Statistics Poland

Tourism statistics in the high tech era

Marek Cierpial-Wolan, Assoc. Prof.
Statistics Poland, University of Rzeszow

Agenda

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

Rapid change in IT

Flood of (big)data

Fierce competition in information market

What should be done?

Data integration

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

- Data integration – the categorical imperative in statistical research
  - ✓ demand for real-time and more disaggregated data that responds to the needs of stakeholders
  - ✓ necessity to reduce the effects of the growing scale and importance of non-sampling errors
- Short-term data integration scenarios in tourism
  - ✓ Big data is complementary to surveys (with/without leading role of surveys)
    - verify and improve of the sampling frame
    - calibrate sampling weights, impute missing data
    - improve the quality of inference (e.g. paradata)
  - ✓ Gradual replacement of surveys by big data in some domains.

Data integration – scenario 1

- Big data is complementary to (sample) surveys (with/without leading role of (sample surveys)

  ✓ Big data can provide the valuable knowledge needed to: **verify and improve of the sampling frame**, calibrate sampling weights, impute missing data

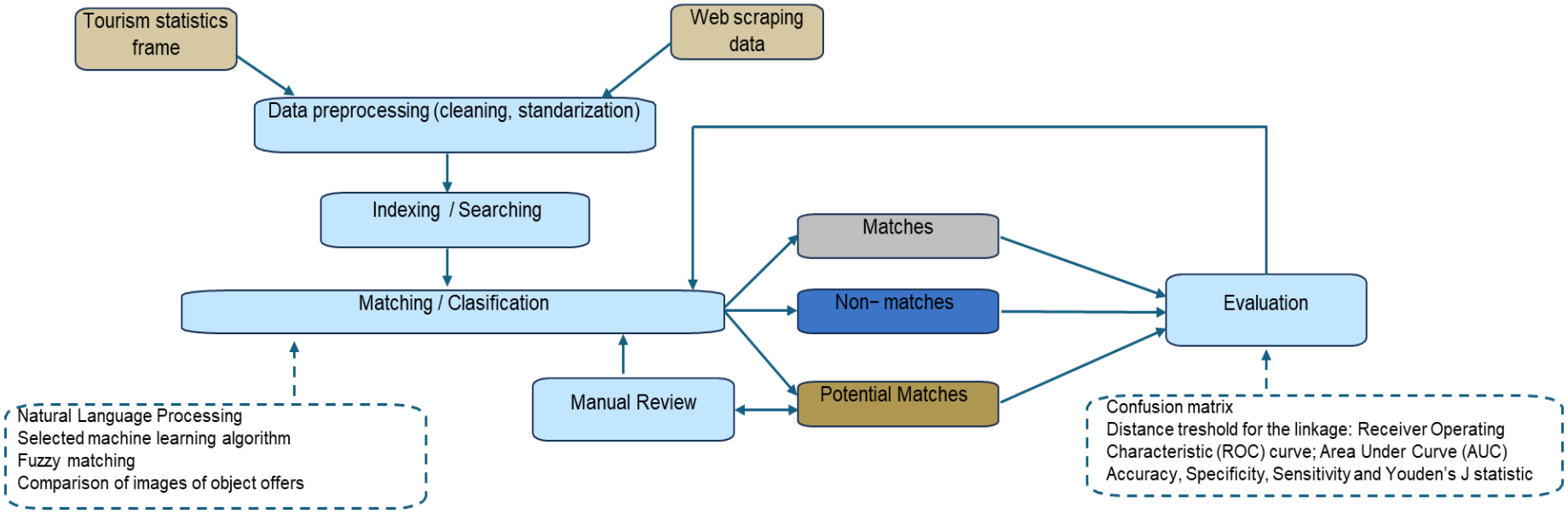
  ✓ Big data technologies can be used to collect and process data (e.g. metadata and paradata) that can improve the quality of inference

Improvement of survey frame

Selected big data sources	Selected administrative registers
Web Scraping	**Register of categorized facilities**
Smart City systems	**Register of non-categorized facilities**
Mobile network operators	**Business register**
Payment/credit card operators	**Geo register**
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

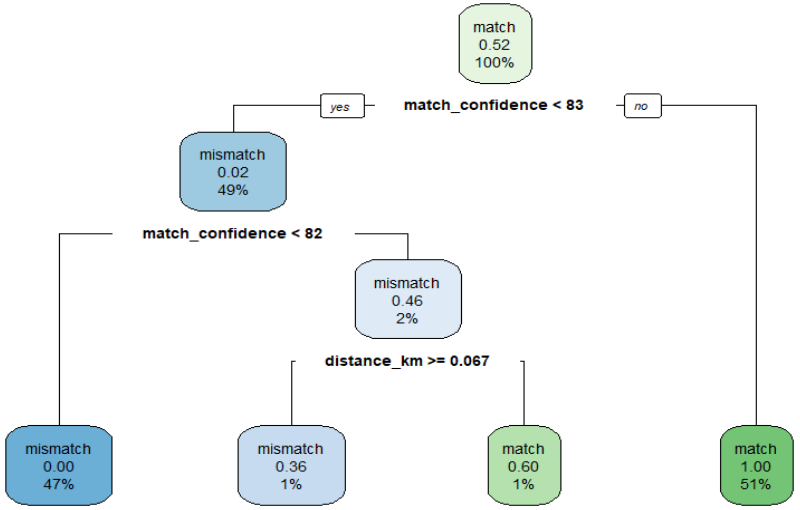
Probabilistic data linkage – simplified diagram



Fuzzy matching with geolocation

- Combination of fuzzy matching and geodetic distances (Vincenty's formula)
- To utilize two matching criteria and find the set of decision rules the decision tree was applied

  - Results
    - Accuracy: 0.9919
    - Specificity: 0.9921
    - Sensitivity: 0.9917
    - Youden's J statistic: 0.9838



Data integration in tourism
scenario 1

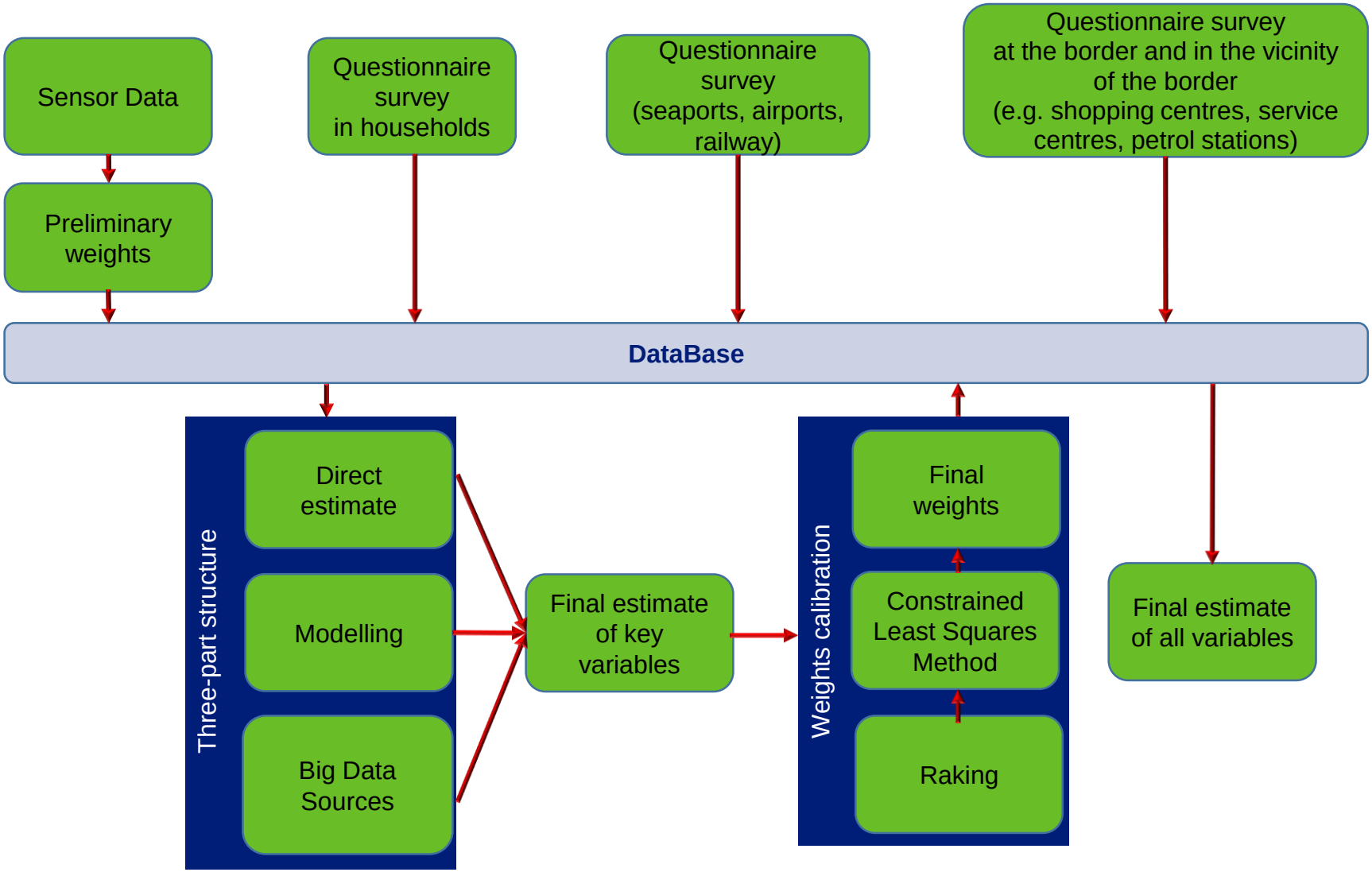
- Big data is complementary to (sample) surveys (with/without leading role of sample surveys)

  ✓ Big data can provide the valuable knowledge needed to: verify and improve of the sampling frame, **calibrate sampling weights, impute missing data**

  ✓ Big data technologies can be used to collect and process data (e.g. metadata and paradata) that can improve the quality of inference

Data integration in tourism

Improvement of final estimate of key variables



eg. Flights schedules

Tourism dashboard

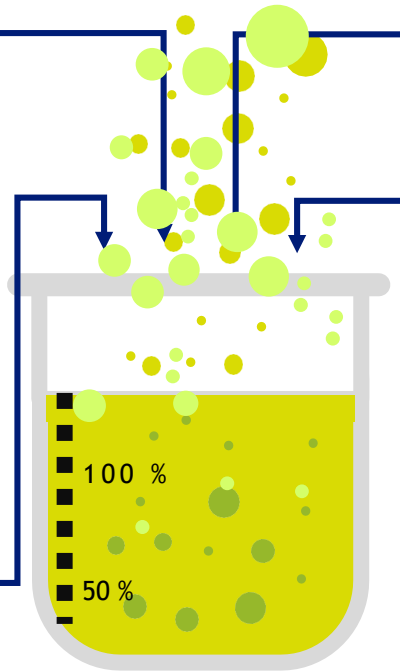
Data sources

Accommodation establishments survey
(10 or more beds)

- tourists
- accommodation
- origin of tourists
- occupancy rate

Survey of tourist travel and expenses

- tourist expenses
- destinations



Data integration

Web Scraping of booking platforms
(all objects)

- accommodations
- tourists' opinions about accommodation establishments

Booking platforms
(less than 10 beds)

- accommodation
- reservations

Booking.com
airbnb
Expedia
Tripadvisor

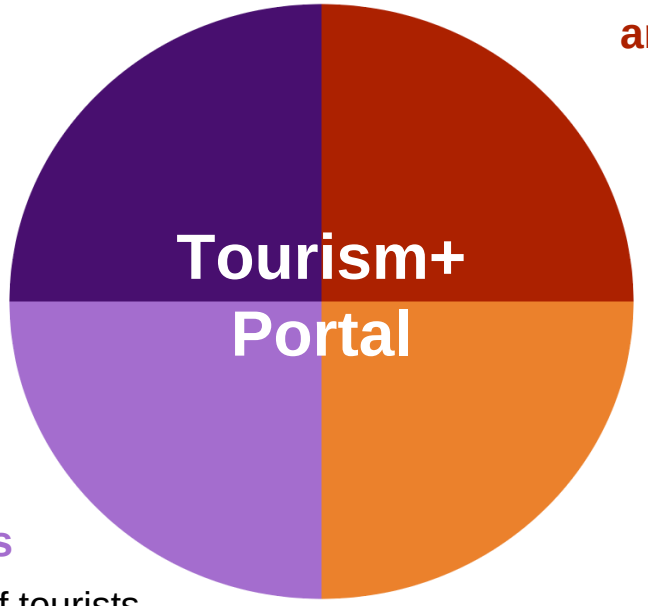
Tourism in Poland

Website: turystyka.stat.gov.pl

Main modules

Accommodation

Research on travel and tourist spending



Experimental Statistics

- Forecast of the number of tourists
- Analysis of tourists' opinions

Regional analysis

- Poland, regions, counties ...

Website: turystyka.stat.gov.pl

Statistics Poland

Data integration in tourism
scenario 2

- Gradual replacement of sample surveys by big data in some domains

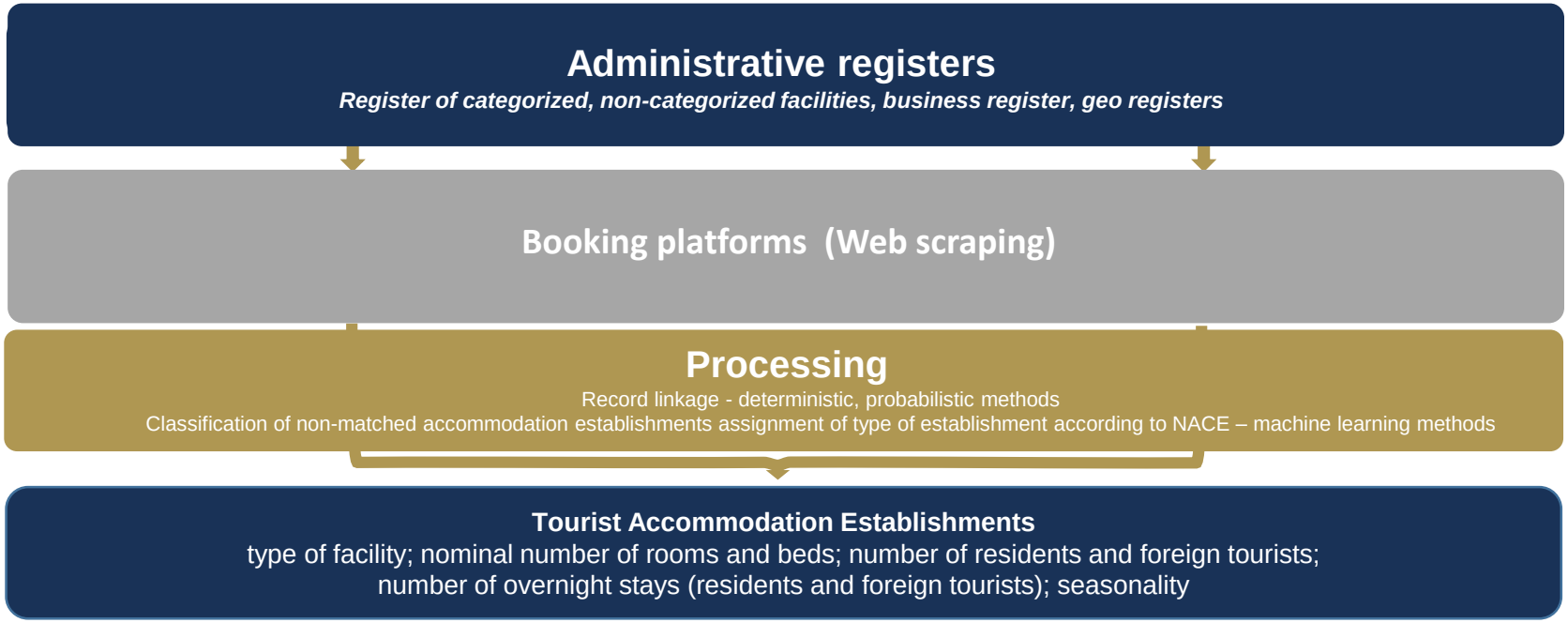
It is not possible to replace sample surveys everywhere

✓ In many fields, especially social life, it is important to accurately define the characteristics of the population not only the overall picture or interdependence of features;

✓ Researchers are not always content to learn about correlational relationships, very useful for forecasting, but less valuable in explaining phenomena.

Data integration in tourism
scenario 2

Tourism accommodation establishment surveys
Input and Outputs

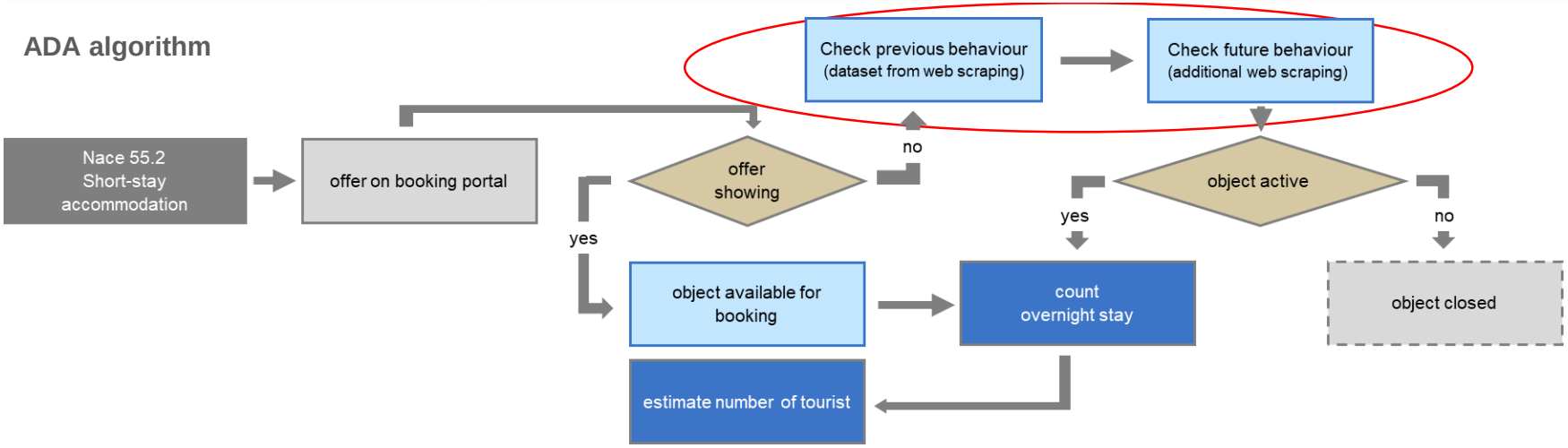


Data integration in tourism
scenario 2

Derive new variables

Variables \ Sources	Web Scraping ML methods	Automatic Report for Establishments (ARE) (NACE 55.1)	Automatic Detection of Activity (ADA)/ARE (other NACE 55)
Type of facility	x	x	x
Nominal number of rooms and beds	x	x	x
Number of residents and foreign tourists		x	
Number of overnight stays (Residents/foreign tourists)	x	x	x
Seasonality (months of activity)	x	x	x

ADA algorithm



Conclusions

Perspectives for tourism statistics:

- In short term, the 2 scenarios presented will prevail:
  - ✓ Big data is complementary to surveys (with/without leading role of 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

Statistics Poland

Thank you for your attention