World - Synthetic Data for an Imaginary Country, Sample, 2023

Development Data Group, Data Analytics Unit

Report generated on: November 1, 2024

Visit our data catalog at: https://nada-demo.ihsn.org/index.php

Identification

SURVEY ID NUMBER
WLD_2023_SYNTH-SVY-EN_v01_M

TITLE
Synthetic Data for an Imaginary Country, Sample, 2023

SUBTITLE
A synthetic hierarchical dataset for simulation and training purposes

COUNTRY

Name	Country code
World	WLD

STUDY TYPE
Synthetic data

SERIES INFORMATION
This dataset is part of a collection of fully synthetic data generated, for training and simulation purposes, for an imaginary middle-income country. The dataset is available in English and French. A full population dataset (~10 million individuals) is also available in English and French as a "synthetic census dataset".

ABSTRACT
The dataset is a relational dataset of 8,000 households households, representing a sample of the population of an imaginary middle-income country. The dataset contains two data files: one with variables at the household level, the other one with variables at the individual level. It includes variables that are typically collected in population censuses (demography, education, occupation, dwelling characteristics, fertility, mortality, and migration) and in household surveys (household expenditure, anthropometric data for children, assets ownership). The data only includes ordinary households (no community households). The dataset was created using REaLTabFormer, a model that leverages deep learning methods. The dataset was created for the purpose of training and simulation and is not intended to be representative of any specific country.

The full-population dataset (with about 10 million individuals) is also distributed as open data.

KIND OF DATA
ssd

UNIT OF ANALYSIS
Household, Individual

Version

VERSION DESCRIPTION
V. 2023-05-01 8K HH EN

VERSION DATE
2023-05-01T04:00:00.000Z

VERSION NOTES
Dataset generated using RealTabFormer (sample of 8,000 households), with post-processing. English version.

Scope

KEYWORDS

Keyword

synthetic data
open data
safe data
demographics
education
mortality
fertility
child malnutrition
labor, employment
housing
dwelling
water and sanitation
household expenditure
migration

Coverage

GEOGRAPHIC COVERAGE
The dataset is a synthetic dataset for an imaginary country. It was created to represent the population of this country by province (equivalent to admin1) and by urban/rural areas of residence.

GEOGRAPHIC UNIT
province (admin1), district (admin2)

UNIVERSE
The dataset is a fully-synthetic dataset representative of the resident population of ordinary households for an imaginary middle-income country.

Producers and sponsors

PRIMARY INVESTIGATORS

Name	Affiliation
Development Data Group, Data Analytics Unit	World Bank

FUNDING AGENCY/SPONSOR

Name	Role
UNHCR-World Bank Joint Data Center on Forced Displacement	Sponsored research work for the development of synthetic data for the purpose of assessing statistical disclosure risk measures.

Sampling

SAMPLING PROCEDURE
The sample size was set to 8,000 households. The fixed number of households to be selected from each enumeration area was set to 25. In a first stage, the number of enumeration areas to be selected in each stratum was calculated, proportional to the size of each stratum (stratification by geo_1 and urban/rural). Then 25 households were randomly selected within each enumeration area. The R script used to draw the sample is provided as an external resource.

RESPONSE RATE
This is a synthetic dataset; the "response rate" is 100%.

WEIGHTING
Sample weights were calculated that take the stratification into account. See the R script provided as an external resource.

Data collection

DATES OF DATA COLLECTION

Start	End
2023	2023

TIME PERIODS

Start date	End date
2023	2023

DATA COLLECTION MODE
other

Questionnaires

QUESTIONNAIRES
The dataset is a synthetic dataset. Although the variables it contains are variables typically collected from sample surveys or population censuses, no questionnaire is available for this dataset. A "fake" questionnaire was however created for the sample dataset extracted from this dataset, to be used as training material.

Data Processing

DATA EDITING
The synthetic data generation process included a set of "validators" (consistency checks, based on which synthetic observation were assessed and rejected/replaced when needed). Also, some post-processing was applied to the data to result in the distributed data files.

METHODOLOGY NOTES
The dataset was generated using REaLTabFormer, a four-level hierarchical generative model. The first-level model is the household composition generator, which generates variables that define each household's composition (household size and basic demographic profile of members, including age and relationship to the head of household). The second-level model is the household-level variables generator, which generates the variables whose values are common to all household members (such as dwelling characteristics) based on the household composition. The third-level model is the household-head generator, which generates observations for the head of the households based on the output of the previous two models. The fourth-level model is the household member generator, which generates data on the household members, excluding the head, for households of size two and above. The household member generator model uses the data generated by the household composition, household-level variables, and household head generator models. This hierarchical model provides relational dependencies within a household that would not be guaranteed if all records were generated independently.

To implement the different models, we adopted a transformer architecture. The household composition generator is a decoder model that generates data from normally distributed noise. The other three models use a sequence-to-sequence model inspired by the application of deep learning to language translation.

More detailed information is available in the Technical Documentation provided as an external PDF document.

Access policy

RESTRICTIONS
The dataset was generated as a fully-synthetic dataset. The model used to create the synthetic observations includes multiple procedures to avoid overfitting and data-copying. Also, the data used for training the model went through processes of sampling and recoding that make it impossible to link a synthetic observation to an actual observation. The dataset is thus safe for dissemination. It can be used with no restriction and is shared as open data.

ACCESS AUTHORITY

Name
World Bank, Microdata Library

LOCATION OF DATA COLLECTION
World Bank Microdata Library

Disclaimer and copyrights

DISCLAIMER
The data are to be used for training or simulation purposes only. It is not intended to be representative of any particular country, and should not be used for inference purpose.

Metadata production

PRODUCERS

Name	Affiliation
OD	World Bank

DATE OF METADATA PRODUCTION
2023-05-01T04:00:00.000Z

DDI DOCUMENT VERSION
1.0 EN

Data Dictionary

Data file	Cases	Variables
WLD_2023_SYNTH-SVY-IND-EN_v01_M	32396	27

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Cases:	32396
Variables:	27

Variables

ID	Name	Label	Question
V1	hid	Unique household identifier	
V2	idno	Person identification number	
V3	relation	Relationship to the head of household	
V4	sex	Sex	
V5	age	Age in years	
V6	age_month	Age in months	
V7	marstat	Marital status	
V8	religion	Religion	
V9	school_attend	School attendance	
V10	educ_attain	Education attainment	
V11	yrs_school	Years of schooling	
V12	literacy	Literacy status	
V13	act_status	Activity status	
V14	labor_force	Labor force status	
V15	occupation	Main occupation	
V16	industry	Industry of main occupation	
V17	migrate_recent	Recent migration	
V18	disability	Has a disability	
V19	blind	Blind	
V20	deaf	Deaf	
V21	mental	Mental disability	
V22	ch_weight	Child weight (for 0 to 59 months)	
V23	ch_height	Child height (for 0 to 59 months)	
V24	children_born	Number of children ever born	
V25	children_surv	Number of surviving children	
V26	births_12m	Number of births in past 12 months	
V27	hhweight	Household weight	

Total: 27

HID: Unique household identifier

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Valid: 32396 Minimum: None Maximum: None Mean: None Standard deviation: None
Type: Discrete Width: 12 Format: character

EDUC_ATTAIN: Education attainment

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 37 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	NIU (not in universe) or no education
1	Less than primary
2	Primary
3	Secondary
4	University

IDNO: Person identification number

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Valid: 32396 Minimum: 1.0 Maximum: 22.0 Mean: 3.087510803802939 Standard deviation: 2.0475208479570877
Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category	Cases	
1		8000	24.7%
2		7273	22.5%
3		6099	18.8%
4		4522	14%
5		2795	8.6%
6		1593	4.9%

7		945	2.9%
8		531	1.6%
9		264	0.8%
10		155	0.5%
11		83	0.3%
12		48	0.1%
13		31	0.1%
14		22	0.1%
15		17	0.1%
16		7	0%
17		5	0%
18		2	0%
19		1	0%
20		1	0%
21		1	0%
22		1	0%

RELATION: Relationship to the head of household

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Valid: 32396 Minimum: 1.0 Maximum: 5.0 Mean: 2.474225212989258 Standard deviation: 1.0295147118123782 Mean (weighted): 2.4726299551289 Standard deviation (weighted): 1.0328668729346
Type: Discrete Width: 14 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category	Cases	Weighted	
1	Head	8000	2501755	24.8%
2	Spouse/partner	5848	1821181	18.1%
3	Child	14047	4344296	43.1%
4	Other relative	4187	1302445	12.9%
5	Non-relative	314	104668	1%

SEX: Sex

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Valid: 32396 Minimum: 1.0 Maximum: 2.0 Mean: 1.505309297444129 Standard deviation:

0.49997952731167966
Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category	Cases	
1	Male	16026	49.5%
2	Female	16370	50.5%

AGE: Age in years

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	

20	
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
57	
58	

59	
60	
61	
62	
63	
64	
65	
66	
67	
68	
69	
70	
71	
72	
73	
74	
75	
76	
77	
78	
79	
80	
81	
82	
83	
84	
85	
86	
87	
88	
89	
90	
91	
92	
93	
94	
95	
97	
98	

99	
100	

AGE_MONTH: Age in months

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	

26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
57	
58	
59	

MARSTAT: Marital status

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 20 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
1	Single/never married
2	Married/in union
3	Divorced/separated
4	Widowed

RELIGION: Religion

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
1	No religion
2	Religion A
3	Religion B
5	Religion D
6	Religion E
7	Other

SCHOOL_ATTEND: School attendance

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 34 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
1	Yes

2	No, never attended
3	No, attended in the past
4	No, not specified if ever attended

YRS_SCHOOL: Years of schooling

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	

LITERACY: Literacy status

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 15 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	Not in universe
1	Yes
2	No

ACT_STATUS: Activity status

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 21 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	NIU (not in universe)
1	Employed
2	Unemployed
3	Inactive

LABOR_FORCE: Labor force status

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 15 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	Not in universe
1	Yes
2	No

OCCUPATION: Main occupation

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 39 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	Not in universe
1	Legislators, senior officials and manag
2	Professionals
3	Technicians and associate professionals
4	Clerks
5	Service workers and shop and market sal
6	Skilled agricultural and fishery worker
7	Crafts and related trades workers
8	Plant and machine operators and assembl
9	Elementary occupations
10	Armed forces
11	Other occupations, unspecified or n.e.c

INDUSTRY: Industry of main occupation

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 39 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	NIU (not in universe)
1	Agriculture, fishing, and forestry
2	Mining and extraction
3	Manufacturing
4	Electricity, gas, water and waste manag
5	Construction
6	Wholesale and retail trade

7	Hotels and restaurants
8	Transportation, storage, and communicat
9	Financial services and insurance
10	Public administration and defense
11	Business services and real estate
12	Education
13	Health and social work
14	Other services
15	Private household services

MIGRATE_RECENT: Recent migration

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 39 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	NIU (not in universe)
10	Same major administrative unit
11	Same major, same minor administrative u
12	Same major, different minor administrat
20	Different major administrative unit
30	Abroad
99	Unknown/missing

DISABILITY: Has a disability

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 14 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	No disability

1	Has disability

BLIND: Blind

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	No
1	Yes

DEAF: Deaf

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	No
1	Yes

MENTAL: Mental disability

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	No

1	Yes

CH_WEIGHT: Child weight (for 0 to 59 months)

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Continuous Width: 10 Format: Numeric

CH_HEIGHT: Child height (for 0 to 59 months)

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Continuous Width: 10 Format: Numeric

CHILDREN_BORN: Number of children ever born

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	

15	
16	
18	

CHILDREN_SURV: Number of surviving children

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	

BIRTHS_12M: Number of births in past 12 months

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Type: Discrete Width: 12 Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
0	
1	
2	

HHWEIGHT: Household weight

Data file: WLD_2023_SYNTH-SVY-IND-EN_v01_M

Overview

Valid: 32396 Minimum: 156.66755315036724 Maximum: 552.2308512356647 Mean: 310.97503364852963
Standard deviation: 70.74169260405742
Type: Continuous Width: 12 Format: Numeric Weighted variable: 1

Download related resources

Questionnaires

Fake questionnaire and survey information

Title	Fake questionnaire and survey information
Date	2023-05
Description	A fake questionnaire and some additional information corresponding to the variables included in the synthetic dataset, intended to be used as training material only. Contains this information in English and French.
Filename	synthetic_survey_questionnaire_info.xlsx

Technical documents

Generating a relational synthetic dataset for an imaginary country - Technical documentation

Title	Generating a relational synthetic dataset for an imaginary country - Technical documentation
Author(s)	Olivier Dupriez and Aivin V. Solatorio
Date	2023-06
Description	A technical description of the process of generating the synthetic dataset.
Filename	synthetic_data_technical_documentation.pdf

Other materials

REaLTabFormer GitHub repository

Title	REaLTabFormer GitHub repository
Author(s)	Aivin V. Solatorio
Description	GitHub repository for the synthetic data generation models, available openly.
Filename	https://github.com/avsolatorio/REaLTabFormer

Sample selection R script

Title	Sample selection R script
Author(s)	Thijs Benschop
Date	2023-05
Description	A R script used to extract the sample of 8,000 households from the full synthetic dataset.
Filename	20230505_draw_sample.R