated_roster$secondary/aggregated_roster$hh_size
aggregated_roster$prop_tertiary<-aggregated_roster$tertiary/aggregated_roster$hh_size
aggregated_roster$prop_educated<-aggregated_roster$prop_primary+aggregated_roster$prop_secondary+aggregated_roster$prop_tertiary
summary(aggregated_roster$prop_primary)
summary(aggregated_roster$prop_secondary)
summary(aggregated_roster$prop_tertiary)
```

Data processing and cleaning

STEP 2: Synthetizing rosters to get one comprehensive datasets with 1 observation per household

- COVID Survey: The roster dataset contained variables with one line per household*type of income source. We synthetized the dataset in order to get for each household total the number of income sources, the proportion of income sources completely lost due to COVID and the proportion of income sources reduced due to COVID.

```
#Income data aggregation per household

income_summary<-income_loss_covid_r1[income_loss_covid_r1$income_source_lastmonths==1,]
income_summary$counting<-rep(1,nrow(income_summary))
income_summary$reduced<-rep(0,nrow(income_summary))
income_summary$no_income<-rep(0,nrow(income_summary))
income_summary$reduced[income_summary$income_evolution==3]<-1
income_summary$no_income[income_summary$income_evolution==4]<-1

income_summary<-income_summary%>%
  group_by(HHID)%>%
  summarise(nb_income=sum(counting),nb_reduced=sum(reduced),nb_noincome=sum(no_income))


income_summary$fq_reduced<-income_summary$nb_reduced/income_summary$nb_income
income_summary$fq_noincome<-income_summary$nb_noincome/income_summary$nb_income
income_summary$total_loss<-rep(NA,nrow(income_summary))
income_summary$reduction<-rep(NA,nrow(income_summary))
```

Multiple correspondence analysis (MCA)

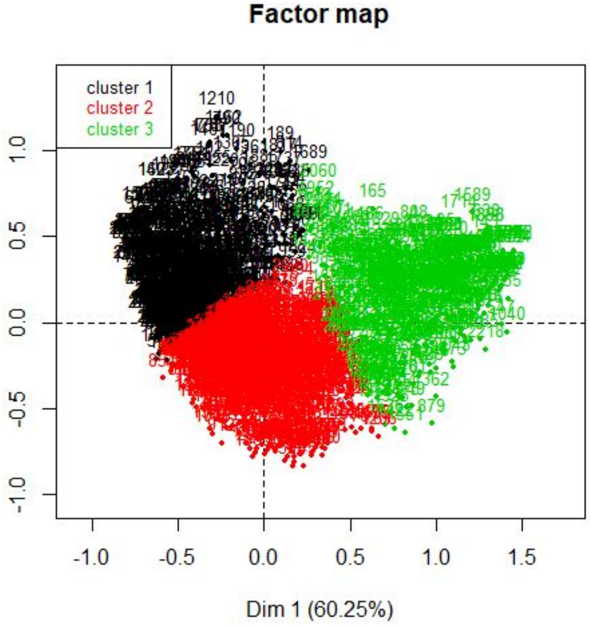
- **Objective** : to segregate households by level of vulnerability

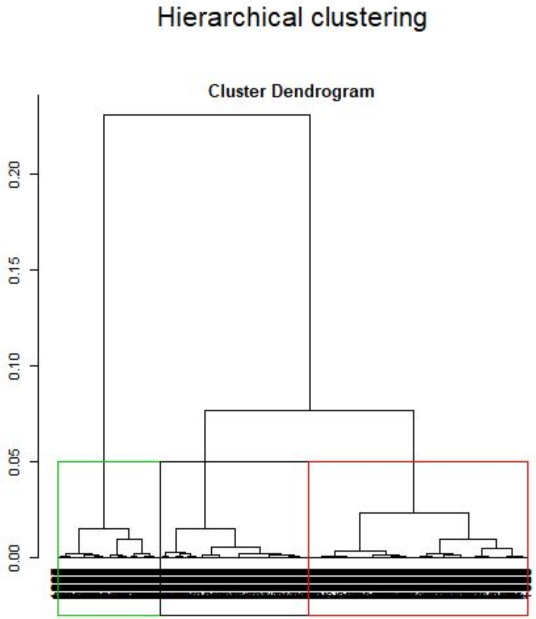
- **Method** : We rely on a MCA analysis (as we used only categorical variables), followed by a hierarchical ascending classification (HAC) consolidated by the k-means method.

- **Variables used for segmentation** :
  - **Housing** : Materials of the walls, floor and roof of the house, access to electricity, water and toilets.
  - **Assets** : Possession of a cellphone, a refrigerator, a motorcycle.
  - **Farming information** : possession of land and crop, and livestock ownership.
  - **Income** : income of the household.
  - **Household composition** : number of persons in the household, education of the household head.

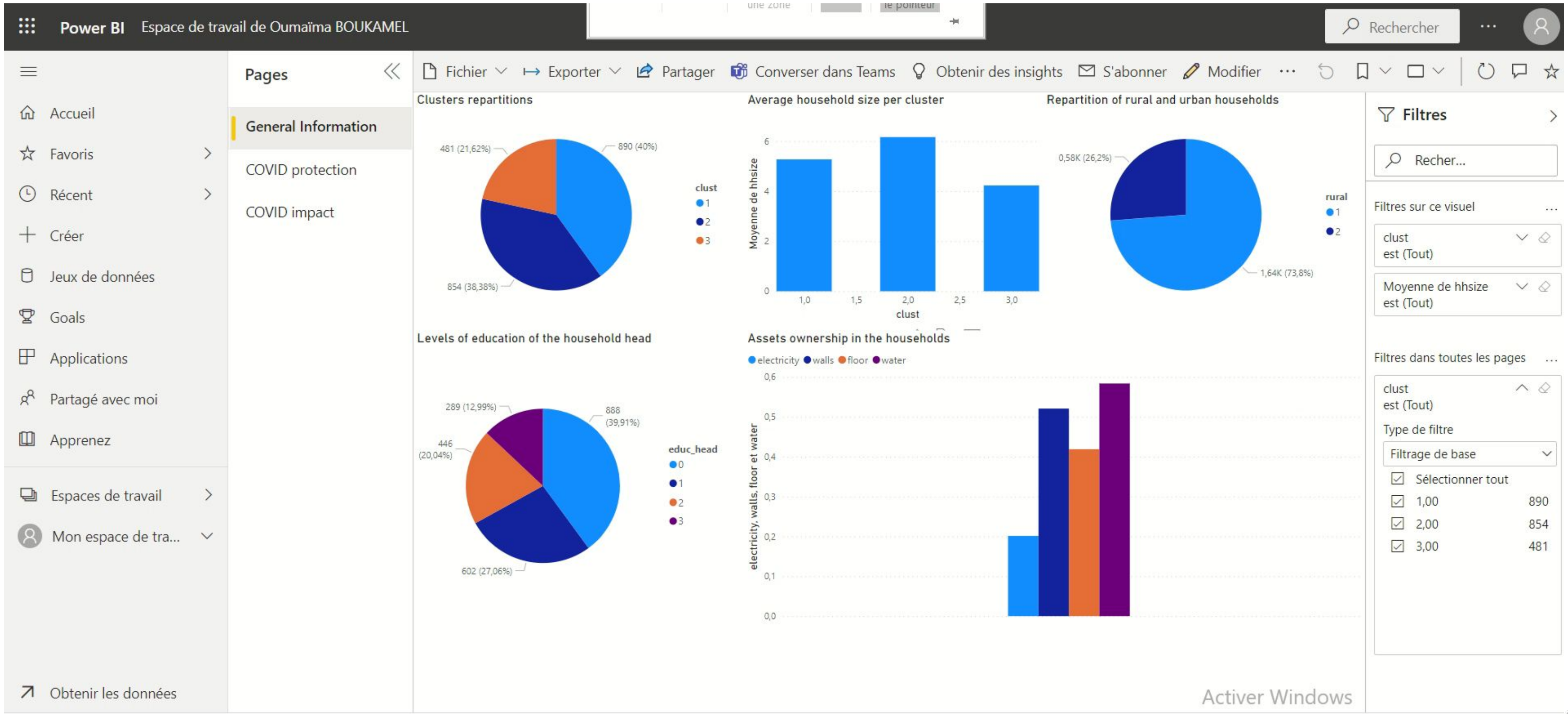
Multiple correspondence analysis (MCA)

- **Findings :** The MCA and the ACH result in the classification of households into 3 distinct groups, which explains 68% of the inter-household variance.
  - Class 1 : Poor rural households
  - Class 2 : Vulnerable rural households
  - Class 3 : Urban, less vulnerable, households





Data visualization per cluster



Data visualization per cluster

STEP 1: Import of the the data cleaning and some processing in power BI through an R script

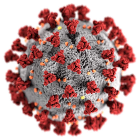
STEP 3: Adding the variable *clust* as a filter so that the user can filter the data per cluster

STEP 2: Building the visualisations on 3 thematics:

- General characteristics of the households
- COVID-19 protection characteristics
- Impact of COVID-19 on the household

Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security: **What are the household profiles that are most likely to face food insecurity due to COVID ?**

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

Methodology: setting-up classification models

- **Naive Bayes (with Rstudio)**

STEP 1: Import and load packages

Import and load the following packages e1071, caTools, caret

STEP 2: Split the dataset in 2 datasets (split ratio = 0.7), using sample.split. One dataset will be the **training** dataset, the other one will be the **test** dataset.

```
split<-sample.split(c(1:nrow(M)),SplitRatio=0.7)
train_cl<-subset(M,split==TRUE)
test_cl<-subset(M,split==FALSE)
```

STEP 3: Scaling of the datasets to « smooth » the data using the function scale

Methodology: setting-up classification models

- **Naive Bayes (with Rstudio)**

STEP 4: Setting seeds (set.seed(120))

STEP 5: Applying the naiveBayes fonction and generating the classifier using the training dataset

```
classifier_cl <- naiveBayes(fs_vulnerability ~ ., data=train_cl)
classifier_cl
```

STEP 6: Predicting on the test data

```
# Predicting on test data
y_pred <- predict(classifier_cl,newdata=test_cl)
```

STEP 7: Model evaluation (using the confusion matrix to compare the predictions with the actual values)

Methodology: setting-up classification models

- **Decision trees (with Rstudio)**

STEP 1: Import and load packages (DAAG, party, rpart, rpart.plot,mlbench, caret, pROC, tree)

STEP 2: Converting the « prediction category » in factors (with as.factor) and setting seeds (set.seed(1234))

STEP 3: Split the dataset in 2 datasets (split ratio = 0.5). One dataset will be the **training** dataset, the other one will be the **test** dataset.

```
ind<-sample(2,nrow(M),replace=T, prob = c(0.5,0.5))
train<- subset(M,ind==1)
test<-subset(M,ind==2)
```

Methodology: setting-up classification models

- **Decision trees (with Rstudio)**

STEP4: Tree classification

```
# Tree classification

tree <-rpart(fs_vulnerability ~., data=train)
rpart.plot(tree,box.palette="blue")

printcp(tree)

rpart(formula = fs_vulnerability ~., data=train)

plotcp(tree)
```

STEP 5: Testing the prediction model on the test data and comparing the outputs to the actual categories

STEP 6: Model evaluation with the confusion matrix (confusionMatrix function)

Methodology: setting-up classification models

- **K-NN (with Rstudio)**

STEP 1: Inputing relevant values to NA as the K-NN model does not work if the data contains empty values

STEP 2: defining a normalization function and run the normalization on the predictor

```
## the normalization function is created

nor <-function(x){(x-min(x)/max(x)-min(x))}

## Run normalization on the predictors

M_norm <- data.frame(lapply(M[,-1],nor))
```

Methodology: setting-up classification models

- **K-NN (with Rstudio)**

STEP 3: Split the dataset in 2 datasets (split ratio = 0.8). One dataset will be the **training** dataset, the other one will be the **test** dataset.

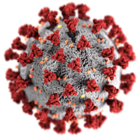
STEP 4: Run the K-NN function

```
##run knn function
pr <- knn(M_train, M_test, cl=M_target_category)
```

STEP 5: Model evaluation with the confusion matrix

Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security: **What are the household profiles that are most likely to face food insecurity due to COVID ?**

Identifying the most vulnerable households towards education: **What are the household profiles in which children are more likely to drop school due to the pandemic ?**

Model 1: Identifying income vulnerability

Defining the categories

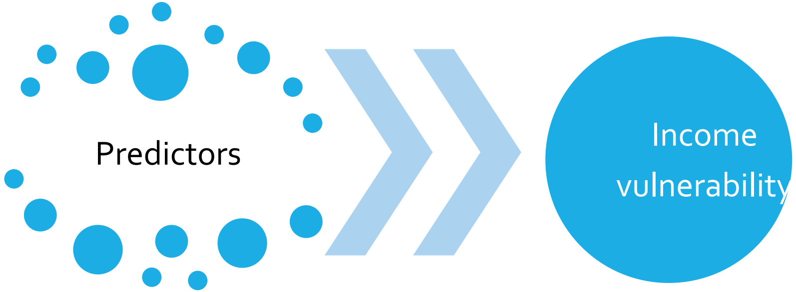
Category	Proportion of income sources lost range	Number of households in this category
The household has lost all their income sources during the pandemic	=1	123
The household has lost less than 50% of their income sources during the pandemic	<0.5	117
The household has lost more than 50% of their income sources during the pandemic	>=0.5	292
The household has lost none of their income sources during the pandemic	=0	1693

The proportion of income sources completely lost was calculated from the income source roster of the High Frequency Phone Survey on COVID-19, that was cleaned and aggregated.

Model 1: Identifying income vulnerability

Within the LSMS dataset we chose the following predictors:

- Rural, roof, floor, walls, toilet,water,rooms,elect,tv,radio,refrigerator,land_tot,land_cultivated, rent, remit, assist, crop, crop_number, cash_crop, sell_crop, fies_mod, fies_sev, hh_size, adulteq, literacy, work, primary_head, secondary_head, tertiary_head



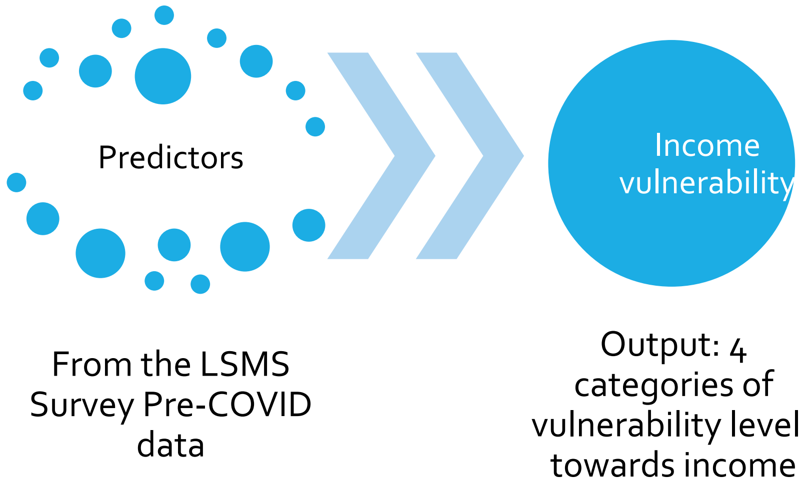
From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels towards income

Model 1: Identifying income vulnerability

We tested 3 classification methodologies in order to select the most performant one:

- **Naives Bayes Classifier**
- **K-NN**



Model 1: Identifying income vulnerability

K-NN Classification results

```
Statistics by Class:

                     Class: The household lost all their income sources
Sensitivity                                                    0.142857
Specificity                                                    0.948113
Pos Pred Value                                                 0.120000
Neg Pred Value                                                 0.957143
Prevalence                                                     0.047191
Detection Rate                                                 0.006742
Detection Prevalence                                           0.056180
Balanced Accuracy                                              0.545485
                     Class: The household lost less than 50% of their income sources
Sensitivity                                                                 0.095238
Specificity                                                                 0.941038
Pos Pred Value                                                              0.074074
Neg Pred Value                                                              0.954545
Prevalence                                                                  0.047191
Detection Rate                                                              0.004494
Detection Prevalence                                                        0.060674
Balanced Accuracy                                                           0.518138
                     Class: The household lost more than 50% of their income sources
Sensitivity                                                                  0.13462
Specificity                                                                  0.89313
Pos Pred Value                                                               0.14286
Neg Pred Value                                                               0.88636
Prevalence                                                                   0.11685
Detection Rate                                                               0.01573
Detection Prevalence                                                         0.11011
Balanced Accuracy                                                            0.51387
                     Class: The household lost no income sources
Sensitivity                                              0.7892
Specificity                                              0.2872
Pos Pred Value                                           0.8052
Neg Pred Value                                           0.2673
Prevalence                                               0.7888
Detection Rate                                           0.6225
Detection Prevalence                                     0.7730
Balanced Accuracy                                        0.5382
```

Naive Bayes classification results

```
Statistics by Class:

                     Class: The household lost all their income sources
Sensitivity                                                     0.07500
Specificity                                                     0.97872
Pos Pred Value                                                  0.90000
Neg Pred Value                                                  0.29299
Prevalence                                                      0.71856
Detection Rate                                                  0.05389
Detection Prevalence                                            0.05988
Balanced Accuracy                                               0.52686
                     Class: The household lost less than 50% of their income sources
Sensitivity                                                                 0.071429
Specificity                                                                 0.953674
Pos Pred Value                                                              0.093750
Neg Pred Value                                                              0.938679
Prevalence                                                                  0.062874
Detection Rate                                                              0.004491
Detection Prevalence                                                        0.047904
Balanced Accuracy                                                           0.512551
                     Class: The household lost more than 50% of their income sources
Sensitivity                                                                 0.250000
Specificity                                                                 0.893939
Pos Pred Value                                                              0.027778
Neg Pred Value                                                              0.989933
Prevalence                                                                  0.011976
Detection Rate                                                              0.002994
Detection Prevalence                                                        0.107784
Balanced Accuracy                                                           0.571970
                     Class: The household lost no income sources
Sensitivity                                              0.8333
Specificity                                              0.2283
Pos Pred Value                                           0.2195
Neg Pred Value                                           0.8403
Prevalence                                               0.2066
Detection Rate                                           0.1722
Detection Prevalence                                     0.7844
Balanced Accuracy                                        0.5308
```

Model 1: Identifying income vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.4332, 0.5102)
K-NN	(0.6031, 0.6938)

We decided to go for the K-NN based on the accuracy confidence interval and based on the comparison of the sensitivity and specificity of the category « The household lost all their income sources » which is the category that we want to determine in priority.

Model 1: Identifying income vulnerability

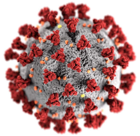
Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.4332, 0.5102)
K-NN 	(0.6031, 0.6938)

We decided to go for the K-NN based on the accuracy confidence interval and based on the comparison of the sensitivity and specificity of the category « The household lost all their income sources » which is the category that we want to determine in priority.

Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security: **What are the household profiles that are most likely to face food insecurity due to COVID ?**

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

Model 2: Identifying food security vulnerability

Figure 4. Actual Example—Calculating a Household CSI Index Score

In the past 7 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to:	Raw Score	Severity Weight	Weighted Score = Frequency X weight
(Add each behavior to the question)			
a. Rely on less preferred and less expensive foods?	5	1	5
b. Borrow food, or rely on help from a friend or relative?	2	2	4
c. Purchase food on credit?	1	2	2
d. Gather wild food, hunt, or harvest immature crops?	0	4	0
e. Consume seed stock held for next season?	0	3	0
f. Send household members to eat elsewhere?	1	2	2
g. Send household members to beg?	0	4	0
h. Limit portion size at mealtimes?	7	1	7
i. Restrict consumption by adults in order for small children to eat?	2	2	4
j. Feed working members at the expense of non-working members?	0	2	0
k. Reduce number of meals eaten in a day?	5	2	10
l. Skip entire days without eating?	0	4	0
TOTAL HOUSEHOLD SCORE	Sum down the totals for each individual strategy		**34**

- This CSI index Score was developed under the framework of collaborative research project, implemented by WFP and CARE in Kenya, with financial support of the UK Department for International Development via WFP, The Bill and Melinda Gates Foundation, and CARE-USA.
- Among the items described on the item described on the left the High Frequency Phone Survey on COVID contains the items a,k,h and l.
- We used this Score definition to set the ponderations of an index we designed in order to assess the food insecurity levels of the households during COVID
- Based on this index we defined 4 categories of households based on their food insecurity level: “Not vulnerable”, “Moderately vulnerable”, “Very vulnerable”, “Severely vulnerable”.

Model 2: Identifying food security vulnerability

Defining the index

Question	Variable	Severity	CSI Index Score equivalent	Ponderation
Were you or any other adult in your household were worried about not having enough food to eat because of lack of money or other resources?	fs_worried	1		1/14
You, or any other adult in your household, were unable to eat healthy and nutritious/preferred foods because of a lack of money or other resources?	fs_healthy	1	a. Rely on less preferred and less expensive food	1/14
You, or any other adult in your household, ate only a few kinds of foods because of a lack of money or other resources?	fs_few	1		1/14
You, or any other adult in your household, skipped meals because of a lack of money or other resources?	fs_skip	2	k. Reduce number of meals eaten in a day	2/14
You, or any other adult in your household, ate less than you thought you should because of a lack of money or other resources?	fs_less	1	h. Limit portion size at meal time	1/14
Your household ran out of food because of a lack of money or other resources?	fs_ranout	2		2/14
You, or any other adult in your household, were hungry but did not eat because there was not enough money or other resources for food?	fs_hungry	2		2/14
You, or any other adult in your household, went without eating for a whole day because of a lack of money or other resources?	fs_day	4	l. Skipped entire days without eathing	4/14

Model 2: Identifying food security vulnerability

Defining the categories

Category	Index range	Number of households in this category
Not vulnerable	Index==0	563
Moderately vulnerable	Index in]0,0.28[	639
Very vulnerable	Index in [0.28, 0,5[	380
Severely vulnerable	Index in >=0,5	643

The categories were defined to ensure that the households who checked an item with a severity score equal to 4 or two items with a severity score equal to 2 (hence with an index superior or equal to 2/7) were in the category very vulnerable or severely vulnerable.

Model 2: Identifying food security vulnerability

Within the LSMS dataset we chose the following predictors:

- Rural, roof, floor, walls, toilet,water,rooms,elect,tv,radio,refrigerator,land_tot,land_cultivated, rent, remit, assist, crop, crop_number, cash_crop, sell_crop, fies_mod, fies_sev, hh_size, adulteq, literacy, work, primary_head, secondary_head, tertiary_head

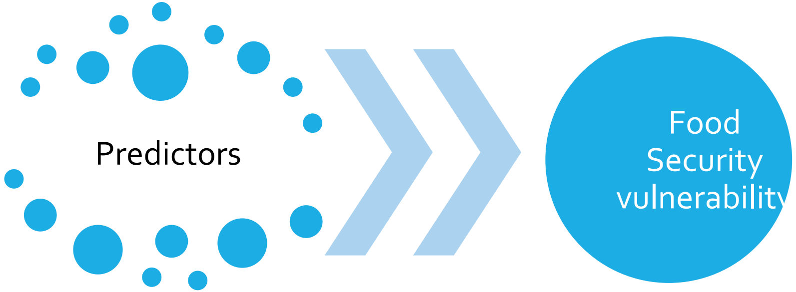
From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels to food insecurity

Model 2: Identifying food security vulnerability

We tested 3 classification methodologies in order to select the most performant one:

- **Naives Bayes Classifier**
- **K-NN**
- **Decision Trees**



From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels to food insecurity

Model 2: Identifying food security vulnerability

Naives Bayes

Statistics by Class:

	Class: Moderately vulnerable	Class: Not vulnerable	Class: Severely vulnerable	Class: Very vulnerable
Sensitivity	0.4984	0.33333	0.6923	0.0000
Specificity	0.6073	0.87875	0.7175	1.0000
Pos Pred Value	0.3383	0.48469	0.5011	NaN
Neg Pred Value	0.7504	0.79393	0.8505	0.8327
Prevalence	0.2871	0.25492	0.2907	0.1673
Detection Rate	0.1431	0.08497	0.2013	0.0000
Detection Prevalence	0.4231	0.17531	0.4016	0.0000
Balanced Accuracy	0.5529	0.60604	0.7049	0.5000

Decision Tree

Statistics by Class:

	Class: Moderately vulnerable	Class: Not vulnerable	Class: Severely vulnerable	Class: Very vulnerable
Sensitivity	0.4984	0.33333	0.6923	0.0000
Specificity	0.6073	0.87875	0.7175	1.0000
Pos Pred Value	0.3383	0.48469	0.5011	NaN
Neg Pred Value	0.7504	0.79393	0.8505	0.8327
Prevalence	0.2871	0.25492	0.2907	0.1673
Detection Rate	0.1431	0.08497	0.2013	0.0000
Detection Prevalence	0.4231	0.17531	0.4016	0.0000
Balanced Accuracy	0.5529	0.60604	0.7049	0.5000

K-NN

Statistics by Class:

	Class: Moderately vulnerable	Class: Not vulnerable	Class: Severely vulnerable	Class: Very vulnerable
Sensitivity	0.4267	0.20000	0.4789	0.34375
Specificity	0.7095	0.74719	0.8092	0.89529
Pos Pred Value	0.4267	0.16667	0.5397	0.35484
Neg Pred Value	0.7095	0.78698	0.7687	0.89062
Prevalence	0.3363	0.20179	0.3184	0.14350
Detection Rate	0.1435	0.04036	0.1525	0.04933
Detection Prevalence	0.3363	0.24215	0.2825	0.13901
Balanced Accuracy	0.5681	0.47360	0.6440	0.61952

Model 2: Identifying food security vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.3345, 0.4091)
K-NN	(0.2536, 0.3792)
Decision trees	(0.4001, 0.459)

Based on the Accuracy CI we decided to go with the Decision tree model.

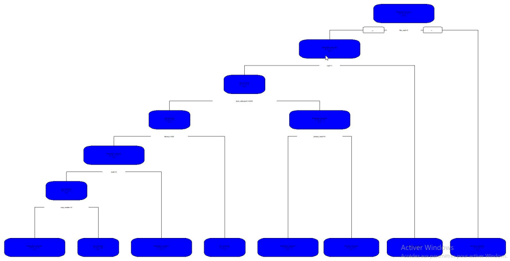
Model 2: Identifying food security vulnerability

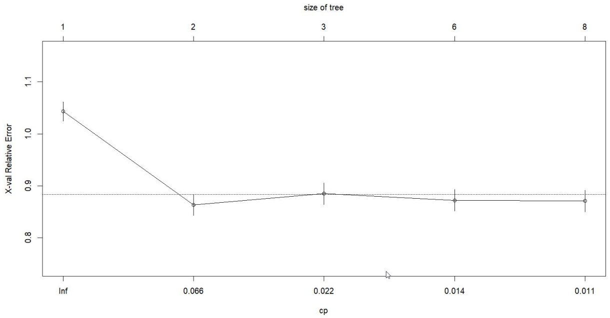
Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.3345, 0.4091)
K-NN	(0.2536, 0.3792)
Decision trees 	(0.4001, 0.459)

Based on the Accuracy CI we decided to go with the Decision tree model.

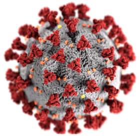
Decision tree visuals





Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic: **What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?**

Identifying the most vulnerable households towards food security: **What are the household profiles that are most likely to face food insecurity due to COVID ?**

Identifying the most vulnerable households towards education: **What are the household profiles in which children are more likely to drop school due to the pandemic ?**

Model 3: Identifying education access vulnerability

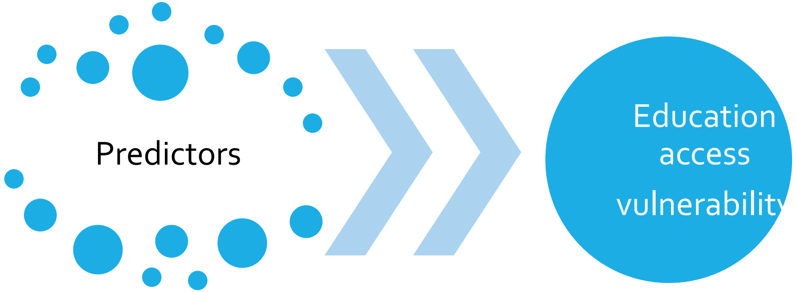
Defining the categories

Category	Value of the variable children_school_covid	Number of households in this category
The children of the households have continued learning activities after the pandemic	=1	1034
The children of the households have stopped learning activities after the pandemic	=2	699

Model 3: Identifying education access vulnerability

Within the LSMS dataset we chose the following predictors:

- Rural, roof, floor, walls, toilet,water,rooms,elect,tv,radio,refrigerator,land_tot,land_cultivated, rent, remit, assist, crop, crop_number, cash_crop, sell_crop, fies_mod, fies_sev, hh_size, adulteq, literacy, work, prop_primary, prop_secondary, prop_tertiary



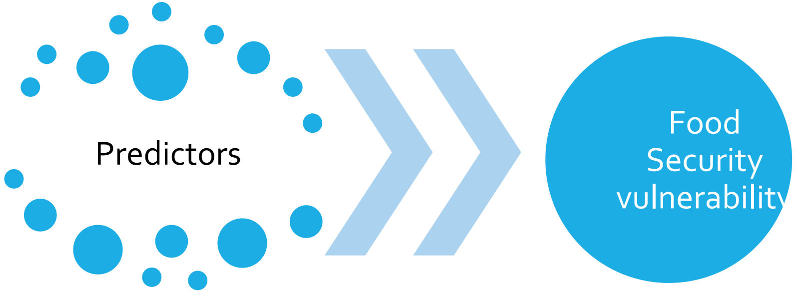
From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels towards education

Model 3: Identifying education access vulnerability

We tested 3 classification methodologies in order to select the most performant one:

- **Naives Bayes Classifier**
- **K-NN**



From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels to food insecurity

Model 3: Identifying education access vulnerability

Naive Bayes

```
    M_test_category
pr     1   2
  1 114  75
  2  92  66

               Accuracy : 0.5187
                 95% CI : (0.4648, 0.5724)
    No Information Rate : 0.5937
    P-Value [Acc > NIR] : 0.9980

                  Kappa : 0.0211

 Mcnemar's Test P-Value : 0.2157

            Sensitivity : 0.5534
            Specificity : 0.4681
         Pos Pred Value : 0.6032
         Neg Pred Value : 0.4177
             Prevalence : 0.5937
         Detection Rate : 0.3285
   Detection Prevalence : 0.5447
      Balanced Accuracy : 0.5107

       'Positive' Class : 1
```

K-NN

```
Confusion Matrix and Statistics

   y_pred
      1   2
  1 156 160
  2  68 136

               Accuracy : 0.5615
                 95% CI : (0.5177, 0.6047)
    No Information Rate : 0.5692
    P-Value [Acc > NIR] : 0.6555

                  Kappa : 0.1485

 Mcnemar's Test P-Value : 1.674e-09

            Sensitivity : 0.6964
            Specificity : 0.4595
         Pos Pred Value : 0.4937
         Neg Pred Value : 0.6667
             Prevalence : 0.4308
         Detection Rate : 0.3000
   Detection Prevalence : 0.6077
      Balanced Accuracy : 0.5779

       'Positive' Class : 1
```

Model 3: Identifying education access vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.5177, 0.6047)
K-NN	(0.4878, 0.5951)

Naive Bayes has a better accuracy CI but K-NN seems to detect better the cases of households whose children has stopped learning during COVID. In the logic of detecting vulnerability this is our priority: we will thus choose the K-NN model.

Model 3: Identifying education access vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.5177, 0.6047)
K-NN 	(0.4878, 0.5951)

Naive Bayes has a better accuracy CI but K-NN seems to detect better the cases of households whose children has stopped learning during COVID. In the logic of detecting vulnerability this is our priority: we will thus choose the K-NN model.

Integrated solution

- Combination of 3 models in order to predict the different categories regarding income, food security and education in which a given household is likely to fall in.
- <u>**Conclusion:**</u>
  - **For income and education access: K-NN model will be used**
  - **For food security: Decision tree model will be used**

Next step: *write an integrated script that takes any socio-economic dataset containing the predictors as arguments and that returns the categories predicted for the household income, education access and food security evolution with COVID-19.*

Application : Context

- TOUTON SA is a company specialized in soft commodities. The sustainability department of TOUTON manages several sustainability projets in sourcing countries (including Uganda, Ghana, Côte d'Ivoire, Kenya, Nigeria and Madagascar) aiming at helping farmers improving their income and livelihoods and requiring large scale data collection.
- TOUTON has collected data on a sample of 304 coffee farmers in Uganda on their livelihoods and agricultural practices. Several variables included in this survey have been used as predictors for our different prediction models.
- Therefore, with the consent of TOUTON SA, we have applied our different models that we developped with open source data to their coffee farmers datasets in order to assess their vulnerability to COVID regarding food security and their access to education.

Application : Cleaning and processing

STEP0: Getting all parties consent to use the data for visualisation only

STEP 1: Retrieving the predictors from the coffee farmer survey in Uganda

STEP 2: Cleaning the data and replacing missing values (using extrapolations)

STEP 3: Import the dataset in the integrated script and applying the 2 predicting models on income, food security and education access to the dataset

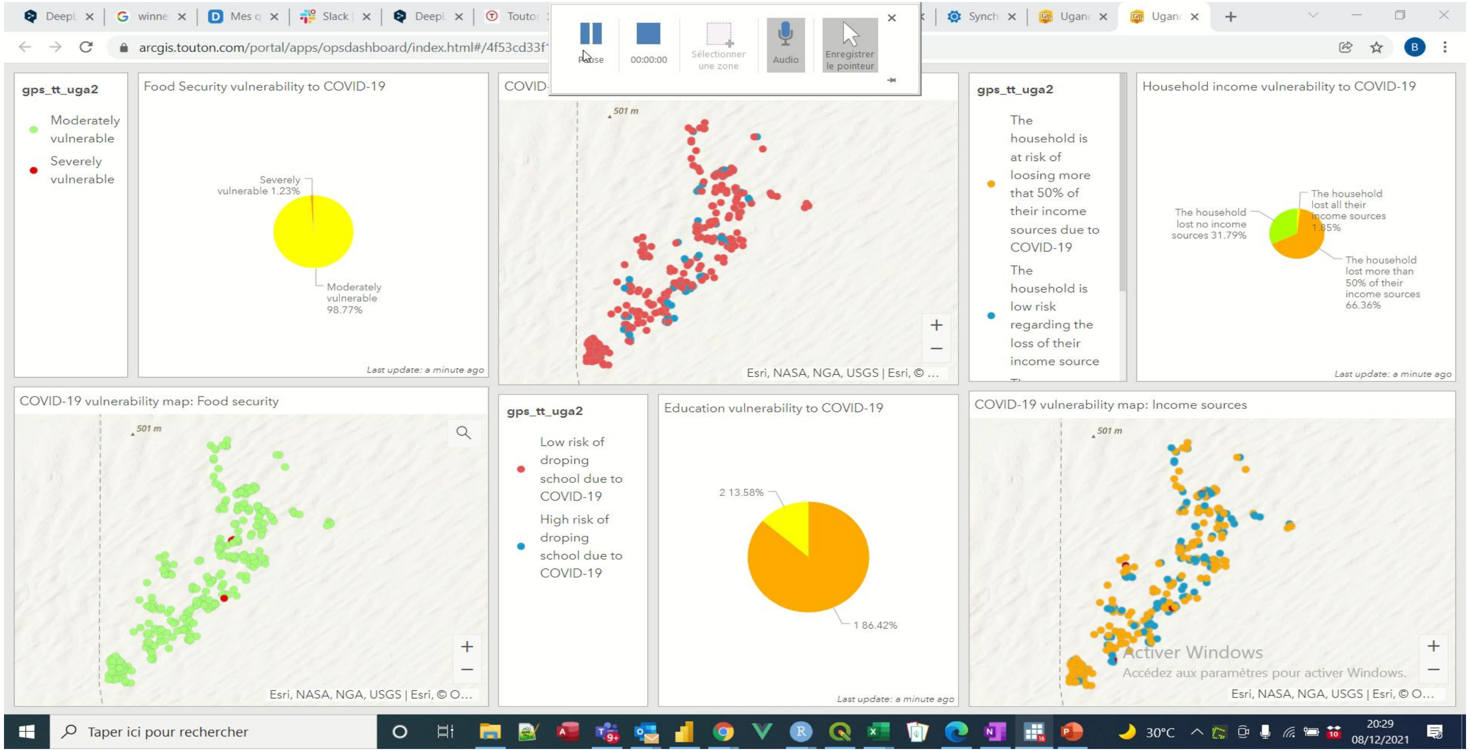
STEP 4: Creating a dataset containing the farmer ID as well as the 3 predictions. This dataset is the prediction dataset.

STEP 5: Merging the geospatial data on farmers with the « prediction dataset ».

STEP 6: Importing the data in Arcgis enterprise

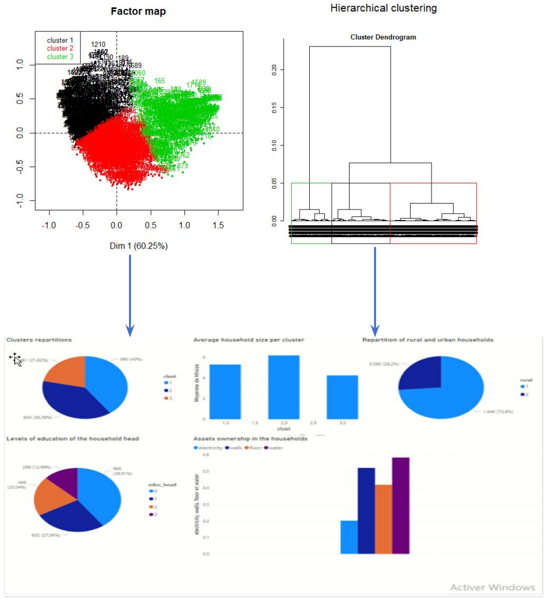
STEP 7: Building a « Vulnerability map dashboard » to visualise the results

Application : Visualizing coffee farmers that are the most vulnerable to COVID consequences

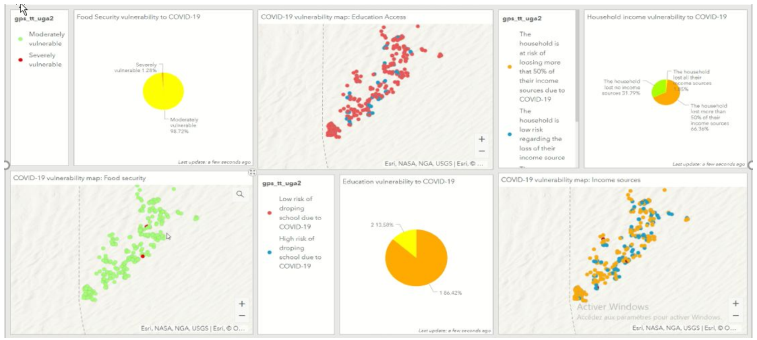


Conclusion: Our solution

A statistical segmentation to better understand the impact of a household socio-economic characteristics on their vulnerability to COVID-19 and their consequences.



A integrated prediction model in order to assess the vulnerability of households to COVID-19 regarding their income, food security and education access



Next steps: data science for vulnerability measurement

I/ Improving the accuracy of reliability of the model and broadening the methodology to more global vulnerability analysis

Why ?: Vulnerability measurement is key in sustainable development: predicting the ability of households to cope with any kinds of shocks.

How ?:

1. Mapping available socio-economic data on households: definition of key predictors based on a factor analysis of socio-economic factors on vulnerability.

2. Collecting data on households that faced a shock (e.g. climatic disaster, drought, pandemics etc.) in order to define more accurate predicted classes.

Next steps: data science for vulnerability measurement

II/ Applying the model in order to build evidence-based and tailor-made programs

STEP 1: Selecting a targeted group for an intervention

STEP 2: Collecting baseline data on the targeted group in order to calculate the different predictors of the model

STEP 3: Running the model on the collected predictors in order to identify the most vulnerable populations on the different project's area of intervention.

STEP 4: Running an impact assessment in order to assess the added value of a vulnerability-based approach for program implementation.

Annex 3: Data references

Data used to train the algorithm:

- LSMS dataset: https://microdata.worldbank.org/index.php/catalog/4183
- High Frequency Phone Survey on COVID-19: https://microdata.worldbank.org/index.php/catalog/3765

Data on which the model was applied:

- Uganda Socio-Economic Survey Coffee farmers: Touton Property

Experience from a Big Data Expert

Vladimir Gonçalves Miranda

Instituto Brasileiro de Geografia e Estatística

Use of web scraped data for price statistics at the Brazilian Institute of Geography and Statistics (IBGE)

Vladimir Miranda – IBGE
vladimir.miranda@ibge.gov.br

Survey Directorate – DPE
Price Indices Coordination – COINP/GPLACON

UNBigDataRegional Hub in Brazil

UN Big data Sources and Analysis webinar

October 10th, 2022

Instituto Brasileiro de Geografia e Estatística IBGE

IBGE

Airfares: automation of collection

(in colaboration with COMEQ/GDP)

Inputs

Outputs

For the CPIs, airfares used to be collected manually on the web by staff at the local units.

Inputs well defined (departure and arrival dates, for a given pair of cities and given profiles of tickets).

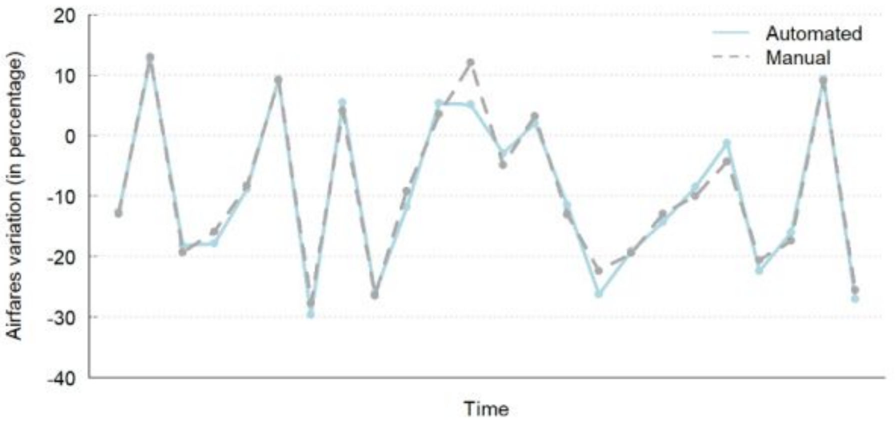
Monopolized marked is a key aspect here.

Airfares

Institudo Brasileiro de Geografia e Estatística IBGE

Scrapers developed in house for the companies in the sample.

Results of the comparison in the analysis phase.



Studies of new data sources and techniques to improve CPI compilation in Brazil, Lincoln Silva et al, paper presented at the Ottawa Group meeting in 2019.

Running in production since january 2020.

Save efforts for the collection of up to 100.000 prices a month.

Ride sharing services: coverage improvement

New challenges for CPI compilers with advent of digital services.

Some results of the last POF (HBS)

Area	IPCA		INPC	
	Taxi	Ride sharing Services	Taxi	Ride sharing Services
BR	**0,21**	**0,21**	**0,16**	**0,15**
AC	0,54	-	0,55	0,07
PA	0,43	-	0,32	-
MA	0,32	0,11	0,41	0,15
CE	0,18	0,15	0,15	0,16
PE	0,30	0,32	0,15	0,28
SE	0,58	0,11	0,53	0,17
BA	0,38	0,30	0,19	0,21
MG	0,24	0,19	0,17	0,16
ES	0,12	0,10	-	0,09
RJ	0,45	0,31	0,20	0,26
SP	0,16	0,20	0,11	0,12
RS	0,26	0,38	0,20	0,27
MS	0,09	0,23	-	0,28
GO	-	0,26	-	0,09
DF	-	0,25	0,11	0,16

Challenges: what to collect, when and how?

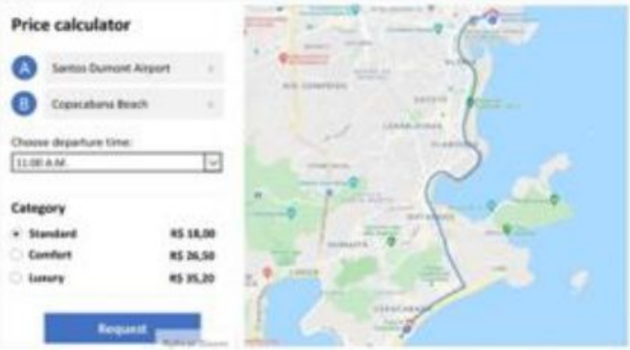
Price components of the service:

- "Rigid" components

  Base rates: per km rates
  Booking fees

- "Flexible" component

  Dynamic multiplier



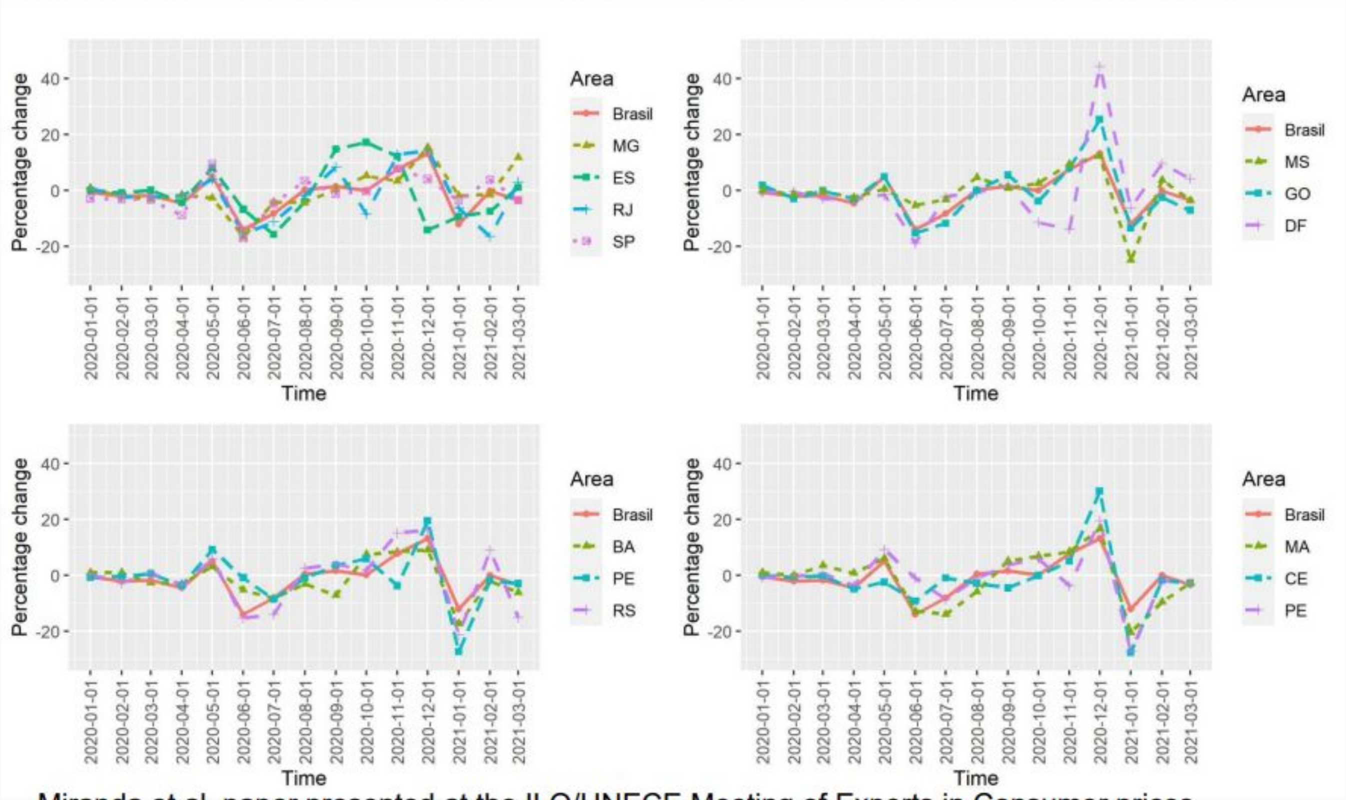
IBGE

Instituto Brasileiro de Geografia e Estatística IBGE

Ride sharing services

Running in production since january 2020.

Results can capture geographical nuances and price dynamics in a timely manner.



Miranda et al, paper presented at the ILO/UNECE Meeting of Experts in Consumer prices Indices Group 2021.

Methodological improvements: Household appliances and electronics

Products attributes and prices can be scraped on web sites

Geladeira/Refrigerador Frost Free cor Inox 310L Electrolux (TF39S) 127V
Marca: Electrolux
24 avaliações de clientes
R$2.804,00
Em até 10x R$ 280,40 sem juros Ver parcelas disponíveis

Total Capacity	24.52 cubic feet
Refrigerator Style	Side-by-Side
Ice Maker	Yes
Lighting Type	LED
Color Finish	Stainless steel

Example of model fit and output:

$$log(\mathrm{Pr}) = \beta_0 + \beta_1\mathrm{Br} + \beta_2\mathrm{Col} + \beta_3\mathrm{Sty} + \beta_4\mathrm{Defr} + \beta_5\mathrm{Cap} + \beta_6\mathrm{Shop}$$

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.592e+00  2.905e-02 226.935  < 2e-16 ***
BrConsul       -1.619e-01  1.486e-02 -10.896  < 2e-16 ***
BrElectrolux   -4.476e-02  1.106e-02  -4.046 5.78e-05 ***
ColInox         1.003e-01  1.126e-02   8.909  < 2e-16 ***
StyDuplex       1.166e-01  1.717e-02   6.791 2.35e-11 ***
StyInverse      2.210e-01  2.212e-02   9.991  < 2e-16 ***
DefrFrost Free  1.615e-01  1.045e-02  15.445  < 2e-16 ***
Cap             2.684e-03  6.284e-05  42.707  < 2e-16 ***
ShopOnline     -1.094e-01  8.593e-03 -12.736  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1001 on 713 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8832
F-statistic: 682.5 on 8 and 713 DF,  p-value: < 2.2e-16
```

Studies of new data sources and techniques to improve CPI compilation in Brazil, Lincoln Silva et al, paper presented at the Ottawa Group meeting in 2019.

Quality adjustment: Household appliances and electronics

IBGE

Instituto Brasileiro de Geografia e Estatística | IBGE

Evolution of products along time. How to get pure price change?

Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n				p_n^{t+3}	p_n^{t+4}

Direct comparison may lead to bias.

$$R_n^{t+3,t+2} = p_n^{t+3}/p_m^{t+2}$$

Quality adjustment: Household appliances and electronics

Use of hedonic regression models to deal with this

$$p = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + \epsilon$$

Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n			$\hat{p}_n^{t+2}$	p_n^{t+3}	p_n^{t+4}

Comparison after the adjustment

$$R_n^{t+3,t+2} = p_n^{t+3} / \hat{p}_n^{t+2}$$

Instituto Brasileiro de Geografia e Estatística IBGE

Other price statistics: ICP program

Make use of prices of a list of a catalogue of products (goods and services) sent to the countries to build the PPP indicators.

110911114.LAC - TV 40 pulgadas, SAMSUNG

Lista Regional : **Si**
Lista Global : **No**

Cantidad de referencia	1
Unidad de medida	Pieza
Marca	SAMSUNG
Tipo	Televisor de pantalla plana LED
Modelo	Especificar
Tamaño de la pantalla	40 / 101 cm
Resolución de pantalla	Full HD 1080p
Conectividad	HDMI, USB, WiFi, Ethernet
Excluir	Modelos 4k o 3D, televisores curvos
Especificar	Marca, Modelo

DESC_COD_PROD_PCI
Detergente en polvo, lavadora, MC / Laundry detergent powder, washing machine, WKB
Limpiador doméstico de uso múltiple, MC / All-purposes household cleaner, WKB
Limpiador doméstico de uso múltiple, MC / All-purposes household cleaner, WKB
Rollo de papel de cocina, SM / Kitchen paper roll, BL
Servilleta de papel, MC / Paper napkins, WKB
Insecticida spray, MC / Insecticide spray, WKB
Velas o candelas, caja, SM / Household candles, box, BL
Detergente de lavavajillas, MC / Dishwashing detergent, WKB
Microondas, MC-B / Microwave oven, WKB-L

Other price statistics : ICP program

For some goods, we have a pilot colecting products. Use of sites search engines for scraping.

Search based on the product descriptions previously inspected

Target list at retailer A
Pineapple (abacaxi) - unit
Chocolate cookies, brand A, 100g
Diet soda, brand B, 2L
Shampoo, brand C, 200ml
...

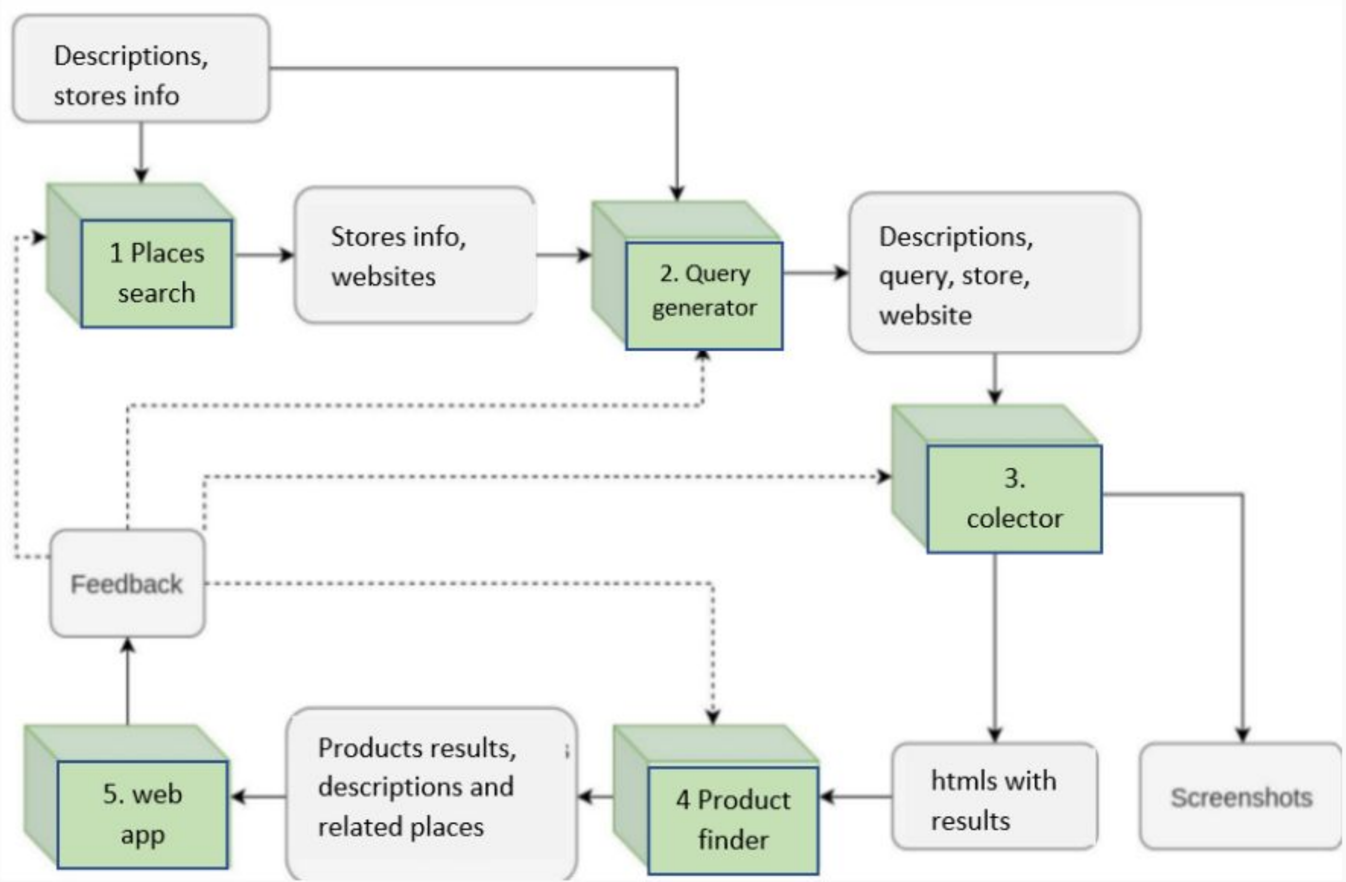
Store	Target Product	Product returned
Retailer A	Abacaxi - Unidade	Abacaxi Perola unidade
Retailer A	Abacaxi - Unidade	Abacaxi desidratado pacote 55g

Use of keywords for products selection and manual validation of the results.

Products of different sectors collected.

IBGE
Instituto Brasileiro de Geografia e Estatística | IBGE

Generic scraper
(in colaboration with UFMG researchers)



IBGE
Instituto Brasileiro de Geografia e Estatística | IBGE

Hotels: increasing the complexity
(in colaboration with COMEQ/GDP)

Important differences

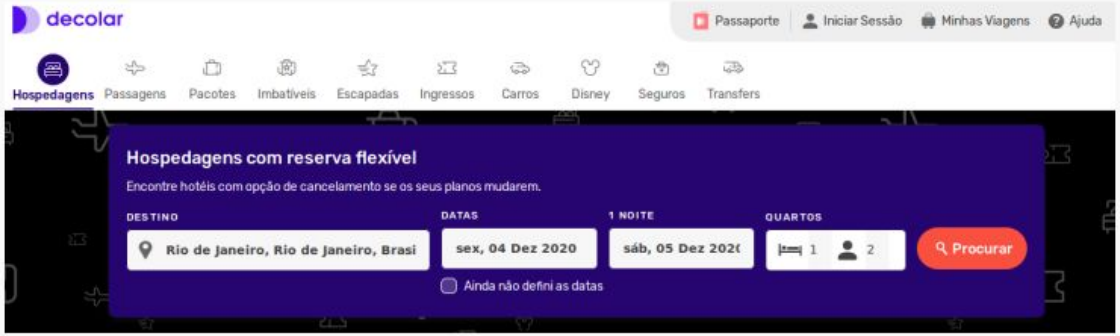
i) Source change (hotels to web sites).

Traditional collection performed during in-person visits to the hotels.

ii) Nonmonopolized marked

Large number of hotels in the samples. Each would have a given site when available.

Possble strategy, use of Booking aggregating sites.



Used cars

Hyundai HB20 usados Niterói - RJ e cidades até 50km (411 ofertas)

Ofertas Relacionadas: Chevrolet Onix | Volkswagen Gol | Fiat Palio | Chevrolet Prisma

Filtro Busca

Localização

Pesquisar em

rio de janeiro

Ver anúncios até

50 km

Veículo

Digite uma marca, modelo ou versão

ex: Fiesta, Nissan

Atenção! Verifique as condições de pagamento e demais informações do veículo diretamente com o anunciante.
Nunca faça depósitos ou pagamentos antes de se certificar da existência do veículo e desconfie de ofertas com o preço muito abaixo do mercado.

HB20 1.0 Vision (BlueAudio) - 2022

R$ 68.990,00

HB20 1.0 Copa do Mundo - 2015

R$ 45.900,00

HB20 1.6 Comfort - 2013

R$ 41.900,00

HB20 1.0 Unique - 2019

R$ 56.900,00

Ordenar: Destaques

1 2 3 4 5 6 7 >

Características

Geral

Transmissão	Automático	**Tração**	4x2	**Final da Placa**
Estacionado em	C-60	**Stock ID**	193990	

Exterior

Faróis	Faróis Halógenos	**Material de aro**	Alumínio

Appealing characteristics:

i) Existence of marketplace sites offer the possibility to use web scraping here also.

ii) Possibility of use of hedonics for quality adjustment.

iii) Also info on new cars.

Instituto Brasileiro de Geografia e Estatística IBGE

IBGE

Thank you for your attention!

vladimir.miranda@ibge.gov.br

Q&A

Do you have additional questions?

un-big-data-hackathon@unmgcy.org

Follow us on:

- Twitter: @unbigdatahackathon
- Instagram: @unbigdatahack
- Facebook: @unbigdatahackathon