# Rethinking tourism statistics : towards an innovative multi-source integrated production system

**Instructions:** Click on the link to access each author's presentation.

**Organiser:** Christophe Demunter

**Chair:** Mariana Kotzeva

**Discussant:** Raúl Figueroa Diaz

**Participants:**

**Alfatihah Reno Maulani:** Utilization and Governance of MPD as a source of data for Official Statistics: Indonesian Experience

**Maria Wiberg:** A joint model for Nordic European countries to compile tourism statistics based on payment transaction data

**Christophe Demunter:** Re-using privately held data obtained from online accommodation platforms to produce new data on short rentals –the European experience

**Marek Cierpial-Wolan:** Effectiveness of selected methods of data integration in tourism statistics

**María Velasco:** The future of Spain's tourism statistics system: integration of traditional statistics and new sources of information

IAOS-ISI 2024
MEXICO CONFERENCE

# Utilization and Governance of MPD as a source of data for Official Statistics: Indonesian Experience

Alfatihah Reno Maulani & Eko Rahmadian
Statistics Indonesia

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

# The Journey

**2016 - 2017**

### Initiation

Ministry of Tourism and Statistics Indonesia collaborate to utilize MPD for Inbound Cross Border Tourism in completing immigration border administrative data

**2018 - 2020**

### Exploration

Maximize MPD utilization with domestic tourism, event analysis, metropolitan statistical area, and outbound tourism.

Expand collaboration: Assist the National Disaster Management Agency with disaster response and guide the Ministry of Transportation to develop MPD OD Matrices

**2021 - 2022**

### Development

Improve accuracy in home/work identification with anchor modeling (AMDA)

Upgrade current scripts for Inbound Cross Border Tourism to reduce processing durations

Secure full collaboration with telco companies nationwide: Telkomsel, Indosat and XL

**2023 - 2024**

### Improvement

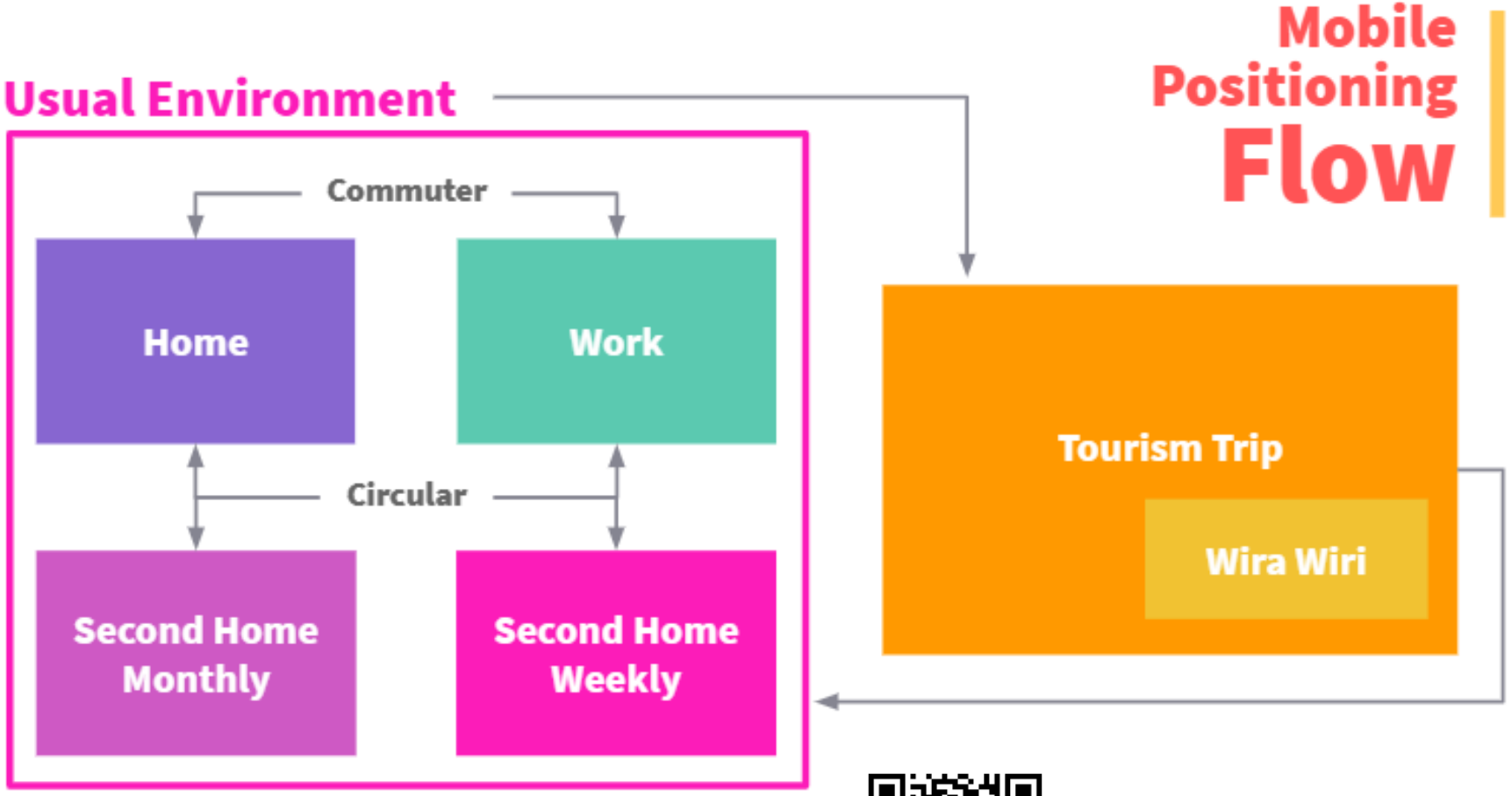
Integrate inbound tourism across Indonesia, with a specific focus on Super Priority Destination

In development, creation on a data pipeline across telcos for QA metrics and generated outputs

IOS-ISI 2024 MEXICO CONFERENCE

# Building Case: Tourism Statistics

The initial purpose of using MPD in Indonesia is for tourism statistics, where the definition of a tourism trip is a trip outside the usual environment.
The basic concept used by Statistics Indonesia in its usual environment is the city where you live/where you do your daily activities. So, the trips we calculate are trips between cities.

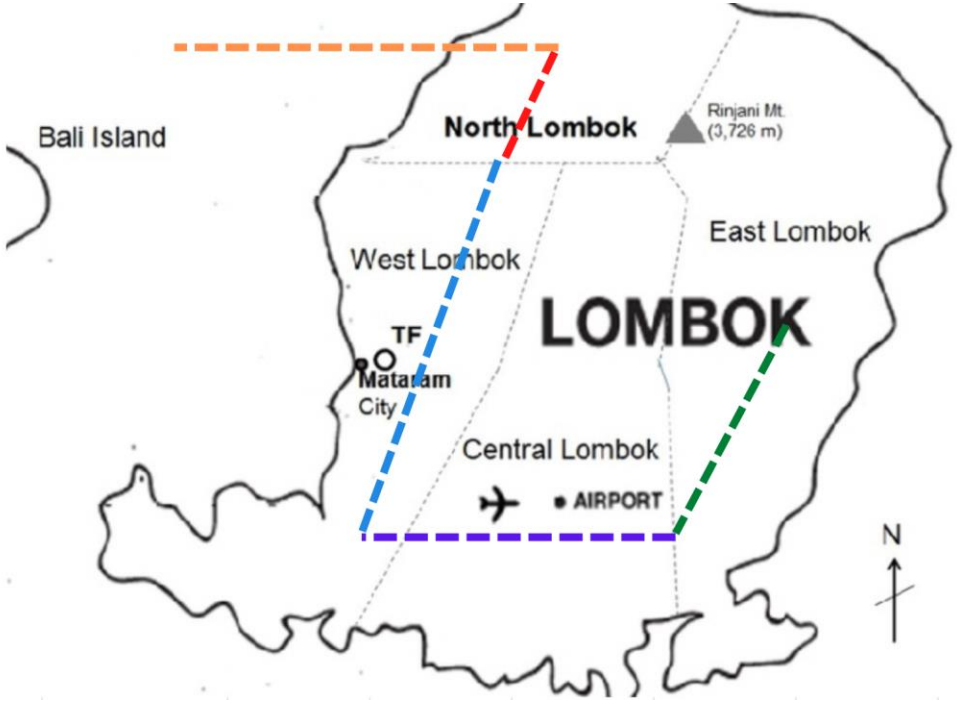



s.id

IOS-ISI 2024 MEXICO CONFERENCE

# Determining Trips: Understanding Subscriber Mobility

| time | longitude | latitude | province | city | district | signal |
|---|---|---|---|---|---|---|
| 2021-05-22 19:27:57 | 106.909 | -6.252 | DKI Jakarta | Kota Jakarta Timur | Makasar | 3G |
| 2021-06-01 4:02:22 | 106.6628333 | -6.120416667 | Banten | Kota Tangerang | Benda | 3G |
| 2021-06-01 6:04:06 | 106.6628333 | -6.120416667 | Banten | Kota Tangerang | Benda | 3G |
| 2021-06-01 7:20:33 | 116.0411778 | -8.463713889 | Nusa Tenggara Barat | Kab. Lombok Utara | Pemenang | 3G |
| 2021-06-01 7:20:48 | 116.0411778 | -8.463713889 | Nusa Tenggara Barat | Kab. Lombok Utara | Pemenang | 3G |
| 2021-06-01 7:25:17 | 116.1602222 | -8.669177778 | Nusa Tenggara Barat | Kab. Lombok Barat | Kuripan | 3G |
| 2021-06-01 7:28:32 | 116.1602222 | -8.669177778 | Nusa Tenggara Barat | Kab. Lombok Barat | Kuripan | 3G |
| 2021-06-01 7:28:54 | 116.30575 | -8.747777778 | Nusa Tenggara Barat | Kab. Lombok Tengah | Praya Tengah | 3G |
| 2021-06-01 7:33:58 | 116.30575 | -8.747777778 | Nusa Tenggara Barat | Kab. Lombok Tengah | Praya Tengah | 3G |
| 2021-06-01 7:34:21 | 116.2739167 | -8.761777778 | Nusa Tenggara Barat | Kab. Lombok Tengah | Pujut | 3G |
| 2021-06-01 12:31:17 | 116.2739167 | -8.761777778 | Nusa Tenggara Barat | Kab. Lombok Tengah | Pujut | 3G |
| 2021-06-01 13:06:02 | 116.5207583 | -8.646419444 | Nusa Tenggara Barat | Kab. Lombok Timur | Selong | 4G |
| 2021-06-03 15:10:36 | 116.5207583 | -8.646419444 | Nusa Tenggara Barat | Kab. Lombok Timur | Selong | 4G |
| 2021-06-03 18:29:17 | 106.9013333 | -6.24925 | DKI Jakarta | Kota Jakarta Timur | Makasar | 3G |




A trip is defined as a journey from one usual environment to another. In each trip, there can be visits to several places, but it is necessary to determine the main destination of the trip.

Please note that there are anomalies in the data, such as fast movers, that could affect your analysis. For that, data cleansing needs to be done.

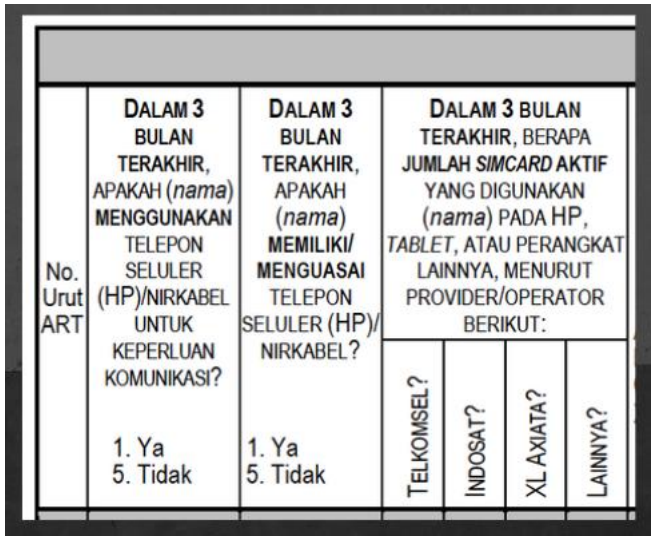
IOS-ISI 2024 MEXICO CONFERENCE

# From Mobile Subscriber to Whole Population

Indonesia's national socio-economic survey in 2020: 78.56% of the population aged 5 years and over, have used cell phones.

**what about those who do not use the phone?**
**what about multiple sim cards?**
**what about other MNO's?**

In the national socio-economic survey, we added questions about the use of communication devices. The results will be used in estimating the actual number with the customer mobility data we obtain from MPD.



| No. Urut ART | DALAM 3 BULAN TERAKHIR, APAKAH (*nama*) MENGGUNAKAN TELEPON SELULER (HP)/NIRKABEL UNTUK KEPERLUAN KOMUNIKASI? 1. Ya 5. Tidak | DALAM 3 BULAN TERAKHIR, APAKAH (*nama*) MEMILIKI/ MENGUASAI TELEPON SELULER (HP)/ NIRKABEL? 1. Ya 5. Tidak | DALAM 3 BULAN TERAKHIR, BERAPA JUMLAH *SIMCARD* AKTIF YANG DIGUNAKAN (*nama*) PADA HP, *TABLET*, ATAU PERANGKAT LAINNYA, MENURUT PROVIDER/OPERATOR BERIKUT: | | | |
|---|---|---|---|---|---|---|
| | | | TELKOMSEL? | INDOSAT? | XL AXIATA? | LAINNYA? |

In the last 3 months:

- have you used a cell phone?
- have you owned/controlled a cell phone?
- how many active SIM card were used on all devices

IOS-ISI 2024 MEXICO CONFERENCE

# Combine Methods: MPD and Digital Survey to replace Traditional Survey

| MPD + Digital Survey | | Traditional Survey |
|---|---|---|
| MPD | Origin | Origin |
| | Destination | Destination |
| | Length of Stay | Length of Stay |
| | Length of Journey | Length of Journey |
| Digital Survey | Purpose of Trip | Purpose of Trip |
| | Demographic | Demographic |
| | Expenditure | Expenditure |
| | Modes of transportation | Modes of transportation |

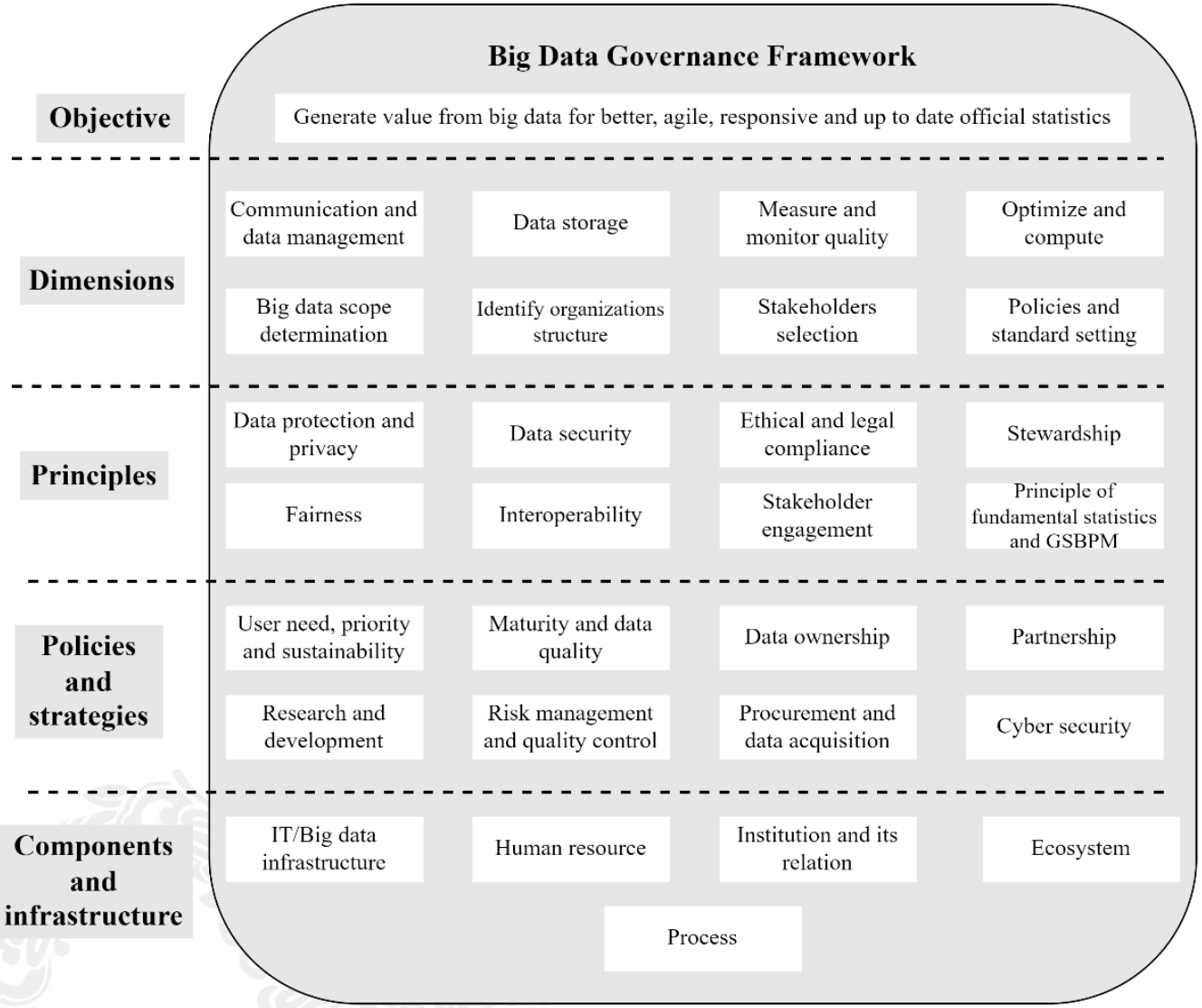
# Comparison between MPD and Traditional Survey

| No | *MPD Approach* | *Household Approach (Conventional)* |
|---|---|---|
| 1 | Estimates for smaller areas: districts/cities | Estimates up to the provincial level |
| 2 | Monthly data period | Yearly data period |
| 3 | No surveyors needed | The number of surveyors needed depends on the number of samples |
| 4 | With a more detailed level of estimation, the required budget is only around IDR 15 billion | To conduct a conventional domestic tourist survey, a budget of around IDR 30-40 billion is needed (based on the 2018 and 2019 conventional domestic tourists survey budget). |
| 5 | It has the potential to be able to generate statistics for domestic tourists monthly | Can only generate statistics for domestic tourists on annually |
| 6 | The solution for collecting data for domestic tourists during the covid-19 pandemic, no door-to-door interviews needed | The Covid-19 pandemic condition limits conventional data collection activities |
| 7 | MPD can be used for other statistical calculations | One activity is only used for one type of statistical calculation |

# Issues & Challenges on the MPD Big Data Governance

| Dimensions | Issues/Challenges |
|---|---|
| Policies and standard setting | The absence/overlapping of law and regulations |
| Data storage | Cost for accessibility is prohibitively high |
| Communication and data management | Inter-ministerial advocacy is required |
| Measure and monitor quality | The nature of big data |
| Identify organizations structure | Urgency to create a special unit for big data analysis |
| Optimize and compute | Methodological challenge |
| Stakeholders selection | Effective stakeholder management |
| Big data scope | Determine the scope or priority |

# Big Data Governance Framework




Many challenges are all intertwined.

Addresses data quality, security, privacy, and ethical standards

Defines the roles and responsibilities of stakeholders (technical and non-technical)

Mitigates potential risks associated with the use of big data

IOS-ISI 2024 MEXICO CONFERENCE

IOS-ISI 2024
MEXICO CONFERENCE

# Thank you

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

# A JOINT MODEL FOR NORDIC EUROPEAN COUNTRIES TO COMPILE TOURISM STATISTICS BASED ON PAYMENT TRANSACTION DATA

**Maria Wiberg**

**The Swedish Agency for Economic and Regional Growth**

IOS-ISI 2024 MEXICO CONFERENCE

# Tourism definition

“ Tourism is a social, cultural and economic phenomenon which entails the movement of people to countries or places outside their usual environment for personal or business/professional purposes.”

***World Tourism Organisation (UNWTO)***

# Tourism statistics 2.0

*Big and expensive project that can give:*

- New data/analysis that we've been missing
- More precise estimates and granularity we didn't think was possible
- Short delivery time for data
- Lower cost in the long term
- Nordic harmonization on regional/municipal level
    - 30 percent higher tourist effects!

IOS-ISI 2024 MEXICO CONFERENCE

# The model

95%

$$C_{tot} = (C_{Visa} + C_{Mastercard}) * \partial_{Nets} * \partial_{PTP} * \partial_{Cash} + C_{Travel\ agencies}$$

# Usual environment

VISA


MASTERCARD

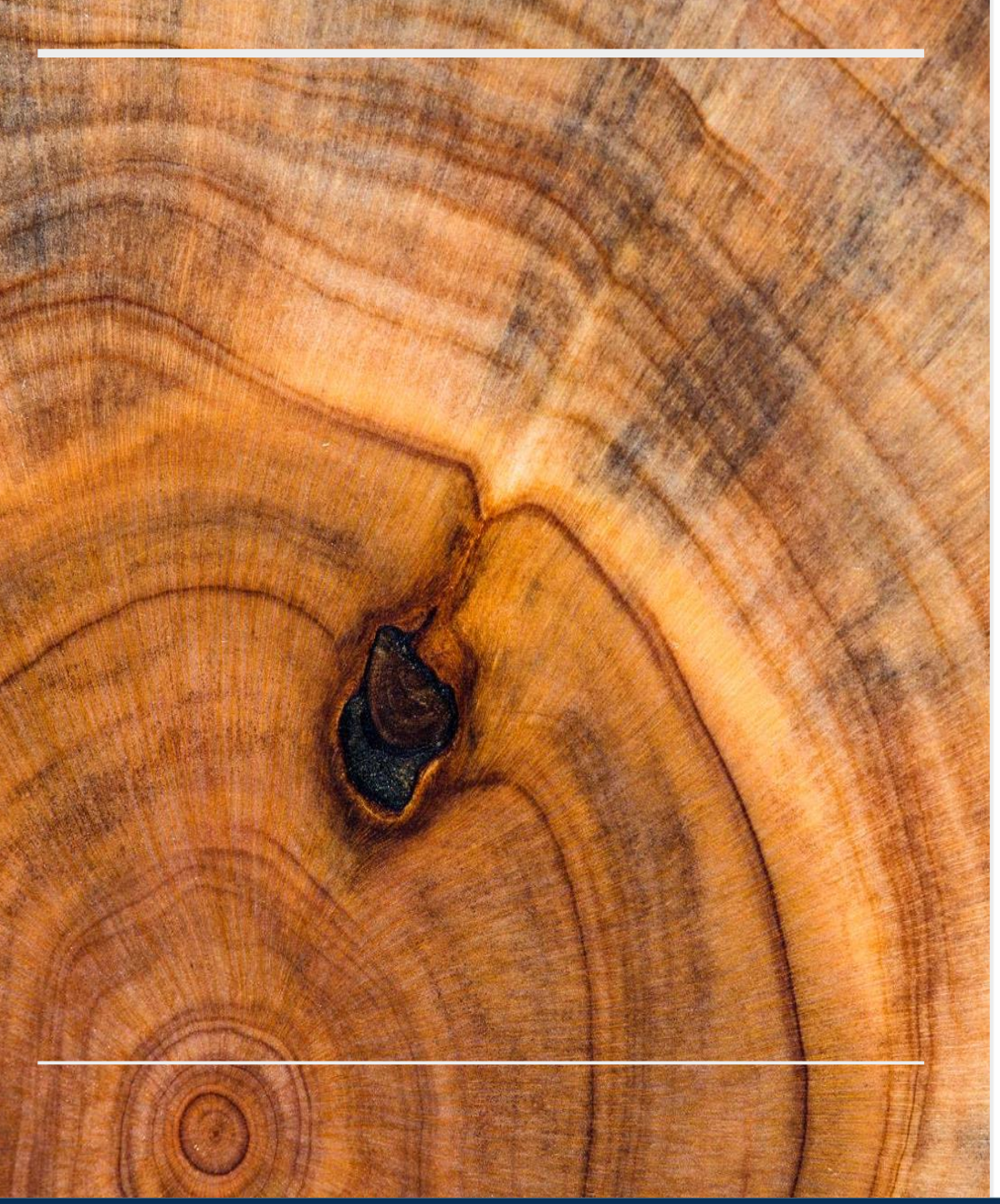
# Possibilities

- Granularity

- Target groups / countries

- Monthly / seasonal data

- Trips



# Key Challenges

- Online / PTP / cash spend

- 5/50 rule

- Improvements of the model

- Import/Export by municipality

# Implementation

- Tourism Satellite Account (TSA) at municipal and regional level
    - Sweden aim to published 2024
    - The other Nordic European countries to start process of acquiring data
    - Big Data partially replace sample surveys

-Surveys at lower costs
    - Smaller sample size
    - Includes no questions about tourist expenses

- Total savings in Sweden: US$ 100K per year

IOS-ISI 2024 MEXICO CONFERENCE

IOS-ISI 2024
MEXICO CONFERENCE

# Thank you

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

IOS-ISI 2024
MEXICO CONFERENCE

Friday 17 May 2024 09:00-10:30

# Rethinking tourism statistics : towards an innovative multi-source integrated production system

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

# Re-using privately held data obtained from online accommodation platforms to produce new data on short rentals – the European experience

*Christophe Demunter, Simon Bley (EUROSTAT)*

IOS-ISI 2024 MEXICO CONFERENCE


# Did you know that?

- In 2022, visitors from **the Americas** spent over **44 million million nights** in short-term rentals in the EU, booked via **platforms**?

  … *this means 120.000 'transatlantic' guests on any given night in the year*

- Their preferred destinations were Italy (10 million), Spain (9 million), France (8 million) and Portugal (5 million)

IOS-ISI 2024
MEXICO CONFERENCE

# Did you know that?

- In 2023, **679 million guest nights** were spent in short-term rentals **booked via Airbnb, Booking, Tripadvisor or Expedia Group**?

  … *nearly 1.9 million tourists per night slept in a bed booked via the platforms*

- 61 million stays were booked in 2023

  … *which corresponds to 117 reservations per minute, or nearly 2 each second!*

IOS-ISI 2024 MEXICO CONFERENCE

# Origins and relevance of Eurostat's platforms project

## Data needs in tourism statistics

- ⇨ Better coverage of short-stay accommodation
- ⇨ Smaller service providers often in grey area and not well covered by tourism registers or surveys
- ⇨ New policy needs for information on this 'new', growing segment of the accommodation sector

## Explore re-using privately held data for statistics

- ⇨ Expensive or not feasible to collect data from many households / small enterprises offering short-stay rentals
- ⇨ Most information available with relatively few platforms, where service provider now leave their 'digital footprint'

IOS-ISI 2024 MEXICO CONFERENCE

# Governance

## Platforms

- ⇨ **Non-disclosure agreements** signed with four major international platforms (in March 2020)
- ⇨ Send very granular data (municipality level) to Eurostat (variables similar to 'traditional' accommodation statistics)
- ⇨ Quarterly transmission, two months after quarter ends
- ⇨ Regular informal bilateral meetings (~ 2 times/year)

## Statistical community (in the EU)

- ⇨ In line with the '**subsidiarity principle**': centralised approach with **Eurostat as single entry point**, in close coordination and cooperation with the Member States
- ⇨ Bilateral (non-disclosure) agreement with each NSI
- ⇨ Coordination via regular meetings


IOS-ISI 2024 MEXICO CONFERENCE

# In a nutshell

Platforms send data aggregated at local level.

eurostat

Eurostat merges, validates, produces output tables.


Data + articles are published on Eurostat's website

Countries receive merged dataset (sum of 4 platforms) and perform additional quality checks.
Use data for their own dissemination

IOS-ISI 2024 MEXICO CONFERENCE

# Methodological topics

## Confidentiality and disclosure

- ⇨ Departure from our comfort zone; different ‘culture’
- ⇨ Minimum number of statistical units (service providers)
- ⇨ Dominance rules or traditional statistical disclosure control overruled by a pragmatic approach (two release waves)

## Concepts and definitions

- ⇨ As closely as possible replicated from traditional tourism statistics, via technical meetings with the platforms
- ⇨ Platforms also provide metadata

## Key challenge: dealing with double counting

- ⇨ No additivity of capacity data: service providers can advertise their listing on more than one platform
- ⇨ No additivity of occupancy data: service providers may be reporting to the NSI, meaning a nights spent is possibly observed twice
- ⇨ But: relevant stand-alone product!!

IOS-ISI 2024 MEXICO CONFERENCE

# Sustainability

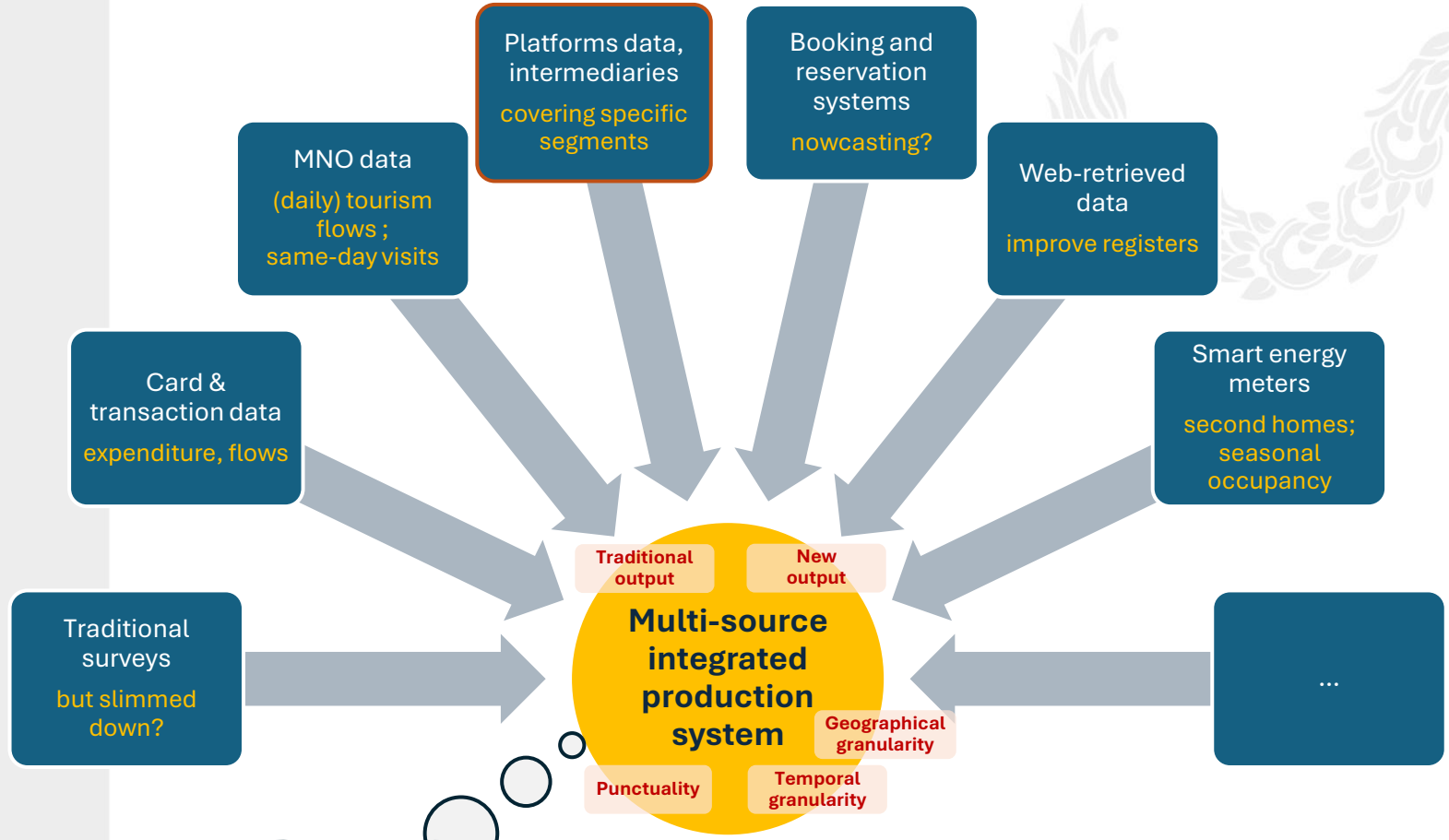
## “So far, so good”

- ⇨ Solid **relations with the data providers**, enabled through a multi-disciplinary and multi-stakeholder approach (lawyers, data scientists, investor relations, …)
- ⇨ Robust **quality assurance framework**
- ⇨ **Regular output** via Eurostat’s usual dissemination channels
- ⇨ Popular product, filling user need & showcasing innovation
- ⇨ Currently published as “**experimental statistics**”, but transition is under discussion

## Continuity of data flows?

- ⇨ Based on non-disclosure agreements, no legal obligation
- ⇨ But: new registration obligations for service providers in the EU and reporting obligations for platforms
- ⇨ But: forthcoming revision of the EU statistical law, including sustainable access to privately held data sources

# Integration




**Risks & constraints**

- ⇨ Access … and continuity of access
- ⇨ Alignment of concepts, definitions
- ⇨ Selectivity bias
- ⇨ Quality, comparability over time
- ⇨ Independence
- ⇨ Skills
- ⇨ Trust

IOS-ISI 2024 MEXICO CONFERENCE

# Take-aways

## The platforms project

- ⇨ The project is just one piece of the tourism statistics puzzle
- ⇨ Successful proof-of-concept of re-using privately held data for statistics, clearly delineated use case
- ⇨ Key success factors: multi-disciplinary & multi-stakeholder approach
- ⇨ Need to invest in mutual understanding of each other's 'business culture' and 'data culture'
- ⇨ Need to give up full control from A to Z but partially rely on reliable partnerships

## Modernisation of tourism statistics

- ⇨ Tourism statistics deal with (the noise in) human mobility
- ⇨ Many new data sources have potential for measuring tourism flows (physical and monetary), but how to unlock?
- ⇨ Relevance of international cooperation, given that many sources are similar across countries

IOS-ISI 2024 MEXICO CONFERENCE

IOS-ISI 2024 MEXICO CONFERENCE

# Thank you

christophe.demunter@ec.europa.eu
simon-johannes.bley@ec.europa.eu

https://ec.europa.eu/eurostat/web/tourism/overview

INEGI

International Statistical Institute

IAOS IMPROVING OFFICIAL STATISTICS

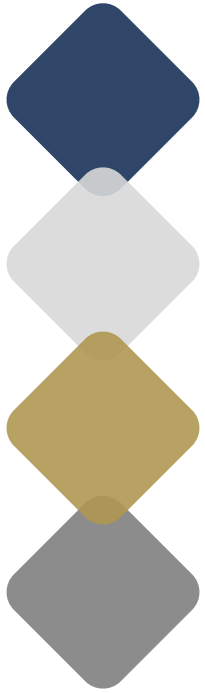
IAOS-ISI 2024
MEXICO CONFERENCE

# Effectiveness of selected methods of data integration in tourism statistics

Marek Cierpial-Wolan, Assoc. Prof., Statistics Poland, University of Rzeszow
Dominik Rozkrut, PhD, Statistics Poland, University of Szczecin

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

# Agenda



1. Background
2. Data sources and data integration
3. Effectiveness of selected methods
4. Conclusions

IOS-ISI 2024
MEXICO CONFERENCE

# Tourism in Hyper-Turbulent Era

**Global security** – the prospects for security is less clear.

**Energy** – the cost and volatility.

**COVID-19 pandemic** – disorganization, economic disruption, lifestyle changes.

**Global migration crisis** – about 300 million migrants worldwide.

**Rapid development of information technology** – BD, AI are constantly changing our world.

**Predatory competition in the information market** – worse information is crowded out by better information.

…

# Challenges of statistics

**Should statistics be a beacon in the contaminated information environment of today's world?**

## Official statistics

- Necessity for faster, more disaggregated and up-to-date information that responds to the needs of stakeholders;
- Quickly detect and estimate changes in contemporary world.

## Statistics – scientific discipline

- Contemporary data analysis, in many cases, goes beyond the traditional understanding of statistics;
- Methodology of statistics as a science must change.

# Data sources

## census survey – sample survey – administrative registers – big data

- **The emergence of big data is changing the approach to data analysis**

- **Advantages:**

  - ✓ huge number of observations,
  - ✓ opportunity to improve the quality of inference, under the growing scale and importance of non-sampling errors;

- **Challenges in statistical inference – the necessity of new conceptual framework in:**

  - ✓ estimation,
  - ✓ hypothesis testing.

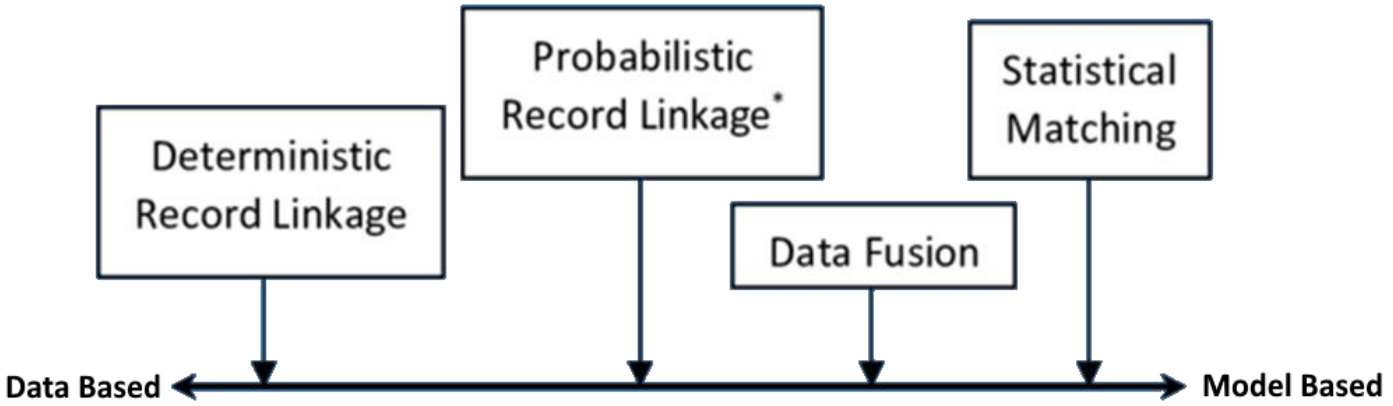
# Data integration

## census survey – sample survey – administrative registers – big data

- Data integration – the general trend observed in statistical research;
- Additional sources of information have been used in sample surveys for years (statistical inference theory, particularly the Bayesian paradigm, sample selection method, where one of the assumptions to have prior knowledge of the population);
- Growing demand for additional information nowadays – reduce the effects of the increasing magnitude and importance of non-sampling errors;
- Big data – methodological challenges for data integration and thus even more sensitivity in terms of output quality assessment.

# Data integration

## Record linkage




*AKA Record Linkage, Probabilistic Record Linkage, Computer Matching, Data Integration, Data Linkage, Data Matching, Deduplication, Duplicate Detection, Entity Extraction, Entity Matching, Entity Reconciliation, Entity Resolution, File Linking, Fuzzy Matching, Information Integration, Object Consolidation, Object Identification, Reference Reconciliation, Re-identification

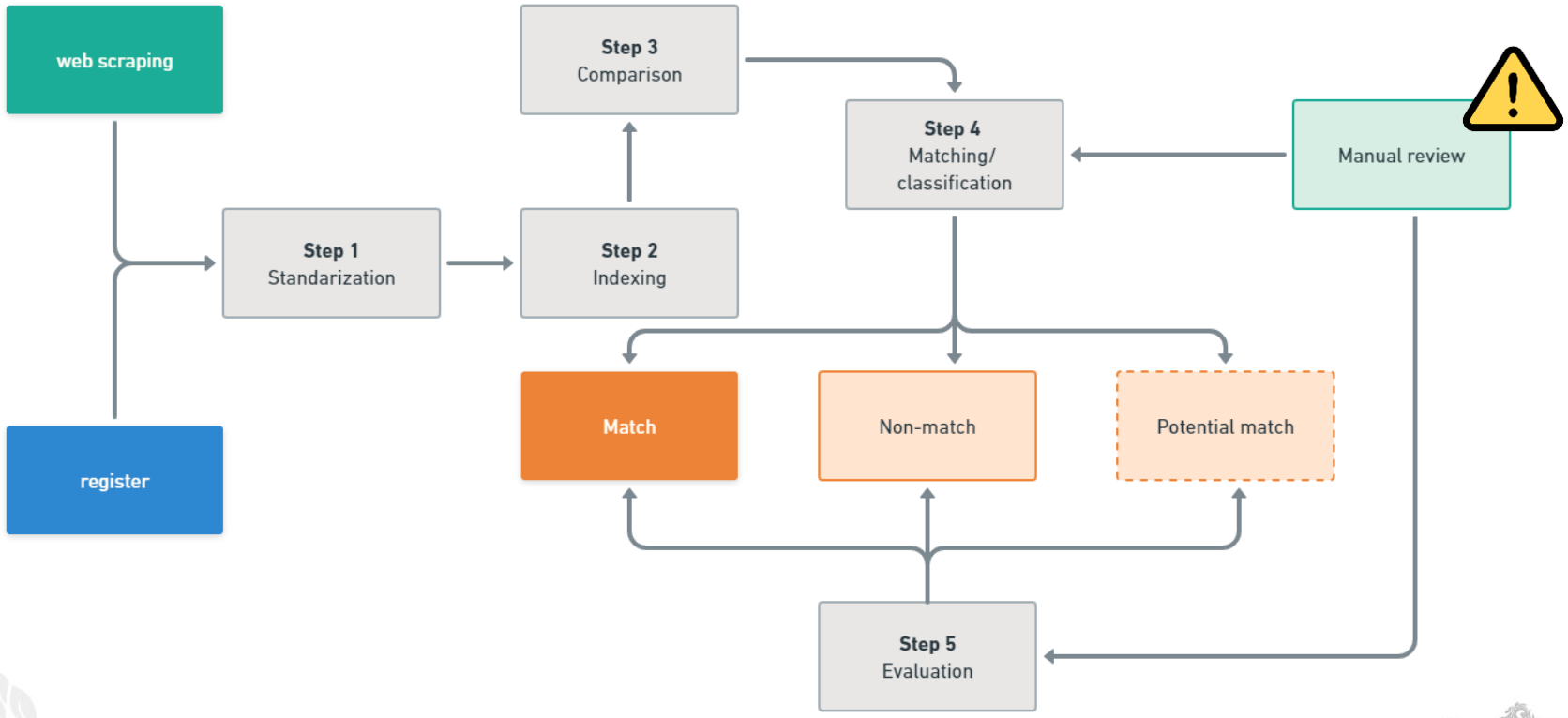
Source: Asher J. et al., An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries (2020)

# Big data sources and administrative registers in tourism

| Selected big data sources | Selected administrative registers |
|---|---|
| Smart City systems | **Register of categorized facilities** |
| **Web Scraping** | **Register of non-categorized facilities** |
| Mobile network operators | **Business register** |
| Payment/credit card operators | **Geo registers** |
| Satellite, drone images | Register of national parks |
| Parking, energy, water meters | Vintage building register |
| Car, bus, road sensors | Register of tourist attractions |

tourism statistics frame

# Probabilistic data linkage – simplified diagram




Source: Christen, P., Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection (2012b)

# Web scraping – analysis of portals


Analysis of 3 most popular short-stay accommodation platforms in Poland:

- **Airbnb**
  - ✓ Online platform for short-term accommodation rentals from private individuals;
  - ✓ Around 67% of offers are 55.2;
  - ✓ No exact address on the portal prior to booking the facility;
  - ✓ Limited to 300 objects per search (an important fact for web scraping purposes).
- **Booking**
  - ✓ One of the most popular digital platforms in Poland for online accommodation bookings; in the country and abroad;
  - ✓ Only about 30% of offers are 55.2;
  - ✓ Exact location of the object;
  - ✓ Limited to 1000 objects per search (an important fact for web scraping purposes).
- **Hotels**
  - ✓ The portal is the part of the Expedia group;
  - ✓ Only about 34% of offers are 55.2;
  - ✓ Exact location of the object (additionally indicated number of rooms);
  - ✓ No limit objects per search (an important fact for web scraping purposes).

airbnb

12 556 objects were scrapped

Booking.com

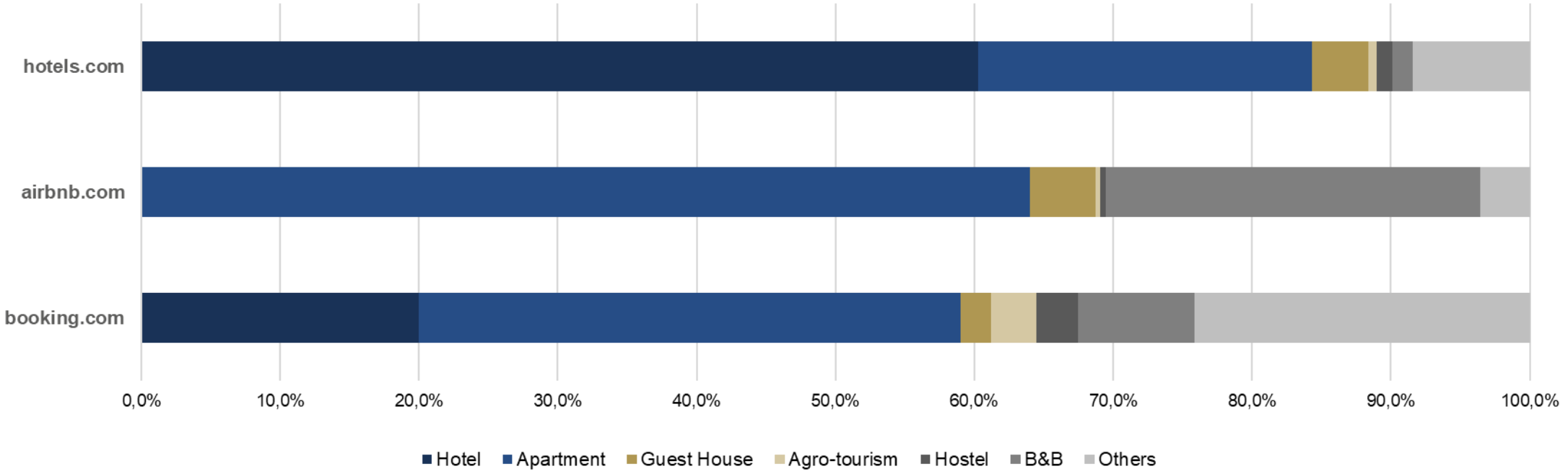
8 884 objects were scrapped

Hotels.com

4 620 objects were scrapped

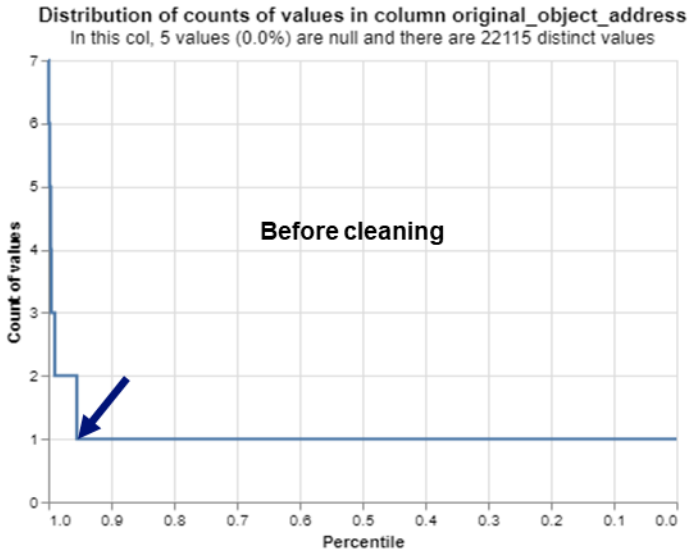
IOS-ISI 2024
MEXICO CONFERENCE

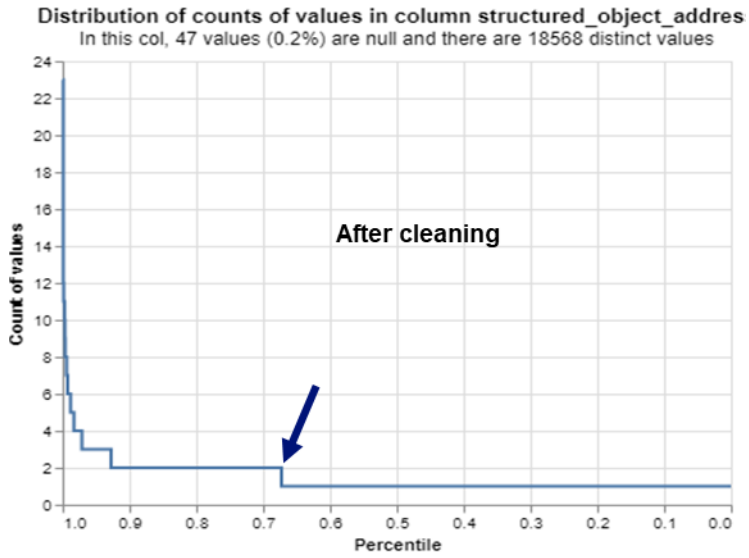
# Booking portals – objects by type



# Cleaning and standardization datasets

- removal of typos, white spaces, html tags, etc.
- automatic parsing of address data into a common structure
- assignment of geographical coordinates to the accommodation facilities




Python programming language

Request/BeautifulSoup library

Pandas library

HERE Maps API

IOS-ISI 2024 MEXICO CONFERENCE

# Data linkage and deduplication methods

1. Splink data linkage
2. Natural Language Processing
3. Selected machine learning algorithms
4. Fuzzy matching
5. Comparison of images of object offers

Data Linkages

# Evaluation of methods

- 1. Confusion matrix
- 2. Receiver operating characteristic (ROC) curve and Area Under Curve (AUC)
- 3. Accuracy, Specificity, Sensitivity and Youden’s J statistic

# Natural Language Processing (NLP)

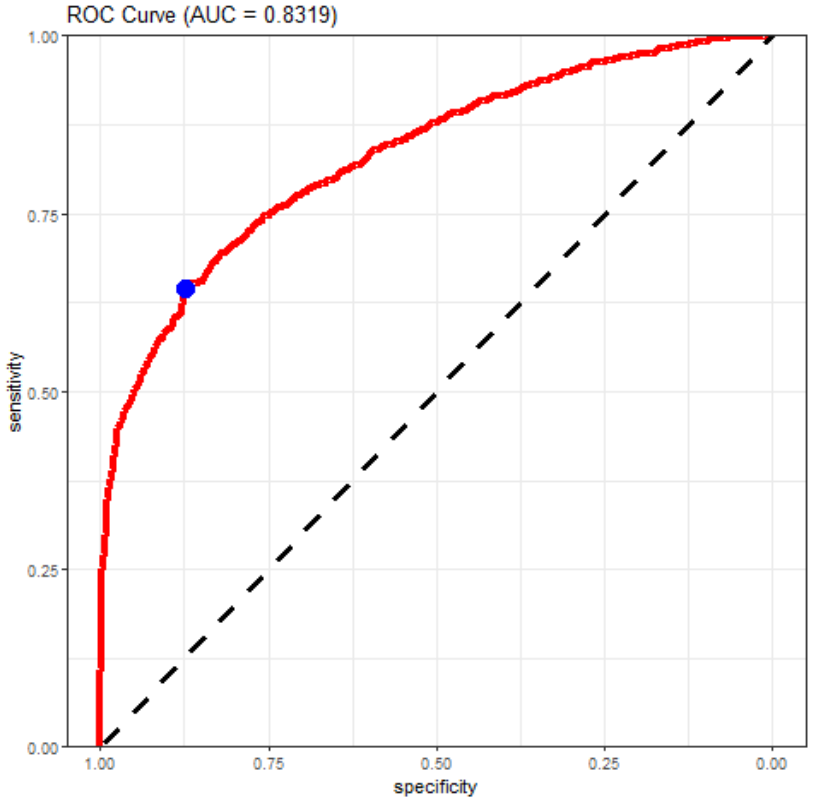
The NLP method with the Faiss library involves the following steps: tokenisation, vectorisation, comparison and similarity assessment, and deduplication decision

- The algorithm begins by transforming the texts in character variables into tokens and then into semantic vectors using the SentenceTransformer library;
- The deduplication process  focuses on evaluating the coincidence between the two strings to determine the degree of their similarity and deciding whether they could be considered duplicates (using Euclidean metrics to calculate the degree of difference between the two vectorised text strings);
- In the context of the NLP processing, the selection of appropriate libraries plays a key role, especially when talking about differences between national languages. For instance, for English, there is a wide range of libraries and models ready to use, which facilitates research and application work.

# Natural Language Processing (NLP)

- Results
  - Accuracy: 0.7621
  - Specificity: 0.6455
  - Sensitivity: 0.8730
  - Youden's J statistic: 0.5184

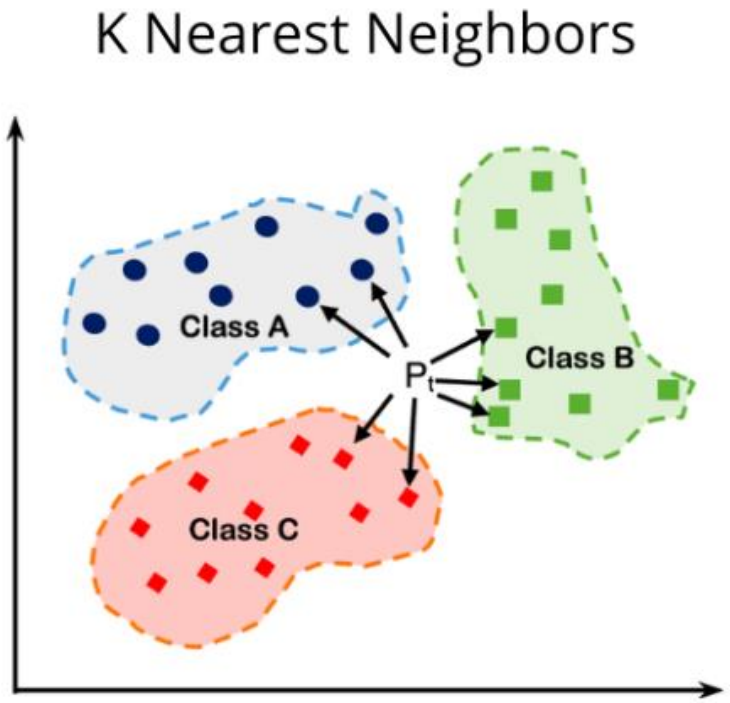
| | | | |
|---|---|---|---|
| Actual | Match | 0.31 (TP) | 0.17 (FN) |
| | Non-match | 0.07 (FP) | 0.45 (TN) |
| | | Match | Non-match |
| | | Predicted | |



# Machine Learning algorithms

ML uses the TF-IDF (Term Frequency-Inverse Document Frequency) and N-gram technique in combination with the K Nearest Neighbors (K-NN) algorithm:

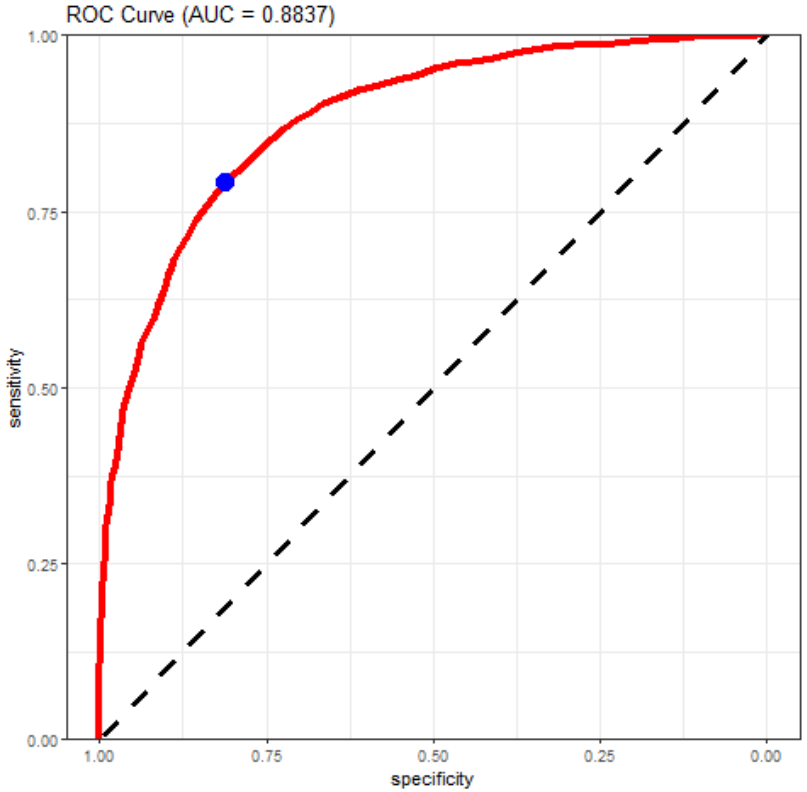
- works by assigning weights to words based on their frequency in the document (TF) and the inverse frequency in the entire document set (IDF),
- the use of N-grams allows contextual information to be taken into account in text analysis, which is useful in the deduplication process where the structure and layout of the data is important,
- the K-NN algorithm is a popular algorithm in machine learning that can be used to find similarity between data.



# Machine Learning algorithms

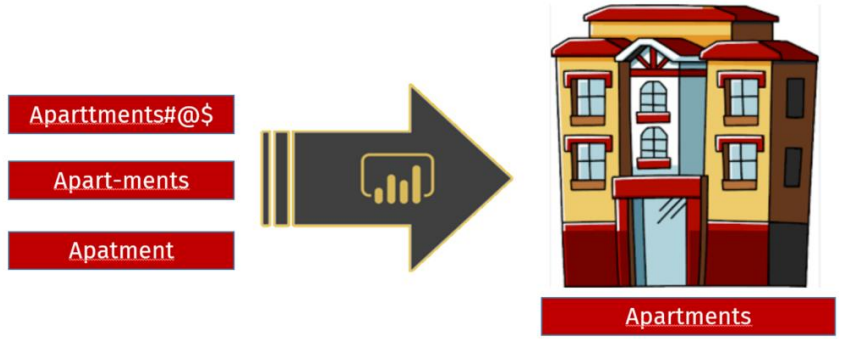
- Results
  - Accuracy: 0.8099
  - Specificity: 0.8121
  - Sensitivity: 0.7909
  - Youden's J statistic: 0.6031

| | | | |
|---|---|---|---|
| Actual | Match | 0.26 (TP) | 0.07 (FN) |
| | Non-match | 0.13 (FP) | 0.55 (TN) |
| | | Match | Non-match |
| | | Predicted | |



# Fuzzy matching

- A technique for finding strings of characters that approximately match a pattern;
- Error-tolerant search that returns records even if the search term contains typos or extra/missing characters;
- Data on the distance between facilities based on geographical coordinate;
- On average, it removes 10.3% of duplicates for web scraping data from various portals.

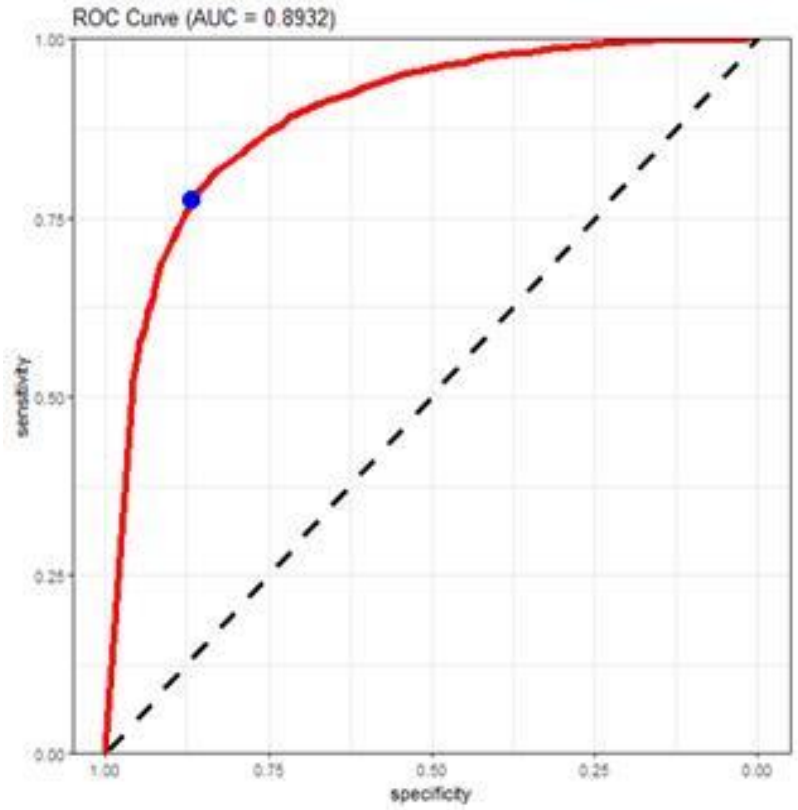



| web_scraping_combined | podo... | fuzzy_match | dystans_km |
|---|---|---|---|
| Arche Hotel Lublin Lublin 20-105 ulica Zamojska 30 | 100 | ARCHE HOTEL LUBLIN Lublin 20-105 ulica Zamojska 30 | 0,002 |
| Hotel Kapitan Szczecin 70-240 ulica Gabriela Narutowicza 17D | 97 | AMW HOTELE SP. Z O.O. HOTEL KAPITAN Szczecin 70-240 ulica Gabriela Narutowicza 17B | 0,004 |
| Pokoje gościnne Truskawkowa 1a Zielona Gora 65-129 ulica Truskawkowa 1 | 82 | Firma Handlowo-Usługowa WOJCIECH SARKA w spadku Zielona Gora 65-129 ulica Truskawkowa 1a | 0,004 |
| Invite Wroclaw 50-502 ulica Hubska 54 | 100 | INVITE SPÓŁKA Z OGRANICZONĄ ODPOWIEDZIALNOŚCIĄ Wroclaw 50-502 ulica Hubska 52/54 | 0,004 |
| Rest Apartments Wroclaw 50-502 ulica Hubska 54 | 79 | INVITE SPÓŁKA Z OGRANICZONĄ ODPOWIEDZIALNOŚCIĄ Wroclaw 50-502 ulica Hubska 52/54 | 0,004 |
| 5D Apartamenty Fredry 2/36 Kołobrzeg 78-100 ulica Aleksandra Fredry 2 | 80 | OŚRODEK WYPOCZYNKOWY "WIGA" GRAŻYNA MIŁOWSKA Kołobrzeg 78-100 ulica Aleksandra Fredry 13 | 0,004 |
| Qubus Hotel Gdańsk Gdansk 80-748 ulica Chmielna 47/52 | 97 | QUBUS HOTEL GDAŃSK-MANAGEMENT SP. Z O.O. Gdansk 80-748 ulica Chmielna 47 | 0,005 |

# Fuzzy matching

- Results for Levenshtein similarity
  - Accuracy: 0.8949
  - Specificity: 0.7757
  - Sensitivity: 0.8681
  - Youden's J statistic: 0.6438

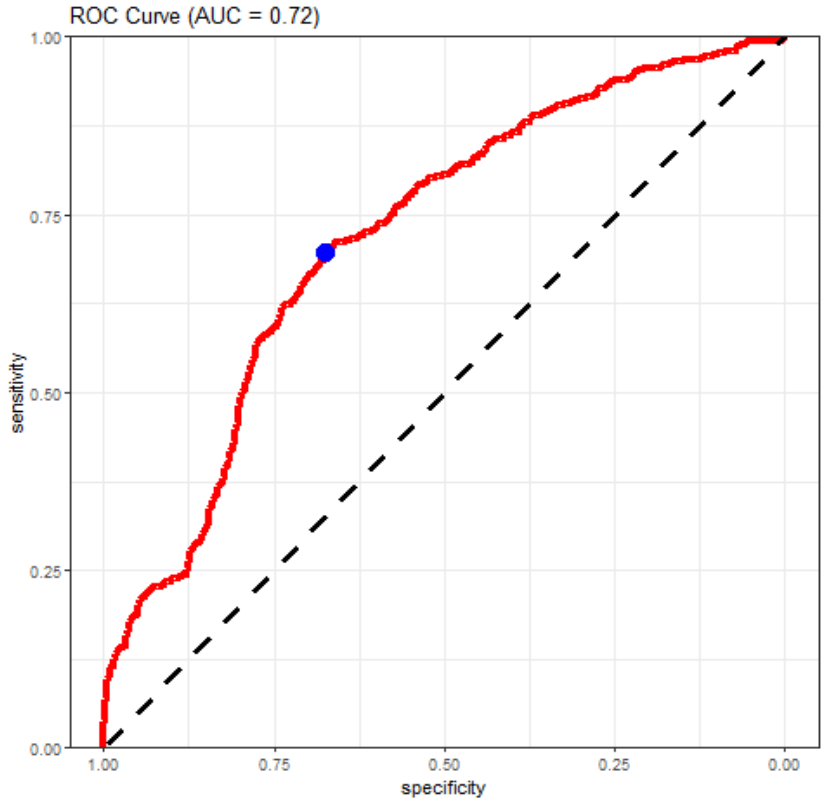
| | | | |
|---|---|---|---|
| Actual | Match | 0.48 (TP) | 0.07 (FN) |
| | Non-match | 0.04 (FP) | 0.41 (TN) |
| | | Match | Non-match |
| | | Predicted | |



# Fuzzy matching

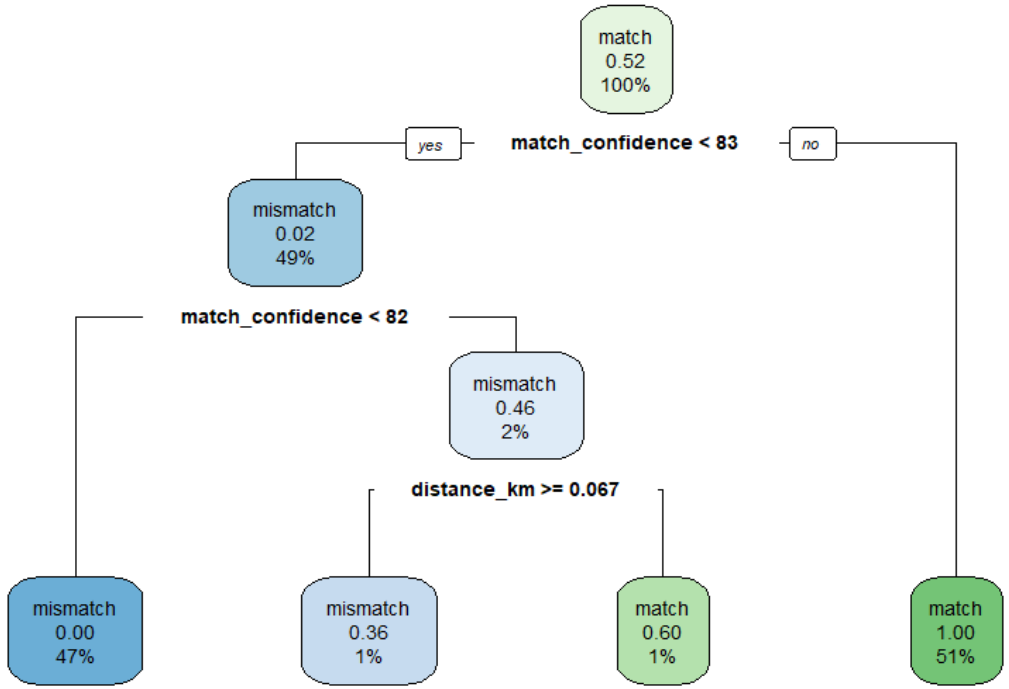
- Results for Jaro-Winkler similarity
  - Accuracy: 0.6803
  - Specificity: 0.6975
  - Sensitivity: 0.6761
  - Youden's J statistic: 0.3736

| | | | |
|---|---|---|---|
| Actual | Match | 0.51 (TP) | 0.25 (FN) |
| | Non-match | 0.07 (FP) | 0.17 (TN) |
| | | Match | Non-match |
| | | Predicted | |



# Fuzzy matching

- To combine two matching criteria and find the set of decision rules we applied the decision tree
- Results
  - Accuracy: 0.9919
  - Specificity: 0.9921
  - Sensitivity: 0.9917
  - Youden's J statistic: 0.9838



# Combining and deduplicating results

- Survey register – 12 149 facilities
- Survey register 55.2 – 6 736 facilities

| Method | Number of matched objects by model | | |
|---|---|---|---|
| | Total | NACE 55.2 | |
| | | in total | in which: perfect matches (1:1) |
| Splink | 0 | 0 | - |
| Natural Language Processing (NLP) | 2 393 | 717 | 0.42% |
| Machine learning algorithms (ML) | 3 157 | 749 | 0.40% |
| Fuzzy matching | 11 274 | 2845 | 3.80% |

| Method | Number of matched objects (model + manual review) | |
|---|---|---|
| | Total | NACE 55.2 |
| | | in total |
| Splink | 0 | 0 |
| Natural Language Processing (NLP) | 2 393 | 954 |
| Machine learning algorithms (ML) | 3 157 | 1525 |
| Fuzzy matching | 11 274 | 3248 |

**3858** newly identified holiday and other short-stay accommodation facilities

# Comparison of images of object offers

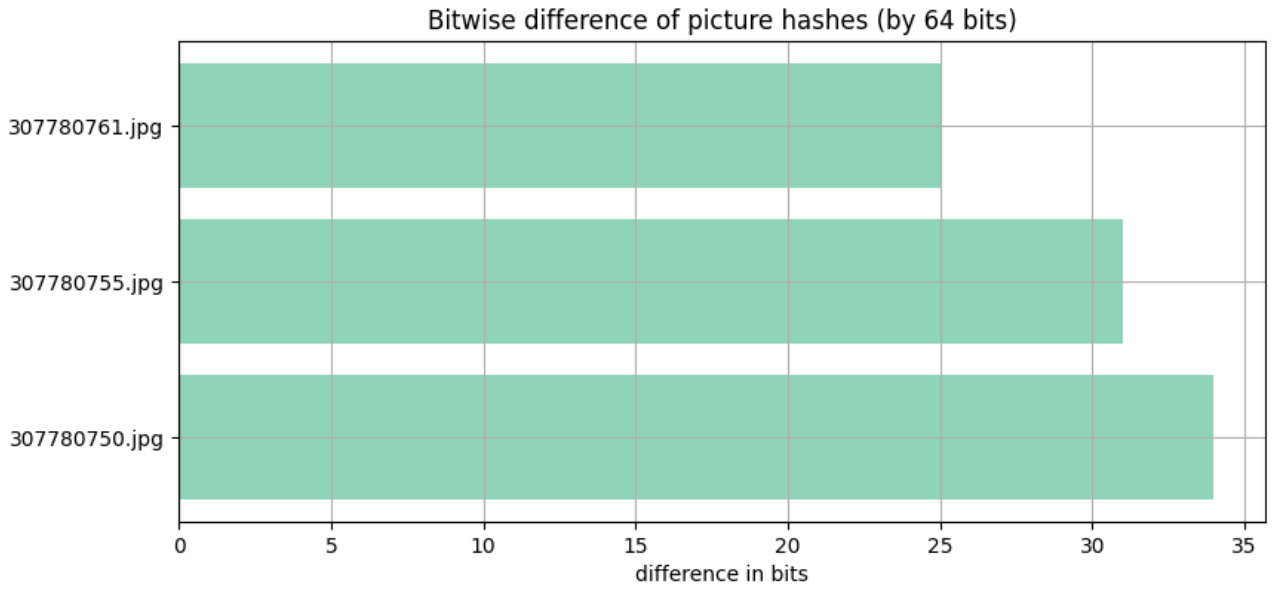
Identifying and comparing images is a very complex topic.
It involves converting an image into digital form and then performing operations that allow obtaining valuable information.

There are a few types of image processing:

- Visualization
- Recognition
- Sharpening and Restoration
- Pattern Recognition
- Retrieval

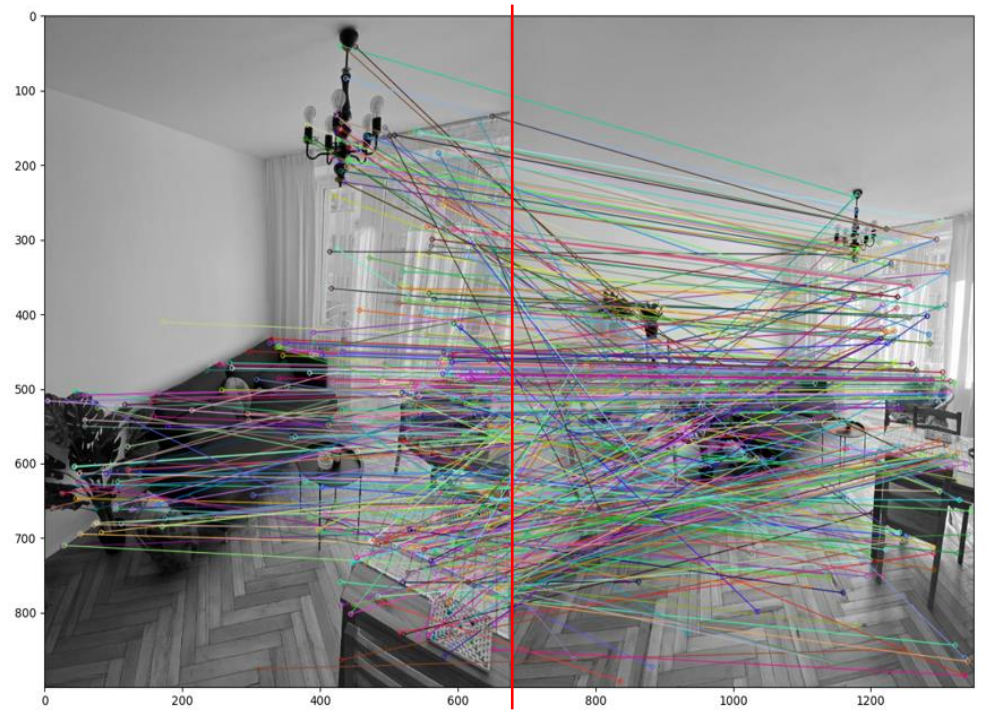
# Comparison of images of object offers – „digital fingerprint”

Assigning a unique hash value to the image ('digital fingerprint')




# Comparison of images of object offers – „digital fingerprint”

The key-point detection


370 good matches from 1786 all matches

# Conclusions

- Three portals considered: Booking.com, Hotels.com and Airbnb.com - Booking.com was chosen due to its largest range of variables corresponding to the tourism survey frame;
- The evaluation of the selected methods of linking and deduplication of data was done using the confusion matrix, the ROC curve and Youden's J statistic;
- The best results were obtained using the Fuzzy Matching method based on Levenshtein similarity combined with Vincenty's formula - this method copes well with arbitrary notation of the names of establishments and can also be used to classify them;
- Improving the quality of the tourism survey frame - the use of web-scraped data resulted in an increase in the number of accommodation establishments classified as NACE 55.1 and NACE 55.2. I (151 new accommodation establishments in Poland, e.g. 1.1% of the total number of accommodation establishments constituting the tourism survey frame in Poland);
- A very promising further research direction is the possibility of using algorithms that compare images. This way, it is possible to combine data from different portals more efficiently (photos become an additional key of correlation).

# Conclusions

## Perspectives for official statistics:

- In short term, the 3 scenarios presented will prevail:
  - ✓ Big data is complementary to sample surveys (with leading role of sample surveys);
  - ✓ Big data is complementary to sample surveys (without leading role of sample surveys);
  - ✓ Gradual replacement of sample surveys by big data in some domains.
- Long-term changes in official statistics in the context of big data depend on:
  - ✓ The pace in terms of developing a coherent theoretical model;
  - ✓ Micro-data access management model and artificial intelligence management model:
    - Societies preferring privacy over technological development (e.g. Europe),
    - Societies prioritizing technological development over privacy (e.g. China, Korea).
- Prerequisites for the use of big data, namely a stable access to such data and a positive assessment of its quality.

IOS-ISI 2024 MEXICO CONFERENCE

# Bibliography:


- Galbraith J.K., (2015). *The end of normal*: *The Great Crisis and the Future of Growth*, Simon & Schuster.
- Gelman A., Stern H. (2006). *The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant*. The American Statistician, Vol. 60, No. 4.
- Kai-Fu Lee (2018). *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin.
- Mayer-Schönberger V., Cukier K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin.
- Szreder M*.,* (2019*). Statistical significance in the era of big data,* The Polish Statistician*,* Vol. 64, No.11*.*
- Szreder M., (2022). *Opportunities and illusions of using large samples in statistical inference,* The Polish Statistician, Vol. 67, No. 8.

IOS-ISI 2024
MEXICO CONFERENCE

# Thank you

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

IAOS-ISI 2024
MEXICO CONFERENCE

# THE FUTURE OF SPANISH TOURISM STATISTICS SYSTEM


Mrs. Belén González belen.gonzalez.olmos@ine.es
Mrs. Marta Sixto marta.sixto.neira@ine.es
Mr. Alfonso Fernández alfonso.fernandez.bes@ine.es
Mrs. Elena González elena.gonzalez.martin@ine.es
Mrs. María Velasco maría.velasco.gimeno@ine.es

INE SPAIN


**Friday 17 May 2024 09:00-10:30**

INEGI
International Statistical Institute
IAOS IMPROVING OFFICIAL STATISTICS

# The future of Spanish tourism statistics system:

1. Traditional tourism statistics: EU Regulation and beyond.
2. Tourism Statistics System today: experimental statistics
3. Tourism Statistics System: planning the near future

IOS-ISI 2024 MEXICO CONFERENCE

# Traditional statistics. EU Regulation and beyond

**Regulation (EU) No 692/2011 European statistics on tourism**

**Annex I: Internal Tourism**

- Occupancy surveys:
- No. of tourism night spent in non-rented accommodation (Inbound Tourism Expenditure Survey)

**Annex II: National Tourism**

Resident Travel Survey (ETR)

**Tourism Satellite Account**

*National Accounts*
*Balance of Payments*

**Prices and revenue management**

- Producer Price Indexes
- Revenue managemet indicators for hotels:

**Inbound Tourism**

- Border Survey (FRONTUR)
- Expenditure Survey (EGATUR)

IOS-ISI 2024 MEXICO CONFERENCE

# New Information Sources

**A. Bank card transaction data**

- Distribution of the expenditure made by foreign visitors in Spain.
- Distribution of expenditure by residents on their visits abroad by country of destination.


**B. Web scraping data**

Measurement of the number of tourism dwellings in Spain and their capacity

**C. Privately held data**

Estimation of tourist accommodation occupancy using data from digital platforms

experimental

https://www.ine.es/en/experimental/experimental_en.htm

**D. MNO data**

Measurement of national and inbound tourism from the position of cell phones

IOS-ISI 2024 MEXICO CONFERENCE

# Experimental statistics

**Available experimental statistics**

Company Demographic Profile

Distribution of expenditure by residents on their visits abroad by country of destination

Estimation of tourist accommodation occupancy using data from digital platforms

Studies on mobility on mobile phone

Multidimensional Quality of Life Indicator (MQLI)

Rental Housing Price Index (RHPI)

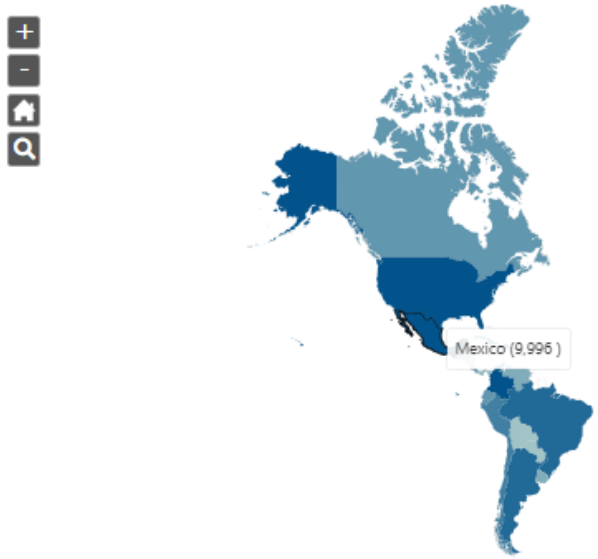
Measurement of Large Company Daily Retail Trade

Measurement of tl Number of Tourist in Spain and their

**Tourists abroad who are residents in Spain in the reference period**

Select the geographical scope: ○ ◉ ○ ○ ○

Period: 2024M02




See Total

**Mexico**


IOS-ISI 2024 MEXICO CONFERENCE

# Bank card transaction data

FOREIGN BANK CARDS

- Bank card transactions made in person at establishments located in Spain (POS Terminal).
- Cash withdrawals at ATMs located in Spain.

Aggregated tables by:
- Card issuing country
- Spanish province (NUTS3)

SPANISH BANK CARDS

- Bank card transactions made in person at establishments located abroad (POS Terminal).
- Cash withdrawals at ATMs located in other countries.

Aggregated tables by country where the transaction takes place.

**NO INDIVIDUAL INFORMATION ON TRANSACTIONS OR CARDHOLDERS**

IOS-ISI 2024 MEXICO CONFERENCE

# Bank card transaction data

Inbound Tourism Expenditure
(in destination)
*EGATUR*

Outbound Tourism Expenditure
(in destination)
*ETR*

SOME DRAWBACKS

EGATUR vs EXPERIMENTAL

- No details on goods and services purchased
- T+32 to T+90

ETR vs EXPERIMENTAL

- No details on goods and services purchased
- No data on region where residents live

# New Information Sources

**A. Bank card transaction data**

- Distribution of the expenditure made by foreign visitors in Spain.
- Distribution of expenditure by residents on their visits abroad by country of destination.


**D. MNO data**

Measurement of national and inbound tourism from the position of cell phones

**B. Web scraping data**

Measurement of the number of tourism dwellings in Spain and their capacity

**C. Privately held data**

Estimation of tourist accommodation occupancy using data from digital platforms

ACCOMMODATION

# Platform data

WEB SCRAPING DATA

- Web scraping (Python).
- Access to all the accommodations extracting their characteristics (name, bed places, licence, subtype, owner, coordinates, …)
- Algorithm to select tourist dwellings from all the accommodations extracted.
- Deduplication algorithm to remove accommodations that are in more than one platform.

Capacity variables.
Subset of NACE 55.2

DATA PROVIDED BY PLATFORMS

- Agreement EUROSTAT with digital platforms to provide data. Also agreements EUROSTAT-NSIs.
- Aggregated data for occupancy variables. Also capacity variables (duplication).
- Quarterly data on monthly bases.

Occupancy variables.
NACE 55.2

**THESE DATA CAN NOT BE MERGED**

IOS-ISI 2024 MEXICO CONFERENCE

# Platform data

**Measurement of the number of tourist dwellings in Spain and their capacity**

Reference periods:
February and August

**Estimation of tourist accommodation occupancy using data from digital platforms**

Reference periods:
Monthly

SOME DRAWBACKS

- No stability to access data

- Can´t be integrated
- Difficult to change agreements
- Capacity variables can´t be used

# Measurement of national and inbound tourism from the position of cell phones

| Inbound Tourism | Outbound Tourism | Internal Tourism |
|---|---|---|
| - Variables: | - Variables: | - Variables: |

SOME DRAWBACKS

- Budget
- Only flows (more variables are required)
- Adaptation of MNOs databases to tourism definitions and needs

**Confidentiality**: INE only receives files with aggregated data

(we don't have access to individual records nor can identify individuals).

Only cells with 30 tourists or more are published

IOS-ISI 2024 MEXICO CONFERENCE

# Tourism Statistics System: planning the future

## Traditional statistics. EU Regulation and beyond

**Regulation (EU) No 692/2011 European statistics on tourism**

**Annex I: Internal Tourism**

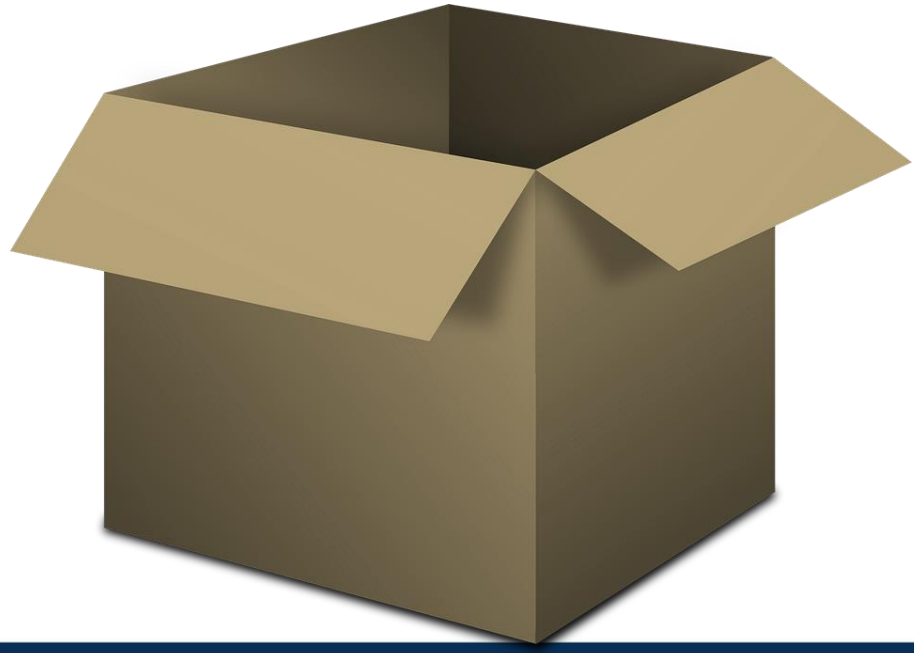
- Occupancy surveys:
- No. of tourism night spent in non-rented accommodation (Inbound Tourism Expenditure Survey)

**Annex II: National Tourism**

Resident Travel Survey (ETR)

**Tourism Satellite Account**

*National Accounts*
*Balance of Payments*



**Prices and revenue management**

- Producer Price Indexes
- Revenue managemet indicators for hotels:

**Inbound Tourism**

- Border Survey (FRONTUR)
- Expenditure Survey (EGATUR)


# Tourism Statistics System: planning the future

**PROS**

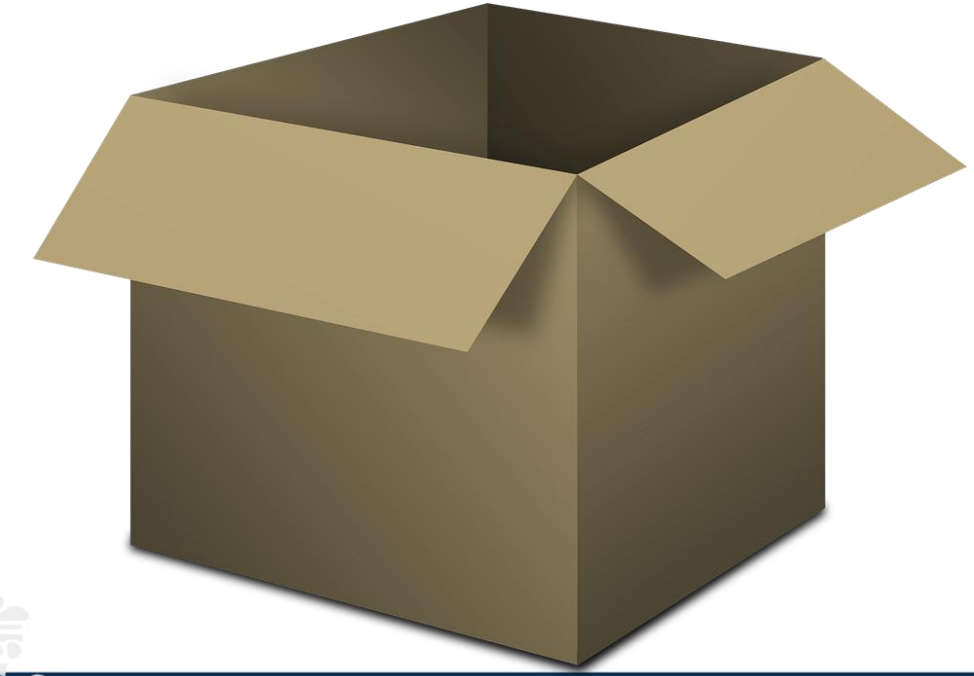
Traditonal surveys and sources.
New sources and methods.

**CONS**

Traditonal surveys and sources.
New sources and methods.

**NEEDS**

- National users
- International Organizations
- Regulations requirements
- Methodological Frameworks



IOS-ISI 2024 MEXICO CONFERENCE

# Border Survey + MNO data

## INTEGRATION: taking the best of both sources

- Coverage and granularity from MNO data
- Traditional survey to collect data of the characteristics of trips and expenditure.
- Administrative registers as reference framework of border movements

### Working Group at INE

- Tourism Stat. Unit
- Data Collection Unit
- Metolodogy Unit
- Sample Design Unit
- TIC Unit

### Close collab INE-MNOs

- Improving experimental datasets
- Defining new datasets
- Testing solutions

IOS-ISI 2024 MEXICO CONFERENCE

# Final remarks

- ✓ The new sources of information are a great opportunity for the producers of statistics since they will allow us to provide much more **frequently, timely and detailed information**.
- ✓ It is important to reach **agreements** with the owners of the information to access the databases.
- ✓ It will be necessary to work together with the owners of the information to develop **new methodologies to integrate the new sources into the traditional statistics**.
- ✓ The **definitions used in tourism surveys should be reviewed** to see if new sources of information can capture the same concepts.
- ✓ Reputational impact of data transfers. **Data confidentiality.**

IOS-ISI 2024 MEXICO CONFERENCE

IOS-ISI 2024
MEXICO CONFERENCE

# Thank you

# Muchas gracias

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS