Statistics Poland

Big data and tourism statistics – challenges and prospects

Marek Cierpiał-Wolan, Assoc. Prof.

Statistical Office in Rzeszów
University of Rzeszów

Agenda

1. Background

2. Some quality aspects

3. Scenarios for using big data in (official) statistics

4. Conclusions

Challenges of statistics

Background

- Global security, Energy, COVID-19, Global migration crisis, Rapid development of IT, Fierce competition on information market;
- Social expectations – high emotional charge

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

- Modern data analysis, in many cases, goes beyond the traditional understanding of statistics;
- Methodology of statistics as a scientific discipline must constanly be changing.

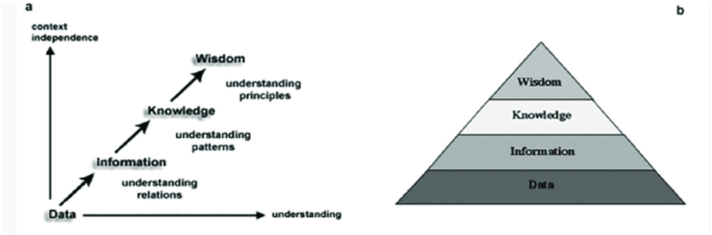
Quality of data in official statistics vs other data producers

- Fierce competition on the information market
  - ✓ Competitors of official statistics - especially companies in the business sector, do not always have to rely on the quality paradigm in their strategy;
- Paradox:
  - ✓ The dominant position of a given entity on the information market should be determined by the quality of the data provided;
  - ✓ Better information is crowded out by worse information;
- Why official statistics must pay special attention to quality issues:
  - ✓ The information system built and coordinated by national statistical institutes is supposed to bring a certain information order in society;
  - ✓ Official statistics have an indirect impact on the living conditions of the population and the operating conditions of businesses.

Can statistics still be a beacon in the contaminated information environment of today’s world?

Quality and errors

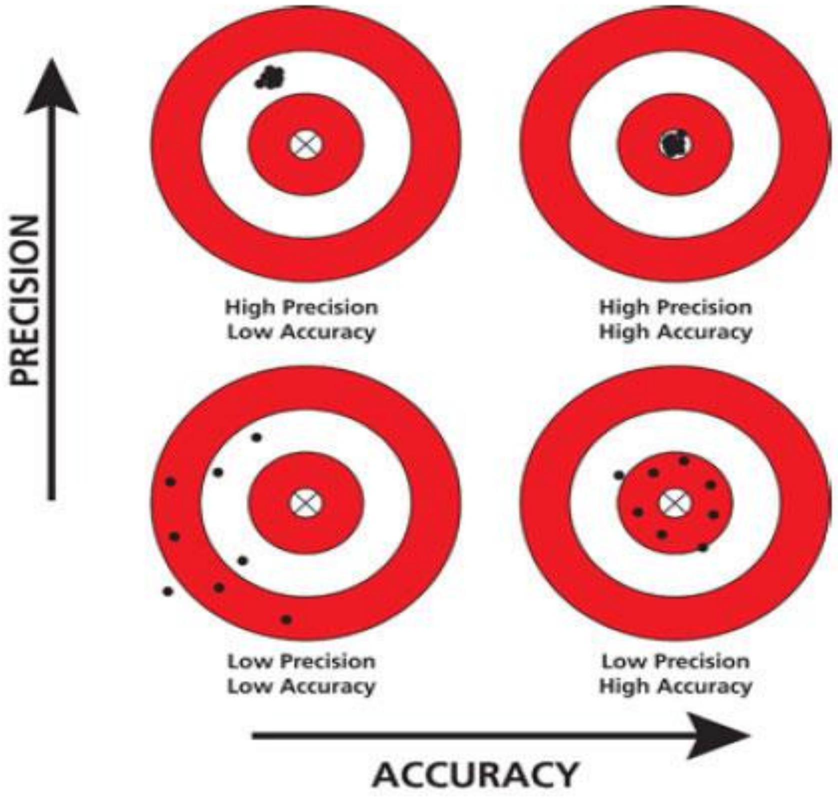
We need to deliver good quality data in real time



- Information quality depends on data quality;
- Main data sources: census surveys, **sample surveys** (most of data), administrative registers, **big data**;
- Error is an inherent part of survey;
- In the surveys there are two basic types of error:
  - ✓ Sampling error;
  - ✓ Non-sampling error.

Assessment of data quality

- **Accuracy –** difference between a survey results and the true value of a characteristic of the population;
- **Precision (reliability) –** indicates how close measure values are to each other**.**



Census survey – sample survey – admistrative registers – BIG DATA
Confusion

- The emergence of big data is changing approach to data analysis;
  - ✓ huge number of observations,
  - ✓ opportunity to improve the quality of inference, under the growing scale and importance of non-sampling errors.

„The idea of sampling loses its meaning when we can use a large number of data" [Mayer-Schönberger, Cukier, 2014, p. 50]

- is it true?

Big data and quality

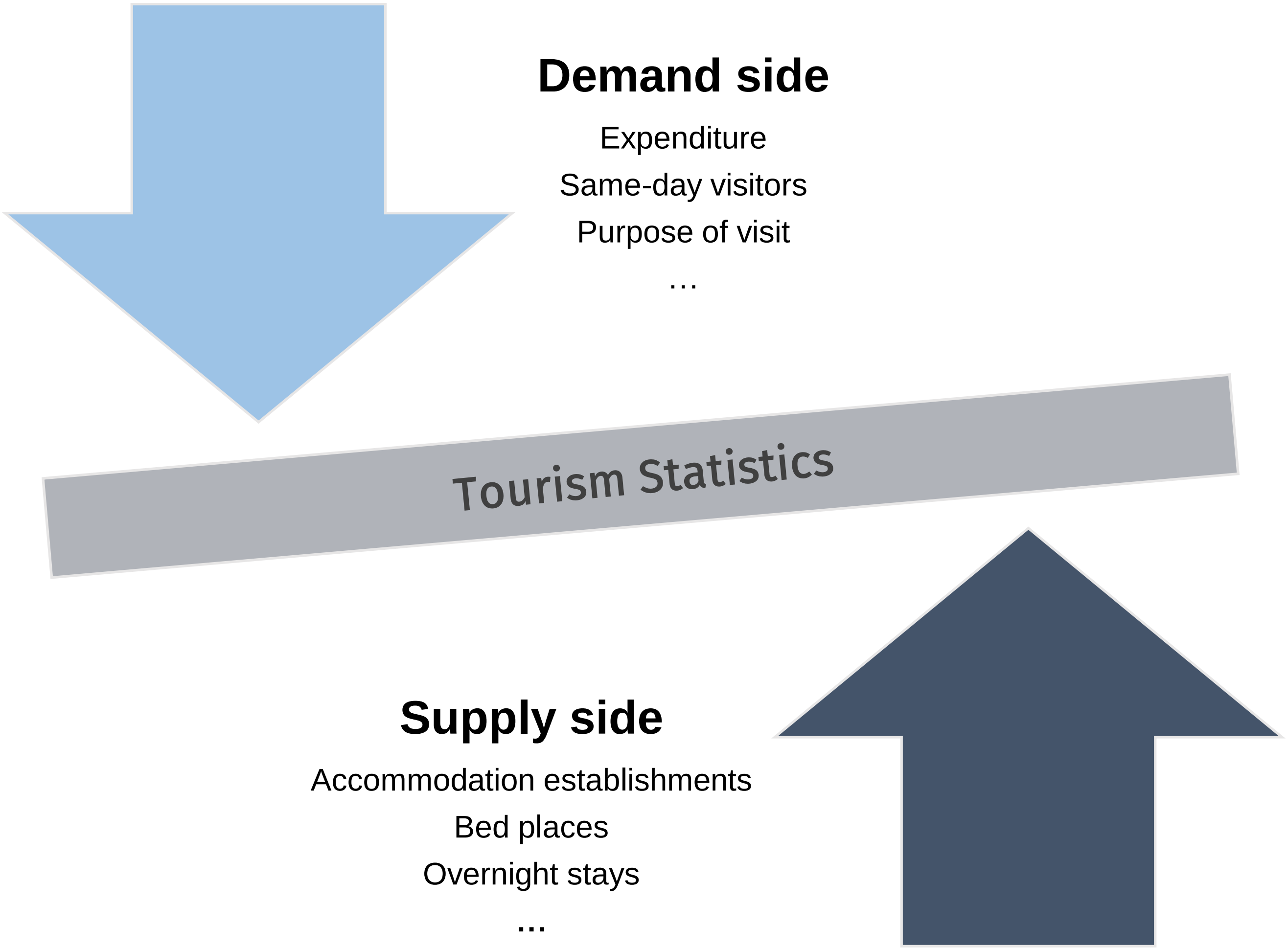
- Increase in the number of observations (mainly thanks to big data):
  - ✓theoretically leads to lower sampling errors (under the condition of randomness of the sample - very unlikely),
  - ✓the problem of non-sampling errors remains (theoretically they should be smaller) – even census survey may be subject to non-sampling errors;
- Increasing failure to adhere the assumptions of the statistical inference model and the rigor of sampling;
- Increasing the risk of erroneous decisions using statistical inference methods;
- Increasing the number of observations in the sample cannot relieve the researcher from the duty to carefully analyze **the quality of the data.**

Data integration

census survey – sample survey – administrative registers – big data

- Untapped potential of big data - methodological challenges for data integration and thus even more sensitivity in terms of output quality assessment;
- Additional sources of information have been used in sample survey for years (statistical inference theory, particularly the Bayesian paradigm, sample selection method where one of the assumption is 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.

Tourism Statistics



Demand side

Expenditure
Same-day visitors
Purpose of visit
…

Supply side

Accommodation establishments
Bed places
Overnight stays
…

Opportunities – scenarios for using big data in (official) statistics (1)

- Big data is complementary to sample surveys (with leading role of sample surveys)

  ✓ Big data can provide the valuable knowledge needed to: impute missing data, verify and improvement of the sampling frame, correct the sample structure using imputation and calibration techniques;

  ✓ Big data technologies can also be used to collect and process data that can improve the quality of inference, such as the metadata and paradata sets.

Improvement of survey frame (a)

Survey frame of accommodation establishments

Register of Hotels and similar accommodation

- obtained from Ministry of Sport and Tourism

Booking platforms (Web scraping)

+ all types of facilities

+ frequently updated

- linking data with a statistical survey

Improvement of survey frame (a)

Nocowanie.pl
krok po kroku

Expedia

airbnb

Booking.com

eholiday.pl
DOBRE NOCLEGI

Hotels.com™

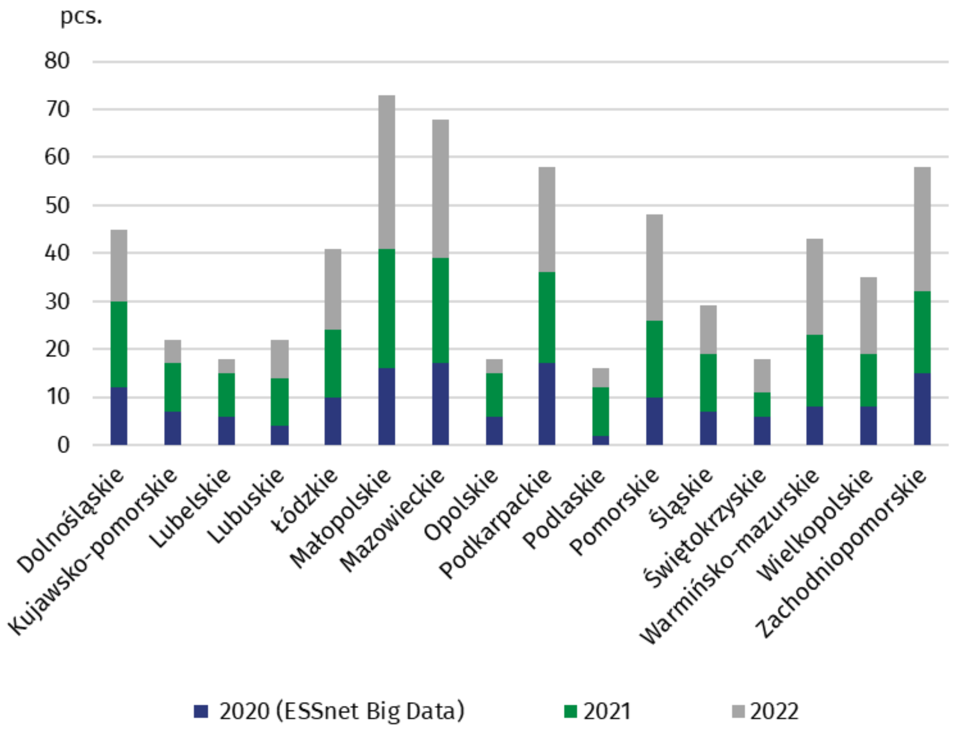
OdkryjPolske.pl
Polski Portal Rezerwacyjny

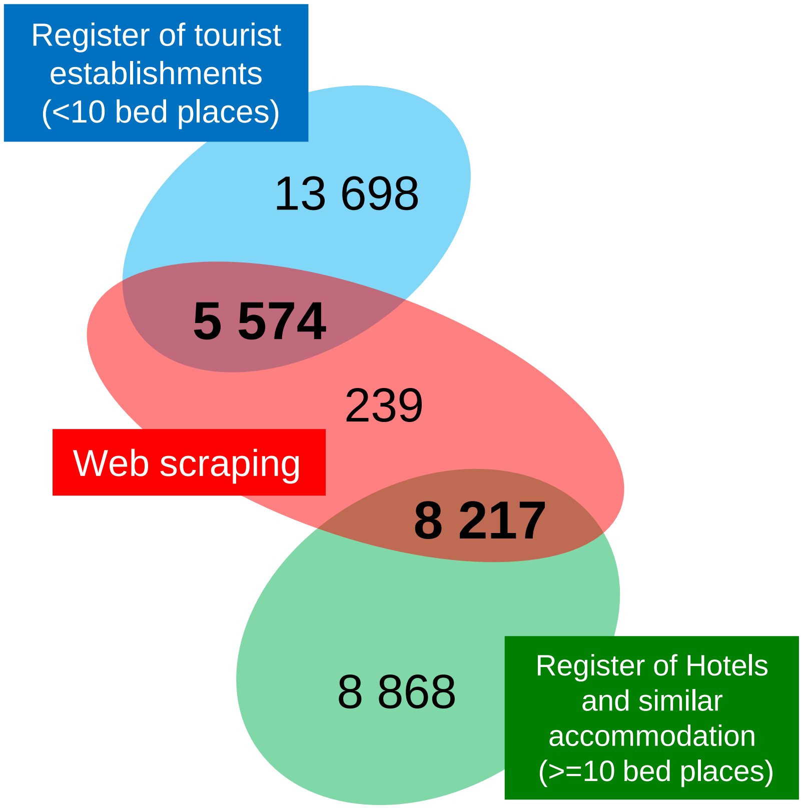
Tripadvisor

Improvement of survey frame (a)

Web scraping – from 2020 around 600 new accommodation establishements (increase by 8%).

New accommodation establishements by regions





Register of tourist establishments (<10 bed places)

13 698

5 574

239

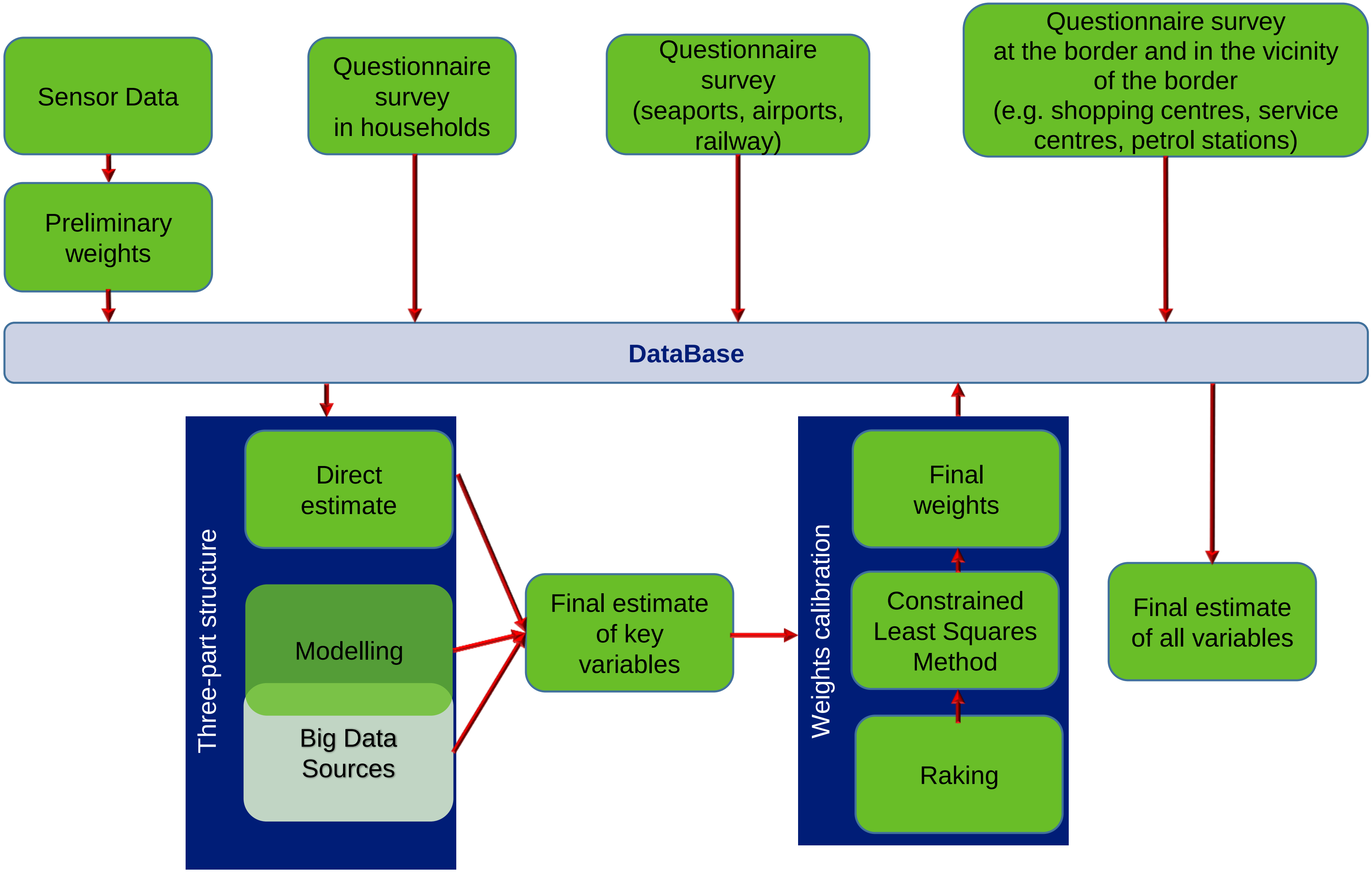
Web scraping

8 217

8 868

Register of Hotels and similar accommodation (>=10 bed places)

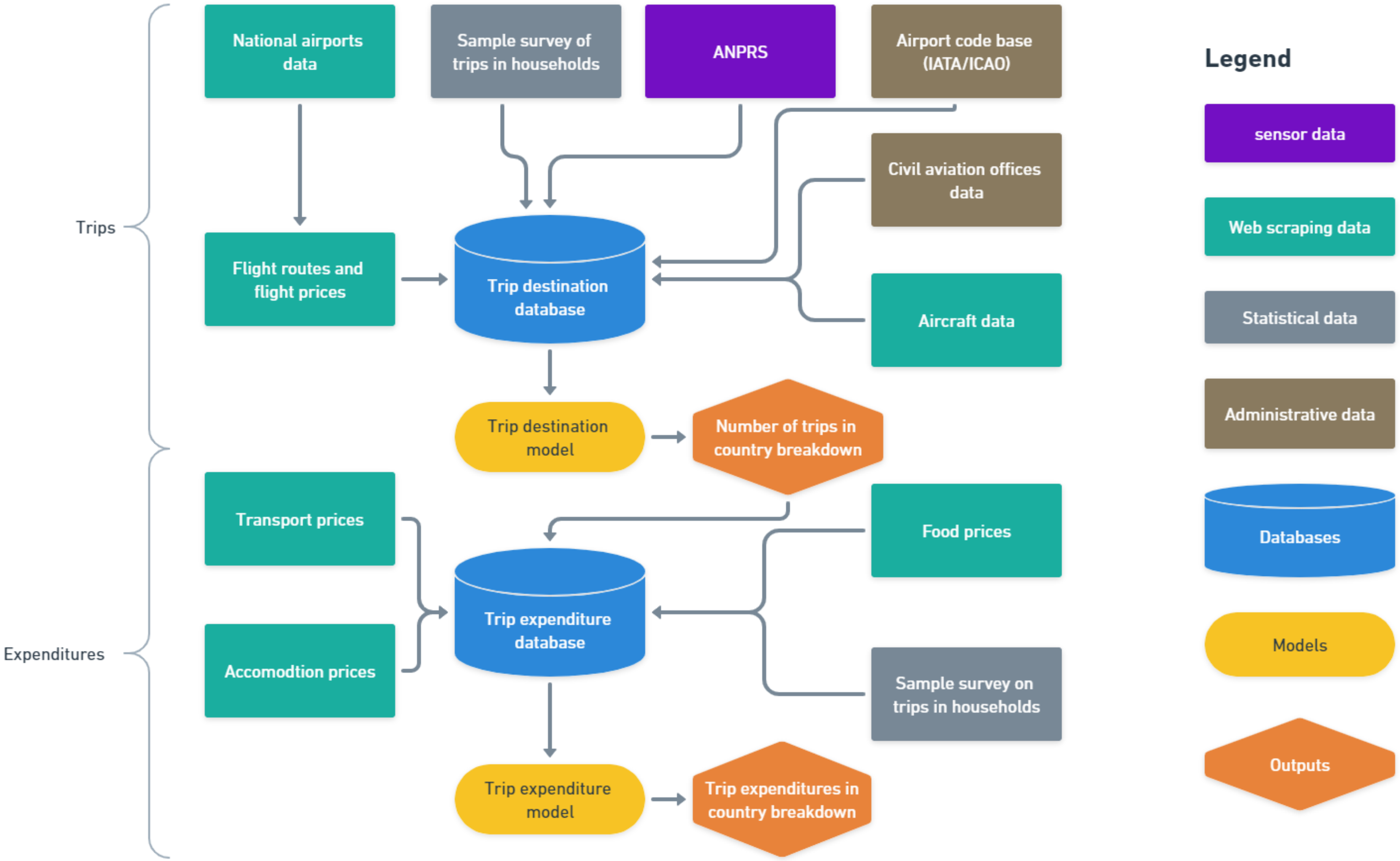
Improvement of final estimate of key variables (b)


Sensor Data
Preliminary weights
Questionnaire survey in households
Questionnaire survey (seaports, airports, railway)
Questionnaire survey at the border and in the vicinity of the border (e.g. shopping centres, service centres, petrol stations)
DataBase
Three-part structure
Direct estimate
Modelling
Big Data Sources
Final estimate of key variables
Weights calibration
Final weights
Constrained Least Squares Method
Raking
Final estimate of all variables

Statistics Poland

Big data as complementary data source

expenditure estimation model



Trips

National airports data

Sample survey of trips in households

ANPRS

Airport code base (IATA/ICAO)

Civil aviation offices data

Flight routes and flight prices

Trip destination database

Aircraft data

Trip destination model

Number of trips in country breakdown

Expenditures

Transport prices

Food prices

Trip expenditure database

Accomodtion prices

Sample survey on trips in households

Trip expenditure model

Trip expenditures in country breakdown

Legend

sensor data

Web scraping data

Statistical data

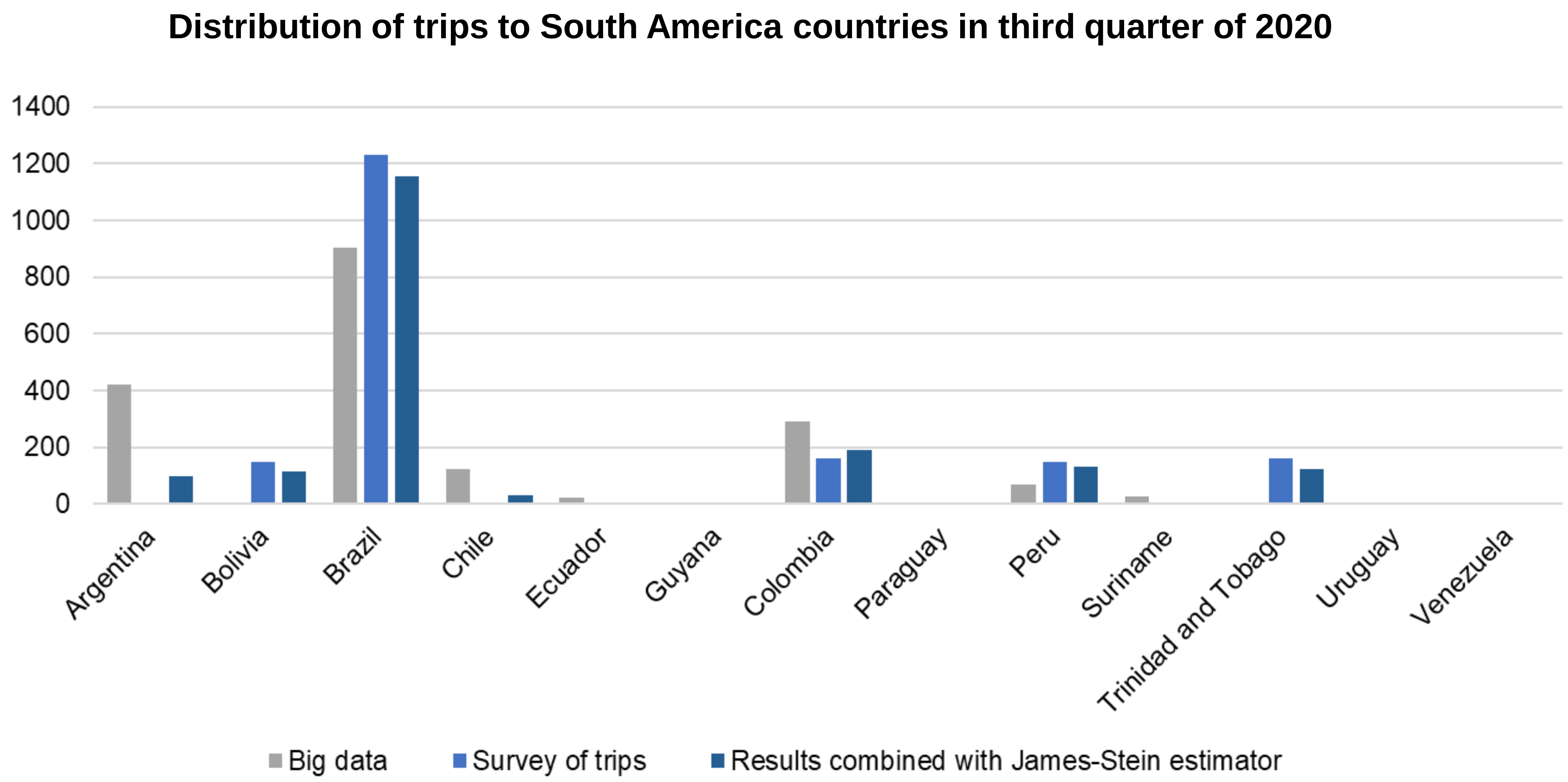
Administrative data

Databases

Models

Outputs

Big data as complementary data source



For 2020 – 20 new countries, total expenditure increased after modelling by almost 18%.

Statistics Poland

Opportunities – scenarios for using big data in (official) statistics (2)

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

Dominant position of NSI in data integration process:

- ✓ free access to micro-data from administrative registers,
- ✓ ability to use many of its own censuses and sample surveys conducted systematically.

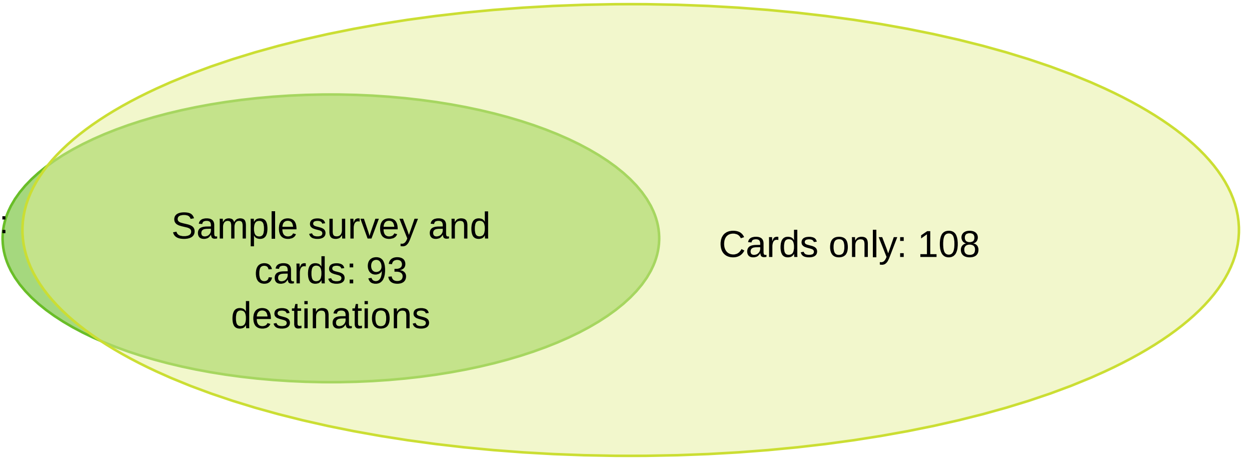
Data integration (2)

sample survey in households + big data (Visa cards)

- Sample of 18 750 households (~ 50 ths. interviewees)
- 0,13 % population
- Includes credit card and cash payments

There is ~18,5 mln active cards
17,8 % population
Assuming 2,7 cards per person ~6,8 mln users

Coverage of destinations



Sample survey only: 2 destinations

Sample survey and cards: 93 destinations

Cards only: 108

- The precision of the estimate of the fraction :

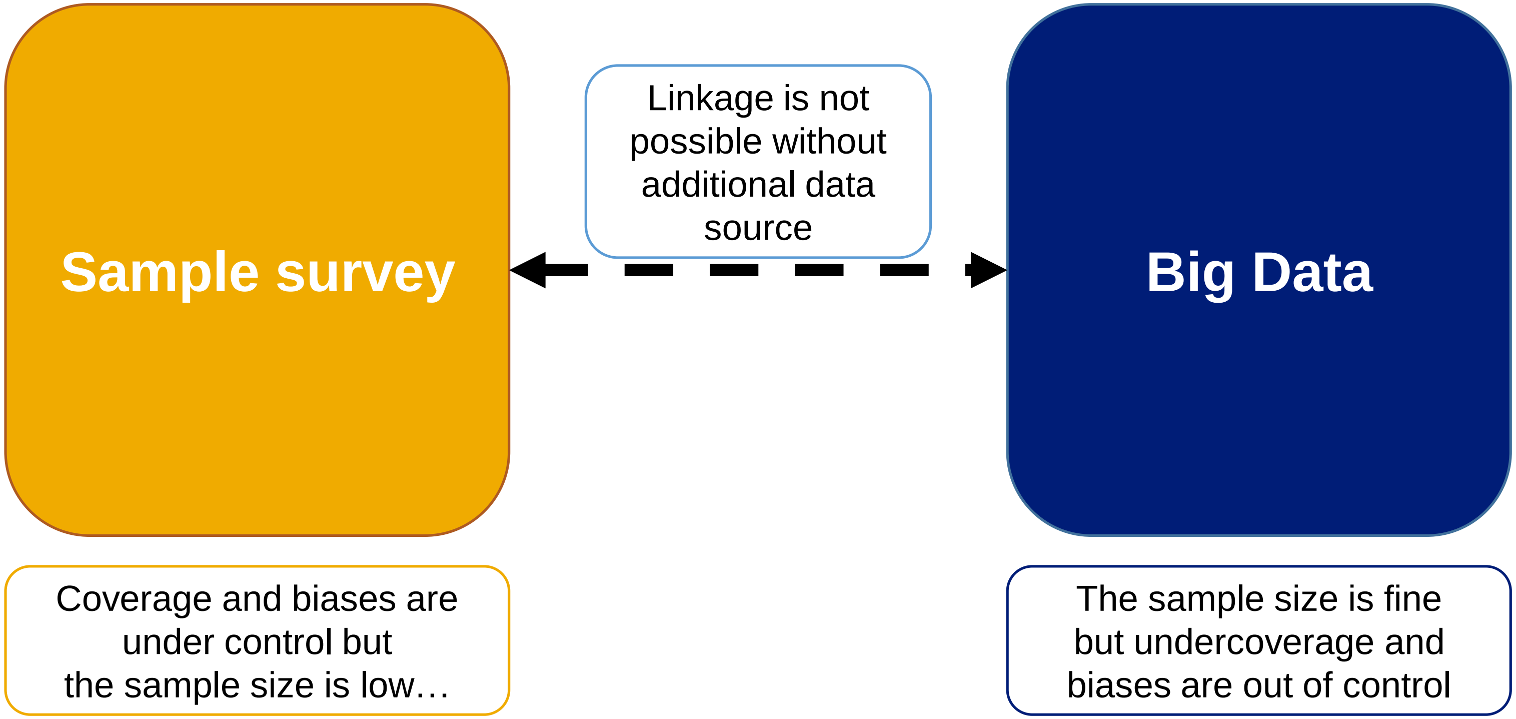
  ✓ sample survey on trips: 0,2270%
  ✓ credit cards operator: 0,0208%

Statistics Poland

Data integration (2)

sample survey – big data

Big Data + Sample Survey



Sample survey

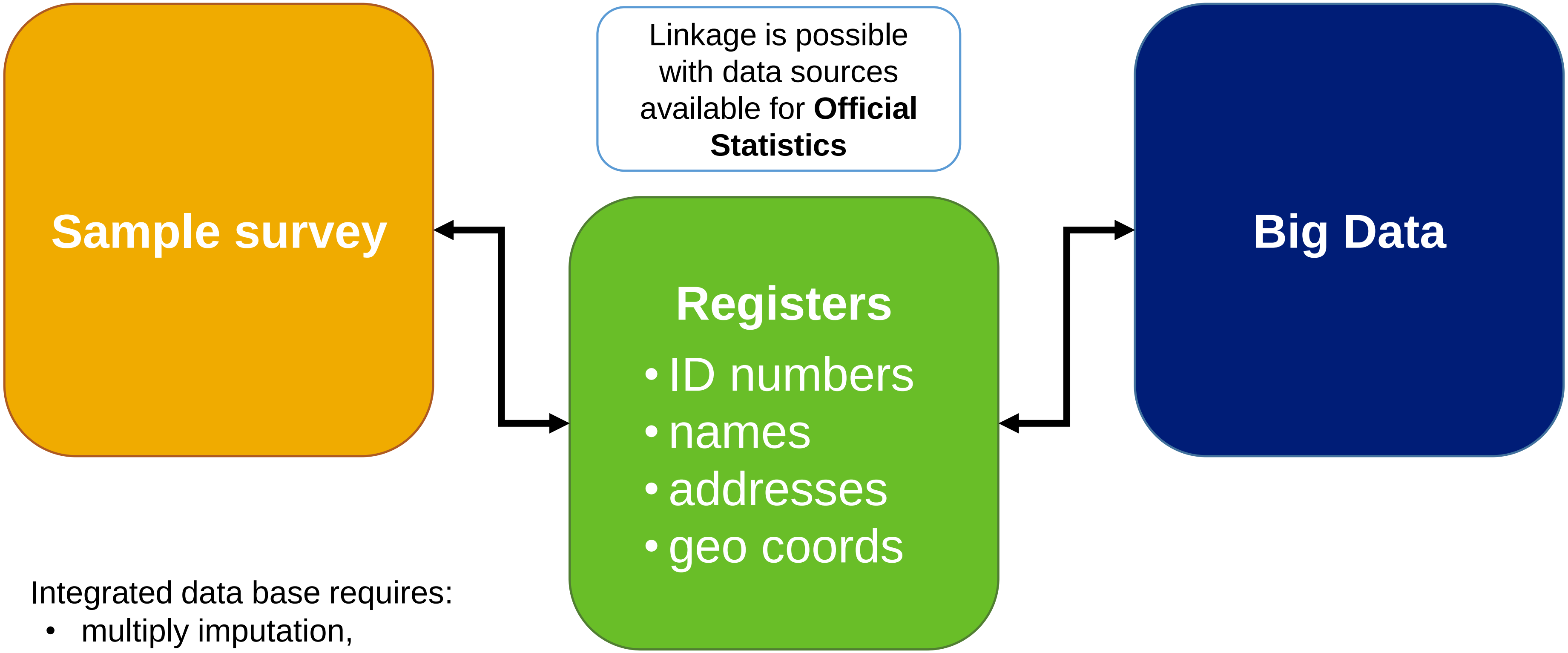
Linkage is not possible without additional data source

Big Data

Coverage and biases are under control but the sample size is low…

The sample size is fine but undercoverage and biases are out of control

Data integration (2)
sample survey – administrative registers – big data



Linkage is possible with data sources available for **Official Statistics**

Integrated data base requires:

- multiply imputation,
- weights assignement and calibration.

Benefits:

- better control of coverage and possible biases,
- wider set of variables and cases,
- statistical interference enabled.

Opportunities – scenarios for using big data in (official) statistics (3)

- 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;

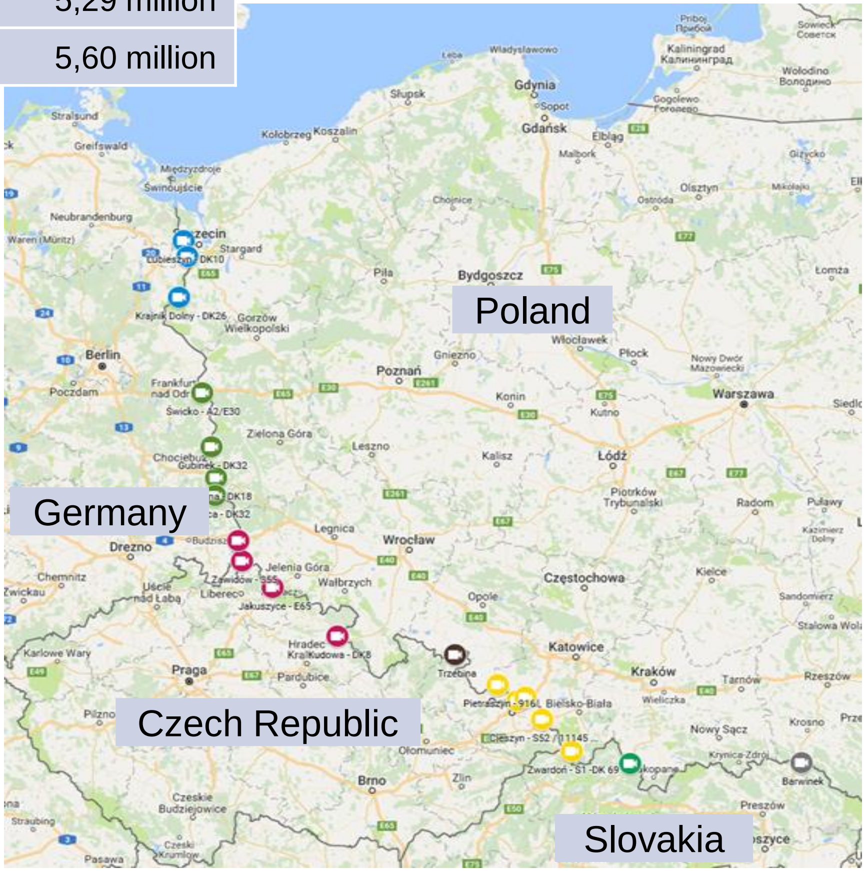
✓ Researchers are not always content to learn about correlational relationships, very useful for forecasting, but less valuable in explaining phenomena.

Sample survey vs. ANPRS (a)

Traffic surveys at the EU’s internal border crossings (vehicles)

Source	IV quarter 2019	I quarter 2020
ANPRS	7,15 million	5,29 million
Traffic intensity survey	7,17 million	5,60 million

Quarterly data from traffic sensors from the **ANPRS** system allowed for the development of the volume of border traffic of vehicles and people on the internal border of the European Union in Poland.



Accommodation establishement survey
Vs.
Registers+Web scraping (b)

- **Input (main data sources)**

  administrative registers (Register of Hotels and similar accommodation) and web scraping of global and regional portals;

- **Processing**

  ✓ Data combining processess

  deterministic, probabilistic record linkage methods

  ✓ Classification of non-matched accommodation establishments

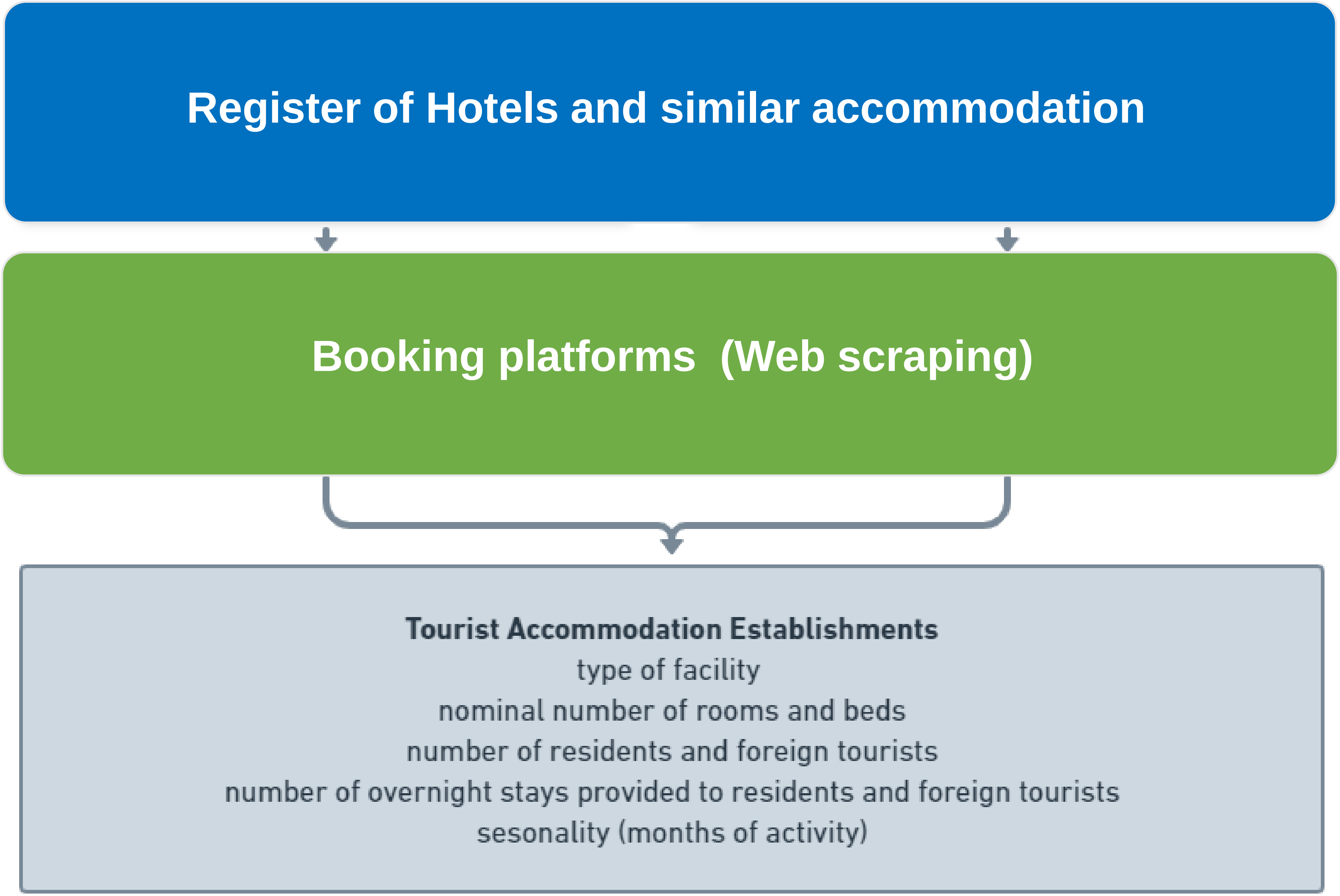
  assignment of type of establishment according to NACE Rev. 2 – machine learning methods;

- **Output - core indicators:**

  type of facility, nominal number of rooms and beds, number of residents and foreign tourists, number of overnight stays provided to residents and foreign tourists, seasonality (months of activity).

Tourism accommodation establishment surveys

Input and Outputs



Register of Hotels and similar accommodation

Booking platforms (Web scraping)

Tourist Accommodation Establishments

type of facility

nominal number of rooms and beds

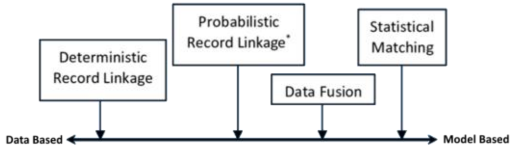
number of residents and foreign tourists

number of overnight stays provided to residents and foreign tourists

sesonality (months of activity)

Data combining processess

Methods used



*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

variable	method
name, address	fuzzy matching
coordinates	distance using Haversine or Vincenty formulas (distance treeshold)
type of establishment	machine learning (decision tree)

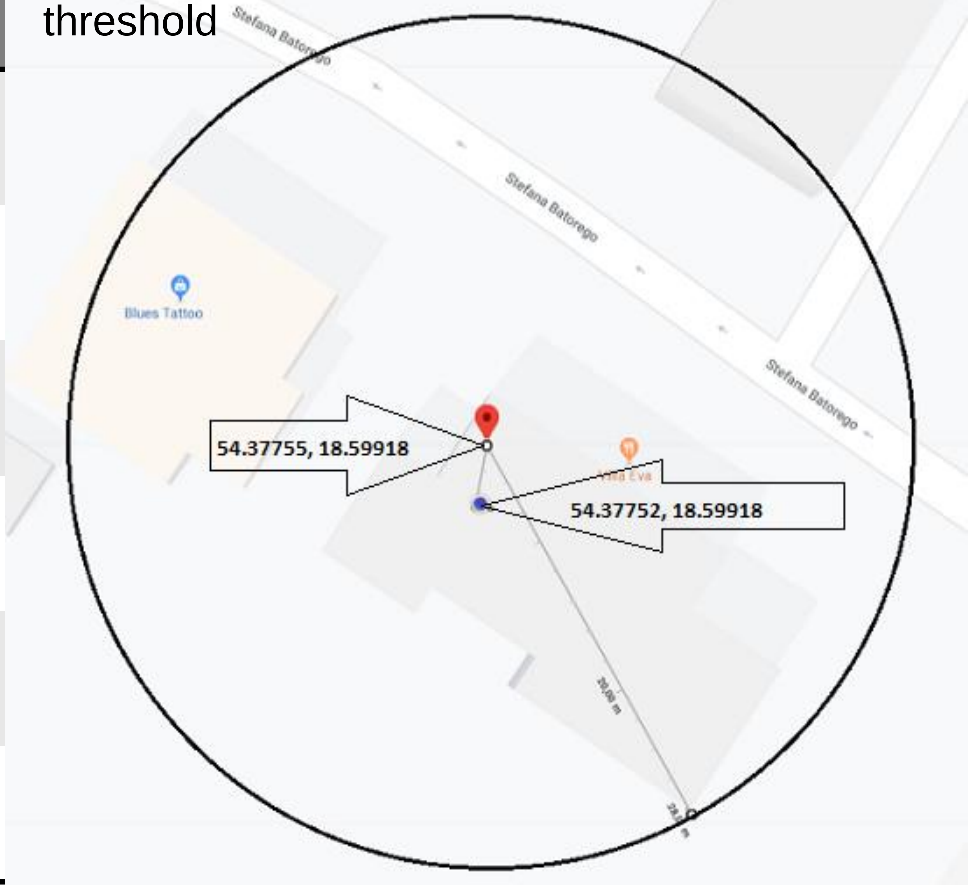
Statistics Poland

Deterministic Record Linkage

Solution?

Threshold	Precision	Sensitivity	Accuracy
30 m	0.99	0.5	0.82
50 m	0.97	0.52	0.82
70 m	0.99	0.55	0.83
100 m	0.98	0.52	0.8
200 m	**0.99**	**0.64**	**0.87**
500 m	0.97	0.6	0.81

distance-based methods with appropriate threshold



Haversine and Vincenty formulas are the **two major formulas used for calculating distances on a sphere and elliptic shape.**

Classification of objects

Process – classify and code

Section I - *Accommodation and food service activities*

NACE

55.1 - Hotels and similar accommodation

55.2 - Holiday and other short-stay accommodation

55.3 - Camping grounds, recreational vehicle parks and trailer parks

55.9 - Other accommodation

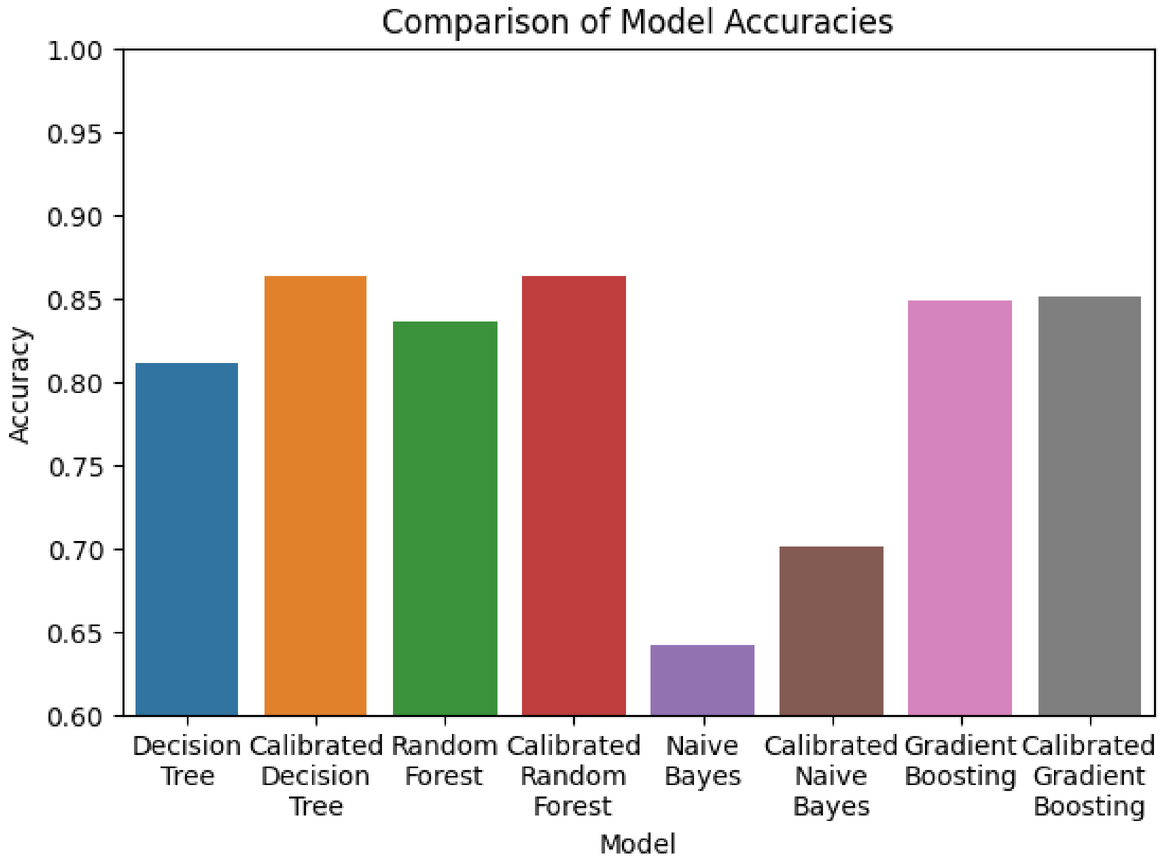
disabled guests
Airport shuttle
family rooms
non smoking
free parking
smoking rooms
free WiFi
Fitness center
WiFi family
coffee making
Private parking
shuttle free
rooms free
rooms Bar
rooms restaurant
service Free
good breakfast
guests Restaurant
Tea coffee
parking non
breakfast free
Room service
making facilities
rooms Facilities
Bar Fantastic
family rooms
parking free
Fantastic breakfast
rooms Airport
categorised

Code	Cleaning services	Parking	…	Restaurant	Breakfast
55.1	1	1		2	2
55.2	2	2		2	0
55.3	0	0		0	0
55.9	2	0		0	0

0 – don’t have; 1 – must have; 2 – might have;

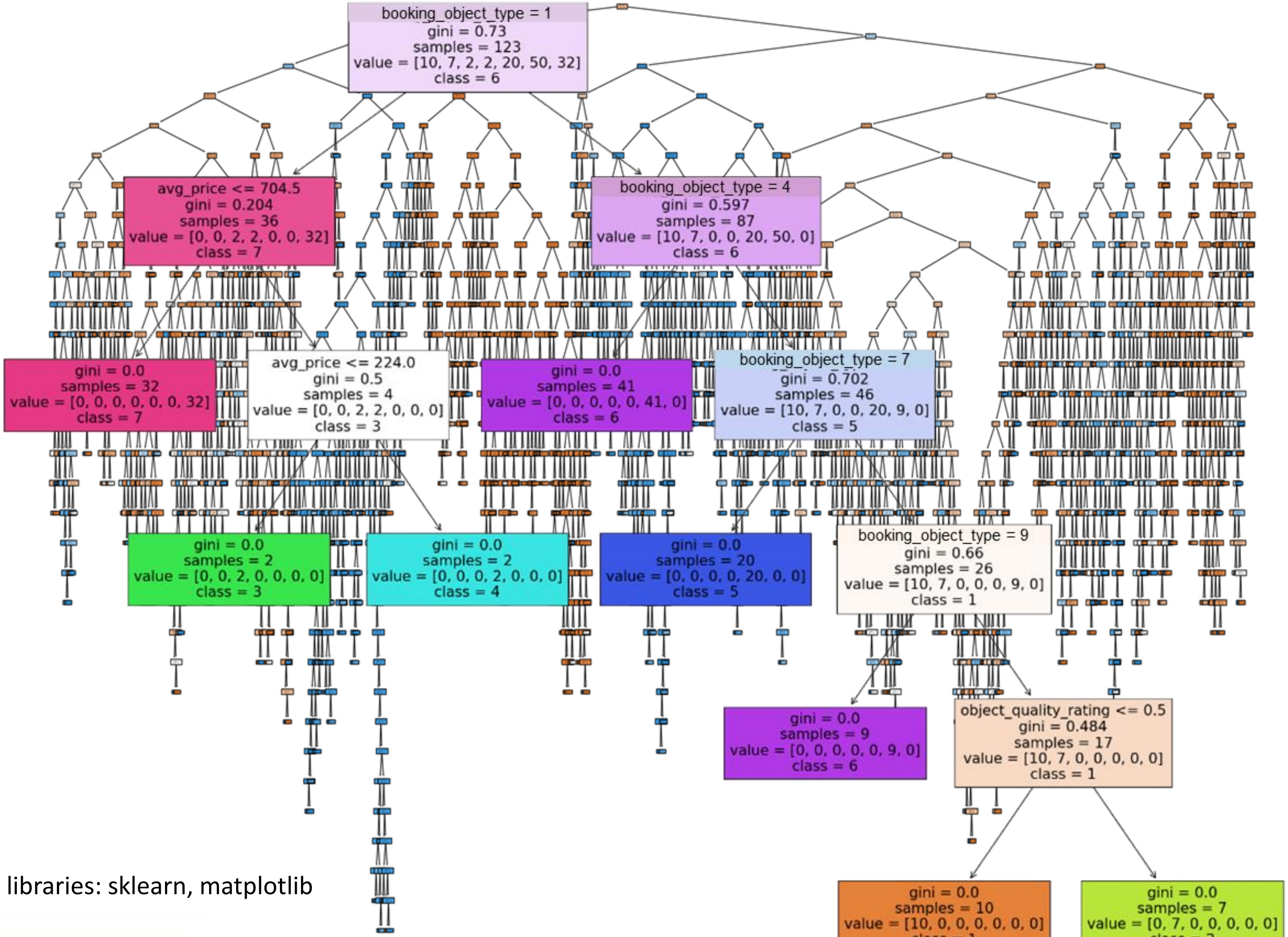
Classification of objects

Comparison of methods



Classification of objects

Decision Tree

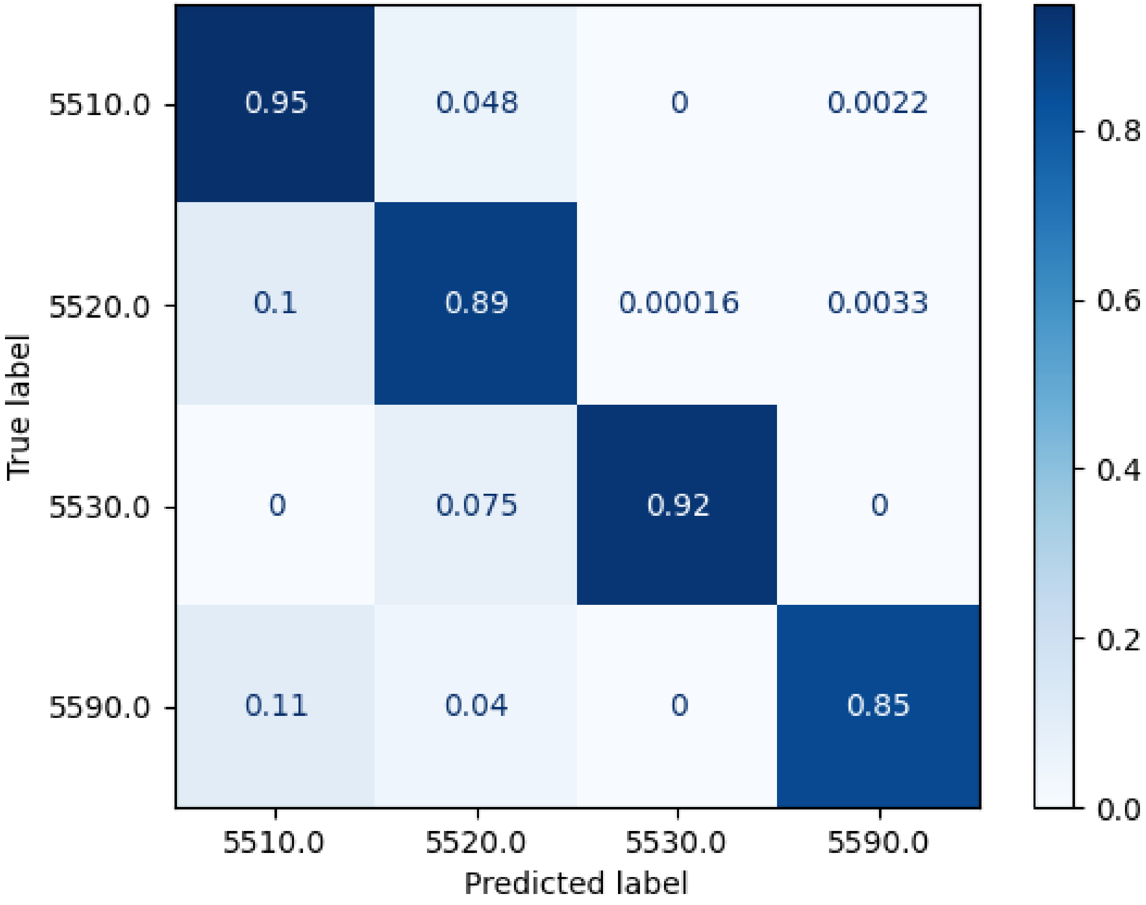


Python libraries: sklearn, matplotlib

Statistics Poland

Classification of objects

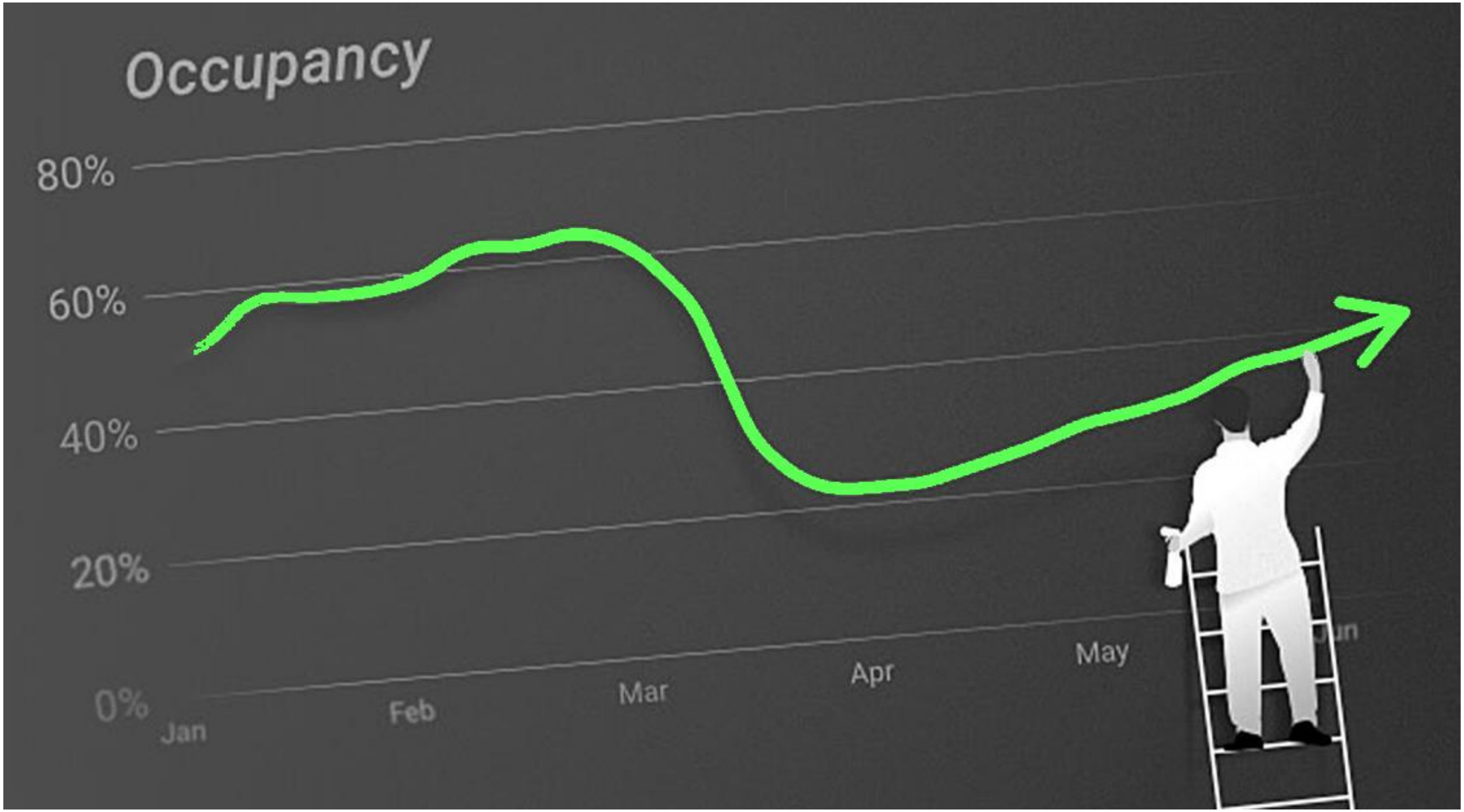
Quality – Confusion Matrix



True label \ Predicted label	5510.0	5520.0	5530.0	5590.0
5510.0	0.95	0.048	0	0.0022
5520.0	0.1	0.89	0.00016	0.0033
5530.0	0	0.075	0.92	0
5590.0	0.11	0.04	0	0.85

Statistics Poland

Derive variables



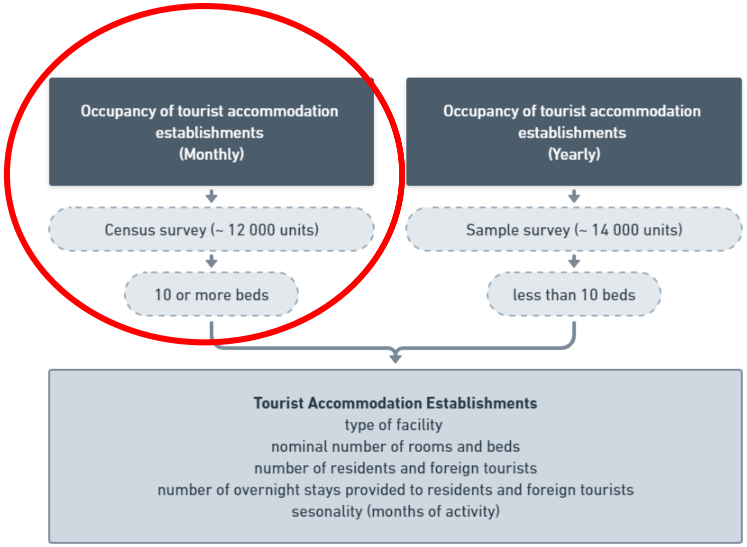
Derive variables

Nace 55.1

AFRE – automatic report intended for accommodation establishments that submit reports as an XML file.

Feedback report contain:

- information on the number of tourists accommodated (arrivals),
- number of nights spent,
- occupancy rate of bed places in each month of the year.



Occupancy of tourist accommodation establishments (Monthly)

Census survey (~ 12 000 units)

10 or more beds

Occupancy of tourist accommodation establishments (Yearly)

Sample survey (~ 14 000 units)

less than 10 beds

Tourist Accommodation Establishments
type of facility
nominal number of rooms and beds
number of residents and foreign tourists
number of overnight stays provided to residents and foreign tourists
sesonality (months of activity)

Derive variables

NACE 55.2

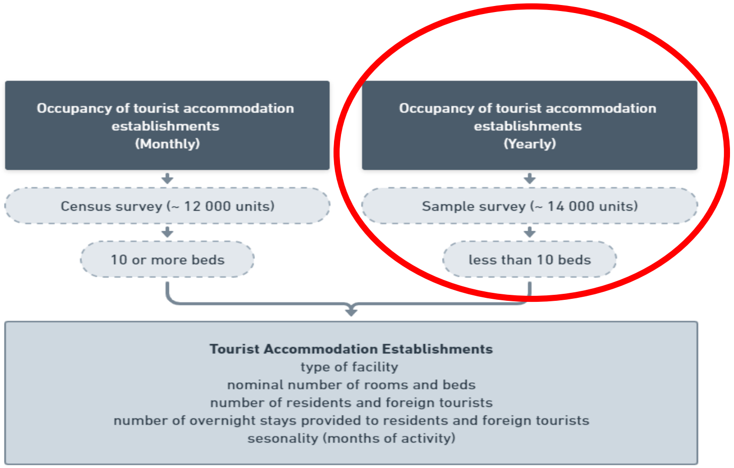
type of facility *classification* **+**

nominal number of rooms and beds *web scraping* **+**

number of residents and foreign tourists **+**

number of overnight stays provided to residents and foreign tourists **-**

sesonality (months of activity) *web scraping* **+**



Occupancy of tourist accommodation establishments (Monthly)

Census survey (~ 12 000 units)

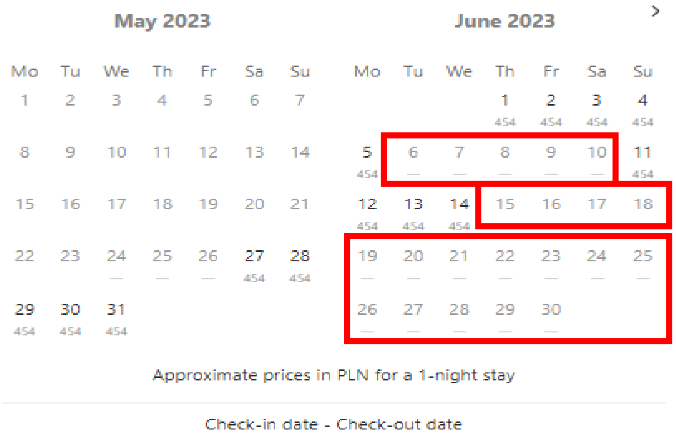
10 or more beds

Occupancy of tourist accommodation establishments (Yearly)

Sample survey (~ 14 000 units)

less than 10 beds

Tourist Accommodation Establishments
type of facility
nominal number of rooms and beds
number of residents and foreign tourists
number of overnight stays provided to residents and foreign tourists
sesonality (months of activity)



May 2023

Mo	Tu	We	Th	Fr	Sa	Su
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 —	27 454	28 454
29 454	30 454	31 454				

June 2023 >

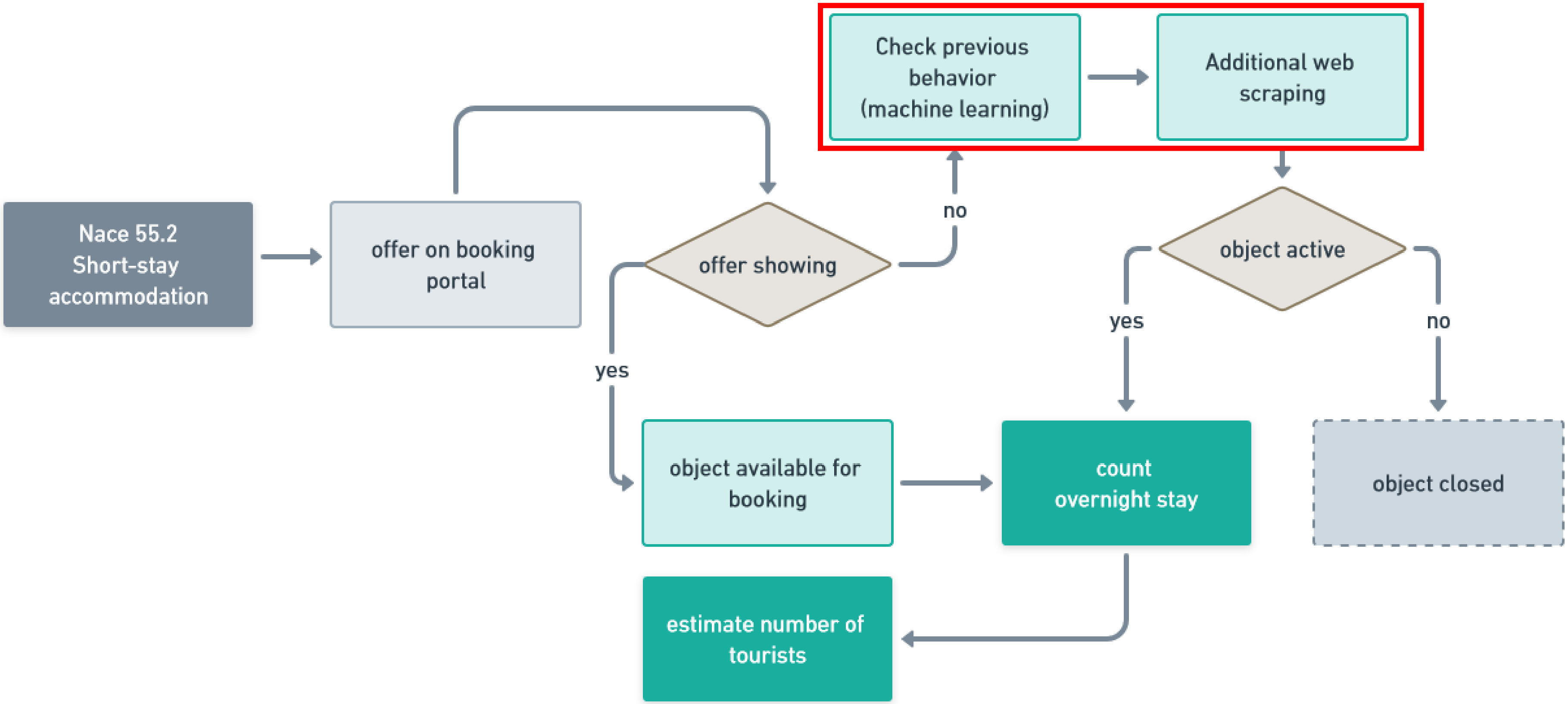
Mo	Tu	We	Th	Fr	Sa	Su
			1 454	2 454	3 454	4 454
5 454	6 —	7 —	8 —	9 —	10 —	11 454
12 454	13 454	14 454	15	16	17	18
19 —	20 —	21 —	22 —	23 —	24 —	25 —
26 —	27 —	28 —	29 —	30 —		

Approximate prices in PLN for a 1-night stay

Check-in date - Check-out date

Derive variables

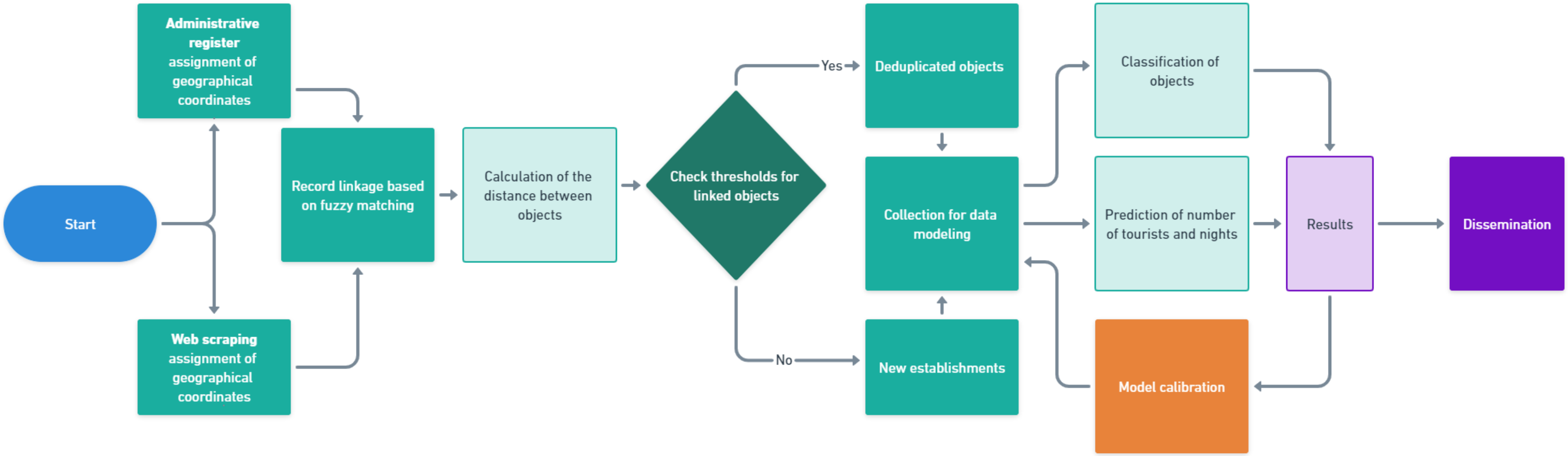
NACE 55.2



Check previous behavior (machine learning)
Additional web scraping
Nace 55.2 Short-stay accommodation
offer on booking portal
offer showing
no
object active
yes
no
yes
object available for booking
count overnight stay
object closed
estimate number of tourists

Replacement of the accommodation survey

Concept - Summary


Start
Administrative register assignment of geographical coordinates
Web scraping assignment of geographical coordinates
Record linkage based on fuzzy matching
Calculation of the distance between objects
Check thresholds for linked objects
Yes
No
Deduplicated objects
New establishments
Collection for data modeling
Classification of objects
Prediction of number of tourists and nights
Model calibration
Results
Dissemination

Statistics Poland

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 models (quality aspects);
  - ✓ Micro-data access management model;
    - ○ Societies preferring privacy over technological development (e.g., Europe),
    - ○ Societies prioritizing technological development over privacy (e.g., China, Korea).
  - ✓ Artificial intelligence management model.

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.

Statistics Poland

Thank you for your attention

Marek Cierpiał-Wolan, Assoc. Prof.

University of Rzeszów
Statistical Office in Rzeszów