# Exact extraction of the largest balanced panel from MIX Market microfinance data using a superset-sum transform

*LUIS EDUARDO MUÑOZ GUERRERO*

**Abstract**: *Balanced panels are a common input to cross-country studies of microfinance institutions (MFIs) based on MIX Market data, yet the step that cuts a balanced panel from an unbalanced reporting history is seldom treated as part of the method. This study treats it as a data-engineering problem with a stated objective, an exact algorithm, verification and measured cost. The reporting history is encoded as a binary institution-by-year matrix. The balanced panel with the most institution-years is a maximum edge biclique; because the year side is short, it is found exactly by counting, for every set of years, the institutions that report in all of them, and a superset-sum transform gives all 2m counts in O(nm + m·2m) time. On the MIX Market file, with 2,627 MFIs reporting the core variables in fiscal years 1999–2019, the optimum is unique: 548 MFIs over 2006–2011, or 3,288 institution-years (20.2% of those available). The transform takes 19.8 ms against 8.71 s for direct enumeration and agrees with it on all 2,097,152 year sets. A pre-specified comparison shows that the panel is not a random draw from the MFIs reporting in those years: its members sit 24.4 within-year percentile points higher in assets and 6.0 points higher in operational self-sufficiency, and Latin American MFIs make up 38.3% of the panel against 18.6% of those excluded. Studies that use balanced MIX panels should state the extrac-tion rule and report the composition it selects.*

**Keywords:** Balanced Panel; Maximum Edge Biclique; Superset-Sum Transform; Reproducible Data Pipeline; Microfinance Institutions; Mix Market; Selection Bias

PhD, Facultad de Ingenierías, Universidad Tecnológica de Pereira, Pereira, Risaralda, Colombia
Correspondence: lemunozg@utp.edu.co · ORCID: 0000-0002-9414-6187

## Introduction

Cross-country research on the performance of microfinance institutions (MFIs) has examined profitability and outreach (Cull et al., 2007, 2009), efficiency (Gutiérrez-Nieto et al., 2007; Hermes et al., 2011), governance (Hartarska, 2005; Mersland & Strøm, 2009), regulation (Hartarska & Nadolnyak, 2007), the macroeconomic context (Ahlin et al., 2011), capital structure (Bogan, 2012) and leverage (Chikalipah, 2019). This work rests on institution-level financial statements that MFIs report themselves, and MIX Market, a public database to which MFIs reported their financial and outreach indicators on a voluntary basis, is a common source of such data (e.g., Bauchet & Morduch, 2010; Chikalipah, 2019). Because reporting is voluntary, the history of each institution in the database is incomplete: institutions enter, leave and skip years.

The extent of the gaps is easy to state. In the release used here, 2,627 MFIs report the four core variables (total assets, gross loan portfolio, number of active borrowers and operational self-sufficiency, OSS) in at least one fiscal year between 1999 and 2019. None reports them in all 21 years, and only 29.5% of the institution-year cells are filled. Methods for unbalanced panels exist (Baltagi & Song, 2006), but a balanced panel, with the same institutions observed in every year, remains a convenient input for estimators, for efficiency frontiers computed year by year and for comparisons over a fixed period. A researcher who wants one must choose the years to keep, and that choice fixes which institutions remain.

When the choice is made by hand, a natural candidate is a window of consecutive years in which many institutions report. Two questions then arise, one of engineering and one of statistics. The engineering question is whether the chosen panel is the largest available and whether anyone else can rebuild it. Choosing a set of institutions and a set of years so that every institution reports in every year, with as many institution-years as possible, is the maximum edge biclique problem, which is NP-complete in general (Peeters, 2003) and hard to approximate (Ambühl et al., 2011). A window chosen by inspection carries no guarantee of being optimal and, unless the selection rule is written down, no record of how it was found. The statistical question is what the panel selects. Any balanced panel keeps only institutions that report persistently, a condition of the kind that the literatures on missing data (Rubin, 1976), selectivity in panels (Verbeek & Nijman, 1992) and survivorship bias (Brown et al., 1992; Elton et al., 1996) treat with caution, and reporting to MIX is known to be non-random (Bauchet & Morduch, 2010).

This study addresses both questions with a systems-engineering approach. The extraction of the panel is specified by requirements with an explicit objective, implemented as a pipeline whose stages are verified and timed, and followed by an audit of what the extraction selects. The study makes four contributions. (i) It shows that, because the number of years m is small, the exact optimum can be computed by evaluating, for all 2m sets of years, the number of institutions that report in every year of the set, using a superset-sum transform at a cost of O(nm + m·2m) time and 8·2m bytes of memory. (ii) It describes a reproducible pipeline, with hash-checked inputs, reporting rules stated as code, property-based tests against brute-force oracles, exhaustive cross-checks on the real data and an analysis plan fixed before the composition

comparison, and it reports the pipeline's measured cost. (iii) It applies the method to MIX Market under four reporting rules and gives the exact optimum, the frontier between panel length and panel width, and the performance of contiguous windows and of a greedy heuristic. (iv) It measures how the institutions of the largest balanced panel differ from those it leaves out.

The rest of the article is organised as follows. The literature review covers MFI performance research, selection in panels, bicliques and subset transforms, and reproducible computation. The methodology section states the requirements, the data, the problem and its algorithm, the verification, the composition analysis, the descriptive analyses added after the plan was fixed and the cost measurement. The results section reports the reporting structure, the optima, the frontier, the cost and the composition audit; the discussion interprets them, turns them into a reporting checklist and states the limitations, and the conclusion closes.

**Literature review**

The review covers work published up to December 2020 in four strands: MFI performance research based on reported financial data; selection, attrition and unbalanced panels; bicliques and subset transforms; and reproducible computational research.

MFI performance research and its data. Morduch (1999) described how microfinance programmes had lent to low-income households with high repayment rates, even without collateral, while for the most part high repayment rates had not translated into profits, and highlighted the measurement of financial sustainability. Cull et al. (2007) examined profitability, loan repayment and cost reduction with high-quality financial information on 124 institutions in 49 countries, and Cull et al. (2009), with a large dataset that includes most of the world's leading MFIs, asked who the lenders and the customers are, how widespread profitability is and how important subsidies are. Efficiency has been studied on its own (Gutiérrez-Nieto et al., 2007) and together with outreach (Hermes et al., 2011), and it has been measured with data envelopment analysis for financial and social outputs, the latter including indicators of impact on women and of poverty reach (Gutiérrez-Nieto et al., 2009). Other work has related performance to the macroeconomic context (Ahlin et al., 2011), to governance (Hartarska, 2005; Mersland & Strøm, 2009) and to regulation: with data on 114 MFIs from 62 countries, Hartarska and Nadolnyak (2007) found that regulatory involvement did not directly affect operational self-sustainability or outreach. Mission drift (Mersland & Strøm, 2010) and, in a global analysis, the repayment of women borrowers (D'Espallier et al., 2011) have also been examined. With data from Eastern Europe and Central Asia, Caudill et al. (2009) estimated that roughly half of the MFIs are able to operate with reduced costs over time, and found that larger MFIs offering deposits, and those receiving lower subsidies, operate more cost-effectively over time; Bogan (2012) linked greater use of grants by large MFIs to lower operational self-sufficiency.

Articles in this journal have used institution-level panels in the same way. Chikalipah (2017) studied the determinants of financial sustainability for 324 MFIs in 33 sub-Saharan African countries over 2003–2014 and found return on assets to be the main one; Chikalipah (2019)

used a MIX data set of 465 MFIs from 37 countries over 1996–2012 and found financial leverage to be negatively associated with profitability; Ranjani and Singh (2019) worked with an unbalanced panel of 211 Indian MFIs; and Alinsunurin (2014) measured the efficiency of Philippine MFIs with data envelopment analysis. Impact evaluations ask a different question, the effect of access to microcredit on households, and their syntheses report modest, non-transformative average effects (Banerjee et al., 2015; Meager, 2019). The data examined here describe lenders, not borrowers, and the questions of this study concern only institution-level data.

Selection, attrition and unbalanced panels. Rubin (1976) gave the conditions under which the process that causes missing data can be ignored: for sampling-distribution inference, the missing data must be missing at random and the observed data observed at random. Verbeek and Nijman (1992) noted that missing observations are the rule rather than the exception in panel data and proposed simple variable-addition and (quasi-)Hausman tests for selectivity bias; Wooldridge (1995) developed selection corrections for panel data models under a conditional mean independence assumption; and Baltagi and Song (2006) surveyed the econometrics of unbalanced panels. In finance, survivorship bias distorts performance studies based on time series (Brown et al., 1992), and mutual funds that disappear tend to be poor performers (Elton et al., 1996), so a sample restricted to survivors overstates average performance.

For microfinance, Bauchet and Morduch (2010) analysed 2,072 MFIs in the MixMarket and Microcredit Summit Campaign databases for 2004–2006 and found systematic biases in which survey institutions respond to and which indicators they report. Poverty-focused institutions were more likely to report to the Microcredit Summit Campaign and less likely to report to MixMarket, and by contrasting results on the relationship between outreach to women and operational self-sufficiency they showed how non-random reporting can influence analyses. A balanced panel adds a further condition to voluntary reporting: the institution must report, under the chosen rule, in every year of the panel. How large the additional selection is, and in which direction, is an empirical question. Comparing the covariate distributions of the retained and the excluded institutions is the natural diagnostic, in the spirit of the balance diagnostics used to compare treated and untreated groups in matched samples (Austin, 2009).

Bicliques, biclustering and subset transforms. A reporting matrix defines a bipartite graph that links each institution to the years in which it reports. A balanced panel is a biclique of this graph, and the largest panel in institution-years is a maximum edge biclique, a problem shown to be NP-complete by Peeters (2003). Dawande et al. (2001) studied bipartite and multipartite clique problems, Hochbaum (1998) the approximation of clique and biclique problems, and Ambühl et al. (2011) obtained a hardness-of-approximation result for maximum edge biclique under a standard complexity assumption. Related problems arise in biclustering, the search for submatrices whose rows and columns share a pattern, surveyed for gene expression data by Madeira and Oliveira (2004); maximal bicliques can be generated by consensus algorithms (Alexe et al., 2004), and in formal concept analysis a maximal all-ones submatrix is a formal concept, a set of objects together with exactly the attributes they share (Ganter & Wille, 1999).

The hardness results concern the general case, in which both sides of the graph can be large. In an annual reporting matrix one side, the years, is short, and functions defined on the subsets of a short side can be transformed for all subsets at once. Kennes (1992) presented fast algorithms for the Möbius transformation of the Boolean lattice, with applications to Dempster–Shafer theory, and Björklund et al. (2007) used the Möbius transform and its inversion to evaluate subset convolutions over a ground set of N elements with a number of operations of order N22N. The superset-sum transform used in this study is the zeta transform of this family, applied to the histogram of reporting patterns.

Reproducible computation and testing. Reproducibility, the possibility of repeating an analysis from the original data and code, has been proposed as a minimum standard for judging scientific claims when independent replication is not possible (Peng, 2011). Practical rules for reproducible computational research (Sandve et al., 2013) and good-enough practices in scientific computing (Wilson et al., 2017) stress recording every step, keeping data and code under version control and avoiding the manual manipulation of data. Property-based testing, in which a program is checked against stated properties on many randomly generated inputs rather than on a few examples, is the approach of QuickCheck for Haskell (Claessen & Hughes, 2000) and is available for Python through Hypothesis (MacIver et al., 2019). The pipeline described below uses NumPy (Harris et al., 2020), pandas (McKinney, 2010) and Matplotlib (Hunter, 2007).

The studies reviewed describe their samples by numbers of institutions, countries and years. The combinatorial literature supplies exact and approximate tools for the problem that underlies the choice of a balanced sample, and the selection literature explains why the retained sample matters. This study brings the strands together: it treats the choice of a balanced panel as an optimisation with an exact solution and treats the solution as a selection whose composition must be reported.

**Methodology**

Figure 1 shows the architecture of the pipeline. Each stage reads the output of the previous one and writes its own to files under version control; a verification layer checks the central computation against independent oracles, and every number in this article is read from the result files by the program that writes the manuscript. The design starts from the requirements in Table 1, each paired with the check that verifies it.

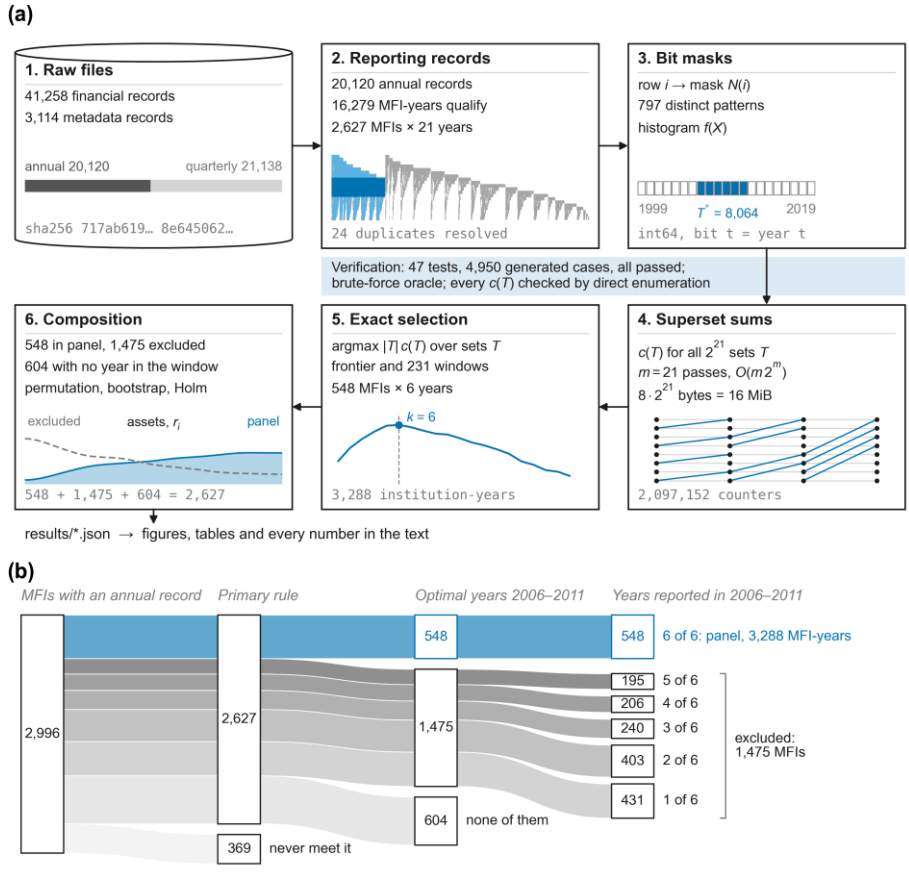


**Figure 1. Architecture of the extraction pipeline and flow of MFIs, primary rule.**

(a) The six stages run in the order of the arrows and write their outputs to version-controlled files. Each box gives the size of its output, read from the result files, and a sketch of what it holds: in Raw files (the cylinder), the annual (dark grey) and quarterly (light grey) records of the financial file and the SHA-256 digests of both files; in Reporting records, the reporting matrix with MFIs across and years down (dark blue: cells of the optimal panel; light blue: other years of its members; grey: other reporters); in Bit masks, the mask of the optimal years, one square per year, blue if the year is in T*, and its value as an integer; in Superset sums, the passes of Algorithm 1 for three years, each blue line an addition g(T) ← g(T) + g(T ∪ {j}) and each grey line a value carried over; in Exact selection, the institution-years k·c*(k) along the frontier, with the optimum; and in Composition, the distributions of the mean within-year percentile of assets in the panel (blue) and among the excluded MFIs (dashed). The shaded layer verifies the superset sums against independent oracles (Table 5), and the arrow at the bottom leads to the result files from which the figures, the tables and every number of the text are written. (b) Where the 2,996 MFIs with an annual record end up. From left to right: all of them; those that meet the primary rule in at least one year, with those that never do below; how these split between all, some and none of the 6 optimal years; and the excluded MFIs by the number of optimal years in which they report. Each node is a box with its number of MFIs, and its height, like the width of each band, is proportional to that number. Every band carries, from the left,

the colour of the group where its MFIs end up: blue for the panel, whose nodes are outlined in blue; grey from dark (all but one of the optimal years) to light (one year) for the excluded MFIs; and the palest greys for the MFIs that report in none of the optimal years, which belong to neither group, and for those that never meet the rule.

Data and system boundary. The input consists of two files of MIX Market data as distributed through the World Bank Data Catalog in December 2020 (MIX Market, 2020): the financial performance dataset in US dollars and the metadata file of the MFIs. The financial file has 41,258 rows, of which 20,120 are annual records (period type ANN) for 2,996 MFIs; the metadata file has 3,114 rows. Before either file is read, its SHA-256 digest is compared with the digest recorded when the file was obtained, so that every later stage refers to the same bytes. Only annual records are used, and the period of a record is the fiscal year that MIX assigns to it. The columns of the reporting matrix are the fiscal years 1999 to 2019 (m = 21), including years with few reporters. Table 2 gives the coverage of the nine variables used below and the values that lie outside their possible range, such as negative OSS or shares of women borrowers above 100%. These values are not corrected; the comparisons below use within-year ranks, in which an extreme value carries no more weight than any other value on the same side of the distribution.

**Table 1. Requirements of the extraction and their verification**

| ID | Requirement | Verification | Reported in |
|---|---|---|---|
| R1 | Balanced: every institution of the panel reports, under the stated rule, in every year of the panel. | The optimum is checked to be an all-ones submatrix of the reporting matrix. | Tables 5 and 7 |
| R2 | Maximal: no balanced panel under the same rule has more institution-years. | Lemma 1; comparison with direct enumeration on all $2^m$ year sets; brute force over all row and column subsets on small matrices. | Methodology; Tables 5 and 7 |
| R3 | Determinate: ties are counted and broken by a stated rule, so the result does not depend on the order of the data. | Count of maximising sets; tie-breaking and closure tested. | Tables 5, 7 and 11 |
| R4 | Traceable: the panel can be rebuilt from the raw files, and every number in the text comes from the result files. | SHA-256 check of the inputs; rules, plan and code under version control; fixed random seeds; manuscript generated from the result files. | Methodology |
| R5 | Bounded cost: time and memory are predictable from *n* and *m* and small for annual data. | Stage timings on the real data; scaling in *m* and *n*; operation counts. | Table 12; Figure 5 |
| R6 | Audited: the composition of the panel is compared with that of the MFIs it excludes. | Pre-specified comparison with permutation tests, bootstrap | Tables 13–16; Figures 6 and 7 |

| | | | |
|---|---|---|---|
| | | intervals and Holm's adjustment. | |

Note: R1–R3 concern the panel itself, R4 its provenance, R5 its cost and R6 what it selects.

**Table 2. Coverage and range of the variables in the annual records**

| Variable | With a value (%) | Zeros | Lowest | Highest | Outside the range |
|---|---|---|---|---|---|
| ***Required by the primary rule*** | | | | | |
| Assets (USD million) | 94.7 | 9 | 0.0 | 3,914,940 | 0 (0) |
| Gross loan portfolio (USD million) | 97.4 | 16 | 0.0 | 3,381,385 | 0 (0) |
| Active borrowers (thousand) | 91.1 | 5 | 0.0 | 8,935 | 0 (0) |
| Operational self-sufficiency (%) | 88.3 | 10 | −4,785 | 8,122 | 23 (2) |
| ***Added by the broad rule*** | | | | | |
| Return on assets (%) | 75.4 | 53 | −746.4 | 208.9 | 38 (1) |
| Portfolio at risk > 30 days (%) | 75.0 | 666 | 0.0 | 711.4 | 9 (3) |
| Women borrowers (%) | 73.5 | 57 | 0.0 | 668.9 | 25 (2) |
| ***Other covariates*** | | | | | |
| Average loan balance / GNI per capita (%) | 88.2 | 21 | 0.0 | 382,753 | 0 (0) |
| Operating expense / loan portfolio (%) | 75.7 | 21 | −133.3 | 2,652 | 4 (0) |

Note: All 20,120 annual records of the financial file, whether or not they satisfy a reporting rule; values in the units of the first column. Zeros: values equal to zero. Outside the range: values below zero, or below −100% for return on assets, or above 100% for portfolio at risk and women borrowers; in parentheses, those within the 3,288 institution-years of the optimal panel (2006–2011). No value is corrected. Post hoc (descriptive): added after the analysis plan was fixed.

Reporting matrix and rules. A reporting rule is a pair (V, τ): the set V of variables that an annual record must contain and the map τ that assigns each record r its period. With Ri the annual records of institution i and D(r) the variables present in r, the rule defines the reporting matrix A(V, τ) ∈ {0,1}n×m, Equation (1):

$$a_{it}(V,\tau) = \mathbf{1}[\,\exists\, r \in R_i : \tau(r) = t,\ V \subseteq D(r)\,],$$
$$V \subseteq V' \ \Rightarrow\ A(V',\tau) \le A(V,\tau)$$
$$\Rightarrow\ \mathrm{OPT}(V',\tau) \le \mathrm{OPT}(V,\tau)$$

where matrices are compared cell by cell: requiring a further variable can only remove ones, and so can only lower the optimum OPT defined below. Under the primary rule V holds total assets, gross loan portfolio, number of active borrowers and OSS, which cover the size, outreach and sustainability dimensions used throughout the studies reviewed above, and τ(r) is the fiscal year of r. When an institution has several annual records for the same period (24 cases

under the primary rule), a record that satisfies the rule is kept, and among several such records the one with the latest as-of date. Three alternative rules serve as sensitivity checks: any annual record counts (any record, V = ∅); the primary variables plus return on assets (ROA), portfolio at risk over 30 days (PAR30) and the percentage of women borrowers must be present (broad); and the primary variables must be present, with τ(r) the calendar year of the record's as-of date (as-of year). The any-record, primary and broad rules thus share τ and nest their V, whereas the as-of rule changes τ. Institutions without any qualifying year are not rows of the matrix. Table 3 summarises the four matrices, and Table 4 lists the notation used below.

**Table 3. Reporting rules and the reporting matrices they produce (fiscal years 1999–2019)**

| Rule: a year counts when the annual record has | MFIs (*n*) | Patterns | Institution-years | Filled cells |
|---|---|---|---|---|
| **Primary:** assets, gross loan portfolio, active borrowers and OSS (fiscal year) | 2,627 | 797 | 16,279 | 29.5% |
| **Any record:** any content (fiscal year) | 2,996 | 676 | 20,096 | 31.9% |
| **Broad:** the primary variables plus ROA, PAR30 and women borrowers (fiscal year) | 2,173 | 814 | 11,532 | 25.3% |
| **As-of year:** the primary variables (calendar year of the as-of date) | 2,627 | 836 | 16,271 | 29.5% |

Note: MFIs are the institutions with at least one reported year (rows of A); patterns are the distinct rows; filled cells are the institution-years over n × m, with m = 21.

**Table 4. Notation**

| Symbol | Meaning | Domain | Eq. |
|---|---|---|---|
| ***Data and problem*** | | | |
| $R_i$, $D(r)$ | Annual records of institution *i*; variables present in record *r* | Sets of records and of variables | 1 |
| $(V, \tau)$ | Reporting rule: variables that a record must contain, and the period $\tau(r)$ of record *r* | $\tau(r) \in Y$ | 1 |
| $a_{it}$, $A$ | Whether institution *i* reports in year *t* under the rule; the reporting matrix | $\{0,1\}$; $\{0,1\}^{n\times m}$ | 1 |
| $\mathbf{1}[\cdot]$ | Indicator: one when the condition holds, zero otherwise | $\{0,1\}$ | 1 |
| OPT | Largest number of institution-years of a balanced panel | $\{0, \ldots, nm\}$ | 1 |
| $\lvert\cdot\rvert$, $\#\{\cdot\}$ | Number of elements of a set | $\{0, 1, \ldots\}$ | 2 |
| $I$, $n$ | Institutions with at least one qualifying year; their number | $n = \lvert I\rvert$ | 2 |
| $Y$, $m$ | Years, numbered 0, …, $m-1$ from the earliest; their number | $m = \lvert Y\rvert$ | 2 |

| | | | |
|---|---|---|---|
| $(S, T)$ | Balanced panel: every institution of $S$ reports in every year of $T$ | $S \subseteq I$, $T \subseteq Y$ | 2 |
| $N(i)$ | Years in which institution $i$ reports, stored as an $m$-bit mask | $N(i) \subseteq Y$ | 3 |
| $S(T)$, $c(T)$ | Institutions that report in every year of $T$; their number | $S(T) \subseteq I$; $\{0, \ldots, n\}$ | 3 |
| $T^*$, $S^*$ | Optimal year set and its institutions, $S^* = S(T^*)$ | $T^* \subseteq Y$, $S^* \subseteq I$ | 9 |
| $k$ | Number of years of a panel, $\lvert T\rvert$ | $\{1, \ldots, m\}$ | 8 |
| $c^*(k)$ | Frontier: largest $c(T)$ over the sets of $k$ years | $\{0, \ldots, n\}$ | – |
| $T_k$, $S_k$, $E_k$ | Best set of $k$ years, its institutions and the institutions excluded from it | As $T^*$, $S^*$ and $E$ | – |
| $\gamma$, $c_\gamma(T)$, $\mathrm{OPT}_\gamma$ | Region or legal status; $c(T)$ and OPT among its institutions | As $c(T)$ and OPT | 16 |
| ***Algorithm and cost*** | | | |
| $X$, $f(X)$ | Reporting pattern; number of institutions whose mask equals $X$ | $X \subseteq Y$; $\{0, \ldots, n\}$ | 4 |
| $j$, $g_j$ | A year, or the bit that marks it; table after $j$ passes, one entry per year set | $\{0, \ldots, m-1\}$; $2^m$ integers | 5 |
| $d$, $d_m$ | Distinct reporting patterns; their number on synthetic instances with $m$ years | $1 \le d \le \min(n, 2^m)$ | 7 |
| $W_{tr}$, $W_{en}$ | Additions of the transform; containment tests of direct enumeration | Positive integers | 7 |
| $G_k$, $\hat{T}$ | Year set of $k$ years left by backward elimination; best set it visits | $G_k \subseteq Y$, $\hat{T} \subseteq Y$ | 8 |
| $t_{tr}$, $t_{en}$ | Time of the exact method (histogram and transform) and of direct enumeration | Seconds | 18 |
| $\kappa_{tr}$, $\kappa_{en}$ | Time per addition, histogram included, and per containment test | Nanoseconds | 18 |
| $R$ | Timed repetitions of a stage after one warm-up run | Positive integers | – |
| ***Composition analysis*** | | | |
| $h_i$ | Years of $T^*$ in which institution $i$ reports | $\{0, \ldots, \lvert T^*\rvert\}$ | 9 |
| $E$ | Excluded institutions: they report in some but not all years of $T^*$ | $E \subseteq I$ | 9 |
| $v$, $x_{it}$ | Continuous covariate; its value for institution $i$ in year $t$ | Real numbers | 10 |
| $P_t$ | Institutions that report in year $t$ and have a value of $v$ | $P_t \subseteq I$ | 10 |
| $\rho_{it}$, $q_{it}$ | Mid-rank of $x_{it}$ within $P_t$; within-year percentile | $[1, \lvert P_t\rvert]$; (0, 100) | 10 |
| $\varphi_t$ | Any strictly increasing transformation of the values of year $t$ | Real functions | 10 |
| $Y_i$, $\bar{r}_i$ | Years of $T^*$ in which $i$ has a value of $v$; mean of its percentiles | $Y_i \subseteq T^*$; (0, 100) | 10 |

| | | | |
|---|---|---|---|
| $S^*_v$, $E_v$ | Institutions of $S^*$ and of $E$ with a value of $v$ in $T^*$ | $S^*_v \subseteq S^*$, $E_v \subseteq E$ | 11 |
| $\Delta_v$ | Difference in mean percentile, panel minus excluded | (−100, 100) points | 11 |
| $B$, $b$ | Number of random permutations of the panel and excluded labels; their index | $B$ = 10,000 | 12 |
| $\Delta^{(b)}$, $\Delta_{obs}$ | $\Delta_v$ after permutation $b$; observed $\Delta_v$ | As $\Delta_v$ | 12 |
| $p$, se($p$) | Permutation $p$-value; its Monte Carlo standard error | [1/($B$ + 1), 1] | 12 |
| $B'$, $\tilde{\Delta}^{(b)}_v$ | Bootstrap resamples, drawn with replacement within $S^*_v$ and within $E_v$; $\Delta_v$ on resample $b$ | $B'$ = 10,000 | 13 |
| $Q_a$, $CI_{95\%}$ | $a$-quantile over the resamples; 95% confidence interval | – | 13 |
| $\ell$, $\pi^S_\ell$, $\pi^E_\ell$ | Category of a categorical variable; its share in $S^*$ and in $E$ | $\pi \in [0, 1]$ | 14 |
| TVD, $p_{TVD}$ | Total variation distance; its permutation $p$-value | TVD ∈ [0, 1] | 14 |
| $TVD^{(b)}$, $TVD_{obs}$ | TVD after permutation $b$; observed TVD | TVD ∈ [0, 1] | 14 |
| $(x)^+$, $\mathcal{K}$ | max($x$, 0); a set of categories | – | 14 |
| $L$, $p_{(j)}$ | Tests in the confirmatory family; the $j$-th smallest of their $p$-values | $L$ = 11 | 15 |
| $p^{Holm}_{(r)}$ | Holm-adjusted $p$-value of the test of rank $r$ | [0, 1] | 15 |
| α | Significance level of the adjusted $p$-values | α = 0.05 | – |
| $S^*_{v,t}$, $E_{v,t}$ | Institutions of $S^*_v$ and of $E_v$ with a value of $v$ in year $t$ | $S^*_{v,t} \subseteq S^*_v$, $E_{v,t} \subseteq E_v$ | 20 |
| $\Delta_{v,t}$, $AUC_{v,t}$ | $\Delta_v$ within year $t$; chance that a panel member has the larger value, ties counting one half | [−50, 50]; [0, 1] | 20 |
| $\psi_{ii't}$ | Score of a panel member $i$ against an excluded MFI $i'$ in year $t$: one if $i$ has the larger value, one half for a tie | {0, ½, 1} | 20 |
| μ, $E_{\mu,v}$ | Reason for exclusion (late entry, early exit, both, gap); institutions of $E_v$ excluded for it | $E_{\mu,v} \subseteq E_v$ | 21 |
| $\bar{r}(G)$ | Mean of $\bar{r}_i$ over a group $G$ of institutions | (0, 100) | 21 |

Note: Eq.: the equation in which the symbol first appears; – marks a symbol defined only in the text. Subscripts tr and en refer to the superset-sum transform and to direct enumeration. Years are also numbered 0, …, m − 1, and a set of years is also written as a mask with bit t for year t. Problem statement. Let I be the set of institutions, Y the set of years and N(i) = {t ∈ Y : ait = 1} the years in which institution i reports. A balanced panel is a pair (S, T), with S ⊆ I and T ⊆ Y, in which every institution of S reports in every year of T. The largest balanced panel maximises the number of institution-years, Equation (2):

$$\max_{S \subseteq I,\ T \subseteq Y} |S|\,|T| \quad \text{subject to } a_{it} = 1 \text{ for all } i \in S,\ t \in T \tag{2}$$

This is the maximum edge biclique problem on the bipartite graph whose edges are the pairs {i, t} with ait = 1. The objective counts institution-years because that is the size of the data set an analysis receives; maximising |S| for a given number of years, or |T| for a given number of institutions, are points on the frontier described below. The problem is hard in general, but one observation reduces it to counting.

Lemma 1. For a fixed year set T, the largest admissible set of institutions is S(T), the set of institutions whose reporting years include T. Any admissible S is contained in S(T), and S(T) is itself admissible, which proves the lemma. Writing c(T) = |S(T)|, the optimum follows from Equation (3):

$$S(T) = \{\, i \in I : N(i) \supseteq T \,\}, \qquad \mathrm{OPT} = \max_{\emptyset \neq T \subseteq Y} |T|\, c(T) \tag{3}$$

The problem thus reduces to computing c(T) for every non-empty T ⊆ Y. With m years there are 2m − 1 such sets (2,097,151 for m = 21); listing them is cheap, and the cost lies in counting, for each T, the institutions whose reporting set contains it.

Superset-sum transform. Each row of A is encoded as an m-bit mask N(i), with bit t set when institution i reports in year t, and f(X) denotes the number of institutions whose mask equals X, the histogram of reporting patterns. Then c(T) is the sum of f over the supersets of T, Equation (4):

$$f(X) = |\{\, i \in I : N(i) = X \,\}|, \qquad c(T) = \sum_{X \supseteq T} f\,(X) \tag{4}$$

The superset-sum, or zeta, transform computes this sum for all T at once. Starting from g0 = f, it processes one bit j at a time, Equation (5):

$$g_{j+1}(T) = \begin{cases} g_j(T) + g_j(T \cup \{j\}) & \text{if } j \notin T, \\ g_j(T) & \text{if } j \in T. \end{cases} \tag{5}$$

After j passes the table holds the partial sums of Equation (6):

$$g_j(T) = \sum_{X \supseteq T,\ X \setminus T \subseteq \{0,\ldots,j-1\}} f\,(X), \qquad g_0 = f, \qquad g_m = c \tag{6}$$

Each pass performs 2m−1 additions on one table of 2m 64-bit integers, and the masks and the histogram take O(nm) time, whereas direct enumeration tests every T against each of the d distinct reporting patterns. Equation (7) counts the operations of each and gives the cost of the exact method:

$$W_{\mathrm{tr}} = m\,2^{m-1}, \qquad W_{\mathrm{en}} = d\,2^{m}, \qquad \frac{W_{\mathrm{en}}}{W_{\mathrm{tr}}} = \frac{2d}{m}, \qquad \text{time} = O(nm + m\,2^{m}), \qquad \text{memory} = 8 \cdot 2^{m} \text{ bytes} \tag{7}$$

For the primary rule, with m = 21 and d = 797, Wtr = 22,020,096 and Wen = 1,671,430,144, a ratio of 75.9, and the table takes 16 MiB. Algorithm 1 lists the procedure. In the implementation

each pass is one vectorised NumPy operation: the table is viewed as an array of shape (2m−j−1, 2, 2j), whose middle index is bit j, and the slice with the bit set is added to the slice without it. Figure 2 follows Algorithm 1 on a matrix of 5 institutions and 3 years, one of the hand-built cases of the tests. Its histogram has two non-zero entries, f({0, 2}) = 4 and f({0, 1, 2}) = 1, and after 3 passes of Equation (5) the table holds c(T) for all 8 sets. The optimum is T = {0, 2}, with all 5 institutions and 10 institution-years. It is not a window of consecutive years: the best windows are the single years 0 and 2, with 5 institution-years each.

Ties and closure. When several year sets reach the maximum, the one with more institutions is chosen, and then the one with the smaller mask; the number of maximising sets is reported with every result. An optimal pair (S(T*), T*) is closed: the years shared by all institutions of S(T*) are exactly T*, because a shared year outside T* could be added without losing any institution and would increase the number of institution-years. In the terms of formal concept analysis the optimum is a formal concept (Ganter & Wille, 1999). Closure is a necessary property of any optimum, and it is checked on every result as a guard against implementation errors.

Algorithm 1. Largest balanced panel by superset sums

**Input:** reporting matrix $A \in \{0,1\}^{n\times m}$. **Output:** an optimal pair $(S^*, T^*)$ and $c(T)$ for every $T$.

1 **for each** institution $i$: $N(i) \leftarrow \Sigma_t\, a_{it}\cdot 2^t$ (mask of reported years)
2 $f \leftarrow$ array of $2^m$ zeros; **for each** $i$: $f[N(i)] \leftarrow f[N(i)] + 1$
3 $g \leftarrow f$
4 **for** $j = 0, 1, \ldots, m-1$:
5 **for each** $T$ with $j \notin T$: $g[T] \leftarrow g[T] + g[T \cup \{j\}]$
6 now $g[T] = c(T)$ for every $T \subseteq \{0, \ldots, m-1\}$
7 $T^* \leftarrow$ argmax over non-empty $T$ of $|T|\cdot g[T]$; ties: larger $g[T]$, then smaller mask
8 $S^* \leftarrow \{i : N(i) \cap T^* = T^*\}$
9 **return** $(S^*, T^*)$ and $g$

Note: Years are numbered 0, …, m − 1 from the earliest; bit t of a mask marks year t, and ∪ and ∩ act on masks as bitwise OR and AND.

Variants and baselines. Three further outputs are read from the same table. The frontier c*(k) = max|T|=k c(T), for k = 1, …, m, gives the largest number of institutions that can be kept in a panel of k years, and k·c*(k) is the best number of institution-years for that length; it is computed over all year sets and over contiguous windows only. The contiguous optimum is the best of the m(m + 1)/2 = 231 windows of consecutive years, the natural choice for a panel picked by hand. The greedy heuristic is backward elimination, Equation (8), with ties broken as above: starting from all years, it removes at each step the year whose removal keeps the most institutions.

$$G_m = Y, \qquad G_{k-1} = \arg\max_{T\subset G_k,\ |T|=k-1} c(T), \qquad \hat{T} = \arg\max_{1\le k\le m}\ k\, c(G_k) \tag{8}$$

Both baselines read the exact counts c(T), so they differ from the exact method only in the sets they examine; their results are reported as fractions of the optimum.

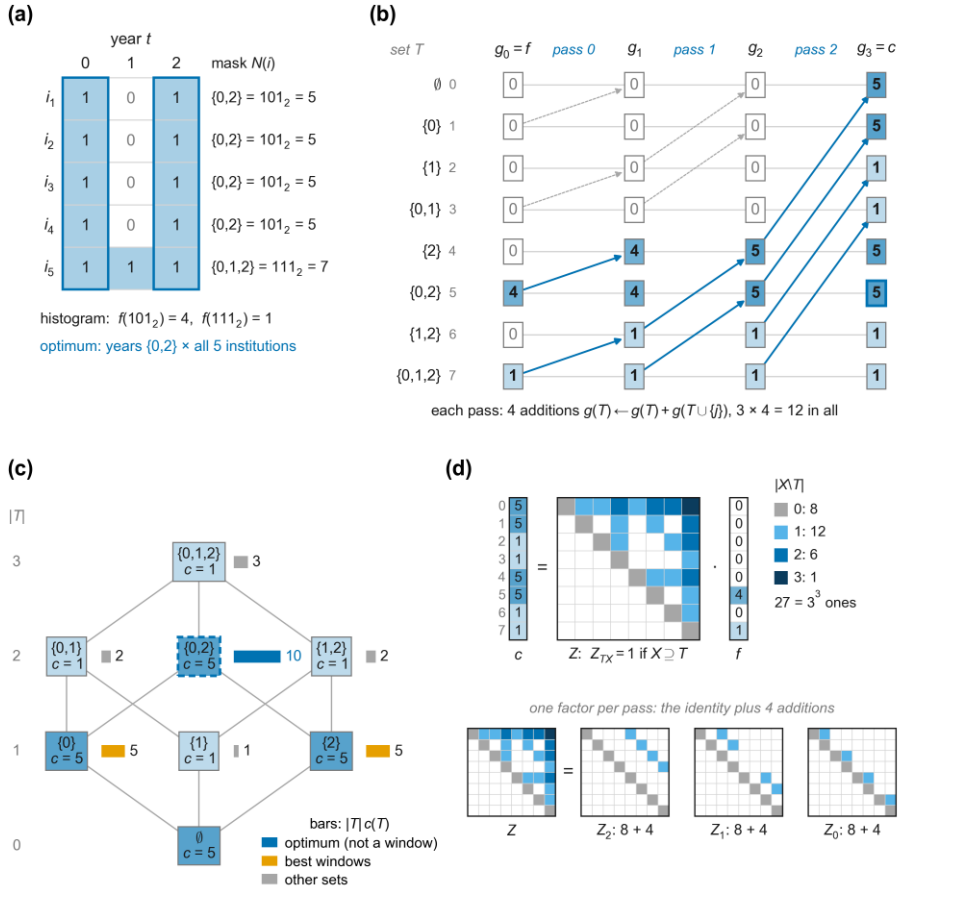


**Figure 2. Worked example of Algorithm 1 on a hand-built matrix with n = 5 institutions and m = 3 years.**

(a) Reporting matrix (shaded cells are reports) and the mask N(i) of each institution, as a set of years, in binary and as an integer; the outlined columns are the years of the optimum, given below with the histogram f of the masks. (b) The table g for each of the 8 sets of years T, from g0 = f to g3 = c. Each arrow is one addition of Equation (5), g(T) ← g(T) + g(T ∪ {j}): solid blue arrows add a non-zero value and change the cell they point to, dashed grey arrows add zero, and grey lines carry the other values over. Each cell is shaded by its value, white for zero and a darker blue for a larger count, so that the counts of f can be followed down to the subsets; non-zero values are in bold, and the cell outlined in blue is the optimum. (c) The lattice of the sets of years by |T|, with lines between sets that differ by one year, c(T) in each box, shaded as in (b), and a bar for the institution-years |T|·c(T) beside it: the optimum, T = {0, 2} (blue, dashed box), gives 10 and is not a window of consecutive years; the best windows, the single years 0 and 2 (orange), give 5 each; other sets are grey. (d) The transform as a matrix product, c = Zf, with ZTX = 1 if X ⊇ T. The vectors c and f, the last and first columns of (b), are shaded as there; each 1 in Z is shaded by the number of years that X adds to T, |X \ T|, and the key counts the ones of each shade. Below, Z as the product of one matrix per pass, the identity plus the additions of that pass.

Verification. The implementation is verified at three levels, listed check by check in Table 5. Property-based tests written with Hypothesis generate random binary matrices and compare the transform, the optimum and the baselines with independent references, among them direct enumeration and a brute-force search over all subsets of rows and columns. Hand-built matrices pin a case in which a year set with a gap beats every window (the matrix of Figure 2), a case in which the greedy heuristic fails, and the tie-breaking rule. On the real data, the transform is compared with direct enumeration on all 2m = 2,097,152 year sets of each rule, and the optimum is checked to be an all-ones submatrix and closed. The statistical routines, and the post hoc routines described below, have tests of their own. Hypothesis runs derandomised, so every run of the suite generates the same cases, and all 47 tests pass.

**Table 5. Verification of the implementation**

| Check | Reference | Inputs | Cases |
|---|---|---|---|
| ***Random matrices (property-based tests)*** | | | |
| Superset sums equal direct enumeration of $c(T)$ | Enumeration over the masks | Random, up to 60 × 10 | 300 |
| After each pass, the table holds the partial sums of Equation (6) | Sums over the supersets that agree above bit $j$ | Random, up to 40 × 7 | 200 |
| $c(T)$ counts the rows of $A$ that contain $T$ | Columns of $A$ checked one by one | Random, up to 60 × 8 | 300 |
| The optimum is the largest all-ones submatrix | Brute force over all row and column subsets | Random, up to 7 × 6 | 400 |
| The reported panel is an all-ones submatrix of the stated size | Direct check on $A$ | Random, up to 40 × 9 | 300 |
| The optimal year set is closed | Intersection of the masks of its institutions | Random, up to 40 × 9 | 300 |
| $c(T)$ does not increase when a year is added to $T$ | Pairs $T$ and $T \cup \{j\}$ | Random, up to 40 × 9 | 200 |
| The frontier does not increase with $k$; the best window and the greedy result never exceed the optimum | Order relations between the outputs | Random, up to 40 × 9 | 200 |
| Number of years in a mask | Count of ones in the binary string | Random integers | 100 |
| ***Hand-built matrices*** | | | |
| A year set with a gap beats every window; the table after each pass | Worked by hand (Figure 2) | 5 × 3 | – |
| Backward elimination stops short of the optimum | Worked by hand | Two blocks of rows | – |
| Ties go to the set with more institutions, then to the smaller mask | Worked by hand | Two tied year sets | – |

| | | | |
|---|---|---|---|
| Bit *t* of a mask is year *t*; brute force on a matrix without ones | Worked by hand | Fixed matrices | – |
| ***Real data, four reporting rules*** | | | |
| Superset sums equal direct enumeration; each optimum is an all-ones submatrix and closed | Enumeration over the reporting patterns; direct check on *A* | All $2^{21}$ year sets of each rule | 8,388,608 |
| ***Statistical routines*** | | | |
| Within-year percentiles with ties and missing values | Worked by hand | Fixed case | – |
| Within-year percentiles average 50 and do not change under a strictly increasing transformation of a year's values | Direct check | Random samples with ties | 200 |
| Holm's adjustment | Worked by hand; never lowers a *p*-value or raises it above one | Random lists of *p*-values | 100 |
| Permutation *p*-value | Exact distribution of a small case; a *p*-value of one without a difference; range of *p* | Random samples | 50 |
| Bootstrap interval | Contains the observed difference | Simulated groups | – |
| TVD and its permutation test | Worked by hand | Fixed cases | – |
| ***Post hoc routines*** | | | |
| Reason for exclusion | Regular expressions over the reporting pattern; worked by hand | Random masks and year sets | 300 |
| Yearly flows | Accounting identities; worked by hand | Random masks | 200 |
| Closed year sets | Closure of each year set | Random matrices | 200 |
| Runs of consecutive years | Groups of ones in the binary string; worked by hand | Random masks | 100 |
| Optimum within a group and the part of it that the optimal years keep | Order relations; direct count | Random masks | 200 |
| Values outside the range | Worked by hand | Fixed case | – |
| Groups ordered by size, equal sizes by name | Worked by hand | Fixed case | – |
| Monte Carlo standard error of a *p*-value and Holm's multipliers; the range of an adjusted *p*-value contains it | Worked by hand; Holm's adjustment recomputed at both ends of the range | Random lists of *p*-values | 200 |

| | | | |
|---|---|---|---|
| The positive parts of a TVD add up to it and to the largest difference over sets of categories; shares without one category | Half the sum of absolute differences; every set of categories; worked by hand | Random distributions | 200 |
| Yearly difference in mean percentile equals 100 (AUC − ½) | All pairs of raw values; a year without one of the groups is rejected | Random samples with ties | 200 |
| The contributions of the reasons for exclusion add up to the difference | Difference of the group means | Random groups | 200 |
| *c*(*T*) is the sum of the group counts, and the optimum at most the sum of the optima within groups | Optimum of each group | Random matrices and groups | 200 |
| Thresholds for adding or removing a year; no one-year change from the optimum reaches them | Smallest count that raises the area; every one-year change | Random counts and matrices | 300 |
| Scaling exponent implied by an operation count | Worked by hand | Fixed grids | – |

Note: Inputs: the largest random matrix (rows × columns) or the kind of input. Cases: inputs generated by Hypothesis in one derandomised run of the suite; – marks rows made only of fixed, hand-built cases, which are not counted in rows that have both. On the real data, the cases are the year sets compared, 221 for each of the four rules. All 47 tests pass with pytest 6.2.1 and Hypothesis 5.43.8.

Composition analysis. The composition analysis was written as a plan and fixed under version control before any statistic on the covariates or the metadata was computed; at that point only the engineering results (matrix sizes, optima and frontiers) had been examined. Let T* be the optimal year set under the primary rule. Counting for each institution the years of T* in which it reports, Equation (9) defines the panel S* = S(T*) and the comparison group E: the MFIs that were active in the panel’s years but did not report in all of them. Institutions with hi = 0 belong to neither group.

$$h_i = |N(i) \cap T^*|, \qquad S^* = \{\, i \in I : h_i = |T^*| \,\}, \\ E = \{\, i \in I : 0 < h_i < |T^*| \,\} \tag{9}$$

Eight continuous covariates are compared: total assets, number of active borrowers, average loan balance per borrower relative to gross national income (GNI) per capita, percentage of women borrowers, OSS, ROA, PAR30 and operating expense relative to the loan portfolio. To compare institutions across years without modelling trends in nominal values, each value is converted into a percentile within its year, Equation (10):

$$q_{it} = \frac{100\,(\rho_{it} - 0.5)}{|P_t|}, \qquad \bar{r}_i = \frac{1}{|Y_i|}\sum_{t \in Y_i} q_{it}\,,$$
$$\frac{1}{|P_t|}\sum_{i \in P_t} q_{it} = 50, \qquad q_{it}(\varphi_t(x_{\cdot t})) = q_{it}(x_{\cdot t}) \tag{10}$$

where ρit is the mid-rank of xit among the institutions Pt that report in year t with a value of the covariate, Yi the years of T* in which i has a value, and φt any strictly increasing transformation of the values x·t of year t, such as a change of currency or of price level. An institution without any value in T* is left out of the comparison for that covariate. The statistic is the difference between the group means, in percentile points, Equation (11):

$$\Delta_v = \frac{1}{|S_v^*|}\sum_{i \in S_v^*} \bar{r}_i \;-\; \frac{1}{|E_v|}\sum_{i \in E_v} \bar{r}_i \tag{11}$$

Its p-value comes from a two-sided permutation test (Ernst, 2004) that reassigns the panel and excluded labels among institutions B = 10,000 times, Equation (12), which also gives the Monte Carlo standard error of p:

$$p = \frac{1 + \#\{\, b \le B : |\Delta^{(b)}| \ge |\Delta_{\text{obs}}| \,\}}{B + 1}, \qquad \text{se}(p) = \sqrt{p\,(1-p)/B} \tag{12}$$

Under the null hypothesis, membership of the panel is independent of the institution's mean percentile: the panel behaves like a random subset, of its size, of the institutions that report in its years. The 95% confidence interval is a percentile bootstrap of institutions stratified by group (Efron & Tibshirani, 1993), Equation (13), with B′ = 10,000 resamples:

$$\text{CI}_{95\%}(\Delta_v) = [\, Q_{0.025},\; Q_{0.975}\,], \qquad Q_a = a\text{-quantile of } \tilde{\Delta}_v^{(1)}, \ldots, \tilde{\Delta}_v^{(B')} \tag{13}$$

For the three categorical variables of the metadata (region, last known legal status and last known profit status, with a missing value, or a status recorded as 'none', treated as the category Unknown) the statistic is the total variation distance (TVD) between the two distributions, Equation (14):

$$\text{TVD} = 1/2 \sum_{\ell} \left|\pi_\ell^S - \pi_\ell^E\right| = \sum_{\ell} \left(\pi_\ell^S - \pi_\ell^E\right)^+ = \max_{\mathcal{K}} \sum_{\ell \in \mathcal{K}} \left(\pi_\ell^S - \pi_\ell^E\right),$$
$$p_{\text{TVD}} = \frac{1 + \#\{\, b \le B : \text{TVD}^{(b)} \ge \text{TVD}_{\text{obs}} \,\}}{B + 1} \tag{14}$$

where πSℓ and πEℓ are the shares of category ℓ in the panel and in the excluded group. Because both sets of shares add up to one, the TVD is also the largest gap in combined share over sets of categories $\mathcal{K}$, reached by those more common in the panel. Its test uses the same permutations as Δv but, since the TVD cannot be negative, only the upper tail. The eight continuous and three categorical comparisons form a confirmatory family of L = 11 tests,

adjusted with the procedure of Holm (1979), Equation (15), where p(1) ≤ … ≤ p(L) are the ordered p-values; a difference is significant when its adjusted p-value is below α = 0.05.

$$p_{(r)}^{\mathrm{Holm}} = \max_{j \le r} \min\{1,\ (L - j + 1)\, p_{(j)}\}, \quad r = 1, \dots, L \tag{15}$$

The plan recorded an expectation rather than a directional hypothesis: in view of the reporting-bias and survivorship literature, the panel was expected to contain larger and more self-sufficient institutions.

Two secondary analyses are descriptive and not adjusted for multiplicity. Along the frontier, for every k = 2, …, 20, the comparison of the eight continuous covariates is repeated with Tk = argmax|T|=k c(T) under the same tie-breaking rule, Sk = S(Tk) and Ek defined as above, using 2,000 bootstrap resamples and only lengths at which both groups have at least 20 institutions. Under each alternative reporting rule, the optimum and the confirmatory comparisons are recomputed without Holm's adjustment. Random numbers come from NumPy's generator seeded with 2020, with one independent stream per analysis.

Post hoc descriptive analyses. After the confirmatory results had been computed, descriptive analyses that the plan did not include were added to help interpret them: the coverage and range of each variable, the reporting flows by year, the cost of each variable that the primary rule requires, the optimum within regions, legal statuses and the countries with the most MFIs, the closed year sets and the overlap of the four optimal panels, the countries and composition of the best panel of each length, the quartiles of the raw covariates, why the excluded MFIs are excluded and what each reason adds to each difference, the yearly differences, the Monte Carlo error of each p-value, the comparison among MFIs of known status, the bound set by the group optima, one-year changes to the optimal years and the cost per operation. They involve no statistical test and leave every pre-specified result unchanged; the notes of the tables that report them say so, and the text marks the others.

Cost measurement and software. Each stage is timed on one thread of an eight-core, sixteen-thread x86-64 desktop processor with 16 GB of memory, as wall-clock time from a high-resolution counter. Each stage runs once to warm up and then R times (R = 20 for the stages after reading, 3 for reading and hashing, and 5 for direct enumeration); the median and the interquartile range are reported. Scaling in m is measured on synthetic instances with n = 2,627 uniformly random masks for m = 8, 10, …, 24 (R = 7), with direct enumeration up to m = 20 (R = 3); scaling in n uses random masks with m = 21 and n from 103 to 106 (R = 7). The software is Python 3.9.1 with NumPy 1.19.4 (Harris et al., 2020), pandas 1.2.0 (McKinney, 2010), openpyxl 3.0.5, Matplotlib 3.3.3 (Hunter, 2007), Hypothesis 5.43.8 and pytest 6.2.1.

**Results**

Reporting structure. Figure 3 shows the reporting matrix under the primary rule. The number of reporting MFIs rises from 122 in 1999 to a peak of 1,237 in 2008 and falls to 566 in 2018; the 2019 column holds only 2 institutions. Reporting is short-lived for many institutions: 366 MFIs report in a single year and 393 in two, while 22 report in 20 of the 21 years and none in

all of them. The 2,627 rows collapse into 797 distinct reporting patterns, and 29.5% of the cells are filled (Table 3). When any annual record counts, the matrix gains 369 institutions and fills 31.9% of its cells.

Entry and exit. Table 6 follows the reporting population year by year, separating new reporters, returns after a gap, last reports and temporary absences. Leaving aside 2018, after which only 2 MFIs report, the year with the most last reports is 2011, the final year of the optimal panel (233 MFIs), and temporary absences peak in the year after (220 MFIs in 2012). A panel extended by one year would keep 451 MFIs if the added year were 2005 and 390 if it were 2012, the panel members that report in that year. Gaps are common: 736 of the 2,627 MFIs, and 183 of the panel members, miss at least one year between two reported years.

Exact optimum. Table 7 gives the exact optimum for each rule. Under the primary rule the optimum is unique: the 6 fiscal years 2006–2011 and the 548 MFIs that report the core variables in all of them, for 3,288 institution-years. The panel holds 20.2% of all institution-years reported under the rule and 20.9% of the institutions. The broad and as-of rules give the same years, with 327 and 541 MFIs; when any annual record counts, the optimum is the 7-year window 2005–2011 with 648 MFIs. Every optimum is the unique maximiser, is closed and coincides with the result of direct enumeration over all 2,097,152 year sets.

The best contiguous window coincides with the exact optimum under all four rules. The greedy heuristic reaches the optimum under the primary and as-of rules but not under the other two: when any record counts it stops at 2008–2015 with 483 MFIs (3,864 institution-years, 85.2% of the optimum), and under the broad rule at 2009–2011 with 548 MFIs (1,644 institution-years, 83.8%). Under the broad rule it trades years for institutions, and under the any-record rule institutions for years, losing institution-years in both cases. The four optimal panels overlap without coinciding: every MFI of the broad panel is also in the primary one, as Equation (1) requires of two nested rules with the same optimal years, and 326 MFIs are in all four panels, out of 702 in at least one.

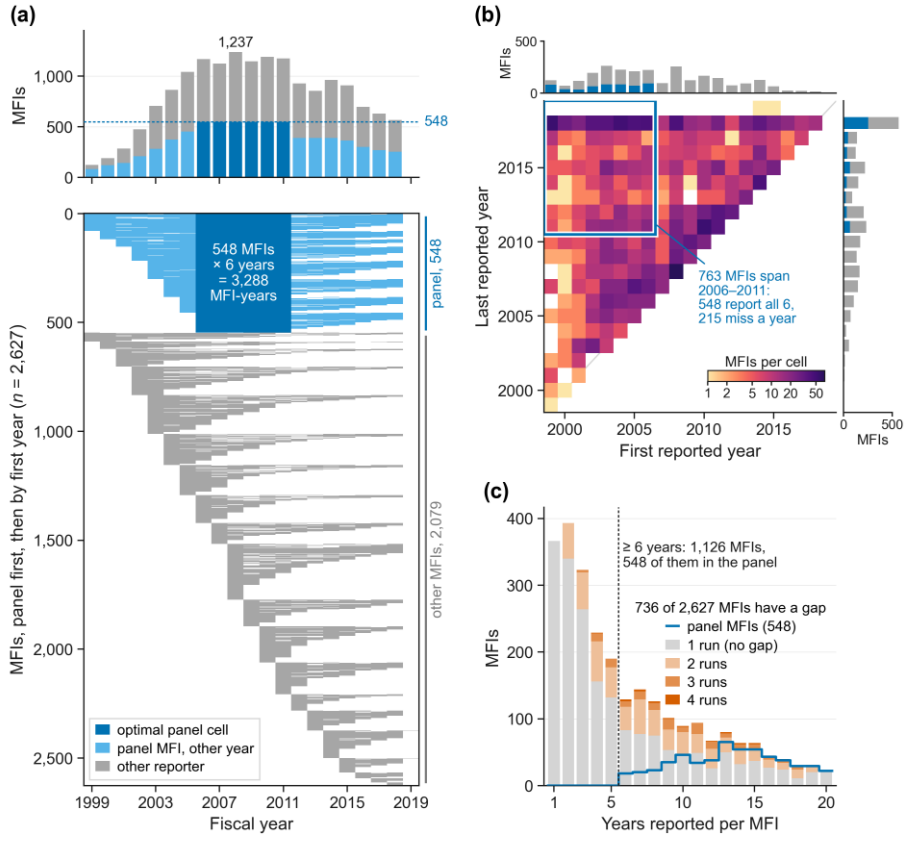


**Figure 3. Reporting structure under the primary rule.**

(a) Reporting matrix, one row per MFI (n = 2,627), panel members first and then by first reported year: dark blue cells form the optimal balanced panel (548 MFIs × 6 years), light blue cells are other years of panel members and grey cells other reporters; the bars at the right mark the two groups of rows. Above, the column sums of the matrix: the MFIs that report in each fiscal year, stacked in the colours of their cells, with the largest total labelled; the dotted line marks the 548 MFIs that report in all the optimal years. (b) MFIs by first and last reported year, on a logarithmic colour scale from light (one MFI) to dark (blank cells hold none), with the MFIs by first year along the top and by last year along the right, on one scale, panel members in blue. The outlined block holds the MFIs whose reports span 2006–2011; those that report in all 6 of these years form the panel. Cells on the diagonal are MFIs that report in a single year. (c) Number of years reported per MFI, stacked by the number of runs of consecutive years in which the MFI reports: one run, with no gap, in light grey, and two to four runs in deepening shades of vermilion. The blue step gives the panel members, and the dotted line separates the MFIs that report in fewer years than the optimum spans.

What each required variable costs. Every variable that a rule requires removes the years in which it is missing, so by Equation (1) dropping a variable can only raise the optimum and requiring one can only lower it. Table 8 recomputes the optimum with one variable removed from the primary rule or one covariate added to it. Dropping the number of active borrowers

would add the most, 10.9% (an optimum of 608 MFIs over the same years); dropping OSS would move the optimum to 2005–2010, with 586 MFIs; and dropping the gross loan portfolio would change nothing. Among the covariates, requiring the share of women borrowers would cost 26.5% of the optimal institution-years and requiring PAR30 19.1%, with a shorter panel, 2006–2010, whereas the average loan balance relative to GNI per capita would cost 0.7%.

**Table 6. Reporting flows by fiscal year, primary rule**

| Year | Reporting | First report | Return after a gap | Last report | Absent | Panel members |
|---|---|---|---|---|---|---|
| 1999 | 122 | 122[a] | 0 | 1 | 0 | 81 |
| 2000 | 191 | 71 | 0 | 3 | 1 | 119 |
| 2001 | 286 | 116 | 0 | 8 | 19 | 141 |
| 2002 | 474 | 195 | 8 | 14 | 18 | 207 |
| 2003 | 706 | 261 | 8 | 55 | 33 | 280 |
| 2004 | 863 | 226 | 19 | 27 | 47 | 374 |
| 2005 | 1,041 | 210 | 24 | 73 | 52 | 451 |
| 2006 | 1,169 | 221 | 18 | 106 | 72 | 548 |
| 2007 | 1,124 | 95 | 30 | 148 | 106 | 548 |
| 2008 | 1,237 | 256 | 63 | 163 | 101 | 548 |
| 2009 | 1,148 | 123 | 45 | 134 | 150 | 548 |
| 2010 | 1,191 | 166 | 67 | 169 | 139 | 548 |
| 2011 | 1,175 | 146 | 82 | 233 | 132 | 548 |
| 2012 | 928 | 74 | 65 | 207 | 220 | 390 |
| 2013 | 855 | 92 | 126 | 87 | 178 | 388 |
| 2014 | 963 | 127 | 120 | 150 | 110 | 388 |
| 2015 | 910 | 66 | 73 | 219 | 79 | 365 |
| 2016 | 698 | 26 | 43 | 125 | 98 | 300 |
| 2017 | 630 | 20 | 71 | 138 | 61 | 270 |
| 2018 | 566 | 14 | 61 | 565[b] | 1 | 253 |
| 2019 | 2 | 0 | 1 | 2[b] | 0 | 0 |

Note: Reporting: MFIs with a record that satisfies the primary rule. First report and last report: the earliest and the latest year in which an MFI reports. Return after a gap: the MFI reports after at least one year without a report. Absent: no report in the year, but reports before and after it. Panel members: MFIs of the optimal panel that report in the year. a All MFIs that report in 1999, the first year of the file. b Includes MFIs that may have continued to report: the file ends with 2019, when only 2 MFIs report. Post hoc (descriptive).

Optimum within groups. The optimal years are a compromise between groups whose reporting peaks at different times. Table 9 gives the optimum computed within each region and legal status, and Table 10 within the ten countries with the most MFIs, together with the part of each group's own optimum that the overall optimal years keep. The optimum for NBFIs, for NGOs and for East Asia and the Pacific coincides with the overall one, whereas the African optimum

is 2005–2008 and that of South Asia 2009–2015; the years 2006–2011 keep 69% of the African optimum. Within Latin America and the Caribbean, and within Bangladesh and Peru, the optimal year set is not a window of consecutive years. The panel includes MFIs from 80 of the 117 countries of the matrix; 35 countries appear among the excluded MFIs but not in the panel. The cost of the compromise is bounded: since the groups partition the MFIs, the optimum cannot exceed the sum of the group optima, Equation (16).

$$c(T) = \sum_{\gamma} c_{\gamma}(T) \;\Rightarrow\; |T|\, c(T) \le \sum_{\gamma} \mathrm{OPT}_{\gamma} \;\Rightarrow\; \mathrm{OPT} \le \sum_{\gamma} \mathrm{OPT}_{\gamma} \tag{16}$$

**Table 7. Exact optimum, greedy result, closed year sets and overlap under each reporting rule**

| Rule | Primary | Any record | Broad | As-of year |
|---|---|---|---|---|
| Optimal years | 2006–2011 | 2005–2011 | 2006–2011 | 2006–2011 |
| MFIs × years | 548 × 6 | 648 × 7 | 327 × 6 | 541 × 6 |
| Institution-years (share of all) | 3,288 (20.2%) | 4,536 (22.6%) | 1,962 (17.0%) | 3,246 (19.9%) |
| Greedy result (share of optimum) | optimal | 3,864 (85.2%) | 1,644 (83.8%) | optimal |
| Closed year sets | 112,396 | 125,658 | 106,456 | 120,424 |
| MFIs also in the primary panel | 548 | 497 | 327 | 532 |
| Countries | 80 | 89 | 65 | 79 |

Note: Share of all: share of all institution-years reported under the rule. Greedy result: institution-years of the panel reached by backward elimination and their share of the optimum; optimal means that it reaches the optimum. Under every rule the best of the 231 windows of consecutive years coincides with the optimum. Each optimum is the unique maximiser, is closed and equals the result of direct enumeration over all 2,097,152 year sets. Closed year sets: year sets with at least one MFI to which no year can be added without losing an MFI. Countries: countries of the MFIs in the optimal panel. The last three rows are post hoc (descriptive).

The overall optimum reaches 89.2% of this bound over the regions, whose optima add up to 3,686 institution-years, and 93.5% over the legal statuses (3,516; post hoc).

Frontier. Figure 4(a) shows the frontier under the primary rule, and Table 11 gives the best year set for every length. The number of MFIs that can be kept falls from 1,237 in a one-year panel to 22 in a 20-year panel, and no institution reports in all 21 years. The best year set is a contiguous window at every length except k = 9, where the set formed by 2006–2011 and 2013–2015 keeps 311 MFIs against 302 for the best window (2006–2014). Under the as-of rule the same happens at k = 9 and k = 12. Figure 4(b) shows the number of institution-years $k \cdot c^*(k)$ under the four rules. It peaks at k = 6 under the primary, broad and as-of rules and at k = 7 when any record counts, and the peak is flat: under the primary rule the 5-year panel 2006–2010

keeps 648 MFIs, 100 more than the optimum, for 48 fewer institution-years (1.5%). Equation (17) gives the condition under which adding a year j to a set T of k years, or removing one, adds institution-years:

$$\begin{aligned}(k+1)\,c(T\cup\{j\}) > k\,c(T) &\Leftrightarrow \frac{c(T\cup\{j\})}{c(T)} > \frac{k}{k+1},\\ (k-1)\,c(T\setminus\{j\}) > k\,c(T) &\Leftrightarrow \frac{c(T\setminus\{j\})}{c(T)} > \frac{k}{k-1}\end{aligned} \tag{17}$$

**Table 8. Optimum under the primary rule with one required variable removed or one covariate added**

| Variables required | MFIs | Institution-years | Optimal years | MFIs × years | Change in optimum |
|---|---|---|---|---|---|
| Primary rule | 2,627 | 16,279 | 2006–2011 | 548 × 6 | – |
| ***Primary rule without*** | | | | | |
| Assets | 2,639 | 16,328 | 2006–2011 | 552 × 6 | +0.7% |
| Gross loan portfolio | 2,629 | 16,292 | 2006–2011 | 548 × 6 | 0.0% |
| Active borrowers | 2,765 | 17,535 | 2006–2011 | 608 × 6 | +10.9% |
| OSS | 2,731 | 17,419 | 2005–2010 | 586 × 6 | +6.9% |
| ***Primary rule with*** | | | | | |
| Loan balance / GNI per capita | 2,619 | 16,052 | 2006–2011 | 544 × 6 | −0.7% |
| ROA | 2,431 | 14,278 | 2006–2011 | 481 × 6 | −12.2% |
| Operating expense / loan portfolio | 2,434 | 14,364 | 2006–2011 | 479 × 6 | −12.6% |
| PAR30 | 2,435 | 14,115 | 2006–2010 | 532 × 5 | −19.1% |
| Women borrowers | 2,475 | 13,974 | 2006–2011 | 403 × 6 | −26.5% |

Note: MFIs and institution-years: rows and filled cells of the reporting matrix under each variant of the rule, fiscal years 1999–2019. Change in optimum: in the optimal number of institution-years, relative to the primary rule. Every optimum in the table is the unique maximiser. Post hoc (descriptive).

From the optimum (post hoc; Figure 4(d)), adding a year would raise the number of institution-years only if at least 470 of the 548 MFIs reported in it, and the best year to add, 2005, keeps 451; removing one would raise it only if at least 658 MFIs reported in the other 5 years, against the 648 of the best 5-year panel.

Figure 4(c) shows how much a window gives up away from the optimum. Only 4 of the 231 windows reach 95% of the optimal number of institution-years, all of them within 2005–2011, and 153 reach less than 60%. A panel confined to the years after the optimum, from 2012 onwards, holds at most 1,944 institution-years (486 MFIs over 2012–2015), 59.1% of the optimum.

**Table 9. Optimum within each region and legal status, primary rule**

| Group | MFIs | Countries | Own optimal years | MFIs × years | In $S^*$ |
|---|---|---|---|---|---|
| ***Region*** | | | | | |
| Africa | 635 | 37 | 2005–2008 | 118 × 4 | 54 (69%) |
| Latin America and the Caribbean | 579 | 24 | 2006–2011, 2013–2015[a] | 149 × 9 | 210 (94%) |
| Eastern Europe and Central Asia | 487 | 24 | 2004–2010 | 89 × 7 | 91 (88%) |
| South Asia | 461 | 7 | 2009–2015 | 87 × 7 | 91 (90%) |
| East Asia and the Pacific | 392 | 15 | 2006–2011 | 69 × 6 | 69 (100%) |
| Middle East and North Africa | 72 | 9 | 2004–2010[b] | 32 × 7 | 33 (88%) |
| ***Legal status (last known)*** | | | | | |
| NBFI | 827 | 92 | 2006–2011 | 200 × 6 | 200 (100%) |
| NGO | 785 | 84 | 2006–2011 | 201 × 6 | 201 (100%) |
| Credit union / cooperative | 556 | 65 | 2006–2010 | 78 × 5 | 58 (89%) |
| Bank | 211 | 70 | 2005–2010 | 78 × 6 | 65 (83%) |
| Rural bank | 151 | 11 | 2005–2007 | 61 × 3 | 20 (66%) |
| Other | 52 | 26 | 2011–2015 | 11 × 5 | 4 (44%) |
| Unknown | 45 | 22 | 2008[c] | 14 × 1 | 0 (0%) |

Note: Own optimum: the largest balanced panel among the MFIs of the group. In S*: MFIs of the group in the overall optimal panel and, in parentheses, the share of the group's own optimum, in institution-years, that the years 2006–2011 keep. The MFI without a recorded region is not shown; its own optimum, 3 institution-years, counts in the sum of Equation (16). a Not a window; the best window is 2006–2015 (134 × 10 = 1,340). b Tied with 2004–2011 (28 × 8). c Tied with 2007–2008 (7 × 2). Post hoc (descriptive).

Cost. Table 12 and Figure 5 report the measured cost. Reading the 41,258-row spreadsheet dominates, with a median of 21.76 s; checking its digest takes 19.6 ms. Once the data are in memory, the records and the matrix take 8.5 ms, the histogram 4.9 ms and the transform 19.8 ms; the optimum with its tie-breaking takes 57.7 ms and the whole frontier 146.2 ms. Direct enumeration over the 797 patterns takes 8.71 s, 352 times as long as the histogram and the transform together. The ratio of times is the product of the ratio of operation counts and that of the time per operation, Equation (18):

$$\frac{t_{\mathrm{en}}}{t_{\mathrm{tr}}} = \frac{\kappa_{\mathrm{en}}}{\kappa_{\mathrm{tr}}} \cdot \frac{W_{\mathrm{en}}}{W_{\mathrm{tr}}} = \frac{\kappa_{\mathrm{en}}}{\kappa_{\mathrm{tr}}} \cdot \frac{2d}{m} \tag{18}$$

**Table 10. Optimum within the ten countries with the most MFIs, primary rule**

| Country | MFIs | Own optimal years | MFIs × years | In $S^*$ |
|---|---|---|---|---|
| India | 229 | 2010–2015 | 51 × 6 | 39 (76%) |
| Russia | 120 | 2008 | 84 × 1 | 5 (36%) |
| Philippines | 113 | 2004–2009 | 36 × 6 | 23 (64%) |
| Mexico | 108 | 2014–2016[a] | 49 × 3 | 18 (73%) |
| Bangladesh | 82 | 2004–2005, 2007–2011, 2013–2017[b] | 19 × 12 | 19 (50%) |
| Indonesia | 75 | 2006–2007 | 43 × 2 | 11 (77%) |
| Peru | 75 | 2005–2011, 2013–2018[c] | 23 × 13 | 42 (84%) |
| Ecuador | 72 | 2006–2018 | 24 × 13 | 33 (63%) |
| China | 69 | 2010–2011 | 38 × 2 | 4 (32%) |
| Ghana | 66 | 2005–2008 | 14 × 4 | 2 (21%) |

Note: Own optimum and In S* as in Table 9. a Tied with 2008–2014 (21 × 7) and 2010–2016 (21 × 7). b Not a window; the best window is 2004–2017 (15 × 14 = 210). c Not a window; the best window is 2005–2018 (20 × 14 = 280). Post hoc (descriptive).

On the real data (post hoc; Figures 5c and 5d), the transform takes 1.123 ns per addition, the histogram included, and direct enumeration 5.208 ns per containment test, so the measured ratio, 352, is the ratio of operation counts, 75.9, times 4.64. Both implementations are vectorised; each enumeration step forms a block of containment tests and a weighted sum, whereas each transform step is a single in-place addition.

On the synthetic instances (Figure 5a), the time of the exact method, histogram and transform together, grows by a factor of about 21.13 per additional year for m ≥ 14, and that of direct enumeration by about 21.03. The operation counts of Equation (7), drawn as dotted lines, imply the slopes of Equation (19):

$$\log_2 \frac{W_{\mathrm{tr}}(m+1)}{W_{\mathrm{tr}}(m)} = 1 + \log_2\left(1+\frac{1}{m}\right) \approx 1 + \frac{1}{m\ln 2}, \qquad \log_2 \frac{W_{\mathrm{en}}(m+1)}{W_{\mathrm{en}}(m)} = 1 + \log_2 \frac{d_{m+1}}{d_m} \tag{19}$$

Fitted over the same values of m (post hoc), they give 1.08 for the transform and 1.02 for direct enumeration, whose number of patterns dm approaches n as m grows, against the measured 1.13 and 1.03. The exact method takes 393.4 ms and 128 MiB at m = 24, whereas direct enumeration already takes 23.70 s at m = 20, against 15.0 ms for the exact method. In a separate run on random masks with m fixed at 21 (Figure 5b), the histogram and the transform together take between 33.4 ms and 37.2 ms as n grows from 103 to 106 rows, while the histogram alone goes from 7.2 ms to 12.7 ms: at this scale the cost is set by the table of 2m entries rather than by the number of institutions.

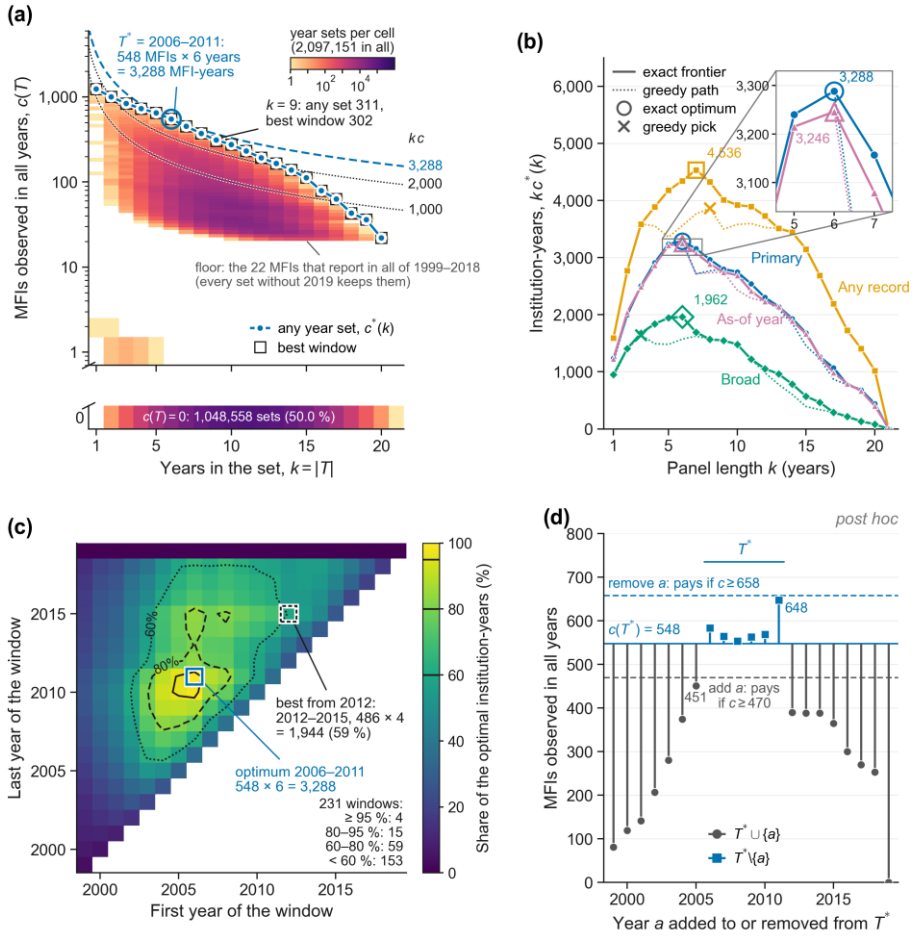


**Figure 4. Frontier between panel length and panel width, primary rule unless stated.**

(a) The 2m − 1 non-empty sets of years T by length k = |T| and by the number of MFIs observed in all their years, c(T), on a logarithmic scale; the colour of each cell gives the number of sets in it, from light (one set) to dark, also on a logarithmic scale, and the strip below the broken axis holds the sets with c(T) = 0. The blue line with dots is the frontier c*(k), the upper edge of the cloud; open squares give the best window of consecutive years, which lies on it except at k = 9, as labelled. Dotted curves join the points of equal institution-years k·c, labelled at the right; the dashed blue one passes through the optimum (open circle) and touches the frontier only there. (b) Balanced institution-years k·c*(k) under the four reporting rules, each named beside its line (solid), and along the path of the greedy heuristic (dotted). Large open markers give the exact optimum of each rule, labelled with its institution-years, and crosses mark the greedy result where it falls short. The inset enlarges the grey box around the peak, where the optima of the primary and as-of rules lie close together. (c) The 231 windows of consecutive years by first and last year, shaded from dark violet to yellow by their institution-years as a share of the optimum, with contours at 60% (dotted), 80% (dashed) and 95% (solid), also marked on the colour bar, and the number of windows in each class at the lower right. The blue square is the

optimum, 2006–2011, and the dashed square the best window after it, 2012–2015. (d) Post hoc: MFIs observed in all years of T* ∪ {a} for each year a outside the optimum (grey circles) and of T* \ {a} for each year inside it (blue squares), drawn from c(T*) = 548 (solid line); the bar at the top spans T*, and the best change of each kind is labelled. The dashed lines are the thresholds of Equation (17), 470 MFIs for adding a year and 658 for removing one; no year reaches its threshold.

**Table 11. Best year set for each panel length k, primary rule**

| *k* | Best year set | MFIs, $c^*(k)$ | Institution-years | Countries |
|---|---|---|---|---|
| 1 | 2008 | 1,237 | 1,237 | 100 |
| 2 | 2006–2007 | 999 | 1,998 | 97 |
| 3 | 2008–2010 | 836 | 2,508 | 94 |
| 4 | 2006–2009 | 728 | 2,912 | 89 |
| 5 | 2006–2010 | 648 | 3,240 | 84 |
| **6** | **2006–2011** | **548** | **3,288** | **80** |
| 7 | 2005–2011 | 451 | 3,157 | 78 |
| 8 | 2004–2011 | 371 | 2,968 | 74 |
| 9 | 2006–2011 and 2013–2015[a] | 311 | 2,799 | 61 |
| 10 | 2006–2015 | 275 | 2,750 | 55 |
| 11 | 2005–2015 | 231 | 2,541 | 54 |
| 12 | 2004–2015 | 192 | 2,304 | 52 |
| 13 | 2006–2018 | 165 | 2,145 | 36 |
| 14 | 2005–2018 | 138 | 1,932 | 36 |
| 15 | 2004–2018 | 112 | 1,680 | 32 |
| 16 | 2003–2018 | 79 | 1,264 | 27 |
| 17 | 2002–2018 | 63 | 1,071 | 26 |
| 18 | 2000–2017[b] | 43 | 774 | 23 |
| 19 | 2000–2018 | 36 | 684 | 18 |
| 20 | 1999–2018 | 22 | 440 | 11 |

Note: c*(k): the largest number of MFIs that report in all years of a set of k years; institution-years are k·c*(k). The optimum, k = 6, is in bold. No MFI reports in all 21 years. Countries: countries of the MFIs in the panel, post hoc (descriptive). a The only length at which the best set is not a window of consecutive years. b Tied with 2001–2018.

Composition of the optimal panel. Table 13 and Figure 6 compare the 548 MFIs of the optimal panel with the 1,475 MFIs that report under the primary rule in at least one but not all of the years 2006–2011; of the latter, 431 report in one of the 6 years and 195 in 5. The other 604 MFIs of the matrix report only outside these years and belong to neither group. The largest differences concern size. The mean within-year percentile of the panel is 60.6 for assets against 36.2 for the excluded MFIs, a difference of 24.4 points (95% CI 21.8 to 26.9), and the difference for active borrowers is 24.8 points. Over the institution-years of 2006–2011, median total assets are USD 10.1 million in the panel against USD 2.8 million among the excluded MFIs (Table 14). The panel also ranks higher in OSS (6.0 points) and ROA (6.2 points) and lower in PAR30

(−3.3 points), and its share of women borrowers is higher (3.9 points). After Holm's adjustment all of these differences are significant at the 5% level, PAR30 only just: its adjusted p-value, 0.0498, is three times the unadjusted 0.0166 (Equation (15)), and over the 95% Monte Carlo range of p (Equation (12); post hoc) it would run from 0.042 to 0.057, so that PAR30, alone in the family, could pass or fail with other permutations. The difference in operating expense relative to the loan portfolio (−3.1 points, adjusted p = 0.0726) is not significant, and the average loan balance relative to GNI per capita shows no clear difference (−2.2 points, p = 0.1320). Within each year (post hoc), the difference in mean percentile between the panel and the excluded MFIs that report is a rescaled probability, Equation (20):

$$\begin{aligned} \Delta_{v,t} &= \frac{1}{|S^{*}_{v,t}|}\sum_{i\in S^{*}_{v,t}} q_{it} - \frac{1}{|E_{v,t}|}\sum_{i\in E_{v,t}} q_{it} \\ &= 100\big(\mathrm{AUC}_{v,t} - 1/2\big), \\ \mathrm{AUC}_{v,t} &= \frac{1}{|S^{*}_{v,t}|\,|E_{v,t}|}\sum_{i\in S^{*}_{v,t}}\sum_{i\prime\in E_{v,t}} \psi_{ii\prime t}\,, \\ \psi_{ii\prime t} &= \mathbf{1}[x_{it} > x_{i\prime t}] + 1/2\ \mathbf{1}[x_{it} = x_{i\prime t}] \end{aligned} \tag{20}$$

**Table 12. Measured cost of each stage on the real data (primary rule)**

| Stage | *R* | Median | Interquartile range |
|---|---|---|---|
| Read spreadsheet (financial file) | 3 | 21.76 s | 21.74–21.98 s |
| SHA-256 check | 3 | 19.6 ms | 19.4–19.7 ms |
| Records and matrix | 20 | 8.5 ms | 8.2–10.3 ms |
| Bit masks | 20 | 0.084 ms | 0.083–0.085 ms |
| Histogram | 20 | 4.9 ms | 4.8–5.2 ms |
| Superset sums (Algorithm 1) | 20 | 19.8 ms | 18.4–21.9 ms |
| Optimum with tie-breaking | 20 | 57.7 ms | 55.9–58.2 ms |
| Frontier, *k* = 1, …, 21 | 20 | 146.2 ms | 141.6–152.8 ms |
| Best contiguous window | 20 | 0.039 ms | 0.038–0.039 ms |
| Greedy elimination | 20 | 0.557 ms | 0.556–0.561 ms |
| Direct enumeration | 5 | 8.71 s | 8.69–8.75 s |

Note: Wall-clock time on one thread; n = 2,627 MFIs, m = 21 years, d = 797 patterns. R is the number of timed repetitions after one warm-up run.

For assets the yearly difference runs from 17.7 points (2007) to 23.3 (2011): in a given year, a panel member drawn at random has more assets than an excluded MFI with probability 0.677 to 0.733. Their mean, 19.9, is below the 24.4 points of Table 13, which give each MFI the same weight, whereas a yearly comparison counts an excluded MFI once for each year in which it reports. For PAR30 the yearly difference is −8.4 and −11.0 points in 2006 and 2007 but 2.1 and 0.9 in 2010 and 2011: the panel ranks lower in portfolio at risk only in 2006–2009.

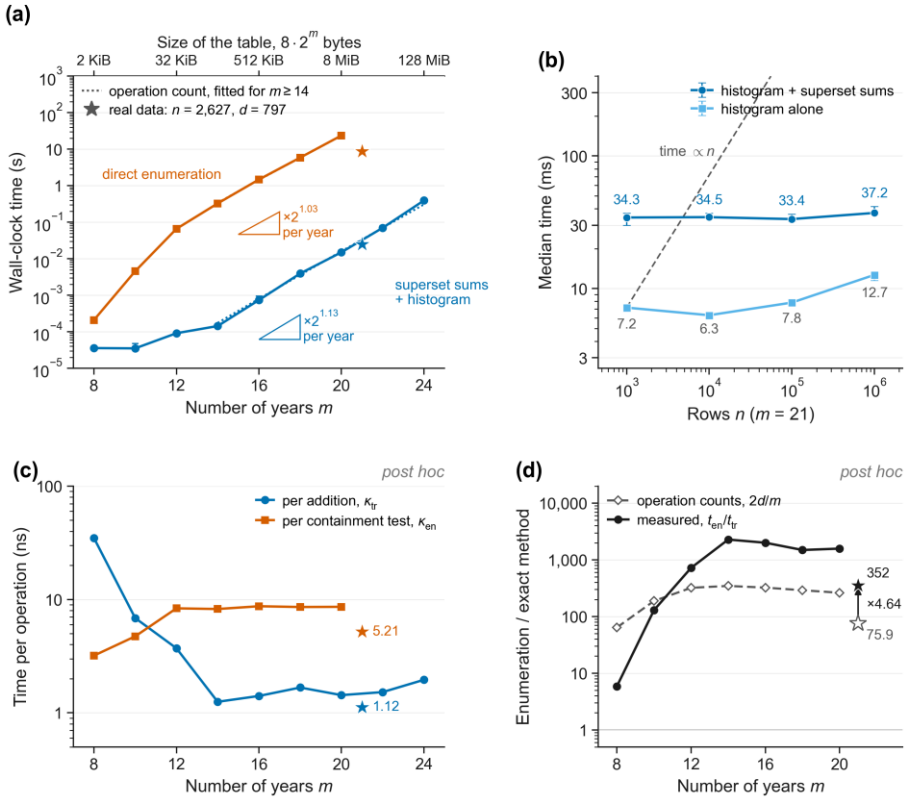


**Figure 5. Measured cost on one thread.**

Points are medians and error bars interquartile ranges, most of them narrower than the markers. (a) Wall-clock time against the number of years m on synthetic instances with n = 2,627 random masks: the exact method, histogram and superset sums together (blue circles), and direct enumeration (vermilion squares, up to m = 20). Dotted lines are the operation counts of Equation (7), additions for the exact method and containment tests for direct enumeration, each scaled to the measured times for m ≥ 14; each triangle spans two years and gives the measured growth per additional year. Stars are the real data (m = 21), and the upper axis gives the memory of the table, eight bytes per set of years. (b) Time against the number of rows n at m = 21, both axes logarithmic: histogram and superset sums together (dark blue circles, medians above) and histogram alone (light blue squares, medians below). The dashed line is a time proportional to n, drawn from the first median of the histogram alone. (c) Post hoc: time per operation, the times of (a) divided by their operation counts: per addition of the exact method, κtr, the histogram included, and per containment test of direct enumeration, κen; stars give the real data. (d) Post hoc: ratio of the times of direct enumeration and the exact method, ten/ttr (filled circles), 2d/m (open diamonds), where d is the number of distinct reporting patterns. On the real data (stars), the arrow from one to the other is the ratio of the times per operation, κen/κtr = 4.64, as in Equation (18); the grey line marks equal times.

The categorical composition differs as well. Latin America and the Caribbean accounts for 38.3% of the panel and 18.6% of the excluded MFIs, whereas Africa accounts for 9.9% and

25.2%; the TVD between the regional distributions is 0.233, by Equation (14) the excess share in the panel of Latin America and the Caribbean together with the Middle East and North Africa, the only regions more common there. By legal status (TVD 0.196), credit unions and cooperatives make up 10.6% of the panel against 22.8% of the excluded group, and rural banks 3.6% against 8.1%, while non-bank financial institutions (NBFIs), NGOs and banks are more common in the panel, and their excess shares add up to the TVD. The profit-status TVD, 0.077 (adjusted p = 0.0195), equals the share of excluded MFIs whose status is unknown, 7.7%, the only category less common in the panel, where it is absent. Among MFIs of known status (post hoc), the non-profit share is 59.1% in the panel against 56.2% among the excluded MFIs, a TVD of 0.029.

**Table 13. Optimal panel against the excluded MFIs, primary rule, 2006–2011**

| Variable | MFIs, panel / excluded | Δ or TVD (95% CI) | *p* | *p*, Holm |
|---|---|---|---|---|
| Assets (USD million) | 548 / 1,475 | 24.4 (21.8, 26.9) | 0.0001 | 0.0011 |
| Active borrowers (thousand) | 548 / 1,475 | 24.8 (22.3, 27.3) | 0.0001 | 0.0011 |
| Average loan balance / GNI per capita (%) | 548 / 1,471 | −2.2 (−4.9, 0.5) | 0.1320 | 0.1320 |
| Women borrowers (%) | 541 / 1,336 | 3.9 (1.2, 6.6) | 0.0071 | 0.0284 |
| Operational self-sufficiency (%) | 548 / 1,475 | 6.0 (3.7, 8.3) | 0.0001 | 0.0011 |
| Return on assets (%) | 548 / 1,345 | 6.2 (3.8, 8.5) | 0.0001 | 0.0011 |
| Portfolio at risk > 30 days (%) | 548 / 1,288 | −3.3 (−5.8, −0.9) | 0.0166 | 0.0498 |
| Operating expense / loan portfolio (%) | 548 / 1,344 | −3.1 (−5.8, −0.4) | 0.0363 | 0.0726 |
| Region | 548 / 1,475 | 0.233 | 0.0001 | 0.0011 |
| Legal status (last known) | 548 / 1,475 | 0.196 | 0.0001 | 0.0011 |
| Profit status (last known) | 548 / 1,475 | 0.077 | 0.0039 | 0.0195 |

Note: MFIs: institutions of each group with at least one value of the variable in 2006–2011. Δ is the difference in mean within-year percentile, panel minus excluded, in percentile points (Equation (11)), with its 95% bootstrap confidence interval (Equation (13), 10,000 resamples); TVD is the total variation distance (Equation (14)). p: from 10,000 permutations, two-sided for Δ and from the upper tail for the TVD; 0.0001 is the smallest attainable value (1/10,001), and the Monte Carlo standard error of p (Equation (12)) is at most 0.0034 (post hoc). Holm: adjusted over 11 tests (Equation (15)).

**Table 14. Medians and quartiles of the continuous covariates, optimal panel and excluded MFIs, primary rule, 2006–2011**

| Variable | Panel, median (Q1, Q3) | *n* | Excluded, median (Q1, Q3) | *n* |
|---|---|---|---|---|
| Assets (USD million) | 10.1 (3.0, 39.4) | 3,288 | 2.8 (0.8, 9.5) | 3,756 |
| Active borrowers (thousand) | 14.8 (4.6, 52.3) | 3,288 | 4.1 (1.0, 14.4) | 3,756 |
| Average loan balance / GNI per capita (%) | 26.1 (12.3, 62.8) | 3,284 | 32.2 (12.0, 75.5) | 3,734 |
| Women borrowers (%) | 66.9 (46.7, 91.7) | 3,004 | 61.0 (40.8, 89.0) | 3,034 |
| Operational self-sufficiency (%) | 114 (102, 130) | 3,288 | 109 (91, 127) | 3,756 |
| Return on assets (%) | 2.4 (0.3, 5.2) | 3,221 | 1.5 (−1.8, 4.5) | 3,266 |
| Portfolio at risk > 30 days (%) | 3.5 (1.3, 7.1) | 3,117 | 4.3 (1.2, 9.6) | 3,040 |
| Operating expense / loan portfolio (%) | 17.8 (12.1, 29.1) | 3,219 | 20.2 (11.8, 37.6) | 3,274 |

Note: Raw values over the institution-years of 2006–2011 in which the variable has a value; n counts these institution-years, out of 3,288 in the panel and 3,756 among the excluded MFIs. Q1 and Q3: first and third quartiles. The medians were pre-specified; the quartiles are post hoc (descriptive).

Why the excluded MFIs are excluded. Table 15 sorts the 1,475 excluded MFIs by their reporting pattern over 2006–2011. Of these, 477 start after 2006 and then report in every year up to 2011, 449 of them without any earlier report in the file; 469 report from 2006 and stop before 2011, 410 of them without reporting again; 311 start late and stop early; and 218 miss a year between two years in which they report. All four groups rank well below the panel in assets and active borrowers. Those that start late and stop early rank lowest, with a mean within-year percentile of 28.9 for assets against 60.6 in the panel, and those that stop early rank highest in portfolio at risk, at 55.6 against 48.0.

Because the four groups partition the excluded MFIs, each difference of Table 13 is the sum of the gaps between the panel and each group, weighted by the group's share, Equation (21):

$$\Delta_v = \sum_{\mu} \frac{|E_{\mu,v}|}{|E_v|} \left(\bar{r}(S_v^*) - \bar{r}(E_{\mu,v})\right), \qquad \bar{r}(G) = \frac{1}{|G|} \sum_{i \in G} \bar{r}_i \tag{21}$$

In this split (post hoc), late entry accounts for 33% of the difference in assets, early exit for 27%, both for 27% and gaps for 13%. Early exit accounts for 74% of the difference in PAR30 and late entry for almost none, whereas late entry accounts for 42% of that in operating expense and early exit for 49% of that in the share of women borrowers.

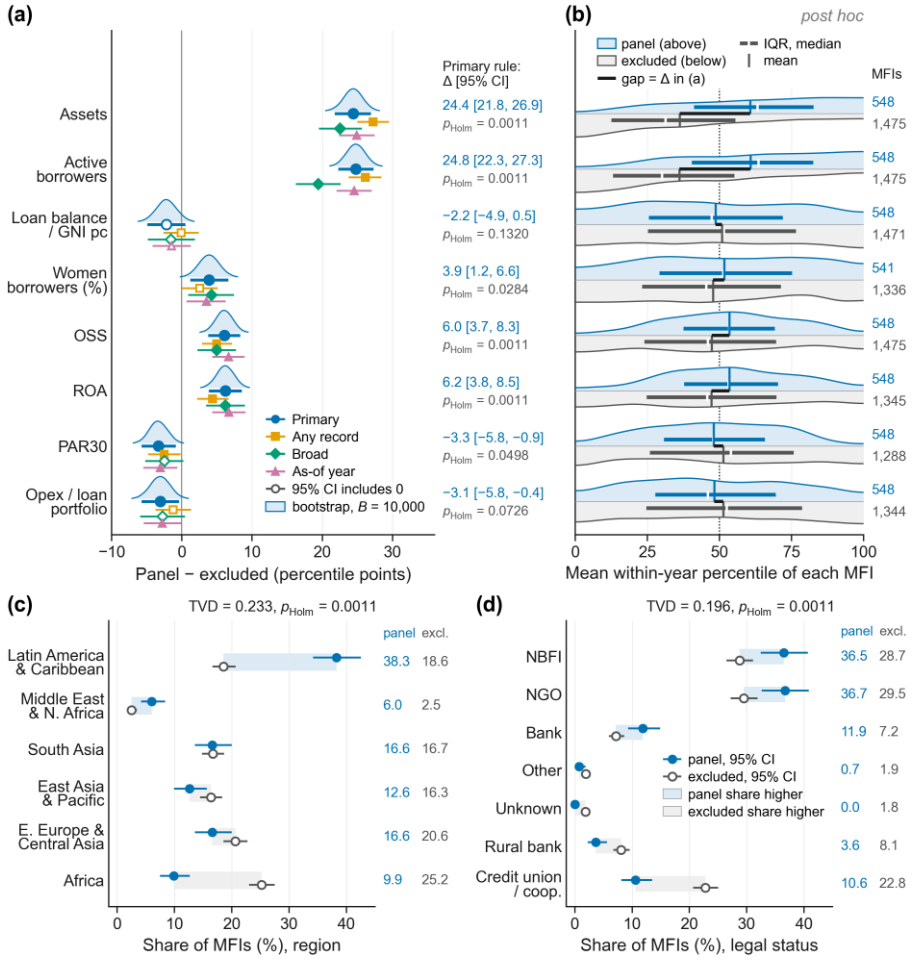


**Figure 6. Composition of the optimal panel: the 548 MFIs of the panel against the 1,475 excluded MFIs.**

(a) Difference in mean within-year percentile (Equation (10)), panel minus excluded, for the eight continuous covariates under the four reporting rules, primary rule on top, with 95% bootstrap intervals; a marker is open when its interval includes zero. Above each primary-rule marker is the distribution of its 10,000 bootstrap replicates (kernel density), and the column gives the primary-rule difference with its interval and the Holm-adjusted p-value of the permutation test. (b) Post hoc: distribution of the mean within-year percentile of each MFI, panel above the baseline and excluded MFIs below (Gaussian kernel densities reflected at both ends of the scale, each pair scaled to its taller member). Bars are interquartile ranges with a white tick at the median, vertical lines are means, and the black segment between the two means is the difference shown in (a); the column gives the number of MFIs with a value, and the dotted line marks the middle of the percentile scale. (c) Region and (d) legal status: share of MFIs in each category in the panel (filled circles) and among the excluded MFIs (open circles), with 95% Clopper–Pearson intervals and the shares in the columns at the right. The shaded bar spans the gap between the two shares, blue where the panel's share is higher and grey where it is lower; the blue bars of each panel add up to the total variation distance (TVD) given above it

with its Holm-adjusted p-value. Categories are ordered by the gap; MFIs of unknown region (none in the panel) are not drawn.

**Table 15. Excluded MFIs by reporting pattern over 2006–2011: number and mean within-year percentile of each variable, primary rule**

| Variable | Panel | Late entry | Early exit | Both | Gap |
|---|---|---|---|---|---|
| Number of MFIs | 548 | 477 | 469 | 311 | 218 |
| Assets | 60.6 | 36.0 | 39.6 | 28.9 | 39.8 |
| Active borrowers | 60.8 | 34.7 | 39.9 | 27.9 | 42.1 |
| Average loan balance / GNI per capita | 48.6 | 49.3 | 53.8 | 49.4 | 49.8 |
| Women borrowers | 51.6 | 50.6 | 45.7 | 45.1 | 50.0 |
| Operational self-sufficiency | 53.4 | 47.5 | 48.2 | 46.5 | 46.3 |
| Return on assets | 53.4 | 47.0 | 46.5 | 47.9 | 48.2 |
| Portfolio at risk > 30 days | 48.0 | 47.9 | 55.6 | 50.8 | 49.7 |
| Operating expense / loan portfolio | 48.2 | 52.3 | 49.6 | 51.8 | 52.1 |

Note: Late entry: reports in every year from a year after 2006 up to 2011. Early exit: reports in every year from 2006 up to a year before 2011 and in none of the later ones. Both: reports in an unbroken run of years that starts after 2006 and ends before 2011. Gap: misses a year between two years in which it reports. Other rows: mean within-year percentile (Equation (10)), averaged over the MFIs of the group with a value of the variable in 2006–2011; each mean rests on at least 213 MFIs. Post hoc (descriptive).

Composition along the frontier and under other rules. Figure 7 repeats the comparison of the continuous covariates along the frontier. The size gap widens as the panel lengthens: the difference in assets grows from 19.1 points for the best two-year panel to 44.7 points for the best 20-year panel, the differences in OSS and ROA grow from 2.5 and 1.3 points to 16.3 and 14.1 points, and the PAR30 difference, −4.0 points at $k = 2$, reaches −12.7 points at $k = 20$. The intervals widen at the long end, where the panels are small (22 MFIs at $k = 20$ against 2,605 excluded), but the trend is visible across the whole range. Not every difference grows: the difference in average loan balance relative to GNI per capita stays between −3.4 and 5.7 points at every length, and the share of women borrowers is lower in the panel at $k = 2$ (−5.3 points) and higher at every other length. Table 16 gives the categorical composition of the best panel at selected lengths. Latin America and the Caribbean accounts for 29.0% of the MFIs in the best one-year panel and 59.1% of the 22 MFIs in the best 20-year panel, and banks for 9.5% and 45.5%; the 20-year panel holds no MFI from Africa or from Eastern Europe and Central Asia and no cooperative.

Under the alternative reporting rules (Figure 6a), the differences in size and profitability keep their sign and magnitude: 22.5 to 27.2 points for assets, 5.0 to 6.6 points for OSS and 4.4 to 6.7 points for ROA, all with $p \leq 0.0031$. The smaller differences are less stable: across the three rules the p-value ranges from 0.025 to 0.140 for PAR30, from 0.017 to 0.063 for the share of women borrowers and from 0.053 to 0.358 for operating expense. The regional TVD ranges

from 0.191 to 0.266 and the legal-status TVD from 0.191 to 0.201, while the profit-status TVD ranges from 0.042 (p = 0.2111) under the broad rule to 0.128 when any record counts.

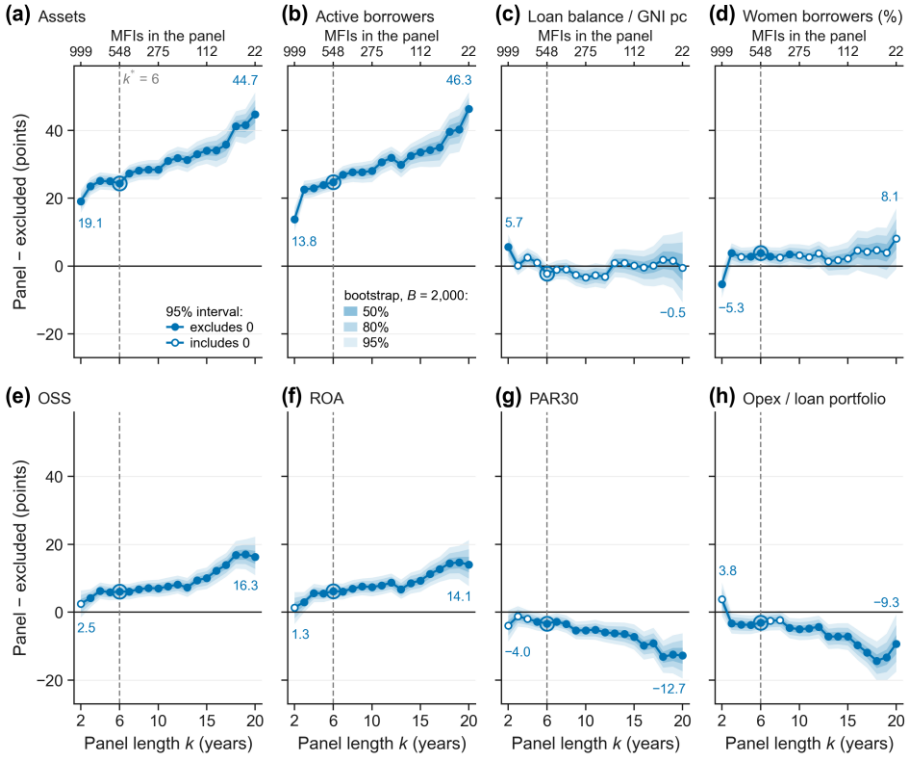


**Figure 7. Composition along the frontier, primary rule.**

Difference in mean within-year percentile between the panel and the excluded MFIs for the best k-year set, k = 2, …, 20, one panel per covariate on a common vertical scale. The line joins the observed differences, with the values at both ends written beside it; the nested bands hold the central 50%, 80% and 95% of 2,000 bootstrap replicates (dark to light), and a marker is open where the 95% band includes zero. The dashed line marks the optimal length, k = 6, and the ring the difference there, the primary-rule value of Figure 6(a). The upper axis gives the number of MFIs in the panel at selected lengths: the panel shrinks as k grows, and the bands widen.

## Discussion

The engineering result is that the largest balanced panel of an annual reporting history can be computed exactly, verified exhaustively and timed, in well under a second once the data are in memory. The hardness of maximum edge biclique (Ambühl et al., 2011; Peeters, 2003) concerns matrices with two large sides; with 21 years, all year sets fit in a table of 16 MiB, and reading the spreadsheet takes about 880 times as long as the histogram and the transform together. The dense table also marks the limit of the method. Its memory grows as 8·2m bytes, 128 MiB at m = 24 and 8 GiB at m = 30. The quarterly records of the same file, 21,138 records for 1,375 MFIs in 50 distinct quarters between the fourth quarter of 2005 and the third quarter

of 2019, would need 8·250 bytes, or 8 PiB, so quarterly or monthly histories need another approach. Because every optimum is closed, the search could then be restricted to the closed sets of periods, the intersections of reporting patterns, which maximal-biclique enumeration and formal concept analysis list (Alexe et al., 2004; Ganter & Wille, 1999); in the annual data only 112,396 of the 221 − 1 non-empty year sets (5.4%) are closed under the primary rule (Table 7). For annual MIX data this is unnecessary.

**Table 16. Categorical composition of the best k-year panel, primary rule (% of MFIs)**

| Panel length *k* | 1 | 2 | 4 | 6 | 8 | 10 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|
| ***Region*** | | | | | | | | |
| Africa | 17.2 | 18.6 | 13.7 | 9.9 | 10.0 | 2.9 | 0.9 | 0.0 |
| East Asia and the Pacific | 12.5 | 13.1 | 12.1 | 12.6 | 13.2 | 8.7 | 11.6 | 9.1 |
| Eastern Europe and Central Asia | 21.9 | 20.0 | 19.0 | 16.6 | 19.7 | 15.3 | 16.1 | 0.0 |
| Latin America and the Caribbean | 29.0 | 28.5 | 33.1 | 38.3 | 33.7 | 48.7 | 44.6 | 59.1 |
| Middle East and North Africa | 4.5 | 4.4 | 5.9 | 6.0 | 7.5 | 4.4 | 5.4 | 13.6 |
| South Asia | 14.8 | 15.3 | 16.2 | 16.6 | 15.9 | 20.0 | 21.4 | 18.2 |
| ***Legal status (last known)*** | | | | | | | | |
| Bank | 9.5 | 9.7 | 12.0 | 11.9 | 14.0 | 15.6 | 17.0 | 45.5 |
| Credit union / cooperative | 18.5 | 15.9 | 12.6 | 10.6 | 8.4 | 8.0 | 5.4 | 0.0 |
| NBFI | 32.4 | 33.1 | 35.2 | 36.5 | 36.1 | 40.4 | 41.1 | 27.3 |
| NGO | 30.7 | 32.3 | 34.5 | 36.7 | 37.5 | 34.2 | 33.0 | 27.3 |
| Other | 1.1 | 0.7 | 0.7 | 0.7 | 0.8 | 0.7 | 1.8 | 0.0 |
| Rural bank | 6.6 | 7.8 | 5.1 | 3.6 | 3.2 | 1.1 | 1.8 | 0.0 |
| Unknown | 1.1 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ***Profit status (last known)*** | | | | | | | | |
| Non-profit | 57.8 | 58.4 | 58.0 | 59.1 | 58.8 | 56.4 | 55.4 | 50.0 |
| Profit | 41.3 | 41.0 | 42.0 | 40.9 | 41.2 | 43.6 | 44.6 | 50.0 |
| Unknown | 0.9 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Note: Share of the MFIs of the best k-year panel in each category; Table 11 gives their number, and k = 6 is the optimal panel. No MFI of unknown region is in any of these panels. Post hoc (descriptive).

In this release the practice of choosing a window of consecutive years would have found the optimum, since the best window and the exact optimum coincide under all four rules. That is a property of these data, not a guarantee: at some lengths the best year set is not a window (k = 9 under the primary rule), the verification includes a matrix in which a set with a gap beats every window, and the greedy heuristic, which reads the same exact counts, falls 14.8% and 16.2% short of the optimum under two of the rules. The exact computation costs little more than the heuristics and replaces an argument about why a window was chosen with a reproducible result

that includes a statement of uniqueness. The flat top of the frontier is informative too: a researcher who values breadth over length can take the 5-year panel, with 100 more institutions for 1.5% fewer institution-years, and say so.

The composition audit shows that the largest balanced panel is a selected group. Its institutions rank about a quarter of the percentile scale higher in size, are more self-sufficient and more profitable, have somewhat lower portfolio at risk, and come more often from Latin America and less often from Africa or the cooperative sector. These are associations between persistent reporting and institutional characteristics; the analysis does not establish why some institutions report persistently. They agree in direction with the reporting biases documented by Bauchet and Morduch (2010) and with the logic of survivorship bias in performance studies (Brown et al., 1992; Elton et al., 1996), and the gaps in size, self-sufficiency, profitability and portfolio at risk grow with the length of the panel. For questions such as the determinants of financial sustainability or the relation between leverage and profitability, examined in this journal by Chikalipah (2017, 2019), the choice between a balanced and an unbalanced panel is therefore not neutral: a long balanced panel describes the persistent reporters, which are larger and more self-sufficient than the other institutions that reported in the same years.

Table 17 turns these results into a reporting checklist for studies that use MIX Market or similar voluntary reporting databases: what to state about a balanced panel, why it matters and where this study states it. The reporting rule, the period definition and the extraction criterion let the panel be rebuilt; the other items tell the reader how much of the reporting population the panel keeps and how it differs from the rest. Where the research question allows, methods for unbalanced panels (Baltagi & Song, 2006) and tests and corrections for selection (Verbeek & Nijman, 1992; Wooldridge, 1995) use the excluded institution-years instead of discarding them.

**Table 17. Reporting checklist for a balanced panel drawn from a voluntary reporting database**

| Item to report | Why it matters | In this study |
|---|---|---|
| **Reporting rule:** the variables that must be present for a year to count | Requiring the share of women borrowers as well would cut the optimal panel by 26.5% of its institution-years | Tables 3 and 8 |
| **Period definition:** fiscal year of the record or calendar year of the as-of date | With the as-of year the optimum covers the same years with 541 MFIs instead of 548 | Table 7 |
| **Extraction criterion,** and whether the optimum is unique | Under the four rules, backward elimination falls up to 16.2% short of the optimum, and a tied optimum would make the panel one choice among several | Table 7 |
| **Size** of the panel relative to the reporting population | The optimal panel keeps 20.2% of the institution-years and 20.9% of the MFIs | Table 7 |
| **Frontier:** the best panel at each length | Shows what a longer or a broader panel would cost when the length is a choice | Table 11; Figure 4 |
| **Composition audit** against the MFIs that reported in the panel's years but were left out | The panel ranks 24.4 percentile points higher in assets, and the gap widens as the panel lengthens | Tables 13–16; Figures 6 and 7 |

| Item to report | Why it matters | In this study |
|---|---|---|
| **Implausible values** and their treatment | Out-of-range values, such as negative OSS, remain in the release | Table 2 |
| **Data release** and file digests | The panel depends on the files used; the digests let others check that theirs are the same | Data availability statement |

Note: Why it matters: what the item changes in this study, under the primary rule unless the row says otherwise.

The study has limitations. It uses one release of the data, distributed in December 2020, and takes MIX's validation and field definitions as given; implausible values (Table 2) are not corrected, although the rank-based comparisons limit their influence. The comparison group is defined relative to the optimal years, and other definitions, such as all institutions that ever reported, would answer a different question. The percentiles are computed within the reporting population of each year, itself a selected group, so the differences describe the panel relative to other reporters, not relative to all MFIs. The categorical variables are last known statuses, which may postdate the panel years. Finally, the timings come from a single machine, although the operation counts and the scaling slopes give a view that does not depend on it.

## Conclusion

Cutting a balanced panel from an unbalanced reporting history is a selection step that can be specified, solved exactly and audited. With annual data the year side of the reporting matrix is short, and a superset-sum transform gives, for every set of years, the number of institutions that report in all of them in $O(nm + m \cdot 2m)$ time; the largest balanced panel follows from one pass over this table. For MIX Market, the optimum under the primary reporting rule is unique: 548 MFIs over 2006–2011, 20.2% of the institution-years available. The superset sums take 19.8 ms once the data are loaded; the optimum agrees with direct enumeration on all 2,097,152 year sets and coincides with the best contiguous window, which a greedy heuristic does not always find. The panel it defines is not representative of the institutions that reported in those years: it holds larger, more self-sufficient and more profitable MFIs, more of them Latin American and fewer African or cooperative, and its lead in size, self-sufficiency and profitability widens as the panel lengthens. Research that uses balanced MIX panels should state how they were extracted and report what they leave out; the frontier and the composition audit described here make both steps routine. Further work could extend the method to longer reporting histories with sparse representations and apply the same audit to other voluntary reporting databases.

## References

Ahlin, C., Lin, J., & Maio, M. (2011). Where does microfinance flourish? Microfinance institution performance in macroeconomic context. Journal of Development Economics, 95(2), 105–120. https://doi.org/10.1016/j.jdeveco.2010.04.004

Alexe, G., Alexe, S., Crama, Y., Foldes, S., Hammer, P. L., & Simeone, B. (2004). Consensus algorithms for the generation of all maximal bicliques. Discrete Applied Mathematics, 145(1), 11–21. https://doi.org/10.1016/j.dam.2003.09.004

Alinsunurin, M. K. (2014). Efficiency of microfinance institutions in the Philippines. Enterprise Development and Microfinance, 25(4), 311–324. https://doi.org/10.3362/1755-1986.2014.031

Ambühl, C., Mastrolilli, M., & Svensson, O. (2011). Inapproximability results for maximum edge biclique, minimum linear arrangement, and sparsest cut. SIAM Journal on Computing, 40(2), 567–596. https://doi.org/10.1137/080729256

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine, 28(25), 3083–3107. https://doi.org/10.1002/sim.3697

Baltagi, B. H., & Song, S. H. (2006). Unbalanced panel data: A survey. Statistical Papers, 47(4), 493–523. https://doi.org/10.1007/s00362-006-0304-0

Banerjee, A., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. American Economic Journal: Applied Economics, 7(1), 1–21. https://doi.org/10.1257/app.20140287

Bauchet, J., & Morduch, J. (2010). Selective knowledge: Reporting biases in microfinance data. Perspectives on Global Development and Technology, 9(3–4), 240–269. https://doi.org/10.1163/156914910x499705

Björklund, A., Husfeldt, T., Kaski, P., & Koivisto, M. (2007). Fourier meets Möbius: Fast subset convolution. In Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing (pp. 67–74). ACM. https://doi.org/10.1145/1250790.1250801

Bogan, V. L. (2012). Capital structure and sustainability: An empirical study of microfinance institutions. Review of Economics and Statistics, 94(4), 1045–1058. https://doi.org/10.1162/rest_a_00223

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. Review of Financial Studies, 5(4), 553–580. https://doi.org/10.1093/rfs/5.4.553

Caudill, S. B., Gropper, D. M., & Hartarska, V. (2009). Which microfinance institutions are becoming more cost effective with time? Evidence from a mixture model. Journal of Money, Credit and Banking, 41(4), 651–672. https://doi.org/10.1111/j.1538-4616.2009.00226.x

Chikalipah, S. (2017). Financial sustainability of microfinance institutions in sub-Saharan Africa: Evidence from GMM estimates. Enterprise Development and Microfinance, 28(3), 182–199. https://doi.org/10.3362/1755-1986.16-00023

Chikalipah, S. (2019). Financial leverage and profitability of microfinance institutions in sub-Saharan Africa. Enterprise Development and Microfinance, 30(1), 4–21. https://doi.org/10.3362/1755-1986.18-00009

Claessen, K., & Hughes, J. (2000). QuickCheck: A lightweight tool for random testing of Haskell programs. In Proceedings of the Fifth ACM SIGPLAN International Conference on Functional Programming (pp. 268–279). ACM. https://doi.org/10.1145/351240.351266

Cull, R., Demirgüç-Kunt, A., & Morduch, J. (2007). Financial performance and outreach: A global analysis of leading microbanks. The Economic Journal, 117(517), F107–F133. https://doi.org/10.1111/j.1468-0297.2007.02017.x

Cull, R., Demirgüç-Kunt, A., & Morduch, J. (2009). Microfinance meets the market. Journal of Economic Perspectives, 23(1), 167–192. https://doi.org/10.1257/jep.23.1.167

Dawande, M., Keskinocak, P., Swaminathan, J. M., & Tayur, S. (2001). On bipartite and multipartite clique problems. Journal of Algorithms, 41(2), 388–403. https://doi.org/10.1006/jagm.2001.1199

D'Espallier, B., Guérin, I., & Mersland, R. (2011). Women and repayment in microfinance: A global analysis. World Development, 39(5), 758–772. https://doi.org/10.1016/j.worlddev.2010.10.008

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Springer. https://doi.org/10.1007/978-1-4899-4541-9

Elton, E. J., Gruber, M. J., & Blake, C. R. (1996). Survivor bias and mutual fund performance. Review of Financial Studies, 9(4), 1097–1120. https://doi.org/10.1093/rfs/9.4.1097

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. Statistical Science, 19(4), 676–685. https://doi.org/10.1214/088342304000000396

Ganter, B., & Wille, R. (1999). Formal concept analysis: Mathematical foundations. Springer. https://doi.org/10.1007/978-3-642-59830-2

Gutiérrez-Nieto, B., Serrano-Cinca, C., & Mar Molinero, C. (2007). Microfinance institutions and efficiency. Omega, 35(2), 131–142. https://doi.org/10.1016/j.omega.2005.04.001

Gutiérrez-Nieto, B., Serrano-Cinca, C., & Mar Molinero, C. (2009). Social efficiency in microfinance institutions. Journal of the Operational Research Society, 60(1), 104–119. https://doi.org/10.1057/palgrave.jors.2602527

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hartarska, V. (2005). Governance and performance of microfinance institutions in Central and Eastern Europe and the Newly Independent States. World Development, 33(10), 1627–1643. https://doi.org/10.1016/j.worlddev.2005.06.001

Hartarska, V., & Nadolnyak, D. (2007). Do regulated microfinance institutions achieve better sustainability and outreach? Cross-country evidence. Applied Economics, 39(10), 1207–1222. https://doi.org/10.1080/00036840500461840

Hermes, N., Lensink, R., & Meesters, A. (2011). Outreach and efficiency of microfinance institutions. World Development, 39(6), 938–948. https://doi.org/10.1016/j.worlddev.2009.10.018

Hochbaum, D. S. (1998). Approximating clique and biclique problems. Journal of Algorithms, 29(1), 174–200. https://doi.org/10.1006/jagm.1998.0964

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kennes, R. (1992). Computational aspects of the Mobius transformation of graphs. IEEE Transactions on Systems, Man, and Cybernetics, 22(2), 201–223. https://doi.org/10.1109/21.148425

MacIver, D. R., Hatfield-Dodds, Z., & Many Other Contributors. (2019). Hypothesis: A new approach to property-based testing. Journal of Open Source Software, 4(43), 1891. https://doi.org/10.21105/joss.01891

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(1), 24–45. https://doi.org/10.1109/TCBB.2004.2

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. American Economic Journal: Applied Economics, 11(1), 57–91. https://doi.org/10.1257/app.20170299

Mersland, R., & Strøm, R. Ø. (2009). Performance and governance in microfinance institutions. Journal of Banking & Finance, 33(4), 662–669. https://doi.org/10.1016/j.jbankfin.2008.11.009

Mersland, R., & Strøm, R. Ø. (2010). Microfinance mission drift? World Development, 38(1), 28–36. https://doi.org/10.1016/j.worlddev.2009.05.006

MIX Market. (2020). MIX Market [Data set]. World Bank Data Catalog. https://datacatalog.worldbank.org/search/dataset/0038647/mix-market

Morduch, J. (1999). The microfinance promise. Journal of Economic Literature, 37(4), 1569–1614. https://doi.org/10.1257/jel.37.4.1569

Peeters, R. (2003). The maximum edge biclique problem is NP-complete. Discrete Applied Mathematics, 131(3), 651–654. https://doi.org/10.1016/S0166-218X(03)00333-0

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Ranjani, K. S., & Singh, P. (2019). Reciprocity, performance, and training expenditure in microfinance institutions: Evidence from South Asia. Enterprise Development and Microfinance, 30(2), 97–116. https://doi.org/10.3362/1755-1986.17-00026

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. PLoS Computational Biology, 9(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Verbeek, M., & Nijman, T. (1992). Testing for selectivity bias in panel data models. International Economic Review, 33(3), 681–703. https://doi.org/10.2307/2527133

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLOS Computational Biology, 13(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. Journal of Econometrics, 68(1), 115–132. https://doi.org/10.1016/0304-4076(94)01645-G