

Measuring representativeness of different data sources connected with short-term statistics

Presented at
the fourth European Establishment Statistics Workshop
in Poznań, Poland

Authors

- Alina Szkop [a.szkop@stat.gov.pl]

Specialist in Methodology and Programming Division

- Mateusz Smektalski [m.smektalski@stat.gov.pl]

IT Specialist in Methodology and Programming Division



Center for Short-term Statistics
Statistical Office Poznan

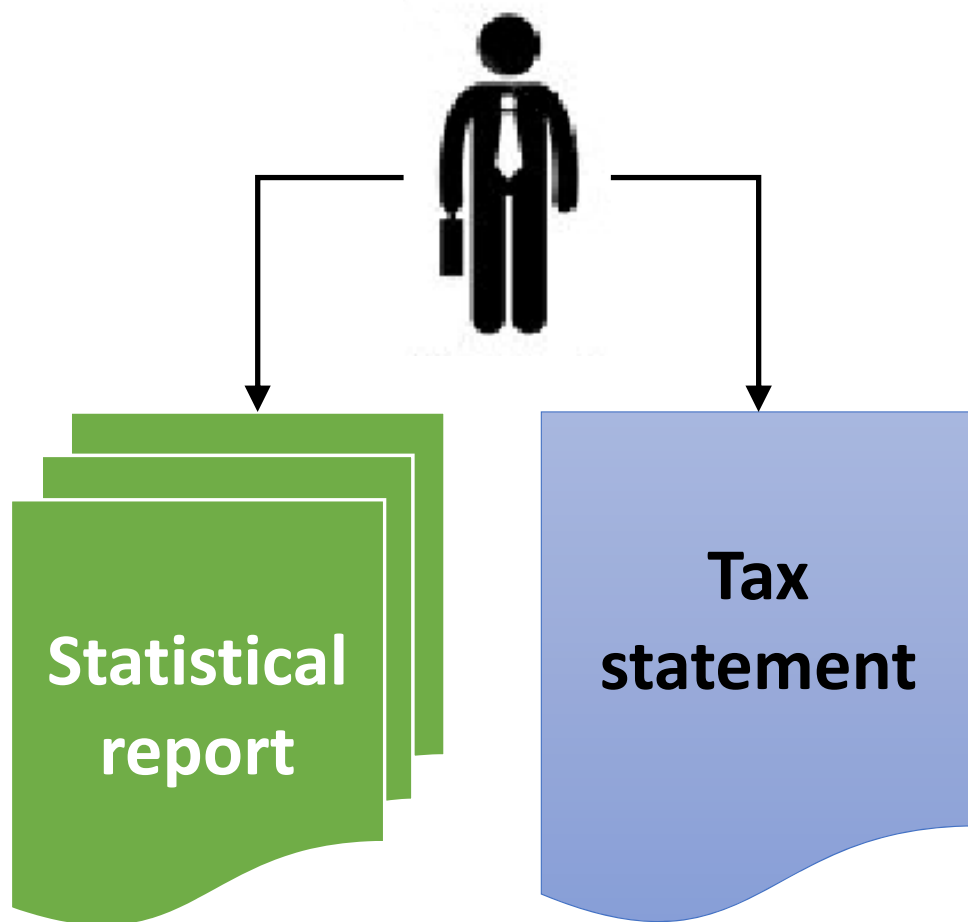
The aim

The present study is aimed at comparing data collected by the Central Statistical Office (CSO) with data collected independently by other units of public administration. The study compares information reported in the monthly DG-1 business survey with data from tax statements submitted to the Ministry of Finance (MF).

Agenda

1. Research methodology
2. Database description
3. Database integration
4. Calculations
 - I. R-indicator
 - II. Coorelation
 - III. Difference clasification
5. Conclusions
6. Further steps

Research methodology



The external perspective in data quality assessment disregards the survey process and any potential errors and focuses on the analysis of data validity and reliability and seeks to determine their “truthfulness” in the process of verification and data validity.

The direct perspective analysis of two independent surveys, in which the same respondents are asked again to answer questions from the basic questionnaire (DG-1) in a control survey (FM)

Database description

Financial Database (FD)

- contains data from tax statements
- revenue, salaries, insurance contributions...
- obtain every year (annual)

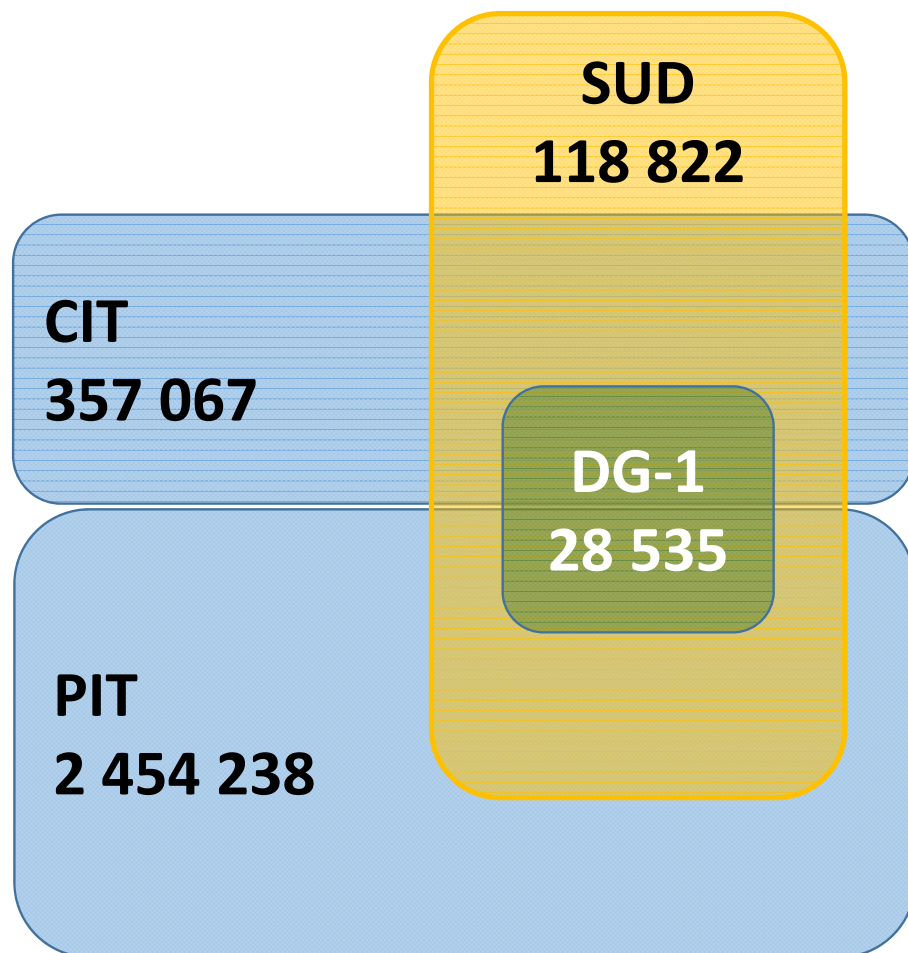
Statistical Unit Database (SUD)

elementary registry of units covered by statistical research

Economic activity report (DG-1)

- contains data from business activity report
- revenue, salaries, prices, sales, working time, transport...
- obtain every month

Database integration



**Financial
Database**

**Statistical Unit
Database**

Statistical survey

[Collection merge details](#)

Measuring representativeness has been devied into 3 steps

I. R-INDICATOR

II. COORELATION

III. DIFFERENCES CLASIFICATION

I. R-INDICATOR

R-indicator

The R Indicator measures the extent to which the response deviates from the representative response:

$$\hat{R} = 1 - 2S_{\hat{\rho}} = 1 - 2 \sqrt{\frac{1}{\sum_{i=1}^n w_i - 1} \sum_{i=1}^n w_i (\hat{\rho}_i - \bar{\bar{\rho}})^2}$$

,where:

w_i – sample design weight for unit i ,

$\hat{\rho}_i$ – response propensities (estimated using a logistic model)

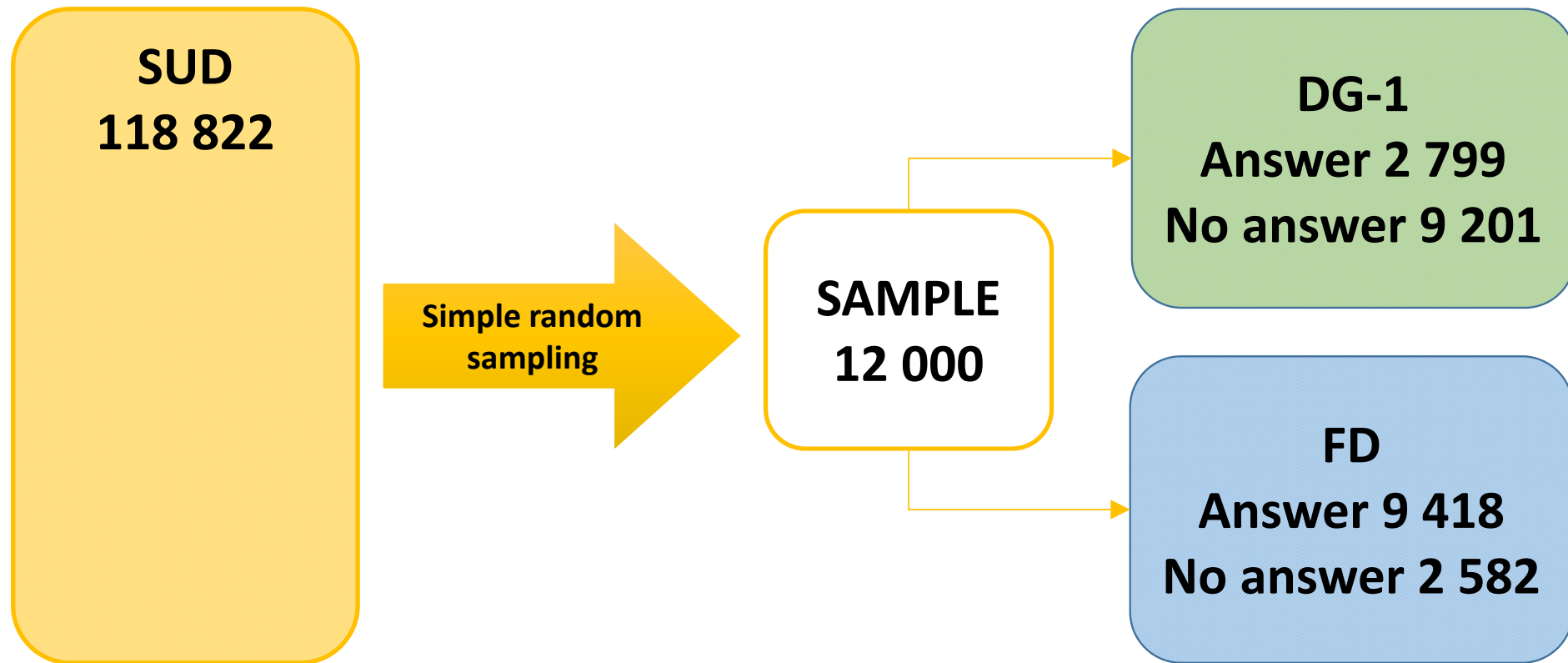
n – number of sample units,

$\bar{\bar{\rho}}$ – mean of estimated values.

R-indicator

1. Random sampling
2. Calculation of the propensity ($\hat{\rho}_i$), we use logit model where
 - dependent variable is response particular unit
 - independent variables are
 - NACE section
 - class size of the unit
 - province
3. Calculation of the R-indicator

R-indicator



R-indicator

Financial Database

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	4753.059	4473.400
SC	4760.253	4710.802
-2 Log L	4751.059	4407.400

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	343.6594	32	<.0001
Score	590.0405	32	<.0001
Wald	339.5388	32	<.0001

Propensity mean: 0.935
R-indicator: 88.3%

DG-1

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	11751.753	7061.198
SC	11758.947	7298.601
-2 Log L	11749.753	6995.198

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4754.5547	32	<.0001
Score	4713.1808	32	<.0001
Wald	1964.1634	32	<.0001

Propensity mean: 0.285
R-indicator: 36.5%

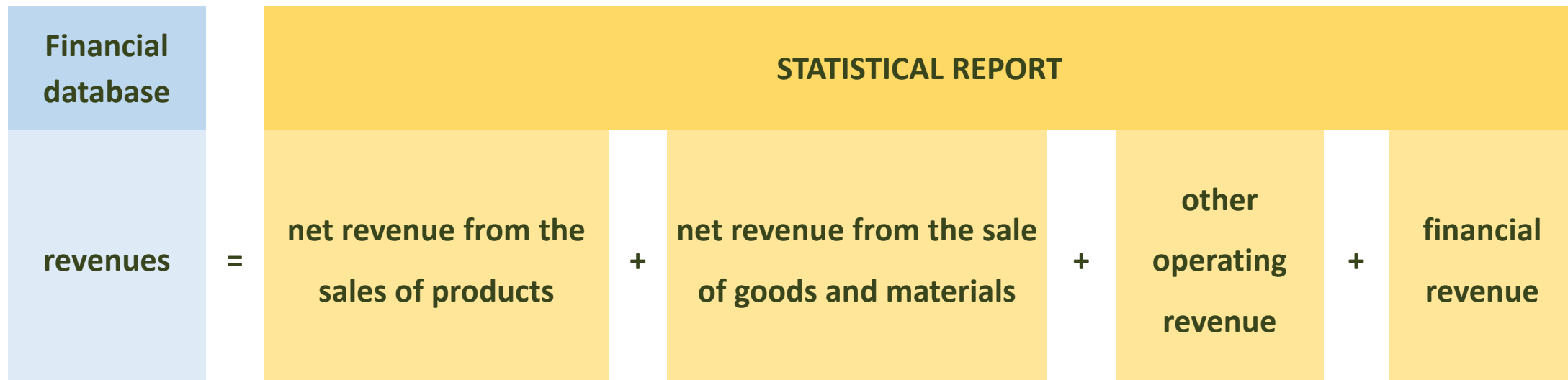
II. CORRELATION

Correlation

There are differences in the definitions of variables.

Statistical report does not contain variable „revenues”.

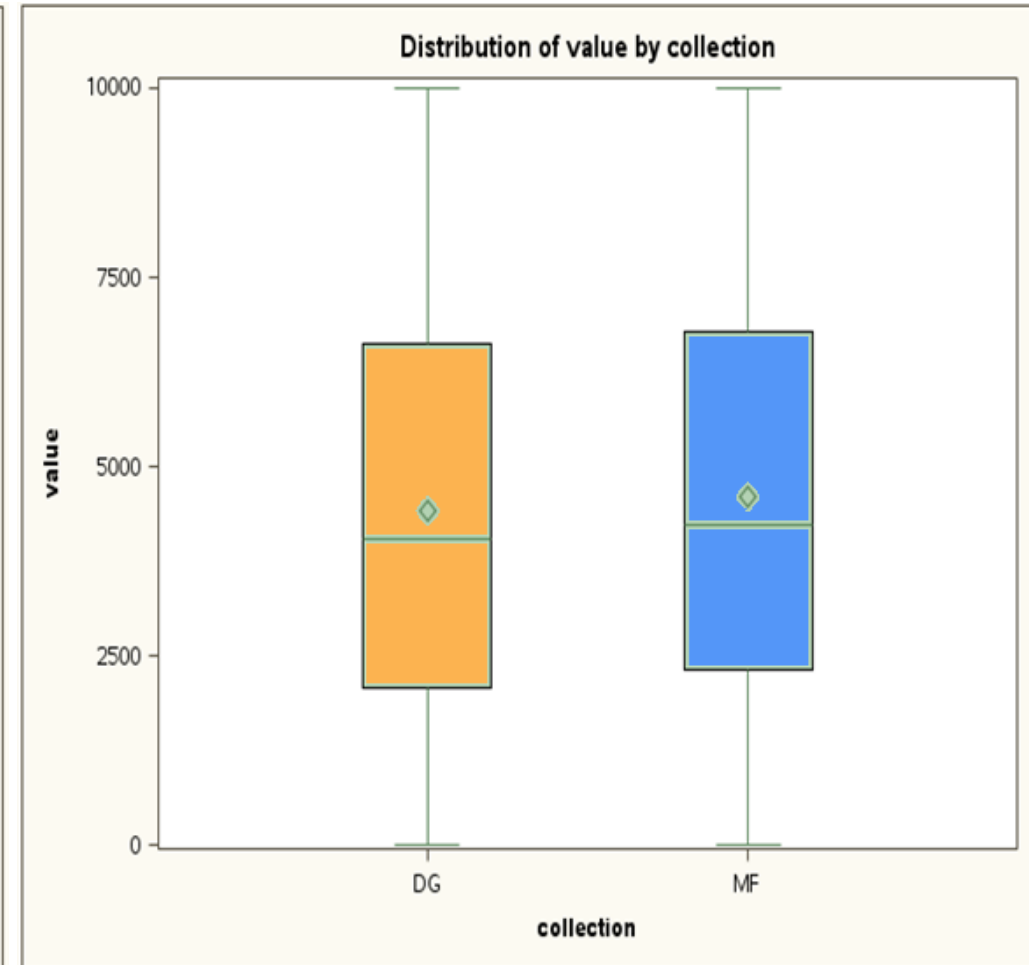
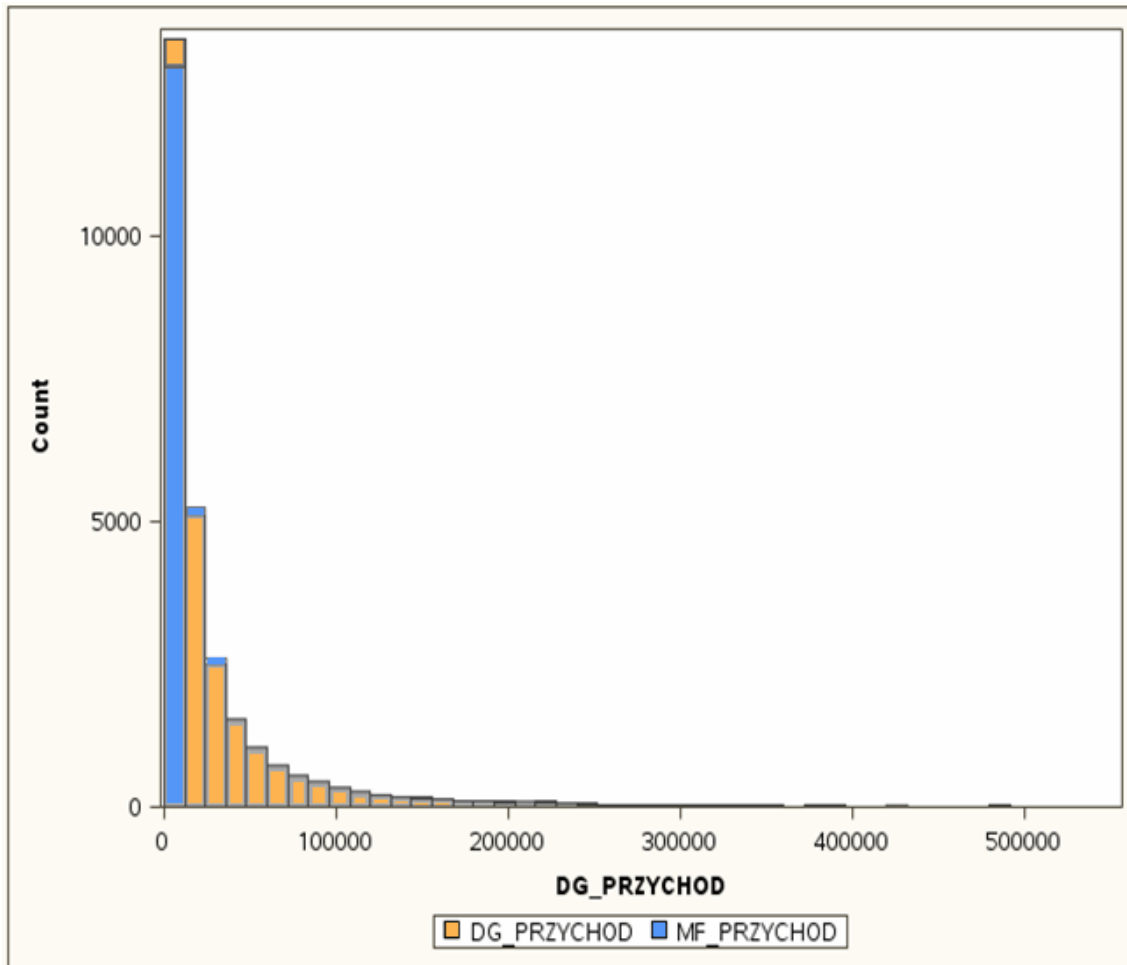
Statistical report has separate kinds of revenues (which we can sum)



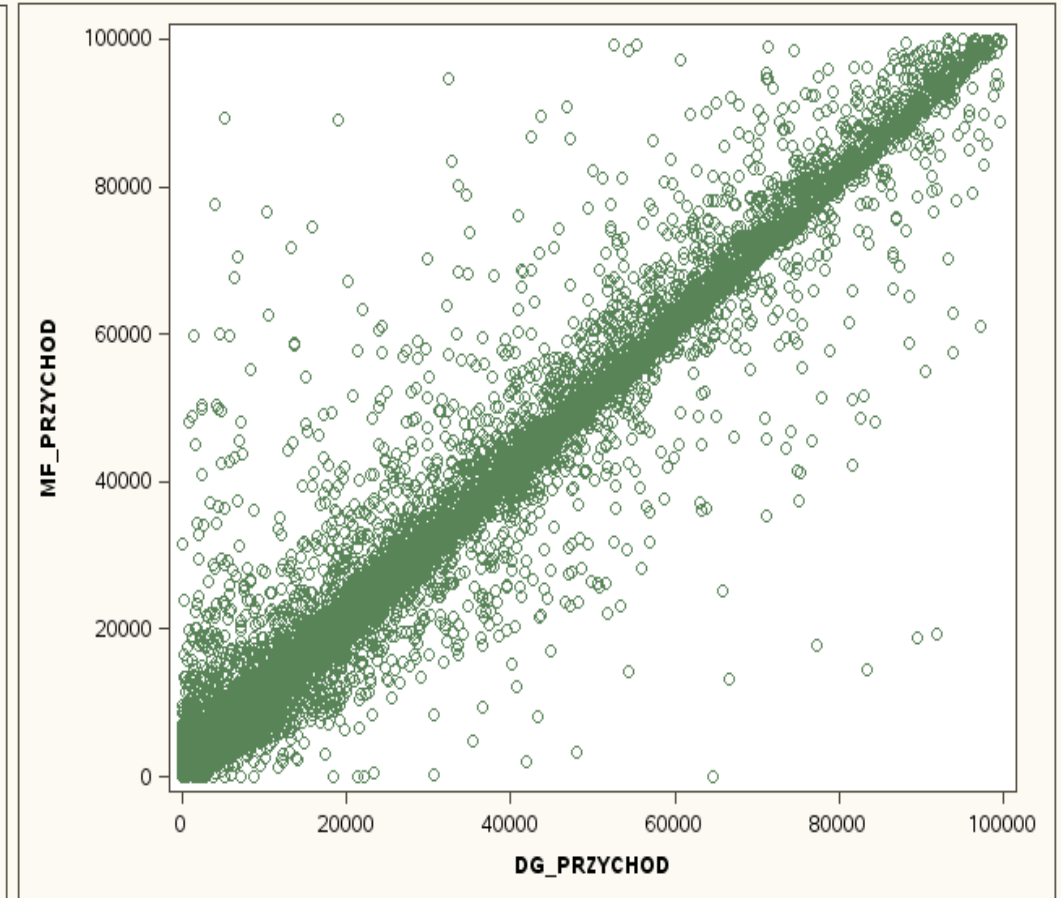
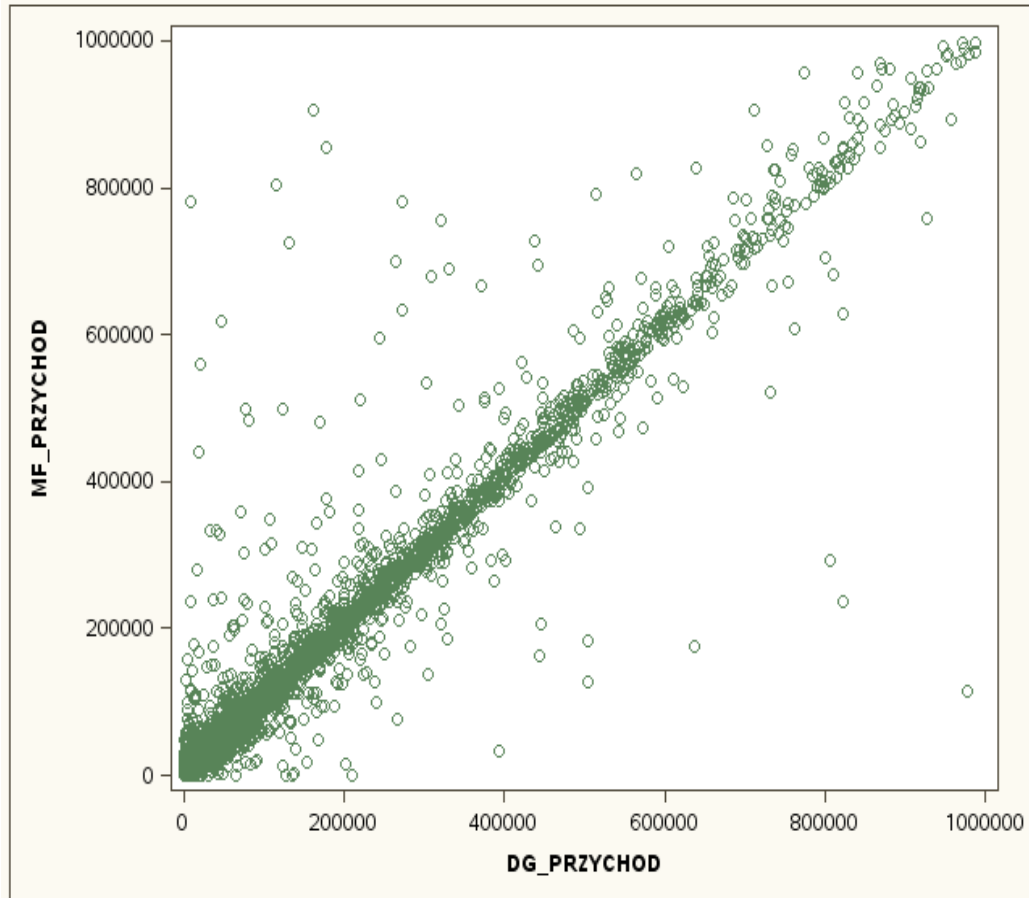
Correlation

Variable	DG_REVENUE	MF_REVENUE	diff_MF_DG
N	28 535	28 535	28535
MEAN	78 823.93	83 262.48	4 438.55
MEDIAN	13 333.90	14 061.60	313.40
MODE	651.00	1 234.70	-
STD EROR MEAN	4 842.47	4 761.08	717.27
SKEWNESS	90.34	76.43	13.24
KURTOSIS	11 278.59	8 503.56	3 348.36

Correlation



Correlation



Correlation

Insights:

- DG_REVENUE and MF_REVENUE are treated as independent variables
- Both variables are not normally distributed
- There are differences in the counts of observations depending on the intervals of the variables
- A scatter plot shows a positive, monotonic (linear) correlation

Two-sample nonparametric methods for independent groups

Correlation

BY SECTION	NUMBER	R	rho	tau	p
A	1	-	-	-	-
B	193	0,98827	0,98494	0,93963	0,0001
C	10606	0,99551	0,98758	0,94343	0,0001
D	346	0,99356	0,99033	0,95743	0,0001
E	825	0,97521	0,98183	0,93402	0,0001
F	2984	0,97179	0,97745	0,92367	0,0001
G	6958	0,98187	0,98924	0,95384	0,0001
H	1392	0,99037	0,97943	0,92868	0,0001
I	744	0,9816	0,94187	0,80956	0,0001
J	698	0,98771	0,9669	0,92122	0,0001
L	862	0,94706	0,96829	0,90493	0,0001
M	1019	0,54728	0,97474	0,9152	0,0001
N	1194	0,73434	0,96551	0,88741	0,0001
R	525	0,99645	0,69711	0,52378	0,0001
S	188	0,98498	0,92257	0,87484	0,0001
TOTAL	28535	0,98899	0,98036	0,9276	0,0001

Null hypothesis:

There is no correlation between datasets

Alternative hypothesis:

There is correlation between datasets

p-value < 5%

The data are sufficiently strong to reject the null hypothesis.

II. DIFFERENCES CLASSIFICATION

Differences classification

The table presents the percentage differences between observations from the FM dataset (CIT/PIT) and the DG-1 dataset were grouped into intervals.

$$\text{difference} = |\text{FM_REVENUE} - \text{DG_REVENUE}|$$

BY SECTION	equal to	0-1%	1-5%	5-10%	10-25%	25-50%	50-100%	1-10x	10-100x	>100x
TOTAL	121	6262	11031	4301	3608	1629	1401	158	13	11
PERCENTAGE	0.42%	21.94%	38.66%	15.07%	12.64%	5.71%	4.91%	0.55%	0.05%	0.04%
CUMULATIVE	121	6383	17414	21715	25323	26952	28353	28511	28524	28535
CUMULATIVE PERCENTAGE	0.42%	22.37%	61.03%	76.10%	88.74%	94.45%	99.36%	99.92%	99.96%	100.00%

Conclusion

- The study of representativeness found a large disproportion between the datasets. The R-indicator for the MF dataset was equal to 0.883, while for the DG-1 dataset only 0.365.
- The comparative analysis revealed a strong correlation of over 90% between the two datasets.
- Over 61% of answers in the DG-1 survey differed from information indicated in tax statements by less than 5%.

Further steps

Consider to

- Use transformation, such as the Box-Cox power transformation to make data normally distributed.
- Compare structures between these two datasets eg. two sample Kolmogorov- Smirnov test, Mann-Whitney U test
- Find out are the differences between revenues stable in time

Koniec

Thank You for your attention