INITIATIVES UNDERTAKEN BY ESCWA FOR CAPTURING DATA THROUGH ALTERNATIVE DATA SOURCES

ROAD SAFETY ANALYTICS USING ALTERNATIVE DATA

▶ **The pilot study aims at understanding the factors that cause lethal car crashes using official data and complimented by alternative data**

Official Data	Consists of <u>digitized</u> police records. They contain the results of questionnaires filled out by the police at crash sites. **This data can be acquired through a request to police officials**
Alternative data	• Data on temperature, visibility, and precipitation from different weather stations. **The data is available online on weather reporting websites**
	• Data on shape of streets and their speed limits **which are available on crowd sourced mapping initiatives on websites (such as open street maps).**
	• Data on road quality. This data can come in the form of digital maintenance records (in spreadsheet format), or as quality ratings set by government specialists for each road type. **This data is available by request from the ministry of transportation or public works**

ROAD SAFETY ANALYTICS USING ALTERNATIVE DATA

The pilot study was attempted in *Lebanon*, *Dubai*, and *Jordan*

Main Challenge	Lack of access to both official and alternative data from governments
Solution	Study was tested using car crash dataset from UK's open data platform and data from open street maps (online)
Methodology	Using Machine Learning algorithms to analyze the data. These algorithms are sets of programming code that go through the data to learn patterns and execute similar processes later on
Methodology Result	Results of the algorithms will provide insights into the factors causing lethal car accidents and how much do alternative data contribute to the overall insight generated

The study will then be implemented on a selected ESCWA member country

ANALYZING AND UNDERSTANDING SYRIAN REFUGEES & HOST COMMUNITIES BEHAVIORS IN LEBANON USING HUMANITARIAN DATA SOURCES

Project aims at understanding refugee characteristics and behaviors by analyzing cell phone records and existing data on refugees in Lebanon

Official Data

Call Detail Records (CDR) that are stored in a database by cell phone service providers. Information such as who the Caller ID, the person receiving the call, call duration, and the location of call are available

Alternative data

- *Vulnerability Assessment of Syrian Refugees in Lebanon (VASYR) by UNHCR* which assesses a representative sample of Syrian refugee families to identify changes and trends in their situation.

- *Labor Force and Household Conditions Survey* which produces estimates of Labor force and living conditions at the national, governorate, and subnational district levels

ANALYZING AND UNDERSTANDING SYRIAN REFUGEES & HOST COMMUNITIES BEHAVIORS IN LEBANON USING HUMANITARIAN DATA SOURCES

To distinguish between refugee and non-refugee, phone numbers provided by the UN to Syrian refugees will be targeted

Main Challenge

Acquiring the data especially given the confidential nature of the data itself:

- At this point, data on CDR by cell providers was requested but has not been received.

- Vulnerability Assessment of Syrian Refugees in Lebanon has been received

ANALYZING AND UNDERSTANDING SYRIAN REFUGEES & HOST COMMUNITIES BEHAVIORS IN LEBANON USING HUMANITARIAN DATA SOURCES

Methodology

- Use CDR to calculate the number of refugees in a certain city/district (based on refugee calls from coverage towers)

- Use indices to predict the educational/labor/economic status of the refugees:

  1. *Social Integration* which measures the extent to which refugees communicate (through phone calls) with non-refugees

  2. *Spatial Integration* which measures the extent to which refugees are divided among areas and populations

  3. *Encounter Index* which measures the extent to which refugees are likely to encounter non-refugees

  4. *Economic Index* which predicts employment based on the time period the refugee is in a certain location during work hours

We expect to provide a snapshot into refugee behavior in Lebanon using a predefined set of metrics and indicators in the near term

YELLOW PAGES

The study aims at utilizing data from Yellow Pages to map business activities Lebanon to the International Standard Industrial Classification of All Economic Activities (ISIC) and enable descriptive analysis

Yellow Pages is an online business directory that gathers and stores information on businesses, their activities, location, and contact information

The data will be acquired through direct contact with the representatives of Yellow Pages

Main Challenge

Use the information available on businesses to properly classify their activities that fall into multiple ISIC categories

YELLOW PAGES

Methodology

- Before requesting the official data from Yellow Pages, proxy data; data similar in nature to the data that will be received; is being used to determine what is needed for effective classification of businesses

- Machine Learning Processes and programming languages are used for classification

We expect to be able to classify businesses according to ISIC and leverage the classified businesses and the additional data (such as geo-locations) to extract insights from descriptive statistics on both the formal and informal economies in near term

WEB SCRAPING

Web scraping is a process employed to **automatically** extract large amounts of data from websites, whereby the data extracted is saved to a local file or to a database in spreadsheet format

USES

- *Helps acquire data from multiple sources in noticeably shorter period of time*

- *Keeps track to any online changes in data*

- *Aids in data archiving*

CHALLENGES

1. Web designs are always evolving, making it harder to label data for scraping
2. Certain websites may prohibit data extraction or use of bots fearing increase in traffic
3. Websites that use CAPTHCHA or similar anti-bot systems

SOLUTIONS

1. Customize and update the script to handle complex web designs
2. Web scrape data during low traffic
3. Unfortunately, no procedure is set to solve such issue

WEB SCRAPING

The ICP in Western Asia Region has piloted web scraping in Bahrain in early January 2020

Aim

Extract <u>Names</u> and <u>Price</u> data for Fast Evolving Technology (FET) Electronics

Methodology

- Online outlets (URLs) in Bahrain were identified and grouped
- Web scraping scripts for these items were coded using *Python* programming language

Piloting Result

Participants in Bahrain were able to implement web scraping to extract Names and Price data from online outlets in their country.

Piloting online data extraction through Web Scraping in the Arab Region will be carried on to reach Kuwait and Qatar respectively

SCANNER DATA

Scanner data is a detailed data on sales of consumer goods obtained by scanning the bar codes for individual products at electronic points of sale in retail outlets.

BENEFITS

- **Provides information on the actual expenditure for all item codes sold**
- **Provides high quality information of actual transactions**
- **Data can be used for several purposes such as PPPs, detailed average prices, and advanced economic analysis**

CHALLENGES

1. Obtaining the approval from the retailers on giving their data to the statistical offices
2. Risk that the retailers might not be able to deliver the necessary data on a timely basis
3. Difficulties with processing the huge amount of data

SOLUTIONS

1. Assure the confidentiality of the received data and its importance to the work of the statistical offices
2. Set agreements with clear rules and conditions and establish automatic data-transfer link with outlets
3. Appropriate IT structure and trained staff need to be available

SCANNER DATA

National Centre for Statistics & Information in Oman has implemented alternative data collection method through scanner data

Aim

- Develop an IT infrastructure to link CPI with the retail outlets' databases
- Establish a link (key) between sold items barcode and CPI items

Methodology

The IT infrastructure and Link (key) required a proprietary (ready-made) software from an outsourced company and a team of IT and Statistics specialists

Result

Oman was able to develop the IT infrastructure to collect scanner data related to items prices

CHALLENGES

1. Convince and establish agreements with retail outlets
2. Retail outlets had different and unique IT frameworks, which means each outlet needed a special IT linking system
3. Identifying and selecting needed items from a pool of products

Benefits

1. Reduced workload and field visits
2. Increased data quality and accuracy
3. Automated and timely data acquisition

THANK YOU