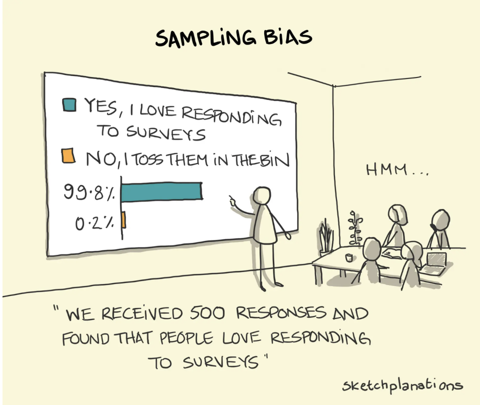
Which Businesses Respond to Surveys? Evidence from Dutch Administrative Data

Jack Fitzgerald

Vrije Universiteit Amsterdam and Tinbergen Institute

November 3, 2025

Sampling Bias in Business Surveys

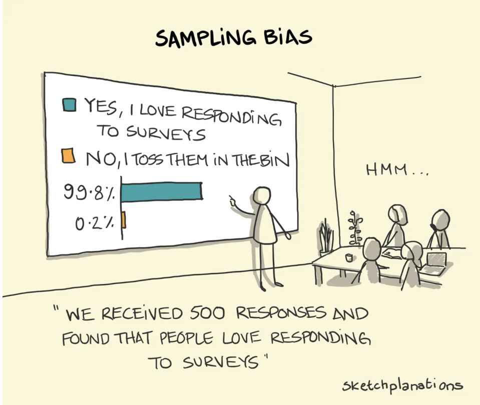


Source: sketchplanations.com/sampling-bias

Everybody thinks that their survey is representative of their population of interest

- **Fundamental problem:** Entities differ in their propensity to answer surveys

Sampling Bias in Business Surveys



Source: sketchplanations.com/sampling-bias

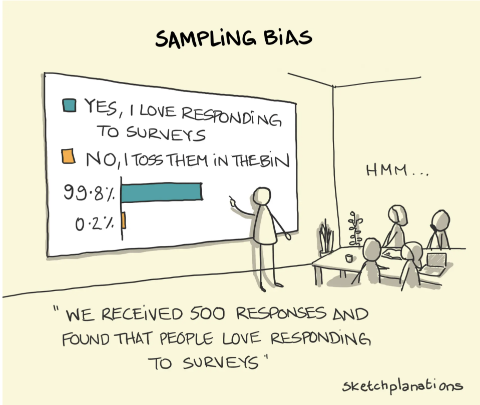
Everybody thinks that their survey is representative of their population of interest

- **Fundamental problem:** Entities differ in their propensity to answer surveys

There are well-established methods for addressing this problem in individual/household surveys

- **Idea:** Calibrate sampling/weights so samples look similar to general population data (e.g., a census)

Sampling Bias in Business Surveys



Source: sketchplanations.com/sampling-bias

Everybody thinks that their survey is representative of their population of interest

- **Fundamental problem:** Entities differ in their propensity to answer surveys

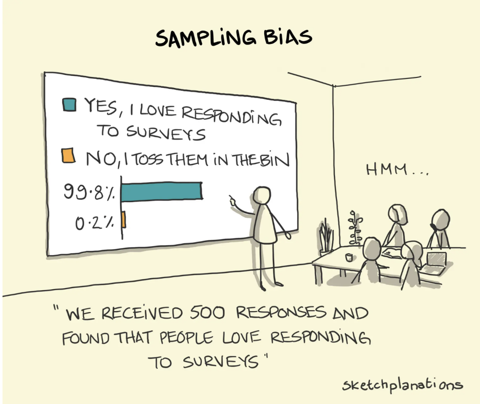
There are well-established methods for addressing this problem in individual/household surveys

- **Idea:** Calibrate sampling/weights so samples look similar to general population data (e.g., a census)

However, similar calibration has historically been impossible for business surveys

- This is because there typically is no 'business census' containing unresponsive businesses

Sampling Bias in Business Surveys



Source: sketchplanations.com/sampling-bias

Everybody thinks that their survey is representative of their population of interest

- **Fundamental problem:** Entities differ in their propensity to answer surveys

There are well-established methods for addressing this problem in individual/household surveys

- **Idea:** Calibrate sampling/weights so samples look similar to general population data (e.g., a census)

However, similar calibration has historically been impossible for business surveys

- This is because there typically is no ‘business census’ containing unresponsive businesses

Important because business surveys are often used in academia and policymaking Literature

This Project

I examine differences between establishments that do and don't respond to business surveys

- I leverage a unique annual register of all establishments in the Netherlands (LISA)

This Project

I examine differences between establishments that do and don't respond to business surveys

- I leverage a unique annual register of all establishments in the Netherlands (LISA)

I find considerable compositional differences between responsive and unresponsive establishments

- The median establishment is a solo enterprise registered to a home address, which is 18 p.p. less likely to respond than the average office

This Project

I examine differences between establishments that do and don't respond to business surveys

- I leverage a unique annual register of all establishments in the Netherlands (LISA)

I find considerable compositional differences between responsive and unresponsive establishments

- The median establishment is a solo enterprise registered to a home address, which is 18 p.p. less likely to respond than the average office
- Establishments with more employees and more fulltime employees are significantly *less* likely to respond to business surveys

This Project

I examine differences between establishments that do and don't respond to business surveys

- I leverage a unique annual register of all establishments in the Netherlands (LISA)

I find considerable compositional differences between responsive and unresponsive establishments

- The median establishment is a solo enterprise registered to a home address, which is 18 p.p. less likely to respond than the average office
- Establishments with more employees and more fulltime employees are significantly *less* likely to respond to business surveys

There's also big sectoral and occupational differences, driven by differences in contact probability

- The highest and lowest sectoral response rates differ by 50 p.p. (8 p.p.) before (after) controlling for contact probability

This Project

I examine differences between establishments that do and don't respond to business surveys

- I leverage a unique annual register of all establishments in the Netherlands (LISA)

I find considerable compositional differences between responsive and unresponsive establishments

- The median establishment is a solo enterprise registered to a home address, which is 18 p.p. less likely to respond than the average office
- Establishments with more employees and more fulltime employees are significantly *less* likely to respond to business surveys

There's also big sectoral and occupational differences, driven by differences in contact probability

- The highest and lowest sectoral response rates differ by 50 p.p. (8 p.p.) before (after) controlling for contact probability
- The difference in response rates between educational facilities and stand locations declines from 53 p.p. to 17 p.p. after controlling for contact probability

This Project

I examine differences between establishments that do and don't respond to business surveys

- I leverage a unique annual register of all establishments in the Netherlands (LISA)

I find considerable compositional differences between responsive and unresponsive establishments

- The median establishment is a solo enterprise registered to a home address, which is 18 p.p. less likely to respond than the average office
- Establishments with more employees and more fulltime employees are significantly *less* likely to respond to business surveys

There's also big sectoral and occupational differences, driven by differences in contact probability

- The highest and lowest sectoral response rates differ by 50 p.p. (8 p.p.) before (after) controlling for contact probability
- The difference in response rates between educational facilities and stand locations declines from 53 p.p. to 17 p.p. after controlling for contact probability

Highlights implementation and generalizability challenges in business surveys, as well as opportunities for improvement

LISA and the Regional Work Registers

Source: lisa.nl

Each year, regional work registers in the Netherlands run *werkgelegenheidsenquêten*

- Surveys ask # male/female fulltime/parttime workers

LISA and the Regional Work Registers

Source: lisa.nl

Each year, regional work registers in the Netherlands run *werkgelegenheidsenquêten*

- Surveys ask # male/female fulltime/parttime workers

The *Landelijk Informatiesysteem van Arbeidsplaatsen (LISA)* aggregates this data from the regional registers each year

- LISA supplements the *werkgelegenheidsenquête* data with administrative records from the *Kamer van Koophandel (KVK)* and *Basisregistratie Addressen en Gebouwen (BAG)* registers

LISA and the Regional Work Registers

Source: lisa.nl

Each year, regional work registers in the Netherlands run *werkgelegenheidsenquêten*

- Surveys ask # male/female fulltime/parttime workers

The *Landelijk Informatiesysteem van Arbeidsplaatsen (LISA)* aggregates this data from the regional registers each year

- LISA supplements the *werkgelegenheidsenquête* data with administrative records from the *Kamer van Koophandel (KVK)* and *Basisregistratie Addressen en Gebouwen (BAG)* registers
- **End result:** Annual panel data on the universe of all establishments in the Netherlands

Data is on *establishments*, rather than *firms*

(Non-)Response

Code	Description	% Establishments in 2022 LISA
1	Data directly from company, statement per branch, obtained in writing, online, or by telephone	19.066%
8	Data directly from company, temporarily no employees	0.011%
11	Data directly from company, statement per branch, through the intervention of a third party authorized by LISA	0.001%
20	Data directly from company total statement, to be allocated to branches	1.205%
30	Data from secondary source per branch (e.g., KVK [recent], annual report, website, press release)	1.407%
40	Data from secondary source total, to be allocated to branches	0.021%
50	Data increased from previous year, VR management module	72.55%
51	Data increased from previous year, other method	3.963%
60	Data imputed, VR management module	0.182%
61	Data imputed, other method	0.103%
72	Data estimated, guesswork	0.029%
73	Data taken directly from previous year	1.462%
76	Data taken directly from following year	0%

In 2022, just over 20% of Dutch establishments responded to the regional *werkgelegenheidsenquêten* (survey type $<= 20$)

(Non-)Response

Code	Description	% Establishments in 2022 LISA
1	Data directly from company, statement per branch, obtained in writing, online, or by telephone	19.066%
8	Data directly from company, temporarily no employees	0.011%
11	Data directly from company, statement per branch, through the intervention of a third party authorized by LISA	0.001%
20	Data directly from company total statement, to be allocated to branches	1.205%
30	Data from secondary source per branch (e.g., KVK [recent], annual report, website, press release)	1.407%
40	Data from secondary source total, to be allocated to branches	0.021%
50	Data increased from previous year, VR management module	72.55%
51	Data increased from previous year, other method	3.963%
60	Data imputed, VR management module	0.182%
61	Data imputed, other method	0.103%
72	Data estimated, guesswork	0.029%
73	Data taken directly from previous year	1.462%
76	Data taken directly from following year	0%

In 2022, just over 20% of Dutch establishments responded to the regional *werkgelegenheidsenquêten* (survey type $<=20$)

- Does not mean 100% were contacted and 20% responded; only a subset are contacted each year

(Non-)Response

Code	Description	% Establishments in 2022 LISA
1	Data directly from company, statement per branch, obtained in writing, online, or by telephone	19.066%
8	Data directly from company, temporarily no employees	0.011%
11	Data directly from company, statement per branch, through the intervention of a third party authorized by LISA	0.001%
20	Data directly from company total statement, to be allocated to branches	1.205%
30	Data from secondary source per branch (e.g., KVK [recent], annual report, website, press release)	1.407%
40	Data from secondary source total, to be allocated to branches	0.021%
50	Data increased from previous year, VR management module	72.55%
51	Data increased from previous year, other method	3.963%
60	Data imputed, VR management module	0.182%
61	Data imputed, other method	0.103%
72	Data estimated, guesswork	0.029%
73	Data taken directly from previous year	1.462%
76	Data taken directly from following year	0%

In 2022, just over 20% of Dutch establishments responded to the regional *werkgelegenheidsenquêten* (survey type $<= 20$)

- Does not mean 100% were contacted and 20% responded; only a subset are contacted each year

Over 72% of establishments' data in the 2022 LISA register were imputed by LISA Standard LISA Imputation

- I introduce a random forest approach which more accurately imputes missing employee headcounts Random Forest Imputation Performance Improvements

Main Variables

Main data source is FIRMBACKBONE employment data (Gerbrands et al. 2025)

- **Focus:** 2022 LISA data (latest year available, first after COVID-19 pandemic)

Main Variables

Main data source is FIRMBACKBONE employment data (Gerbrands et al. 2025)

- **Focus:** 2022 LISA data (latest year available, first after COVID-19 pandemic)
- Restricted to establishments with known sampling classes that are observed in the 2021 LISA ($>$ 1.4M establishments)

Main Variables

Main data source is FIRMBACKBONE employment data (Gerbrands et al. 2025)

- **Focus:** 2022 LISA data (latest year available, first after COVID-19 pandemic)
- Restricted to establishments with known sampling classes that are observed in the 2021 LISA ($>$ 1.4M establishments)

Surveyed data (observed for responsive establishments, imputed for unresponsive):

- Employee headcounts (regional *werkgelegenheidsenquêten* and LISA/RF imputations)
- Proportions of employees that are fulltime and female (computed from employee headcounts)

Main Variables

Main data source is FIRMBACKBONE employment data (Gerbrands et al. 2025)

- **Focus:** 2022 LISA data (latest year available, first after COVID-19 pandemic)
- Restricted to establishments with known sampling classes that are observed in the 2021 LISA ($>$ 1.4M establishments)

Surveyed data (observed for responsive establishments, imputed for unresponsive):

- Employee headcounts (regional *werkgelegenheidsenquêten* and LISA/RF imputations)
- Proportions of employees that are fulltime and female (computed from employee headcounts)

Linked administrative data (always observed):

- Establishment surface area and facility zoning function (Kadaster BAG)
- Year founded and 2008 SBI sector code (KVK)

Descriptives

	P0.5	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5	Mean	SD	*N*
Employees, 2021, LISA Imputation	1	1	1	1	1	1	2	5	12	60	120	4.853	49.72	1433393
Fulltime Employees, 2021, LISA Imputation	0	0	0	0	1	1	1	4	10	51	103	4.169	46.157	1433393
Employees, 2022, LISA Imputation	1	1	1	1	1	1	2	6	12	62	121	4.954	51.129	1433393
Employees, 2022, Random Forest Imputation	1	1	1	1	3	4	5	8	14	64	123	7.399	48.938	1433393
Fulltime Employees, 2022, LISA Imputation	0	0	0	0	1	1	1	5	10	53	106	4.273	47.862	1433393
Fulltime Employees, 2022, Random Forest Imputation	0	0	1	1	3	4	5	7	12	55	109	6.786	45.429	1433393
Female Employees, 2022, LISA Imputation	0	0	0	0	0	0	1	2	6	28	55	2.23	28.214	1433393
Female Employees, 2022, Random Forest Imputation	0	0	0	0	1	1	3	5	7	29	56	3.375	27.754	1433393
Proportion Employees Fulltime, 2022, LISA Imputation	0	0	0	0	1	1	1	1	1	1	1	0.814	0.366	1433239
Proportion Employees Fulltime, 2022, Random Forest Imputation	0	0	0.615	0.778	1	1	1	1	1	1	1	0.935	0.185	1433239
Proportion Employees Female, 2022, LISA Imputation	0	0	0	0	0	0	1	1	1	1	1	0.365	0.436	1433239
Proportion Employees Female, 2022, Random Forest Imputation	0	0	0	0	0.25	0.333	0.667	0.8	1	1	1	0.415	0.28	1433239
Establishment Surface Area, m^2	17	28	55	70	100	140	233	695	1700	9160.16	16481	625.924	4784.077	1433393
Year Founded	2000	2001	2007	2010	2015	2018	2020	2022	2022	2022	2023	2016.775	4.803	1433071

Unbounded continuous variables exhibit extreme skew

- For firm size measures, about as much variation between the 0.5th and 99th percentiles as there is between the 99th and 99.5th percentiles
- I address this by analyzing all unbounded continuous variables using robust regression

Descriptives

	P0.5	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5	Mean	SD	N
Employees, 2021, LISA Imputation	1	1	1	1	1	1	2	5	12	60	120	4.853	49.72	1433393
Fulltime Employees, 2021, LISA Imputation	0	0	0	0	1	1	1	4	10	51	103	4.169	46.157	1433393
Employees, 2022, LISA Imputation	1	1	1	1	1	1	2	6	12	62	121	4.954	51.129	1433393
Employees, 2022, Random Forest Imputation	1	1	1	1	3	4	5	8	14	64	123	7.399	48.938	1433393
Fulltime Employees, 2022, LISA Imputation	0	0	0	0	1	1	1	5	10	53	106	4.273	47.862	1433393
Fulltime Employees, 2022, Random Forest Imputation	0	0	1	1	3	4	5	7	12	55	109	6.786	45.429	1433393
Female Employees, 2022, LISA Imputation	0	0	0	0	0	0	1	2	6	28	55	2.23	28.214	1433393
Female Employees, 2022, Random Forest Imputation	0	0	0	0	1	1	3	5	7	29	56	3.375	27.754	1433393
Proportion Employees Fulltime, 2022, LISA Imputation	0	0	0	0	1	1	1	1	1	1	1	0.814	0.366	1433239
Proportion Employees Fulltime, 2022, Random Forest Imputation	0	0	0.615	0.778	1	1	1	1	1	1	1	0.935	0.185	1433239
Proportion Employees Female, 2022, LISA Imputation	0	0	0	0	0	0	1	1	1	1	1	0.365	0.436	1433239
Proportion Employees Female, 2022, Random Forest Imputation	0	0	0	0	0.25	0.333	0.667	0.8	1	1	1	0.415	0.28	1433239
Establishment Surface Area, m^2	17	28	55	70	100	140	233	695	1700	9160.16	16481	625.924	4784.077	1433393
Year Founded	2000	2001	2007	2010	2015	2018	2020	2022	2022	2022	2023	2016.775	4.803	1433071

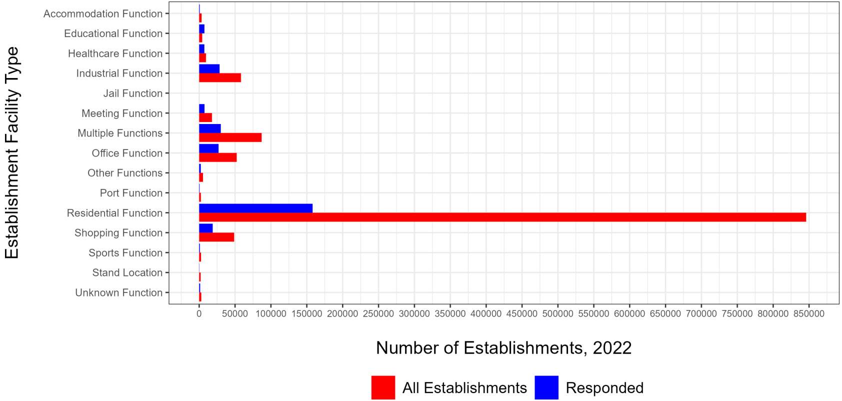
Unbounded continuous variables exhibit extreme skew

- For firm size measures, about as much variation between the 0.5th and 99th percentiles as there is between the 99th and 99.5th percentiles
- I address this by analyzing all unbounded continuous variables using robust regression

The median establishment is a solo enterprise

- Not unique to Dutch context; similar patterns in the U.S. (Conway et al. 2018)

Facility Types



By far most common type of facility is residential, reflecting small enterprises that register at an owner's home address Sectors Regions

Research Questions

Two key questions:

1. How do characteristics differ between establishments who do and do not respond to business surveys?

Research Questions

Two key questions:

1. How do characteristics differ between establishments who do and do not respond to business surveys?
2. *Conditional on being contacted for a business survey*, how are the characteristics of responsive and unresponsive establishments expected to differ?

Research Questions

Two key questions:

1. How do characteristics differ between establishments who do and do not respond to business surveys?
2. *Conditional on being contacted for a business survey*, how are the characteristics of responsive and unresponsive establishments expected to differ?

(1) can be answered using an *unconditional* difference in expectations; letting $Y(R)$ denote potential outcomes by establishment responsiveness...

$$\delta_U = \mathbb{E}\left[Y(1) - Y(0)\right] \tag{1}$$

Research Questions

Two key questions:

1. How do characteristics differ between establishments who do and do not respond to business surveys?
2. *Conditional on being contacted for a business survey*, how are the characteristics of responsive and unresponsive establishments expected to differ?

(1) can be answered using an *unconditional* difference in expectations; letting $Y(R)$ denote potential outcomes by establishment responsiveness...

$$\delta_U = \mathbb{E}\left[Y(1) - Y(0)\right] \tag{1}$$

(2) is a difference in *conditional* expectations; letting C indicate contact, we can write this as

$$\delta_C = \mathbb{E}\left[Y(1) - Y(0) \mid C = c\right] \tag{2}$$

Research Questions

Two key questions:

1. How do characteristics differ between establishments who do and do not respond to business surveys?
2. *Conditional on being contacted for a business survey*, how are the characteristics of responsive and unresponsive establishments expected to differ?

(1) can be answered using an *unconditional* difference in expectations; letting $Y(R)$ denote potential outcomes by establishment responsiveness...

$$\delta_U = \mathbb{E}\left[Y(1) - Y(0)\right] \tag{1}$$

(2) is a difference in *conditional* expectations; letting C indicate contact, we can write this as

$$\delta_C = \mathbb{E}\left[Y(1) - Y(0) \mid C = c\right] \tag{2}$$

(1) is useful for understanding the generalizability of existing business surveys, whereas (2) can help plan for future business surveys

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Determinants of contact probability

1. **KVK size class:** Establishment contact probabilities systematically vary depending on whether they have 1, 2-9, 10-99, or 100+ fulltime employees in the prior year

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Determinants of contact probability

1. **KVK size class:** Establishment contact probabilities systematically vary depending on whether they have 1, 2-9, 10-99, or 100+ fulltime employees in the prior year
2. **Having 5+ employees in the prior year's survey:** May be manually contacted a second time

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Determinants of contact probability

1. **KVK size class:** Establishment contact probabilities systematically vary depending on whether they have 1, 2-9, 10-99, or 100+ fulltime employees in the prior year
2. **Having 5+ employees in the prior year's survey:** May be manually contacted a second time
3. **Prior year survey type:** May be manually contacted if known to be responsive

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Determinants of contact probability

1. **KVK size class:** Establishment contact probabilities systematically vary depending on whether they have 1, 2-9, 10-99, or 100+ fulltime employees in the prior year
2. **Having 5+ employees in the prior year's survey:** May be manually contacted a second time
3. **Prior year survey type:** May be manually contacted if known to be responsive
4. **Responding by email in the prior year's survey:** Virtually always contacted again by email

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Determinants of contact probability

1. **KVK size class:** Establishment contact probabilities systematically vary depending on whether they have 1, 2-9, 10-99, or 100+ fulltime employees in the prior year
2. **Having 5+ employees in the prior year's survey:** May be manually contacted a second time
3. **Prior year survey type:** May be manually contacted if known to be responsive
4. **Responding by email in the prior year's survey:** Virtually always contacted again by email
5. **Regional registers:** Decide how many establishments to survey and how contact is conducted

Controlling for Contact

If I knew which establishments were contacted for the regional *werkgelegenheidsenquêten*, then conditioning on contact would be trivial

- **Problem:** Regional business registers delete which establishments are contacted after six months
- **Solution:** Conditional on stratification/exception variables, surveys are randomly distributed

Determinants of contact probability

1. **KVK size class:** Establishment contact probabilities systematically vary depending on whether they have 1, 2-9, 10-99, or 100+ fulltime employees in the prior year
2. **Having 5+ employees in the prior year's survey:** May be manually contacted a second time
3. **Prior year survey type:** May be manually contacted if known to be responsive
4. **Responding by email in the prior year's survey:** Virtually always contacted again by email
5. **Regional registers:** Decide how many establishments to survey and how contact is conducted

Thanks to conditional randomization, controlling for all of these determinants in matrix X renders $\mathbb{E}[Y(1) - Y(0) \mid X = x]$ an unbiased estimator for $\mathbb{E}[Y(1) - Y(0) \mid C = c]$

Contact Determinants and Response Probability

As expected, systematic contact determinants are associated with higher response probability

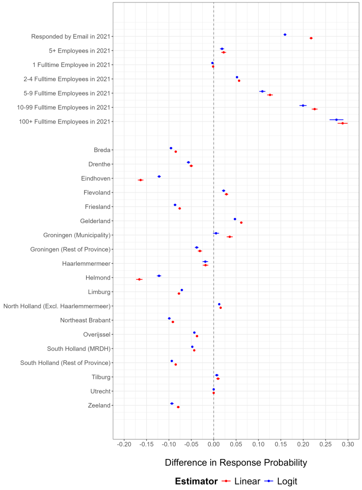
- Responding by email in the previous year's survey yields nearly 16 p.p. bump in response probability

Contact Determinants and Response Probability

As expected, systematic contact determinants are associated with higher response probability

- Responding by email in the previous year's survey yields nearly 16 p.p. bump in response probability
- Response probabilities also steadily rise with size classes up to 28 p.p.

Contact Determinants and Response Probability



As expected, systematic contact determinants are associated with higher response probability

- Responding by email in the previous year's survey yields nearly 16 p.p. bump in response probability
- Response probabilities also steadily rise with size classes up to 28 p.p.
- Considerable heterogeneity across regions (differences in response probability up to 17 p.p.)

Contact Determinants and Response Probability

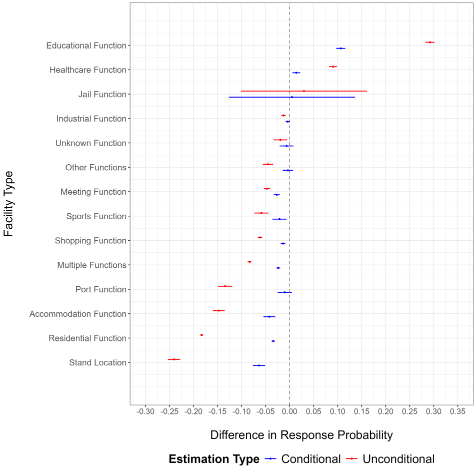
As expected, systematic contact determinants are associated with higher response probability

- Responding by email in the previous year's survey yields nearly 16 p.p. bump in response probability
- Response probabilities also steadily rise with size classes up to 28 p.p.
- Considerable heterogeneity across regions (differences in response probability up to 17 p.p.)

Because these specifications control for prior year survey type, these response rate differences should only be explained by differences in contact probability

- Controlling for these determinants yields unbiased estimates of contact-conditional differences in response probability

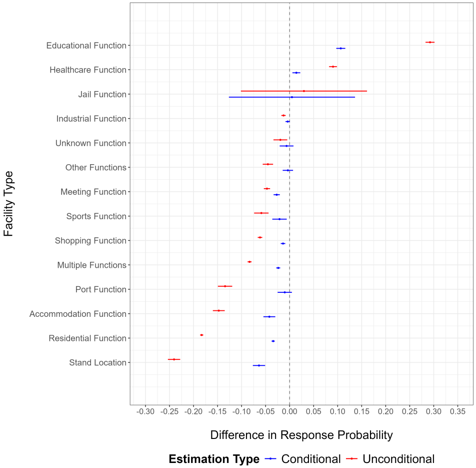
Facility Types



Compared to offices:

- Educational facilities are significantly more likely to be responsive, both conditionally and unconditionally

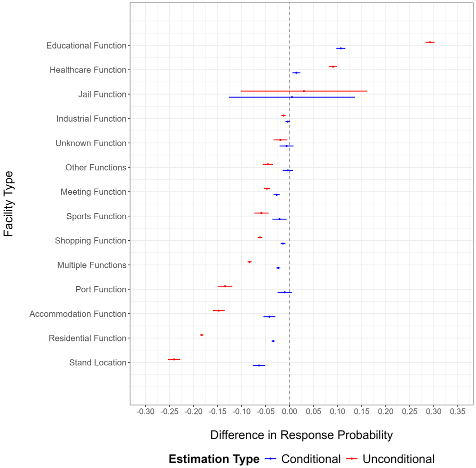
Facility Types



Compared to offices:

- Educational facilities are significantly more likely to be responsive, both conditionally and unconditionally
- Least responsive establishments include hotels, stands, and residential establishments

Facility Types



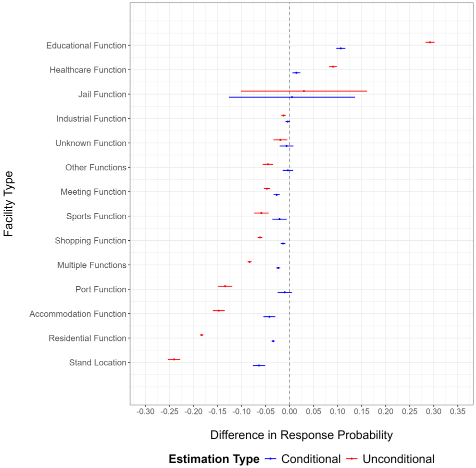
Compared to offices:

- Educational facilities are significantly more likely to be responsive, both conditionally and unconditionally
- Least responsive establishments include hotels, stands, and residential establishments

Residential addresses are 18 p.p. less likely to respond than the average office

- Remember, residential properties are by far the most common type!

Facility Types



Compared to offices:

- Educational facilities are significantly more likely to be responsive, both conditionally and unconditionally
- Least responsive establishments include hotels, stands, and residential establishments

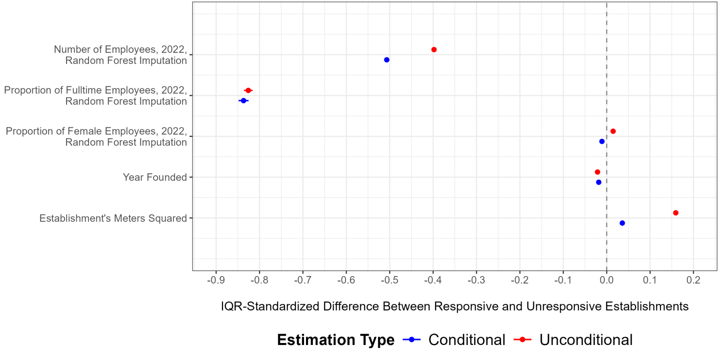
Residential addresses are 18 p.p. less likely to respond than the average office

- Remember, residential properties are by far the most common type!

Controlling for contact probability collapses these estimates by over two thirds

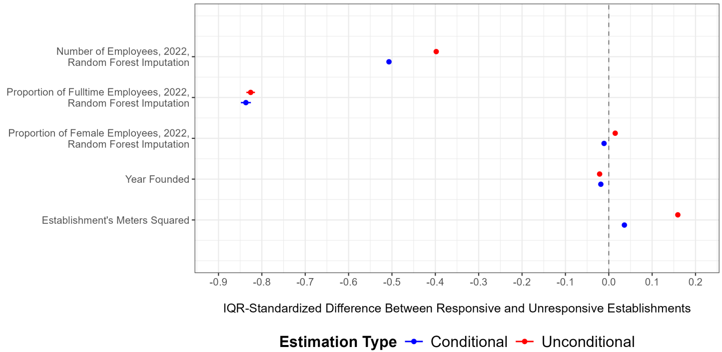
- Implies much of this representativeness gap is driven by differences in contact probability

Continuous Characteristics



Responsive establishments have ~ 2 fewer workers and and are more composed of parttime workers (~ 15 p.p.); consistent with responsive establishments being more likely to efficiently divide labor

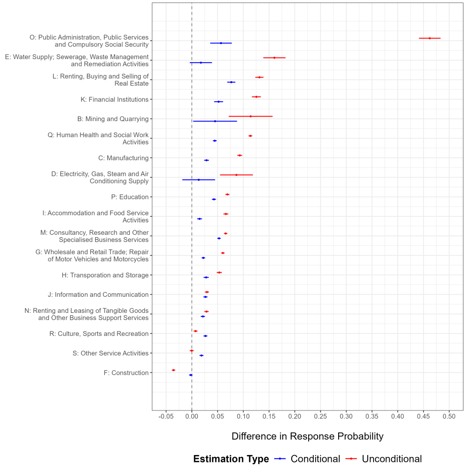
Continuous Characteristics



Responsive establishments have ~ 2 fewer workers and and are more composed of parttime workers (~ 15 p.p.); consistent with responsive establishments being more likely to efficiently divide labor

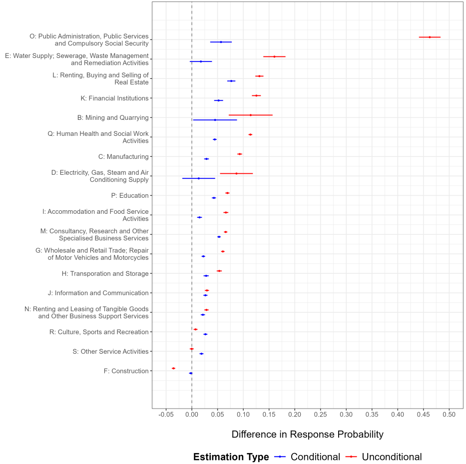
- Some measurable difference in literal establishment size, but only by $\sim 10m^2$

Sectors



Huge overrepresentation of white-collar industries in responsive establishments

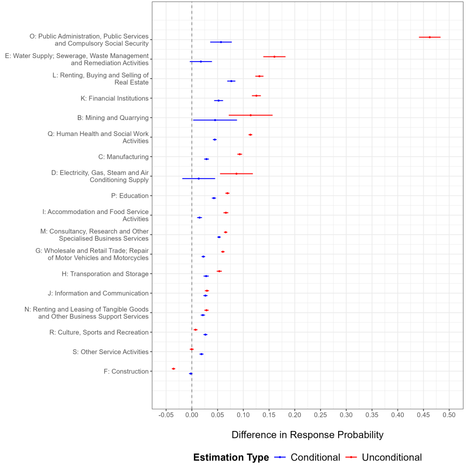
Sectors



Huge overrepresentation of white-collar industries in responsive establishments

- Three of top four include public administration, real estate, and financial institutions

Sectors



Huge overrepresentation of white-collar industries in responsive establishments

- Three of top four include public administration, real estate, and financial institutions

Again, controlling for contact probability attenuates estimated gaps in contact probability by up to 84%

- 11 of the 19 sectoral response rate advantages attenuate by more than half after controlling for contact probability

Conclusion

People running/analyzing business surveys need to recognize that their sample is likely unrepresentative

- Responsive establishments are likely smaller, have larger shares of parttime workers, are concentrated in white-collar industries, and are relatively unlikely to be solo enterprises

Conclusion

People running/analyzing business surveys need to recognize that their sample is likely unrepresentative

- Responsive establishments are likely smaller, have larger shares of parttime workers, are concentrated in white-collar industries, and are relatively unlikely to be solo enterprises

Don't despair! You can use *conditional* differences in response probability to your advantage

Conclusion

People running/analyzing business surveys need to recognize that their sample is likely unrepresentative

- Responsive establishments are likely smaller, have larger shares of parttime workers, are concentrated in white-collar industries, and are relatively unlikely to be solo enterprises

Don't despair! You can use *conditional* differences in response probability to your advantage

- If you want a more representative sample, contact more conditionally underrepresented establishments (my estimates imply good upside here)

Conclusion

People running/analyzing business surveys need to recognize that their sample is likely unrepresentative

- Responsive establishments are likely smaller, have larger shares of parttime workers, are concentrated in white-collar industries, and are relatively unlikely to be solo enterprises

Don't despair! You can use *conditional* differences in response probability to your advantage

- If you want a more representative sample, contact more conditionally underrepresented establishments (my estimates imply good upside here)
- If you're resource-constrained and need to maximize response rates, now you know which establishments are most likely to respond

Conclusion

People running/analyzing business surveys need to recognize that their sample is likely unrepresentative

- Responsive establishments are likely smaller, have larger shares of parttime workers, are concentrated in white-collar industries, and are relatively unlikely to be solo enterprises

Don't despair! You can use *conditional* differences in response probability to your advantage

- If you want a more representative sample, contact more conditionally underrepresented establishments (my estimates imply good upside here)
- If you're resource-constrained and need to maximize response rates, now you know which establishments are most likely to respond

Future directions: Survey weighting for business surveys

References I

Altig, D., J. M. Barrero, N. Bloom, S. J. Davis, B. Meyer, and N. Parker (2022).
Surveying business uncertainty.
Journal of Econometrics 231(1), 282–303.

Baker, H. K., J. C. Singleton, and E. T. Veit (2011).
Survey research in corporate finance: Bridging the gap between theory and practice.
Oxford University Press.

Clar, M., J.-C. Duque, and R. Moreno (2007).
Forecasting business and consumer surveys indicators – A time-series models competition.
Applied Economics 39(20), 2565–2580.

Collins, D. (2001).
The relationship between business confidence surveys and stock market performance.
Investment Analysts Journal 30(54), 9–17.

References II

Conway, M., J. Klein, B. Robles, and M. Weindorf (2018, Dec).
2018 report on nonemployer firms: Findings from the 2017 Small Business Credit Survey.

Gerbrands, P., W. H. J. Hassink, D. L. Oberski, R. Schilpzand, and A. van Witteloostuijn (2025).
FIRMBACKBONE employment data on all Dutch entities.
Dataset V1, DataverseNL.

Hansson, J., P. Jansson, and M. Löf (2005).
Business survey data: Do they help in forecasting GDP growth?
International Journal of Forecasting 21(2), 377–389.

Karanja, E., A. Sharma, and I. Salama (2020).
What does MIS survey research reveal about diversity and representativeness in the MIS field? A content analysis approach.
Scientometrics 122(3), 1583–1628.

References III

Kent Baker, H. and T. K. Mukherjee (2007).
Survey research in finance: Views from journal editors.
International Journal of Managerial Finance 3(1), 11–25.

Klein, L. R. and S. Özmucur (2010).
The use of consumer and business surveys in forecasting.
Economic Modelling 27(6), 1453–1462.

Snijkers, G., G. Haraldsen, J. Jones, and D. K. Willimack (2013).
Designing and conducting business surveys.
John Wiley & Sons.

Wright, M. N. and A. Ziegler (2017).
ranger: A fast implementation of random forests for high dimensional data in C++ and R.
Journal of Statistical Software 77(1), 1–17.

References IV

Zimmermann, K. F. (1999).
Analysis of business surveys, Chapter 9, pp. 369–399.
John Wiley & Sons, Ltd.

Why Should We Care?

Business surveys are routinely used to survey firm expectations and financial management practices which are unobservable in administrative data (e.g., see Zimmermann 1999; Collins 2001; Hansson, Jansson, & Löf 2005; Baker & Mukherjee 2007; Clar, Duque, & Moreno 2007; Klein & Özmucur 2010; Baker, Singleton, & Veit 2011; Snijkers et al. 2013; Altig et al. 2022)

- Often used for understanding and forecasting market and economic outcomes

Many (sub)disciplines in management and finance rely heavily on business surveys

- 32-41% of empirical research publications in management information systems use survey data, of which over 41% target firms as the primary unit of analysis (Karanja, Sharma, & Salama 2020)

If the businesses who respond to these surveys are unrepresentative, then the surveys may yield misleading generalizations on firms, markets, and the economy

- But if we know *how* and *why* the surveys are unrepresentative, then we can leverage this information to correct sampling biases and improve survey design Back

LISA's Standard Imputation Procedure

For each employee headcount of (1) fulltime females, (2) parttime females, (3) fulltime males, and (4) parttime males...

- Within each combination of SBI code groups (A-B, C-F, G-I, H-N, O-P, Q, and S-U) and firm size classes (2-4, 5-49, and 50+)...
  1. Find the average growth rate of the relevant employee headcount between year t and $t-1$ for responsive establishments
  2. For nonresponsive establishments with the same combination of SBI code group and firm size class, obtain the relevant employee headcount for year t by multiplying that headcount from year $t-1$ by the relevant growth rate and rounding

The described procedure in the LISA handbook also provides for the possibility of further stratification by COROP region and some exceptions

- I ignore these because the LISA register doesn't note when these deviations have been applied

TLDR: Employee headcounts for unresponsive establishments are imputed using sector-size growth trends of responsive establishments Back

A Novel Random Forest Imputation Strategy

For each of the four employee headcounts...

1. On a randomly-selected half of the responsive establishments, fit a random forest regression model to predict the relevant employee headcount
   - `ranger` package in R (Wright & Ziegler 2017), using SBI codes, COROP regions, and the previous year's four employee headcounts and survey type as features
2. Predict the relevant headcount of unresponsive establishments using the random forest model estimated in (1) and round

Primary cost is computational power; you realistically need 64GB of RAM to run everything Back

Performance Differences Between Imputation Algorithms

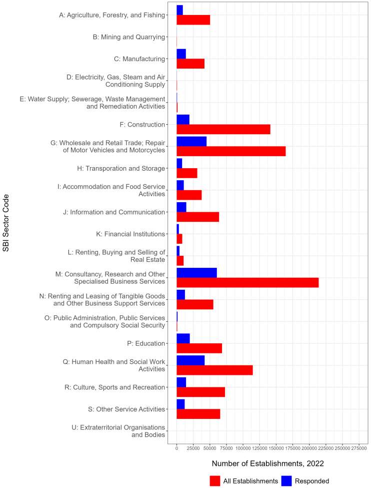
	Same Variable, LISA Imputation	Same Variable, 2021 LISA	Same Variable, Random Forest Imputation
# Employees	0.575	1.003	1.211
	(0.046)	(0.042)	(0.102)
Relative MSPE	1	0.133	0.37
Relative MAD	1	0.376	0.444
% Employees Female	0.815	0.826	0.996
	(0.002)	(0.002)	(0.002)
Relative MSPE	1	0.979	0.871
Relative MAD	1	0.973	1.171
% Employees Fulltime	0.455	0.429	0.831
	(0.007)	(0.006)	(0.006)
Relative MSPE	1	1.098	0.526
Relative MAD	1	1.04	0.825

In hold-out test data, the LISA imputation method significantly underestimates all firm composition measures of interest to my study

- Performs poorly even compared to a simple carryover imputation

Compared to LISA imputation, my random forest imputation achieves global out-of-sample improvements on both slope and fit Back

SBI Sectors

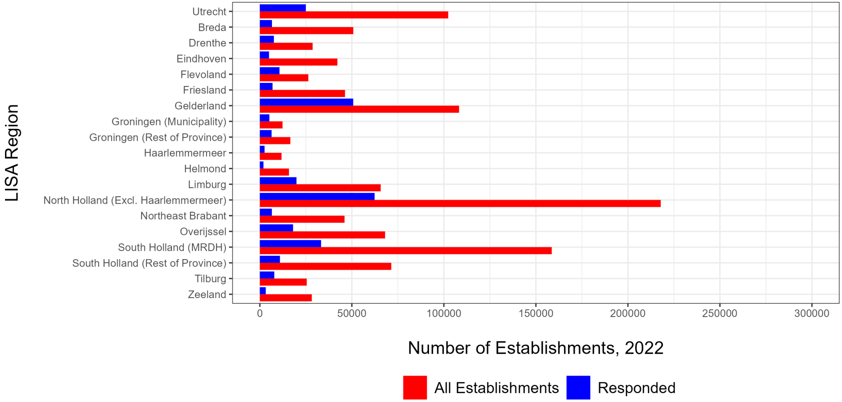


For sparsity, I focus on one-digit SBI sectors (KVK classification)

- Most common sectors are consulting, retail, and construction
- Due to high response rates, the healthcare sector is also well-represented in responsive establishments

Back

LISA Regions



LISA regions are represented in the data roughly according to local population Back