ifcs2022.fep.up.pt

IFCS 2022 Classification and Data Science in the Digital Age

A Review on Official Survey Item Classification for Mixed-Mode Effects Adjustment

Afshin Ashofteh and Pedro Campos

Statistics Portugal, Porto University, Nova University of Lisbon (Nova IMS)

19 – 23 July 2022
Porto, Portugal

What do we propose in this research?

- A systematic review that maps keyword identification search, databases, and bibliometric analysis.
- A meta-analysis of two databases (Scopus and WOS) to identify the PRISMA flow diagram and to characterize the articles, author coauthorship analysis, as well as the Keywords occurrence over the years.

Outline

This presentation is structured as follows

- Motivation
- Methodology
- Content analysis
- Results
- Main conclusions
- The main research gaps
- Future works

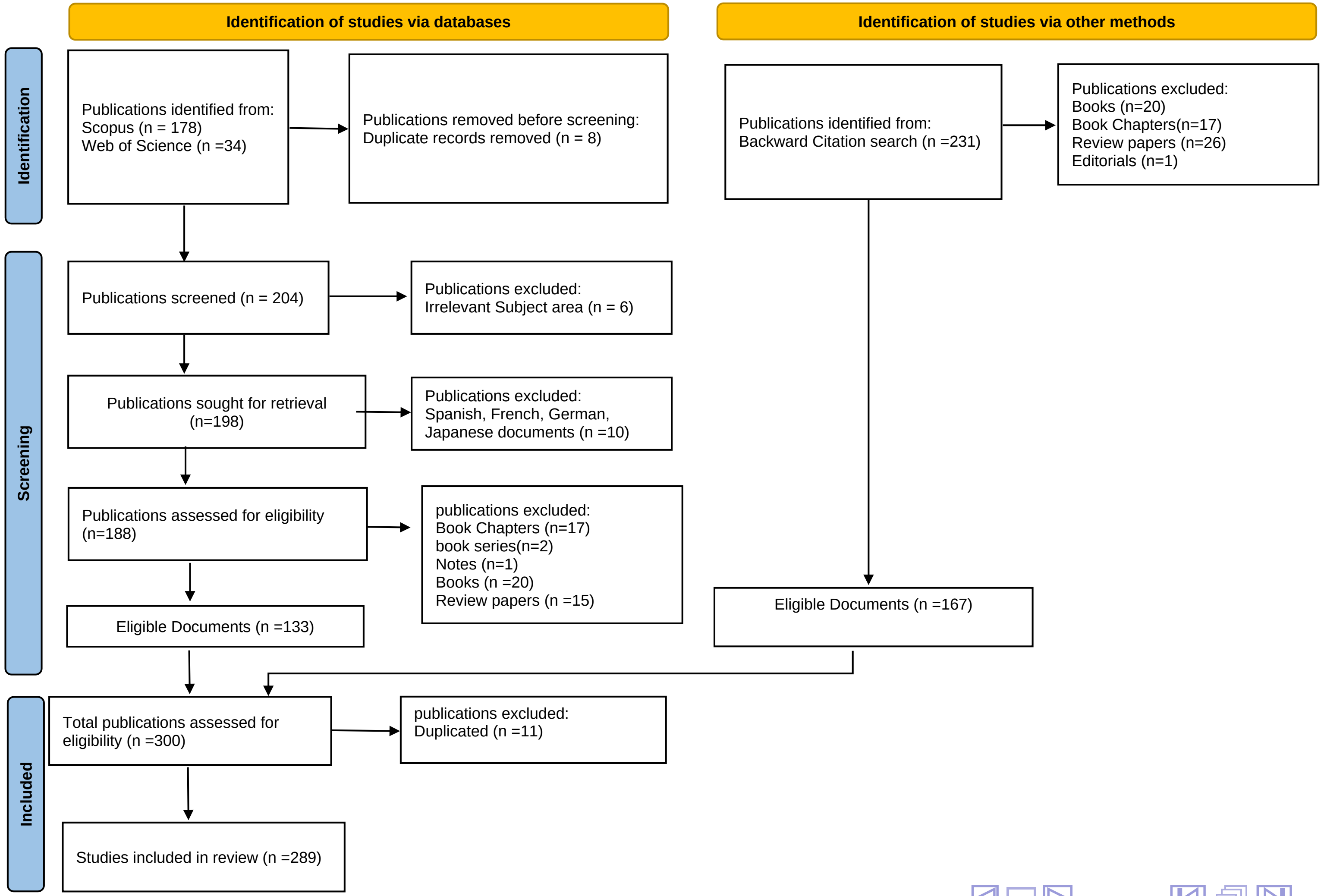
Motivation

- To response to the challenge of the development, production, and dissemination of official statistics at the time of the COVID-19 pandemic.
- To investigate the methodological and practical choices of National Statistics Institutes (NSIs) for survey collection without requiring direct contact with interviewing staff (i.e., Remote survey data collection)
- To use COVID-19’s lock down as an opportunity to study the mixed mode effects, nonresponse and coverage errors of different survey modes.

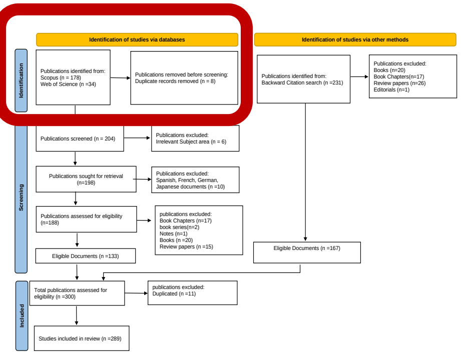
Methodology

PRISMA

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses method.

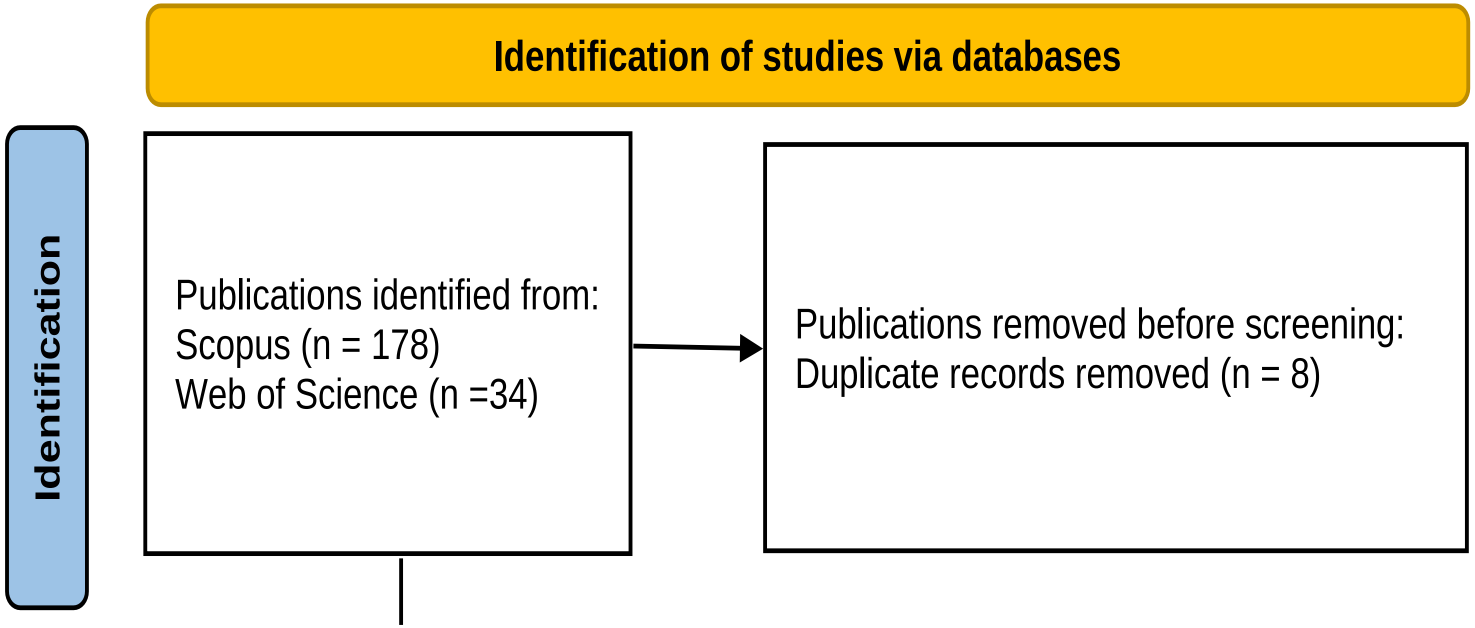


Methodology



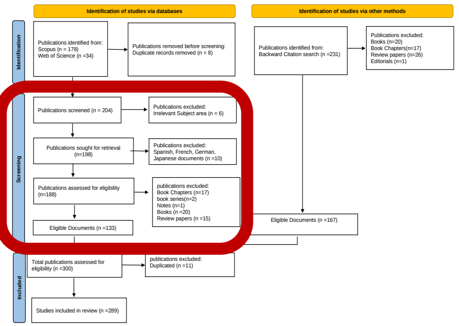
PRISMA

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses method.



Search keywords

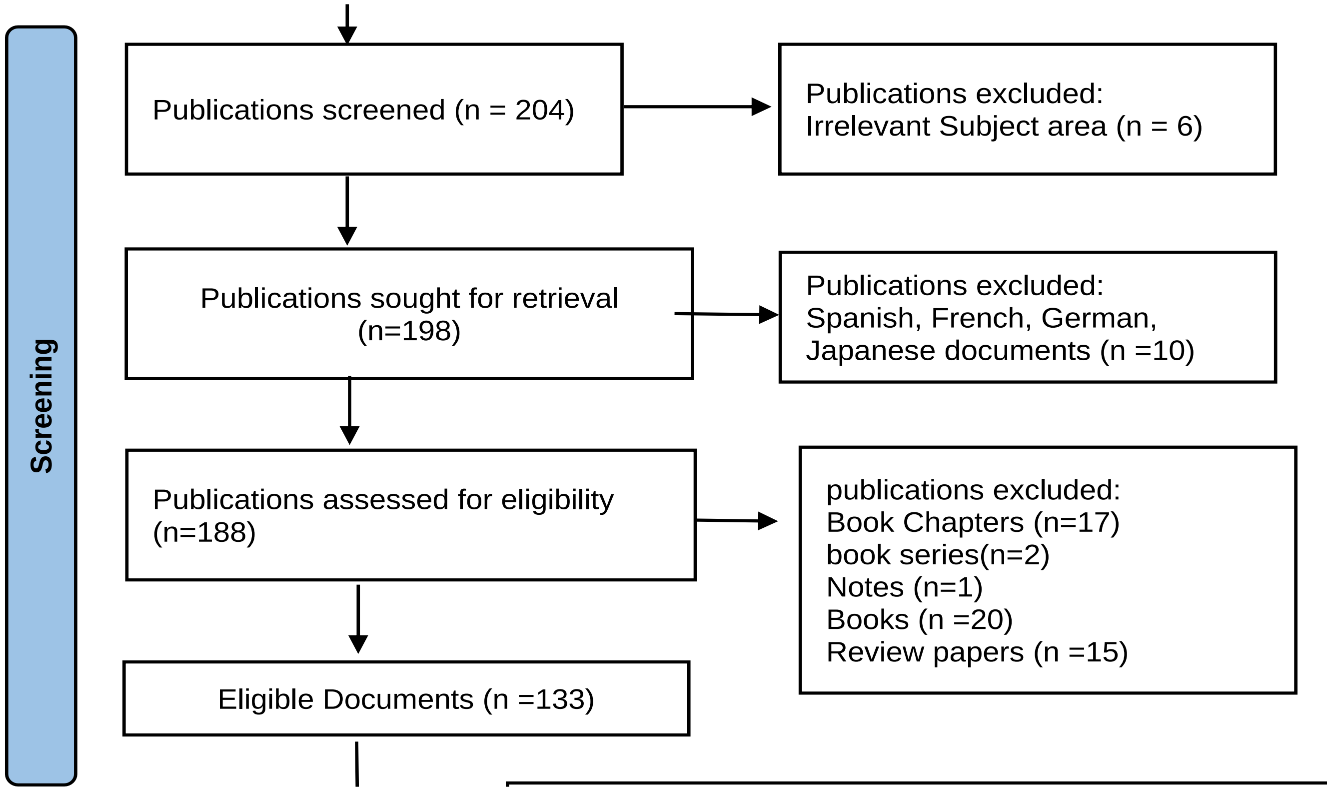
((***mixed-mode**** OR "***Mode effect****") AND (***weighting*** OR ***weight**** OR ***classification***) AND ("***Measurement error****" OR "***Non-response bias***" OR "***Data quality***" OR "***response rate****") AND (***capi*** OR "***Computer Assisted Personal Interview****" OR ***cawi*** OR "***AssistedWeb Interview****" OR ***cati*** OR "***Computer Assisted Telephone Interview****“ OR "***web survey****" OR "***mail survey****" OR "***telephone survey****"))

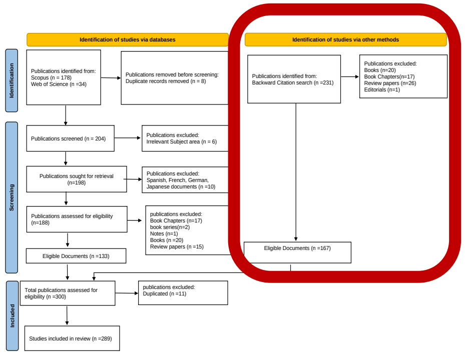


Methodology

PRISMA

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses method.

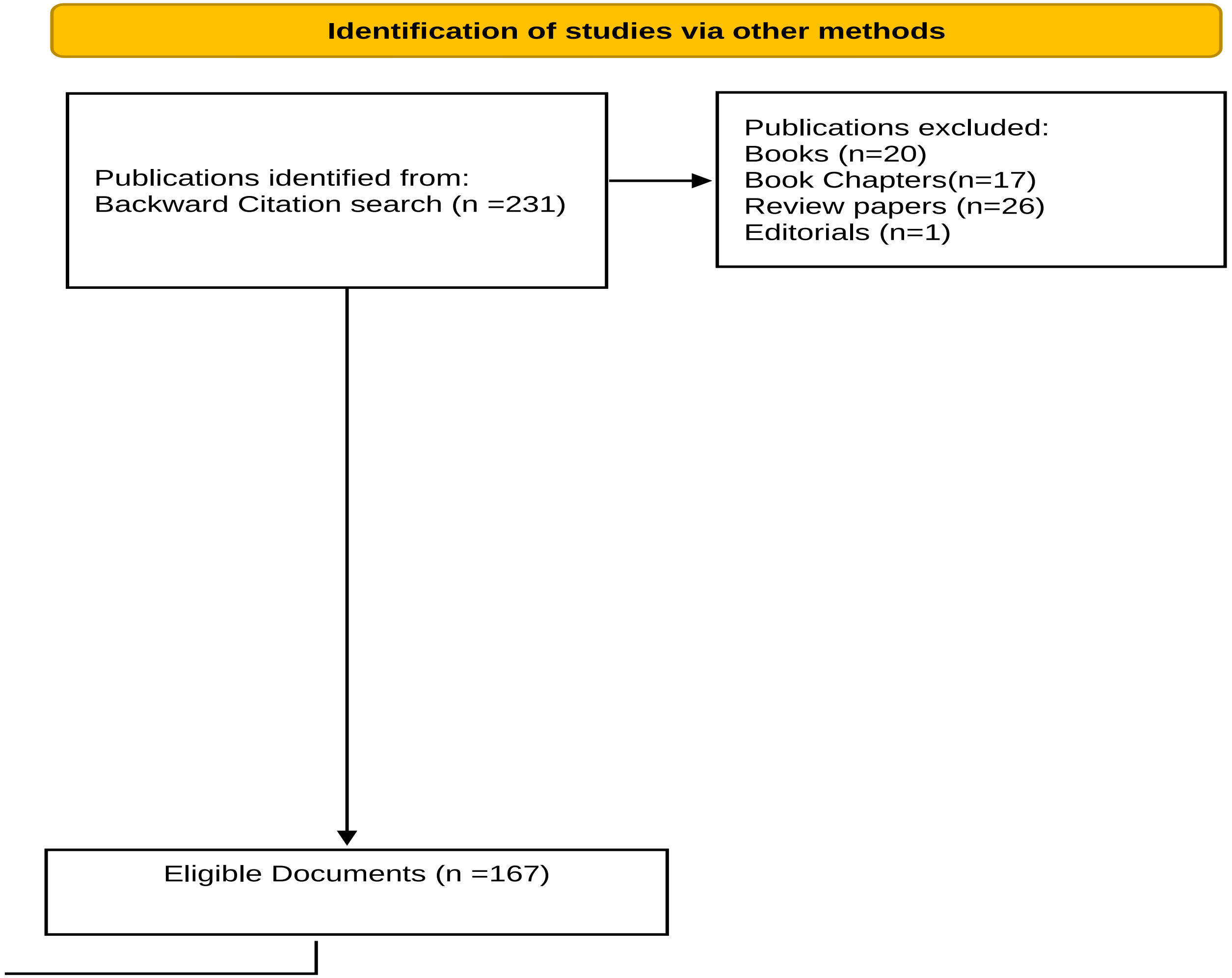


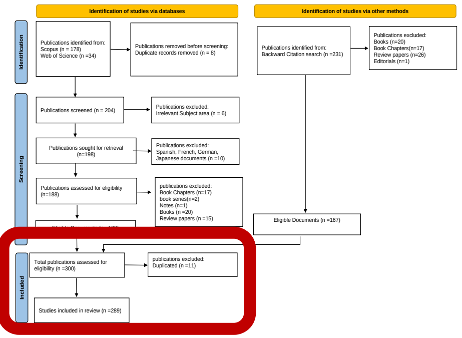


Methodology

PRISMA

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses method.

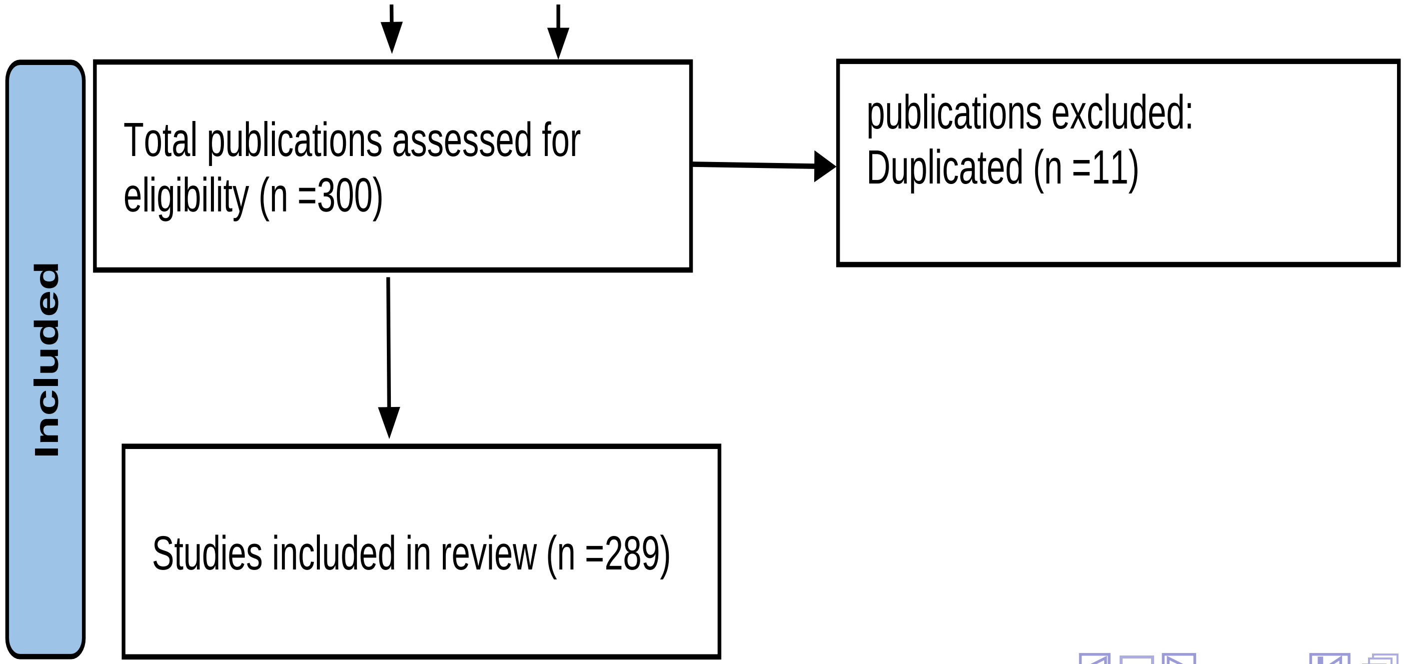




PRISMA

Methodology

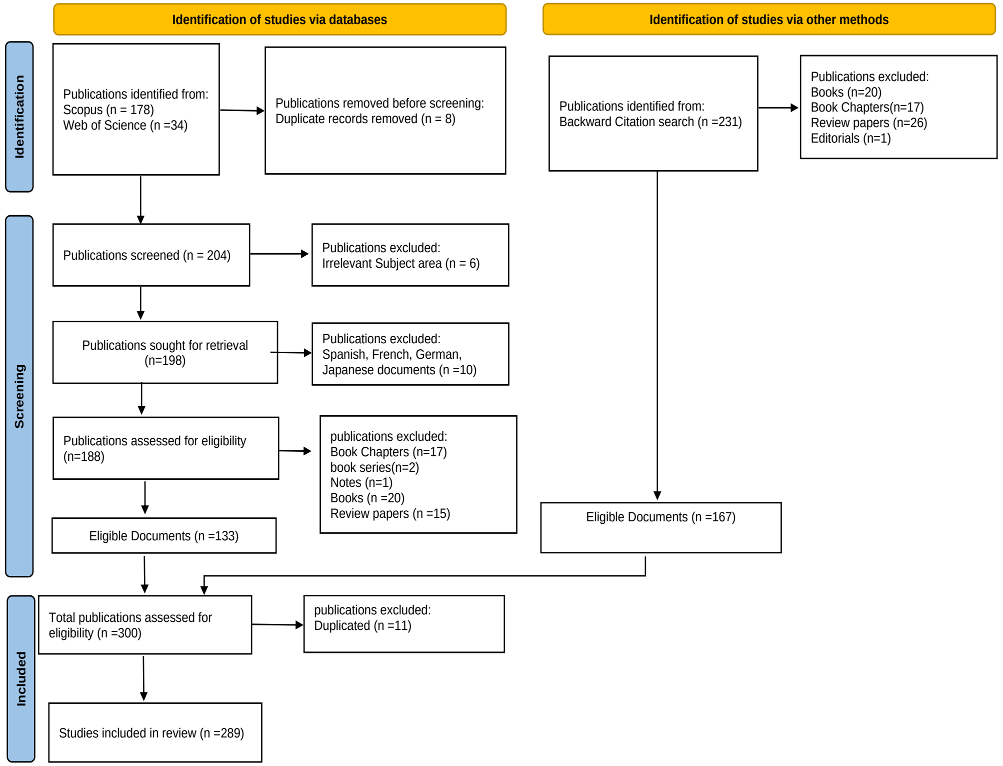
The Preferred Reporting Items for Systematic Reviews and Meta-Analyses method.



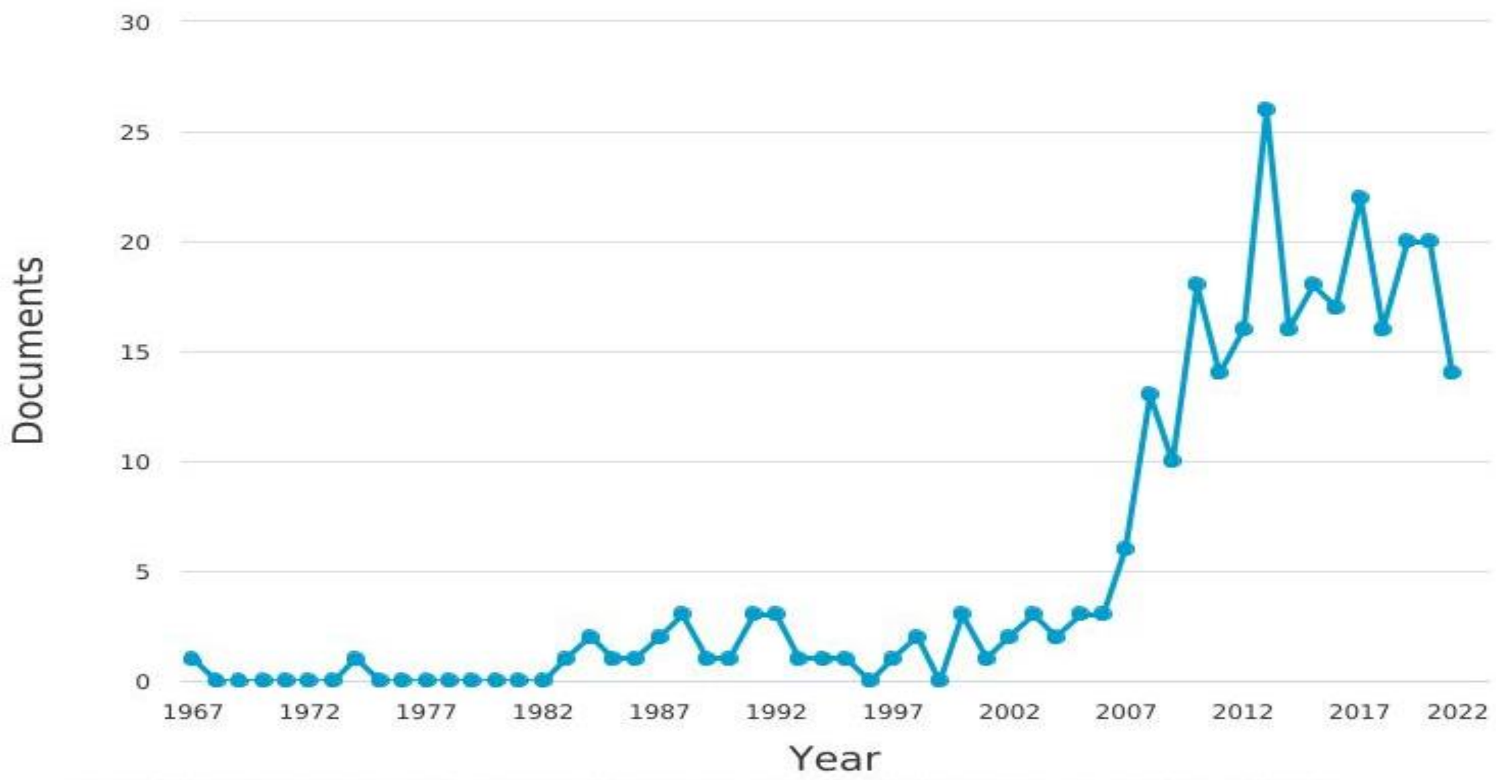
Methodology

Exclusion:

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses method.



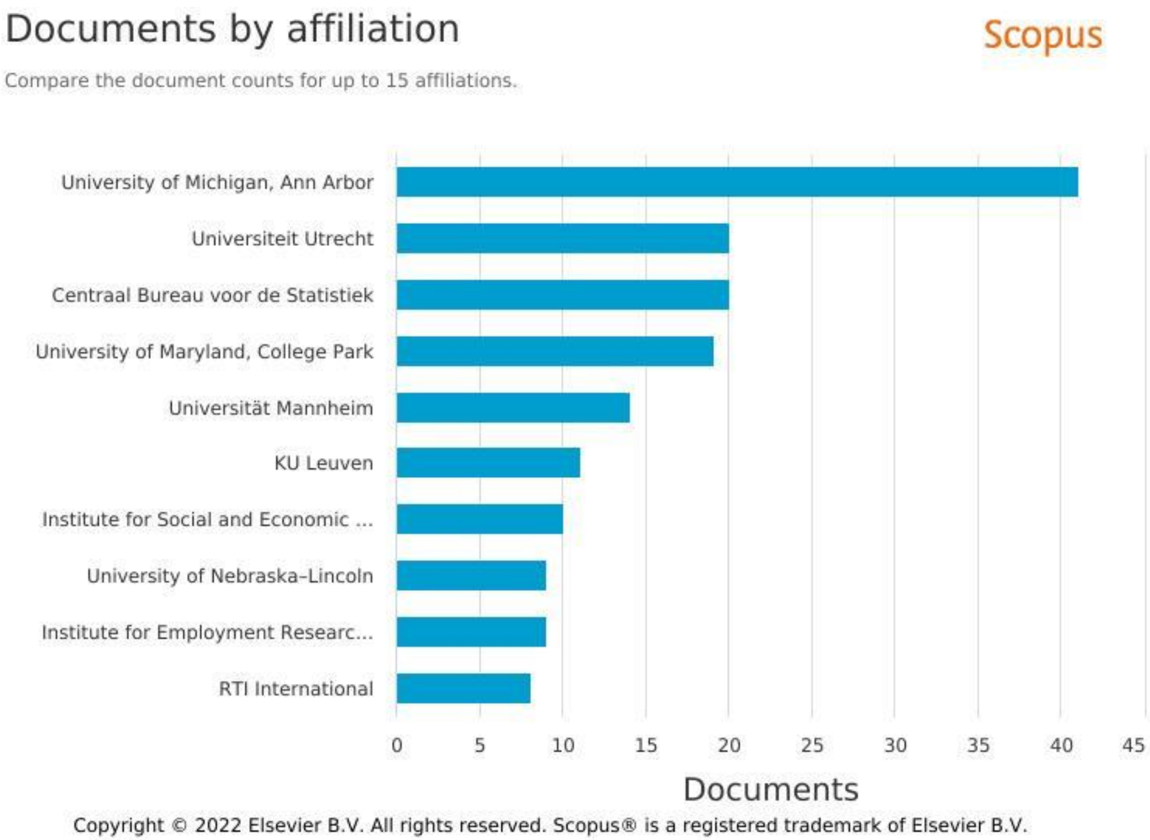
Content analysis

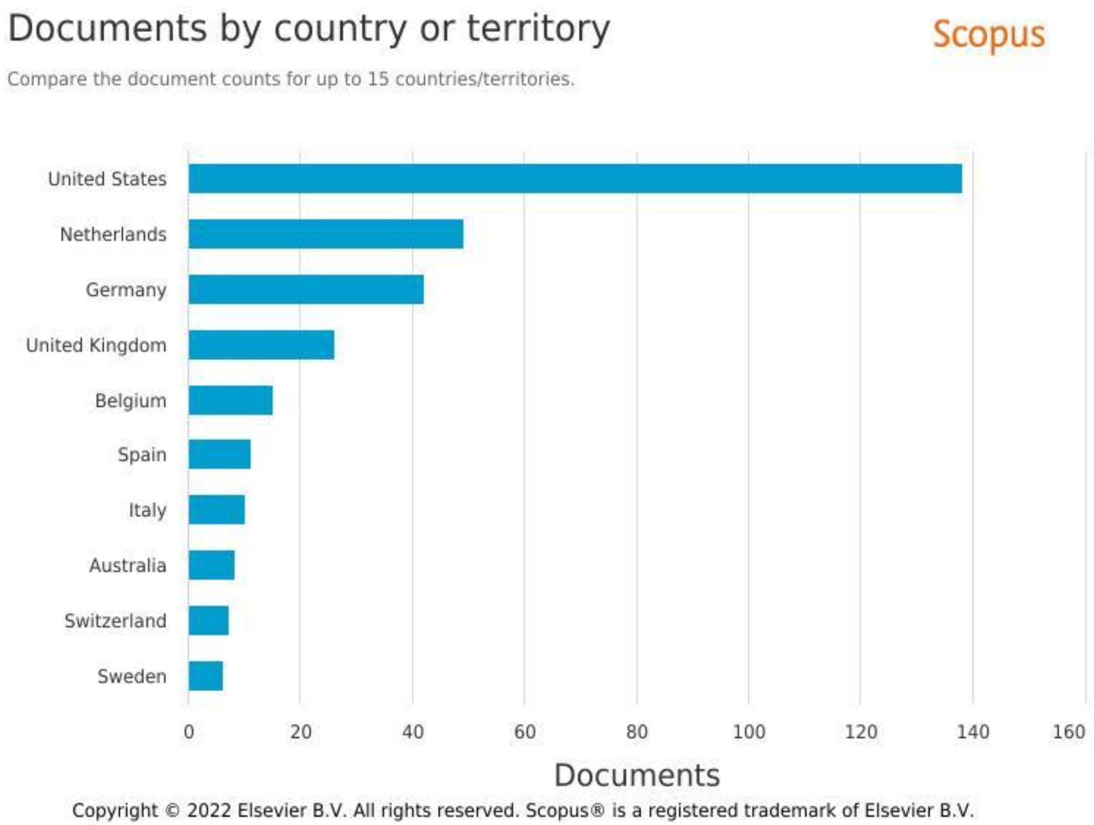


Content analysis

Journals	# Doc
Journal Of Official Statistics	36
Public Opinion Quarterly	30
Survey Research Methods	23
Journal Of Survey Statistics And Methodology	19
Social Science Computer Review	16
Sociological Methods And Research	10
Journal Of The Royal Statistical Society Series A Statistics In Society	9
Survey Methodology	8
Field Methods	7
Health Services Research	5
Quality And Quantity	5
Social Science Research	5
International Journal Of Social Research Methodology	4
International Statistical Review	4
Journal Of Cross Cultural Psychology	4

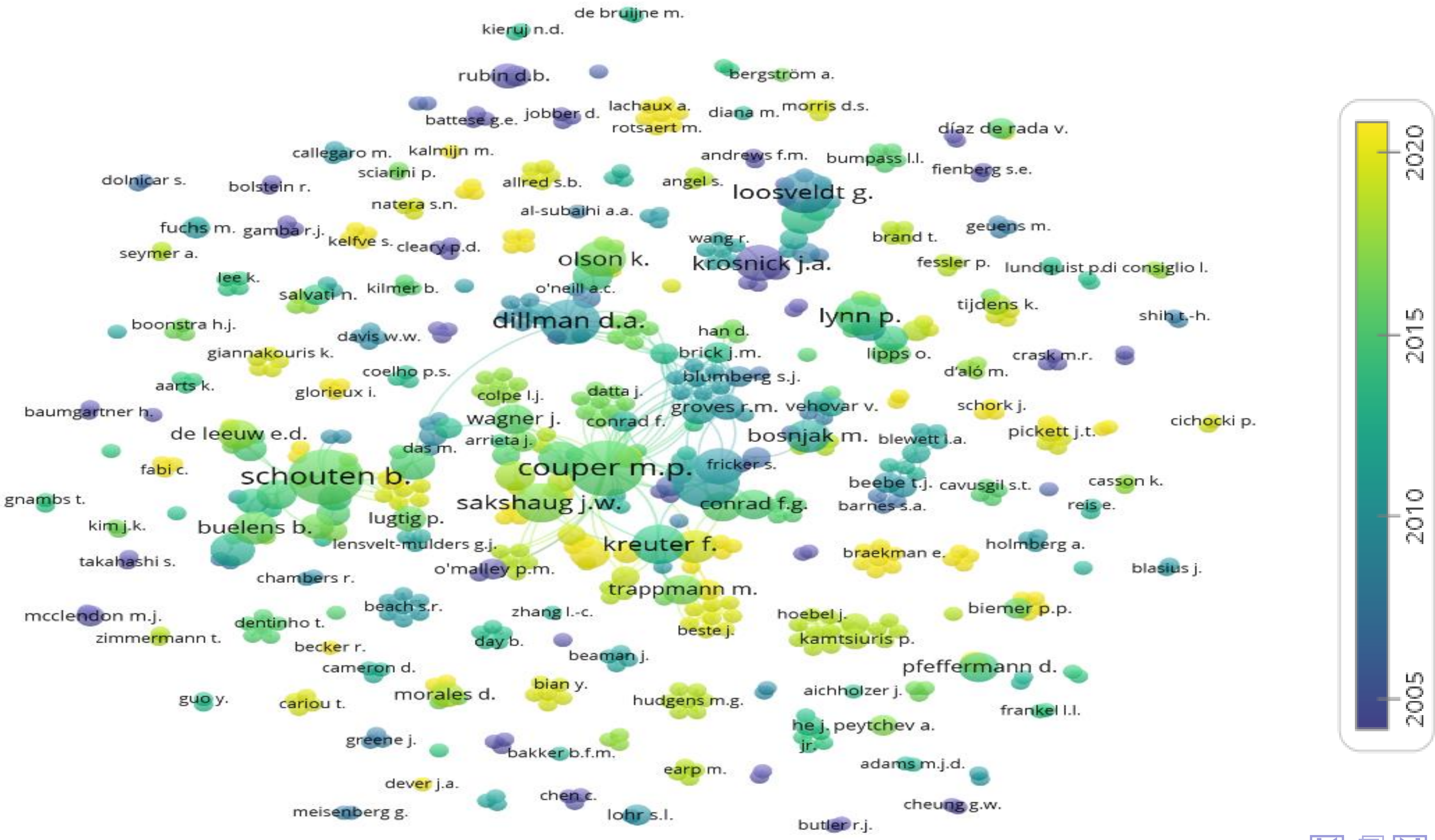
Content analysis





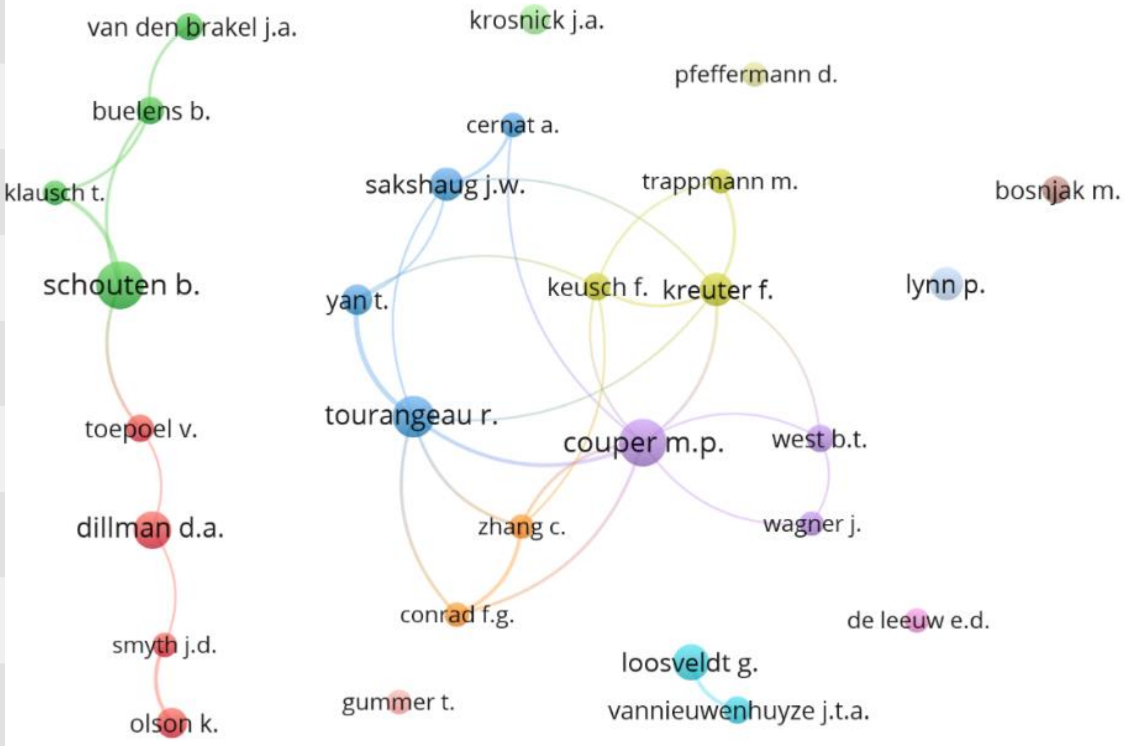
Content analysis

Author co-author density visualization analysis



Content analysis

No	Authors	Year	Cited/Year
1	Rosenbaum	1983	400
2	Tourangeau	2007	103
3	Shih	2008	51
4	Kreuter	2008	48
5	Rubin D.B.	2005	45
6	Yeager	2011	44
7	Manfreda	2008	41
8	Galesic	2009	38
9	Krosnick	1991	38
10	Winship	1994	36



Content analysis

Application of keywords over years.

Before 2010

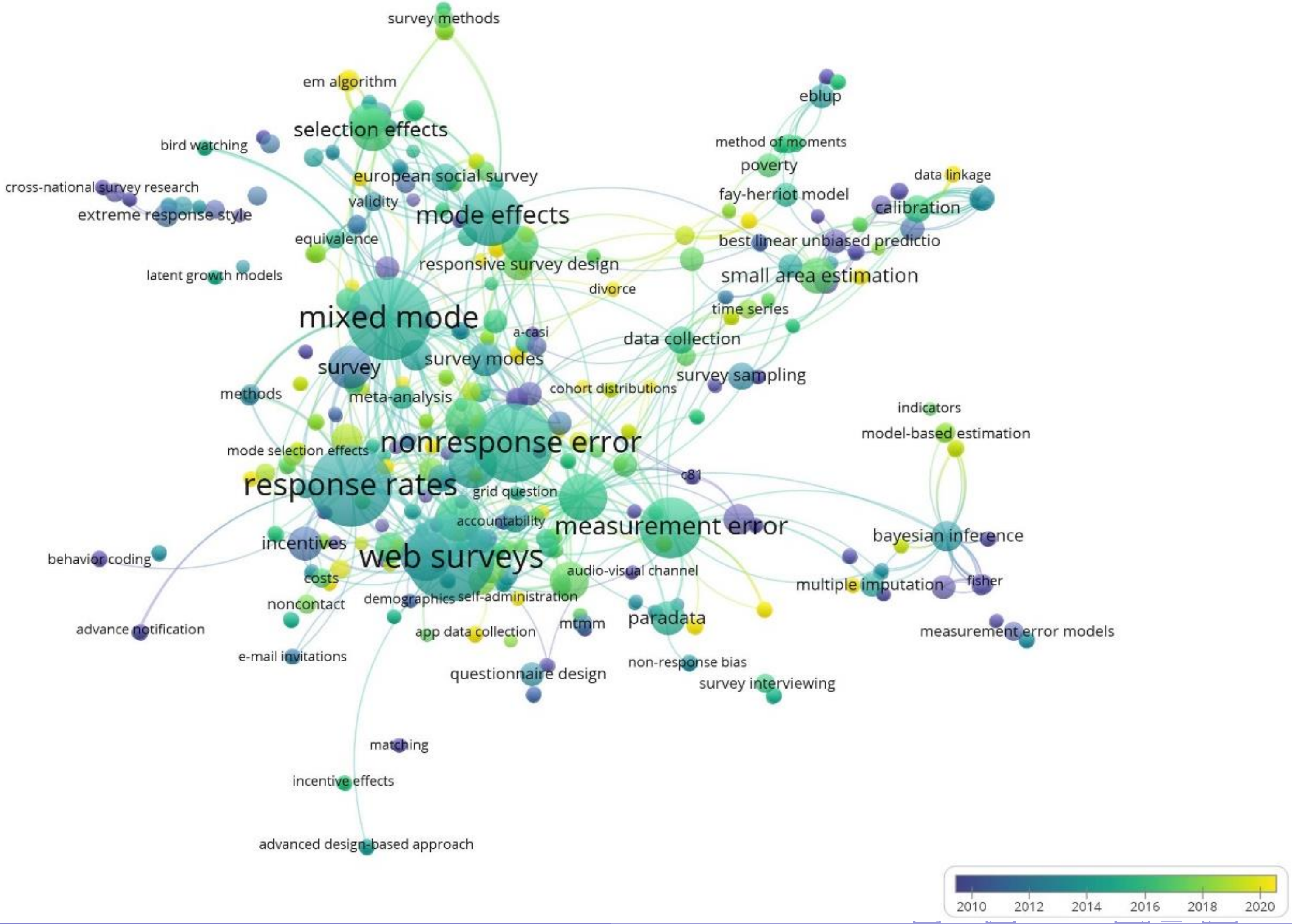
Measurement error - Fisher – Survey Methods– Extreme response - Mail & Phone surveys – Paradata.

2010-2018

Mixed mode – Web survey – Mode effects – nonresponse and measurement errors.

2018-2022

Data linkage – EM algorithm – Responsive survey design – App data collection – Small area estimation.



Results

The problem:

Mixed-mode effects/Nonresponse/Measurement error.

Solutions:

Design weighting to find sampling weights / Nonresponse weighting adjustment / Calibration/Classification:

Estimation of the number of units from the population represented by a specific sample unit and using auxiliary information to deal with nonresponse bias with the following characterstics:

(i) It must be available for all sample units;

(ii) Its population total must be known.

IFCS 2022

Results

- The categorical variables from the demographic information of nonrespondents such as education level, age, income, location, language, and marital status could help the survey methodologists to categorize the target population and recognize the best sequence of the modes.
- Re-interview design and inverse regression estimator (IREG) are among the best approaches to improve measurement bias by using related auxiliary information.

Klausch, T., Schouten, B., Buelens, B., & van den Brakel, J. (2017). Adjusting measurement bias in sequential mixed-mode surveys using re-interview data. *Journal of Survey Statistics and Methodology*, *5*(4), 409–432. https://doi.org/10.1093/jssam/smx022

Results

For categorical variables:

(1) It is important to improve inference in cases where mixed-mode effects are combined with measurement errors caused by primary data collection on categorical variables and socio-demographic information. On one side that the categorical variables are collected with the help of responders (primary data), the survey mode has a strong impact on answering behaviors and answering conditions. Respondents might evaluate some of the new categorical variables as sensitive information or privacy intrusive. They may not be willing to share these personal data by telephone or technological devices, which are necessary for statistical classification.

Results

For categorical variables:

(2) For NSIs, also the new data collection channels are costly and redesign of the survey estimation methodology is time consuming.

(3) the categorical variables should be available in sampling frames (secondary data) and the coverage error is the main concern.

(4) when mixed-mode effects are combined with measurement errors in survey sampling, such as proxy surveys, in which sampled units respond not only for themselves but also for other sampled units, intricate circumstances in subsequent inference might arise. *

* Pfeffermann, D., & Preminger, A. (2021). Estimation Under Mode Effects and Proxy Surveys, Accounting for Non-ignorable Nonresponse. *Sankhya A*, *83*(2), 779–813. https://doi.org/10.1007/s13171-020-00229-w

Results

Re-interview design and inverse regression estimator (IREG):

- the best approaches to improve measurement bias by using related categorical (auxiliary) information.
- The focus of this approach is on the weights of estimators rather than the bias from the measurements.
- the measurement error model is

$$y_{i,m} = u_i + b_m + \varepsilon_{i,m}$$

$y_{i,m}$ the measurement obtained from unit i through mode m

u_i as the observed value for respondent i,

b_m an additive mode-dependent measurement bias,

$\varepsilon_{i,m}$ a mode-dependent measurement variance with an expected value equal to zero.

Results

Re-interview design and inverse regression estimator (IREG):

- the best approaches to improve measurement bias by using related categorical (auxiliary) information.
- The focus of this approach is on the weights of estimators rather than the bias from the measurements.
- the measurement error model is

$$y_{i,m} = u_i + b_m + \varepsilon_{i,m}$$

$y_{i,m}$ the measurement obtained from unit i through mode m

u_i as the observed value for respondent i,

b_m an additive mode-dependent measurement bias,

$\varepsilon_{i,m}$ a mode-dependent measurement variance with an expected value equal to zero.

Results

Re-interview design and inverse regression estimator (IREG):

- If we consider two different modes m and $\acute{m}$, then the differential measurement error between these two modes is given by

$$y_{i,m} - y_{i,\acute{m}} = (b_m - b_{\acute{m}}) + (\varepsilon_{i,m} - \varepsilon_{i,\acute{m}})$$

- The expected value of $(b_m - b_{\acute{m}})$ is the differential measurement bias.
- Now, consider $\hat{t}_y$ as an estimation of the total of variable y according to its observations in different modes y_{i,m'_n} then

$$\hat{t}_y = \sum_{i=1} \omega_i y_{i,m}$$

- ω_i is a survey weight assigned to unit i
- n the number of respondents.

Results

Re-interview design and inverse regression estimator (IREG):

- Now we have a combination of all previouse equations and taking the expectation over the measurement error model (the first equation), we would have

$$E\left(\hat{t}_y\right) = E\left(\sum_{i=1}^{n} \omega_i y_{i,m}\right) = \sum_{i=1}^{n} \omega_i u_{i,m} + \sum_{i=1}^{n} b_m \omega_i K_{i,m} + \sum_{i=1}^{n} \omega_i K_{i,m} E(\varepsilon_{i,m})$$

- with $K_{i,m} = 1$ if unit i responded through mode m, and zero otherwise.

Since $E\left(\varepsilon_{i,m}\right) = 0$

- $E\left(\hat{t}_y\right) = E\left(\sum_{i=1}^{n} \omega_i y_{i,m}\right) = \sum_{i=1}^{n} \omega_i u_{i,m} + \boldsymbol{\sum_{i=1}^{n} \omega_i K_{i,m} b_m}$ (*)
- stating that the expected total of the survey estimates for Y consists of the estimated true total of U, plus true total of b_m from data collected through mode m.
- Since b_m is an unobserved mode-dependent measurement bias, $\sum_{i=1}^{n} \omega_i K_{i,m} b_m$ in equation (*) indicate the existence of an **unknown mode-dependent biases** for estimation of t_y.

IFCS 2022

Results

Re-interview design and inverse regression estimator (IREG):

there is an unknown measurement bias in sequential mixed-mode designs that different estimators might adjust.

How?

By auxiliary information and categorical data obtained via a re-interview design or a sub-set of respondents to the first stage of a sequential mixed-mode survey.

Results

Re-interview design and inverse regression estimator (IREG):

Klausch et al. (2017) propose six different estimators and show that the performance of the estimators strongly depends on the accurate measurement error model. However, they emphasize that an inverse version of the regression estimator (IREG) performs exceptionally well under all considered scenarios.

$$y_i^{m_j} = \hat{\beta}_0 + \hat{\beta}_1 y_i^{m_b}$$

$$\hat{y}^{ireg}_{r_{mm}} = \frac{1}{(\hat{N}_{m_1}+\hat{N}_{m_2})}\left(\sum_{i=1}^{n_{m_b}} d_i y_i^{m_b} + \sum_{i=1}^{m_j} d_i \left(\hat{y}_{re}^{m_b} - \frac{1}{\hat{\beta}_1}\left(\hat{y}_{re}^{m_j} - y_i^{m_j}\right)\right)\right) b,\ j = 1,2; b \neq j$$

Results

So, what we need is auxiliary information in these repeated measurement experiments or mixed-mode re-interviews

Why? to construct regression estimators that correct for mode-dependent selection effects and distinguish mode-specific coverage biases, mode-specific nonresponse biases, and mode-specific relative measurement biases

The result? Good errors adjustment and construct optimal weighting models.

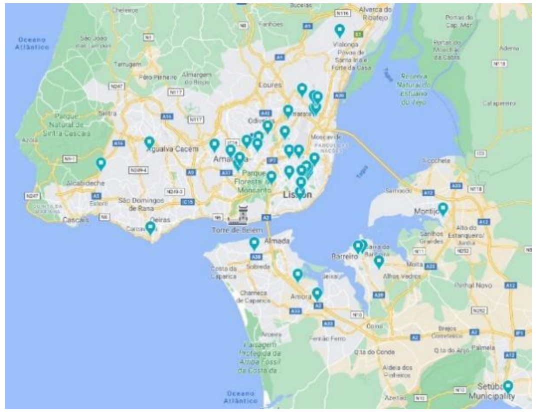
All problems are solved? No! The unknown measurement bias that we have seen in Equations might not be constant along with different editions of our longitudinal mixed-mode surveys as the composition of the mode mixture might change over time.

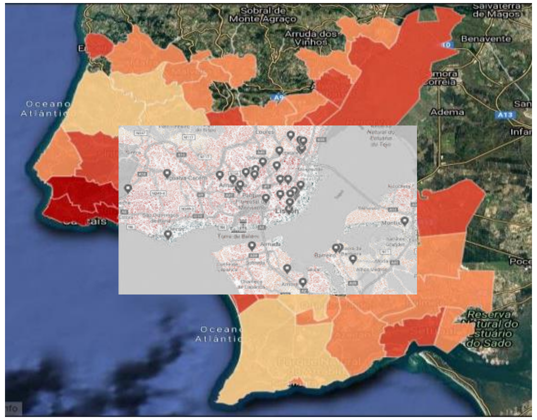
Example

Survey	Labor Force Survey[4]				Household Information and Communication Technology Use Survey (IUTICF)[5]						Survey on Income and Life Condition (SILC)[6]		
Mode	CAPI		CATI		CAPI		CATI		CAWI		CAPI		CATI
Year	2011	2021	2011	2021	2011	2021	2011	2021	2011	2021	2011	2021	2021

ICOT

Área Metropolitana de Lisboa (AML)





Main conclusions

Main concerns:

- nonresponse in CATI surveys and data quality.
- The coverage error: the population of interest are not necessarily in the sampling frame of CATI with telephone
- Lack of enough trained interviewers
- Lack of interest from some new responders, especially in rotation-scheme surveys

Main Solutions:

- Applying a proper statistical classification of survey items and responders to control the nonresponse rates and coverage error risk.
- Calibration of modes by identifying the population subgroups, using categorical variables such as gender, regions, age groups, etc.
- Smoothing the initial weights and recalculating the weights based on a pre-definition of limits between the initial and final weights

IFCS 2022

Main conclusions

- After identifying the categories, applying a sequential mixed-mode design started with CAWI as the cheapest mode supported by an initial postal mail or telephone contact and possible cash incentive.
- With a lag, follow up the non-respondents with giving them a choice between CAPI and CATI according to their specific classification group and demographic information, such as education level, age, income, location, language, and marital status.
- Correcting the mode effects with inverse regression estimator (IREG) and auxiliary information.

Results:

- reduce the cost
- increase the survey accuracy

The main research gaps

- This study showed that sample frames might need updates for necessary categorical information, which are based on choices made several years ago.
- Additionally, more research studies seem necessary for ethics concerns, privacy regulations, and standards for using categorical variables and classification information in social mixed-mode surveys and official statistics.

The following reference might be interesting for following up the new technologies in Official Statistics productions:

Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. Stat. J. IAOS, **37**(3), 771–789, (2021), doi: 10.3233/SJI-200674.