UNSD's Big Data Pilot Project Initiatives

UNECE Workshop on Statistical Data Collection

Washington, DC 29 April – 1 May 2015

United Nations Statistics Division

Nancy Snyder, Statistician, International Merchandise Trade Statistics Section

Using Big Data for Official Statistics: UNSD Pilot Projects

- **Linking postal data to trade statistics**
  - ***Costa Rica***
  - **Universal Postal Union**

- **Mobile positioning data for tourism & migration statistics**
  - ***Estonia***
  - ***Positium***
  - ***Oman***

- **Detailed global merchandise trade statistics**

UN COMTRADE

Costa Rica: Study of correlation between postal data and linked trade-business statistics

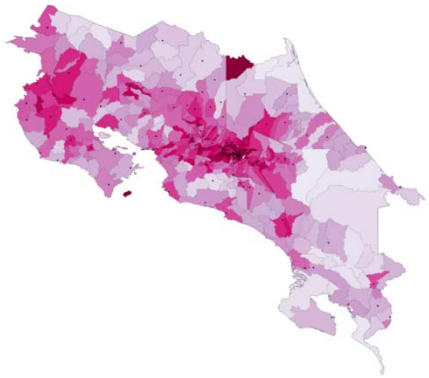
CUSTOMS RECORDS:

- Value & quantity of exports and imports
- Commodity
- Country of import/export
- Date
- Weight
- Identity of importing/exporting company

BUSINESS INFORMATION:

From the Business Register, Business surveys:

- ID#
- Name, address
- Unit (establishment, enterprise, etc.)
- Industry sector
- Employees, revenues, etc.



POSTAL DATA:

Local and international:

- Name
- Exact address
- Weight
- Date

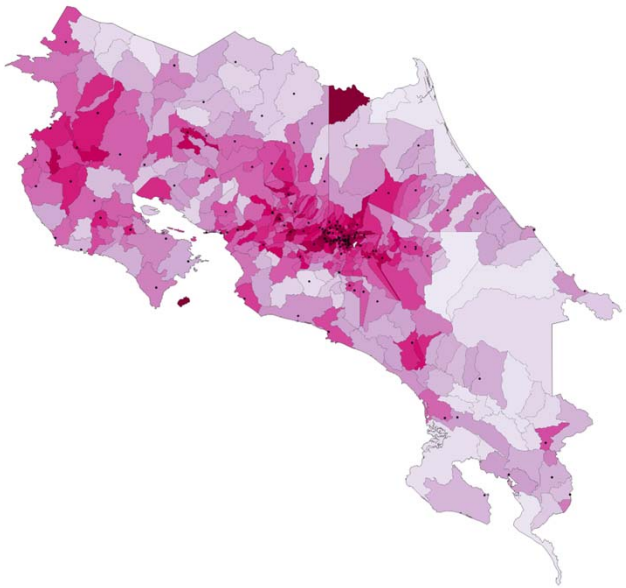
Costa Rica: Results - Trade by Enterprise Analysis

	Main activity	All Business Register		Exporting enterprise		
		No. of firms	No. of employees	No. of firms	No. of employees	Value of exports (US$)
A	**Agriculture, forestry and fishing**	**3,791**	**87,982**	**294**	**40,283**	**216,690,448**
01	Crop and animal production	3,608	85,069	279	38,698	210,146,426
02	Forestry and logging	132	1,567	8	688	5,289,339
C	**Manufacturing**	**4,257**	**144,706**	**598**	**106,429**	**2,554,099,671**
10	Manufacture of food products	1,078	48,328	137	34,661	447,071,957
14	Manufacture of wearing apparel	559	7,949	25	5,022	42,878,988
25	Manufacture of fabricated metal products	441	6,043	39	2,497	42,358,757
26	Manufacture of computer, electronics	40	6,531	21	6,369	787,049,178
27	Manufacture of electrical equipment	37	4,477	20	4,204	143,662,034
G+H	**Wholesale and retail trade; Transport, warehousing, and support activities**	**18,668**	**201,935**	**739**	**55,190**	**387,741,850**
46	Wholesale trade, except motor vehicles	2,471	58,966	591	30,086	349,309,112
47	Retail trade, except of motor vehicles	11,663	88,411	60	17,370	25,826,992
	Other activities	**19,277**	**318,907**	**186**	**33,792**	**63,137,508**
	Hotel and Restaurants	4,892				
	Professional and scientific services	2,472				
	Social and Health services	1,871				
	Unkown activity	**2,988**	**15,771**	**71**	**864**	**24,375,886**
	Non-matches	-	-	-	-	**131,689,763**
	TOTAL	**48,981**	**769,301**	**1,894**	**236,794**	**3,378,826,643**

Costa Rica: Merging Trade & Business Statistics with Postal Data

The input data for this project are:

- UPU Data (international postal data)
- National postal data
- International trade data
- Business Registers
- Foreign direct investment (FDI) data



Local postal data:

- 149 postal offices, each placed in a district.
- There are districts with several post offices in them, so only 115 unique district post offices
- Consider how much population each postal office is associated with
- Link with population and poverty estimates

Costa Rica: Merging Trade & Business Statistics with Postal Data

UPU data:

Annual **worldwide surveys of postal operators and regulators** on employment, financial results, volume of letters, express, small packets and parcels.

UPU "Big Data":

UPU collects Electronic Data Interchange **(EDI) messages** based on **real-time scans of individual postal exchange between designated operators in more than 150 countries.**

- At collection
- Arrival
- Departure from outward office
- When handed over to Customs
- Final delivery
- etc.

Pilot projects in several countries are attaching customs declarations to postal tracking information by UPU, which greatly expands the variables available (such as sender/receiver, HS codes, values, customs duties, etc.)

Case 1. Estimating trends of overall international trade or trade in specific products/partner countries

- Is postal trade a good proxy for trade volume? In general, postal trade can be quite small, a few percentages of total trade, however, its use as leading indicators which are expected to be available very quickly and can provide estimations of the trend for the month or the year.
- Analyze business-to-business shipments (as possible proxy for manufactured products) vs. business-to-consumer
- Analysis of differences/correlation between postal trade volume and existing commodity trading data by mode of transport
- Can we use postal data to estimate size of e-commerce in overall trade?
- Calculation of freight and insurance services through postal shipments.

Case Studies: Merging Trade & Business Statistics with Postal Data

Case 2. Data Mining of integrated micro-data

- Develop transportation-related statistics
- Volume of international trade at local or sub-national regional level
- Speed of international trade by country source/destination, from initial shipment to final delivery to end customer at local/sub-national location
- If there is sufficient information on businesses with FDI, (such as address, etc.), the volume of international trade as shipped via postal channels and identification of trading partners of businesses with an FDI relationship

Case 3. Estimating selected SDG targets & indicators

- Trade data could be used to proxy consumption/production of certain product types covered by SDG targets (e.g., agricultural commodities, pharmaceuticals, mobile phones, manufactured goods, high-tech, labor-intensive, etc.)
- Trade/postal data can be used to nowcast/forecast business cycles for macroeconomic variables such as GDP and employment
- Real-time proxy indicators of socio-economic behaviour (similar to mobile phone for poverty estimates)

Official Statistics Applications of Mobile Phone Data

Statistical Applications of Mobile Phone Data

What is mobile phone data?

Active Positioning Data

- the collector of the statistics contacts phone owners and asks information about their location, themselves and their behavior
- pinging by mobile network operator or through GPS, downloadable software and apps
- generally requires the consent of the mobile phone user
- very accurate information on movement, location, behavior (purchasing, expenses, travel/mobility)

Passive Positioning Data

- data obtained from phone use information automatically recorded in the systems of phone operators (e.g., Call Detail Record, Erlang (circuit load), Anonymous Bulk Location Data, etc.)
- huge masses of data on all phone users
- relatively cost-effective
- concerns are: protection of privacy, difficulties in obtaining the data from operators and lack of characteristics included in the data

Using Mobile Phone Data for Tourism Statistics

- Sample size may be larger than traditional surveys
- Cheaper and lower burden than surveys
- Call Detail Record (CDR) data include the following information about a phone call:

Subscriber ID TIME Location Duration COST

Use for tourism statistics:

- Need to define the place of residence, the usual environment and destination and transit countries
- Domestic tourism: based on “home anchor point”, according to repeated occurrence in the same cell (night-time)
- International tourism: based on roaming data broken down by country of origin
- Requires historic time series and same subscriber ID in domestic and roaming data

CAVEATS:

- Roaming data from non-tourists in border areas of neighboring countries
- Some travelers may purchase local subscriber identity module (SIM) cards and thus not be covered through mobile phone records.

Statistical Applications of Mobile Phone Data

Mobile phone data can also be used for Transportation, Traffic, Commuting, Population and Mobility Statistics

- Daily commuting patterns (based on time of day)
- Number and duration of commuting trips
- Breakdown by geography, demography
- Distances travelled, routes taken
- Traffic patterns
- Changes of residency (internal migration) based on "home anchors"
- Temporary population
- Cross-border migration

Challenges & Applications of Using Mobile Phone Data

- *Access* is a challenge - not only from mobile network operators but also privacy concerns, legal issues, etc.
- *Technological barriers*, including handling large data files of mobile operators
- *Methodological issues* and *quality concerns*

Applications:

•**Estonia** – the central bank has been using mobile positioning data calibrated with accommodation & travel statistics for tourism statistics since 2009

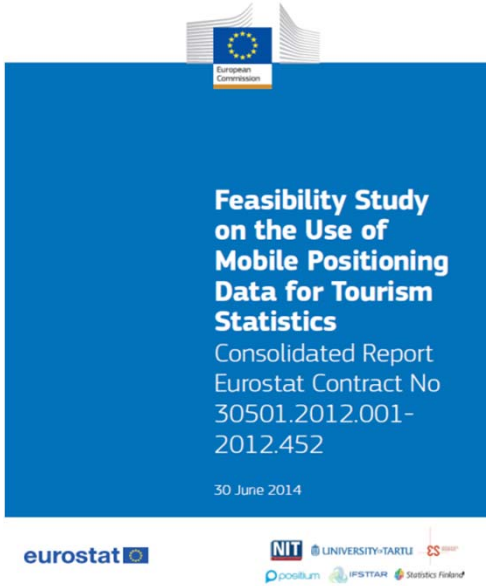
•Estonia partnered with **Positium**, a company providing tools to extract and analyze data from mobile network operators globally.

•**the Netherlands** – "Time patterns, geospatial clustering" Statistics Netherlands

•**Czech Republic** – Czech Tourism

•**Ireland** – "Mobile data for tourism Statistics" The Central Statistics Office Ireland

•2015 – **Oman** to utilize national ID in all digital transactions; working with UNSD to develop a pilot project using mobile phone data tourism, social and population statistics



Eurostat's **Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics**
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

http://positium.com/ positium

The UNECE Sandbox

A web environment for the storage and analysis of large-scale datasets and collaboration to test the feasibility of remote access and processing

http://www1.unece.org/stat/platform/display/bigdata/Sandbox

UNECE Sandbox

- The Sandbox is first example of a shared international statistical Big Data research capability
- It provides a technical platform to load Big Data sets and tools, with the goal of exploring the tools and methods needed for statistical production and the feasibility of producing Big Data-derived statistics and replicating outputs across countries.
- Was created with support from the Central Statistics Office of Ireland and the Irish Centre for High-End Computing
- 38 participants from 18 among national statistics institutes and international organizations

UNECE Sandbox: UN Comtrade Data

Exploring what can be done analyzing detailed international merchandise trade data from UN Comtrade, the global repository of official trade statistics

UN Comtrade Database | Extract data beta | Legacy | Data availability | Metadata & reference

1. Frequency

Annual | Monthly

2. Classification

HS: As reported | 92 | 96 | 02 | 07 | 12

SITC: As reported * | Rev. 1 | Rev. 2 | Rev. 3 | Rev. 4

BEC: BEC

3. Select desired data

Periods (year): × 2014 — All or a valid period. Up to 5 may be selected.

Reporters: × All — All or a valid reporter. Up to 5 may be selected.

Partners: × World — World, All, or a valid reporter. Up to 5 may be

Trade flows: × All — All or select multiple trade flows.

Period	Trade Flow	Reporter	Partner	Commodity Code	Trade Value (US$)	Netweight (kg)	Qty Unit	Qty	Flag
2014	Import	Barbados	World	TOTAL	$1,740,471,483	0	No Quantity	0	0
2014	Export	Barbados	World	TOTAL	$480,753,438	0	No Quantity	0	0
2014	Re-Export	Barbados	World	TOTAL	$197,316,697	0	No Quantity	0	0
2014	Import	Philippines	World	TOTAL	$67,718,868,519	0	No Quantity	0	0
2014	Export	Philippines	World	TOTAL	$61,809,755,230	0	No Quantity	0	0
2014	Import	Benin	World	TOTAL	$3,596,078,234	0	No Quantity	0	0
2014	Export	Benin	World	TOTAL	$951,000,010	0	No Quantity	0	0

http://comtrade.un.org/data/

Proposed analysis:

Network Analysis of Regional Value Chains

- The full trade matrix of about 190 by 190 trading partners shows about 22,000 bilateral flows at the total trade level.
- For each of the more than 5,000 HS commodities there will be a large number of bilateral flows (more at the aggregated commodity classes and a little less for the very detailed HS commodities); which results in about 100 million bilateral flows for each year of data.
- Countries differ in the number of trading partners and the value of the trade flows; both of these measures are seen as measures of strength in a network.

Proposed analysis:

Network Analysis of Regional Value Chains (cont.)

- The network analysis will explore for which groups of countries and for which combination of commodities the strength of trade is relatively high, which would then point to regional value chain networks.
- This project will be executed with the support of Statistics Italy, Statistics Netherlands, the Politecnico University of Milan, the MIT research center, OECD and UNSD;
- It will first replicate the *Network Analysis of World Trade* and some subsequent work done by Prof. Lucia Tajoli and others
- Project is open for further volunteers.

Thank You

Nancy Snyder
snydern@un.org
United Nations Statistics Division