Patterns of Nonresponse and Imputation in the IFO Business Survey

Chunyan Jiang
Qian Feng

January 20, 2026

Outline

1 Introduction
- Background
- Definition
- Data Overview
- Two Main Questions
- Task

2 Descriptive Analysis

3 Quantitative Analysis

4 Imputation

5 Summary

Background

- Monthly survey of German firms on their current situation, recent developments, and their plans and expectations for the near future.
- 2 Core questions: current situation (BS) and business expectations (BE).
- BS and BE form the ifo Business Climate Index - a key leading indicator for the German economy and financial markets.
- Participation is voluntary: unit nonresponse and item nonresponse can occur.

- **Unit nonresponse**: A respondent does not participate in the survey at all; the entire questionnaire is missing.
  - Example: a firm does not respond to the monthly survey.
- **Item nonresponse**: A respondent participates but leaves one or more questions unanswered.
  - Example: a firm returns the questionnaire but skips the question on expectations (BE).

Data Overview

Dataset

- Monthly panel of manufacturing firms (1991–2024)
- $\sim$8,200 firms and $\sim$1,065,000 observations

Key Variables Used in Analysis

- *Time*: calendar time, month indicators (August, December)
- *Firm characteristics*: region (west/east), sector
- *Participation behavior*: participation number, participation length
- *Main survey outcomes*:
  - BS (Business Situation)
  - BE (Business Expectations)
- *Derived variables*: 6-month average BS, 6-month average BE

Two Main Questions

Business Situation (BS)

We characterize our **current** business situation as:

- 1: good
- 2: satisfactory
- 3: poor

Business Expectation (BE)

We **expect** our business situation to:

- 1: become more favorable
- 2: remain roughly the same
- 3: become less favorable

Task

- Conduct descriptive analyses of unit nonresponse.
- Conduct quantitative analyses of unit nonresponse.
- Compare and evaluate different imputation methods for the two core survey variables, BS (Business Situation) and BE (Business Expectations).

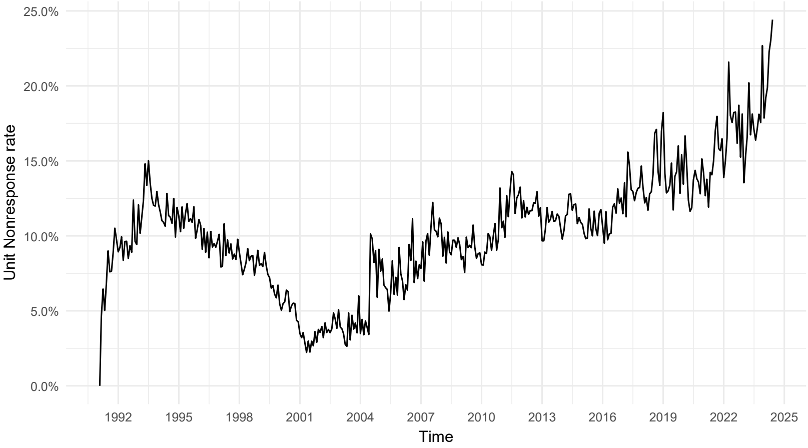
Outline

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

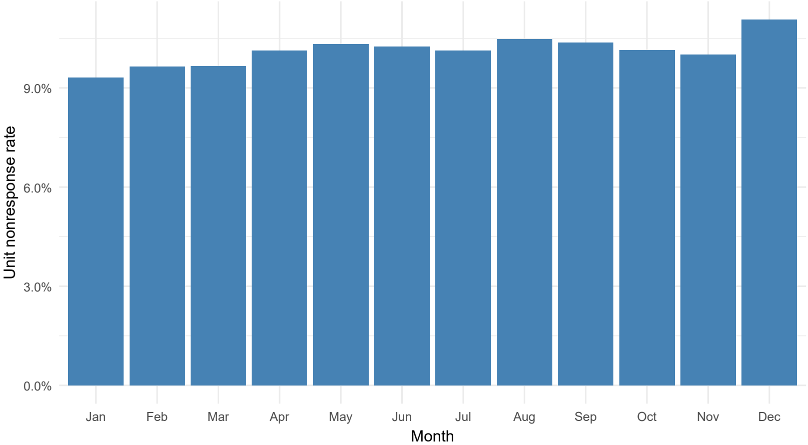
1. Introduction
2. Descriptive Analysis
   - Time-Related Variables
   - Firm Characteristics
   - Survey-Related Variables
   - Two Main Questions
3. Quantitative Analysis
4. Imputation
5. Summary
6. Extension

Time-Related Variables

Unit Non-response by Time



Unit Non-response by Month



Summary

Time Patterns

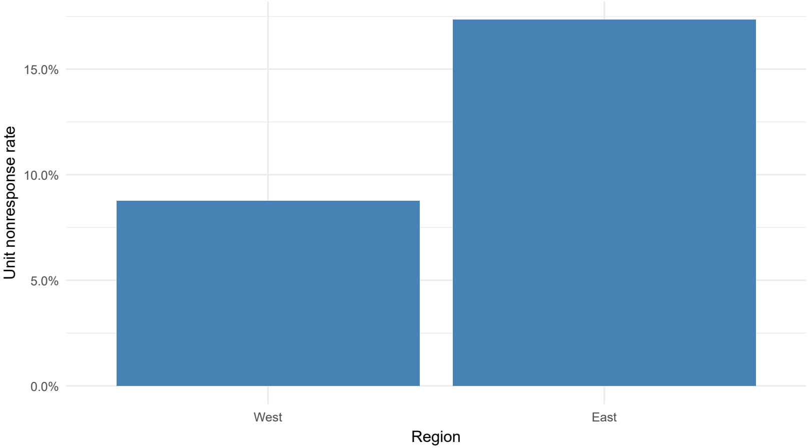
- Unit nonresponse varies over time and may follow a **nonlinear pattern**.
- August and December show **slightly higher** unit nonresponse rates.

Next Steps

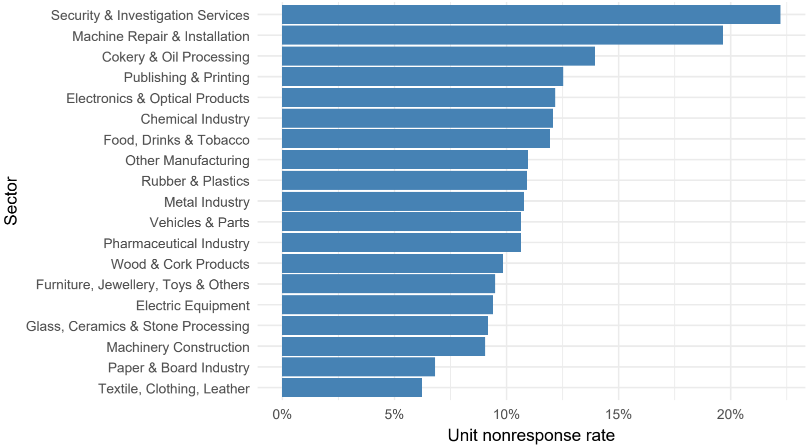
- A formal statistical model is needed to test whether these observed differences are **statistically significant**.

Firm Characteristics

Unit Non-response by Region



Unit Non-response by Sector



Summary

Key Finding

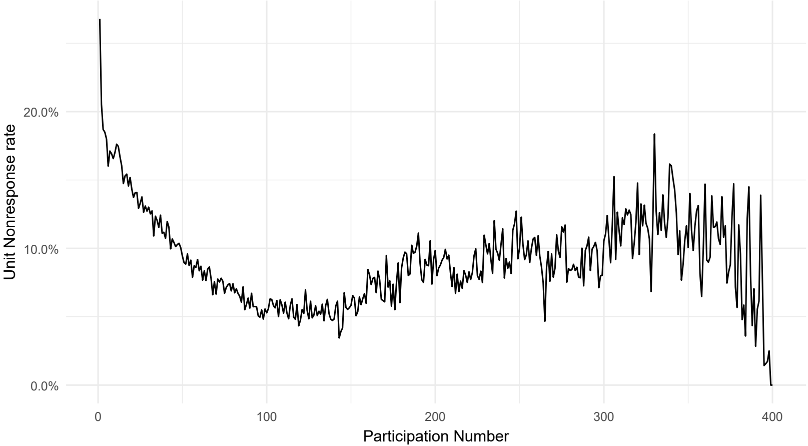
- The unit nonresponse rate exhibits variation across different geographical regions and across different sectors within the manufacturing industry.

Next Steps

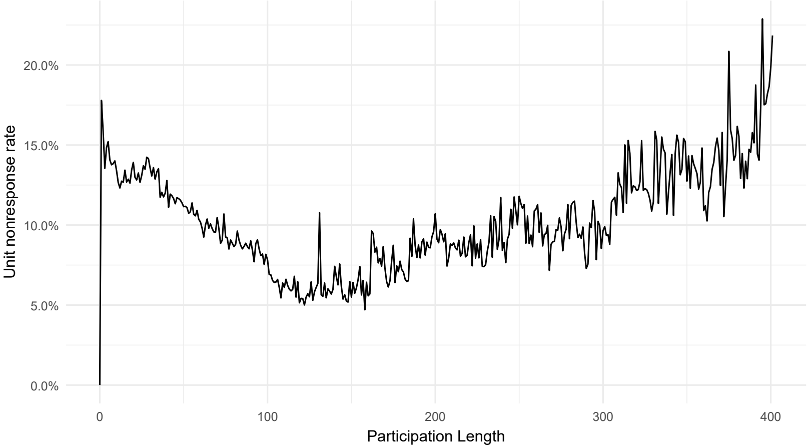
- These differences should be further examined using quantitative models to quantify their **statistical significance**.

Survey-Related Variables

Participation Number: The total number of times a company has participated in the survey up to the current period.



Participation Length: The number of months that have elapsed since a company first participated in the survey.



Summary

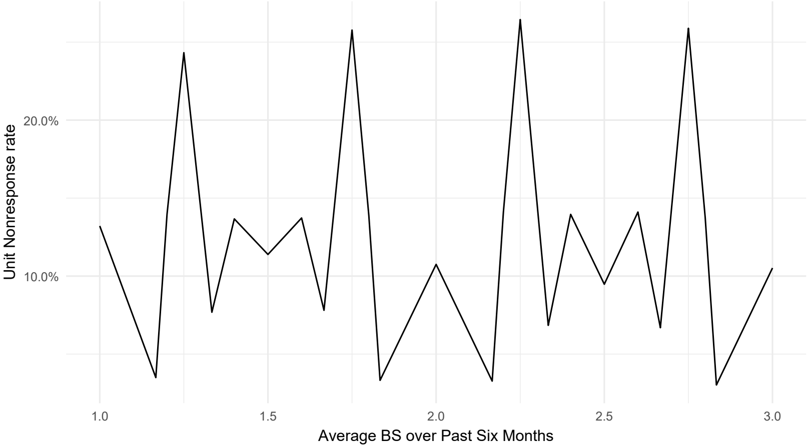
Key Findings

- The relationship between participation number and unit nonresponse is **nonlinear**.
- The relationship between participation length and unit nonresponse is **nonlinear**.

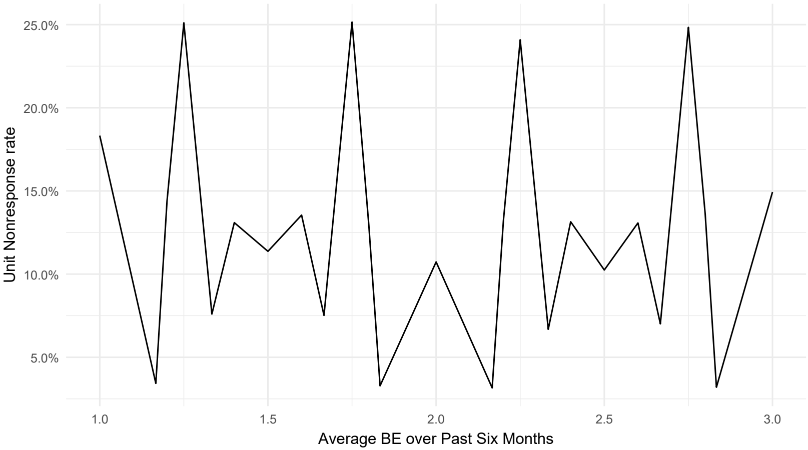
Next Steps

- The nonlinear pattern need to be evaluated within a **quantitative modeling framework**.

Two Main Questions


Unit Nonresponse rate
20.0%
10.0%
1.0
1.5
2.0
2.5
3.0
Average BS over Past Six Months

Unit Non-response by Average BE past 6 months



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Key Findings

- The relationship between average Business Situation (BS) over the past six months and unit nonresponse is **nonlinear**.
- The relationship between average Business Expectation (BE) over the past six months and unit nonresponse is **nonlinear**.

Next Steps

- These relationships should be examined within a **quantitative model** to assess statistical significance and potential nonlinear effects.

Outline

1. Introduction
2. Descriptive Analysis
3. Quantitative Analysis
   - Overview
   - Cluster-robust GAM
   - Model
   - Alternative
4. Imputation
5. Summary
6. Extension

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- **Target variable**: Whether the observation is a unit nonresponse (1 = unit nonresponse, 0 = no unit nonresponse).
- **Clustering variable**: Company ID.
- **Covariates**:
  - Region (West / East)
  - Sector
  - Indicator for December
  - Indicator for August
  - Calendar time
  - Participation number
  - Participation length
  - Average BS over the past 6 months
  - Average BE over the past 6 months
- **Model**: Cluster-robust GAM.

Cluster-robust GAM

- Suitable for repeated observations per company
- Apply a sandwich estimator (`vcovCL`) to a standard GAM
- Standard-error adjustment for within-company correlation
- Point estimates unchanged

Model Specification

$$P = P(\text{Unit Nonresponse} = 1 \mid X)$$

$$\begin{aligned}
\text{logit}(P) = \beta_0 &+ \beta_1 \,\text{August} + \beta_2 \,\text{December} \\
&+ \beta_3 \,\text{East} + \boldsymbol{\beta}_S^{\mathsf{T}} \text{Sector} \\
&+ s(\text{Calendar Time}) \\
&+ s(\text{Participation number}) + s(\text{Participation length}) \\
&+ s(\text{Average Business Situation (past 6 months)}) \\
&+ s(\text{Average Business Expectations (past 6 months)})
\end{aligned}$$

$s(\cdot)$ denotes penalized spline smooth functions.
East $= 1$ for East region (reference: West).

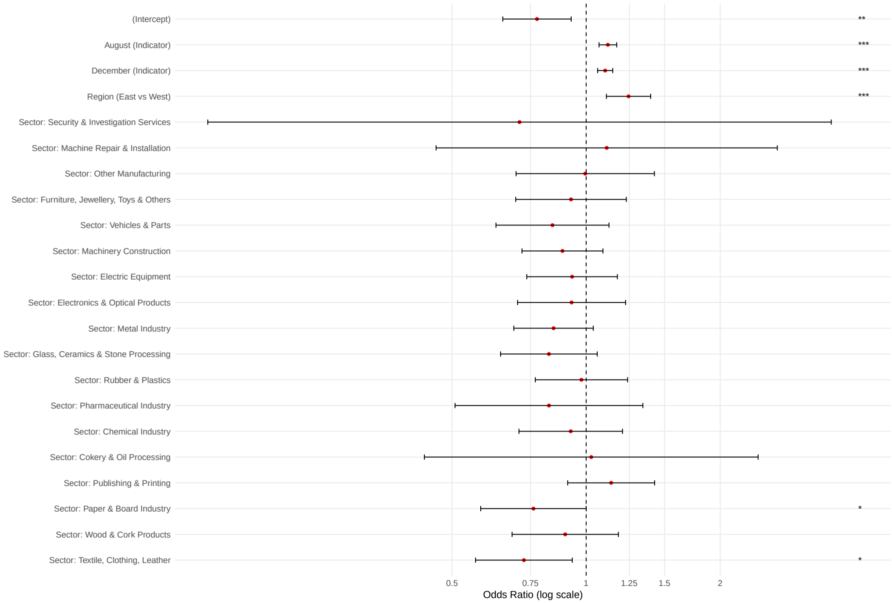
Sector is a categorical variable with the reference category: Food, Drinks & Tobacco.

Model Fit

- Influenced by many unobserved and random factors

Model fit	**Value**
Deviance explained	19.3%

Parameter


(Intercept)
August (Indicator)
December (Indicator)
Region (East vs West)
Sector: Security & Investigation Services
Sector: Machine Repair & Installation
Sector: Other Manufacturing
Sector: Furniture, Jewellery, Toys & Others
Sector: Vehicles & Parts
Sector: Machinery Construction
Sector: Electric Equipment
Sector: Electronics & Optical Products
Sector: Metal Industry
Sector: Glass, Ceramics & Stone Processing
Sector: Rubber & Plastics
Sector: Pharmaceutical Industry
Sector: Chemical Industry
Sector: Cokery & Oil Processing
Sector: Publishing & Printing
Sector: Paper & Board Industry
Sector: Wood & Cork Products
Sector: Textile, Clothing, Leather
**
***
***
***
*
*
0.5
0.75
1
1.25
1.5
2
Odds Ratio (log scale)

Interpretation

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- **Geography:** Eastern firms show higher nonresponse.
  - Smaller firms, lower administrative capacity, or different attitudes toward surveys.
- **Seasonality:** Peaks in December and August.
  - Year-end workload, Christmas holiday and summer holidays reduce survey participation.

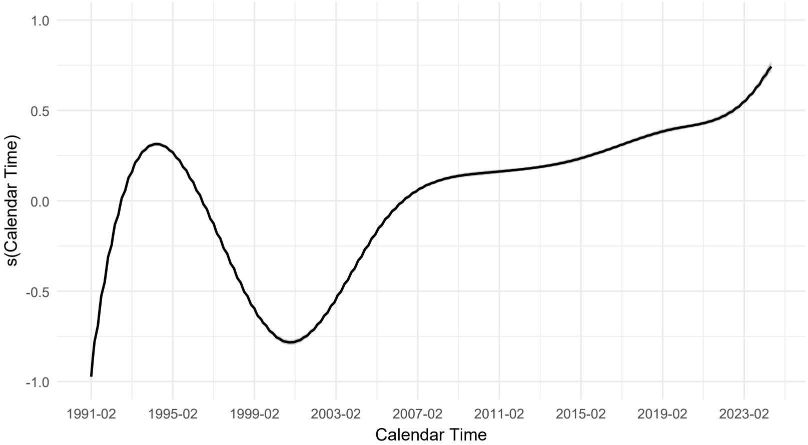
- Most other sectors show no significant difference from the reference sector
- Textile, clothing, leather and Paper & Board Industry exhibit significantly lower odds of nonresponse than Food, drinks and tobacco Industry

Approximate significance of smooth terms

- Strong nonlinear effects.

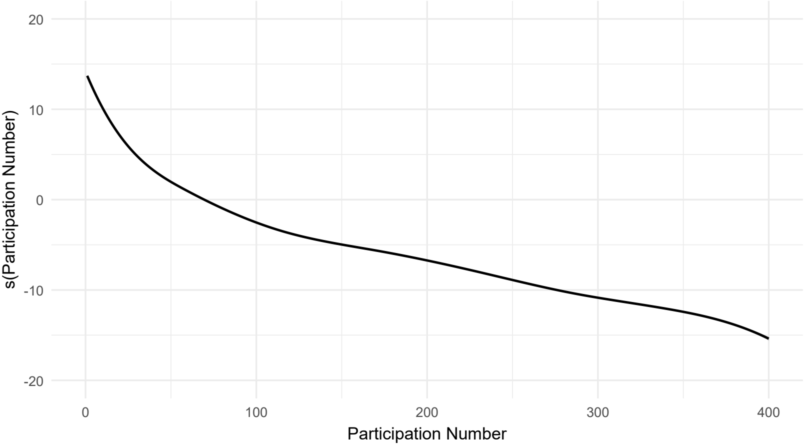
Smooth term	edf	Ref.df	Chi.sq	p-value
s(Calendar Time)	8.999	9	34 763	<2e–16 ***
s(Participation Number)	8.993	9	210 572	<2e–16 ***
s(Participation Length)	8.997	9	189 931	<2e–16 ***
s(Average BS over Past Six Months)	8.999	9	2 216	<2e–16 ***
s(Average BE over Past Six Months)	8.999	9	3 408	<2e–16 ***

Spline of Calendar Time

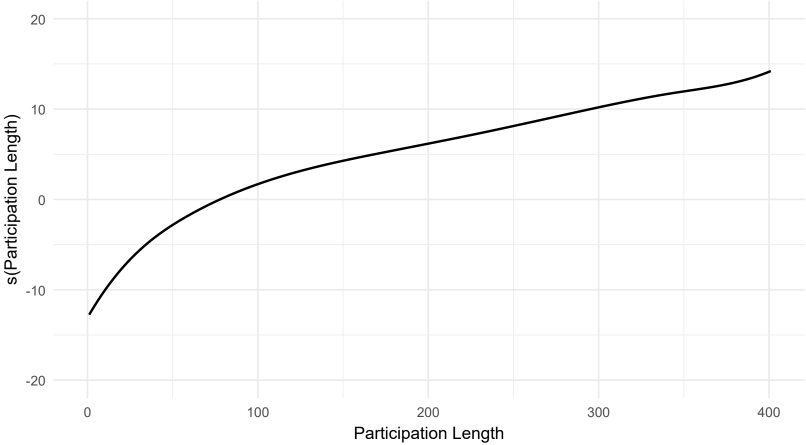


Spline of Participation Number

- More past participations $\rightarrow$ lower nonresponse risk
- More participation $\rightarrow$ greater familiarity $\rightarrow$ higher response willingness.

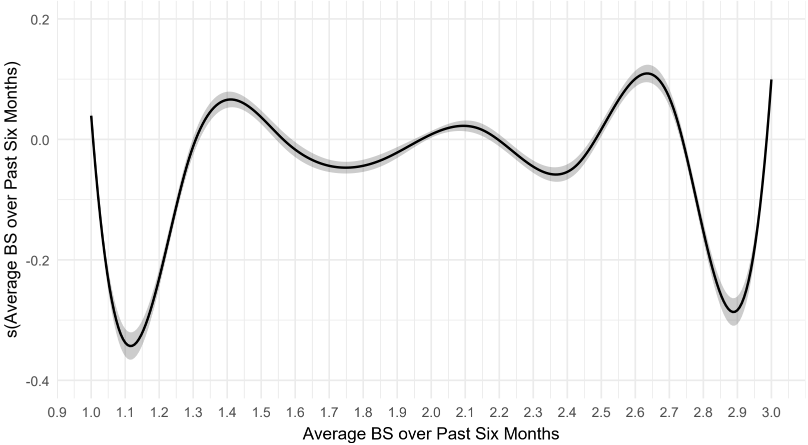


- Longer participation history → higher risk of unit nonresponse
- Survey fatigue



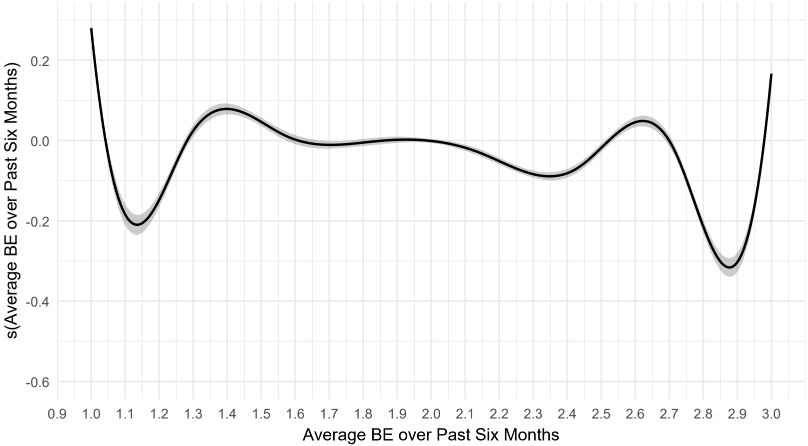
Spline of Average BS over Past Six Months

- Higher response likelihood when business conditions are close to "Good" or "Bad"
- Motivation driven by sharing success or signaling distress



Spline of Average BE over Past Six Months

- Higher response likelihood when business conditions are close to "Becoming more favorable" or "Becoming less favorable"
- Motivation driven by sharing optimistic outlook or signaling upcoming challenges



Alternative Modelling

1. **Model Formulation**
   - **Logistic regression:**
     $$p_{it} = P(y_{it} = 1 \mid X_{it})$$
   - **Discrete-time recurrent-event hazard model:**
     $$h_{itr} = P(Y_{itr} = 1 \mid \text{at risk for event } r,\ \text{history},\ X_{it})$$
     $$\text{logit}(h_{itr}) = \alpha_r(t) + X_{it}^\top \beta + u_i, \qquad u_i \sim N(0, \sigma_u^2)$$

2. **Interpretation (p vs. hazard)**
   - p_{it}: probability of being in the event state (e.g., missing) at time t.
   - h_{itr}: probability that the r-th event occurs at time t, given it has not yet occurred (hazard).

3. **Modelling within-individual correlation**
   - **Logistic regression:** does not structurally model correlation; cluster-robust SE typically used.
   - **Hazard model:** correlation modelled via a random effect (frailty) u_i.

Outline

1 Introduction

2 Descriptive Analysis

3 Quantitative Analysis

4 Imputation
- Methodology
- BS
- BE
- Conclusion

5 Summary

6 Extension

- **LOCF (Last Observation Carried Forward)**
  - Imputes missing values by carrying forward the most recent observed value of the same firm.
  - Simple and fast, but cannot capture changes in firms' business conditions.
- **Markov Chain**
  - Uses observed transition probabilities between BS/BE states to impute the next value.
  - In our analysis we apply a **homogeneous** Markov Chain — transition probabilities are identical across firms and time, whereas Random Forest introduces **heterogeneity** through covariates.
- **Random Forest**
  - Imputes missing BS/BE using covariates.
  - Introduces heterogeneity through covariate-dependent predictions.

- **Target variables**: Business Situation (BS) and Business Expectations (BE).
- **Covariates**:
  - Company ID
  - Region (West / East)
  - Sector
  - Online vs. Offline questionnaire
  - Indicator for August
  - Indicator for December
  - Lagged BS values (lags 1–4)
  - Lagged BE values (lags 1–4)
  - Calendar time
- **Model tuning**:
  - We tune the Random Forest via a **grid search** over three hyperparameters:
    - `mtry`: number of variables sampled at each split (candidates: $\sqrt{p}$, $p/3$, $p/2$)
    - `ntree`: number of trees (candidates: 200, 350, 500)
    - `nodesize`: minimum terminal node size (candidates: 1, 3)

- **Motivation for Simulation**
  - In both BS and BE, **1–6-period consecutive missingness** accounts for more than **95%** of all missing patterns.
  - Therefore, our simulation focuses on replicating these dominant consecutive-missing structures.
- **Simulated Missingness Patterns**
  - Data is restricted to 2014-2024 in order to reduce simulation time.
  - For each gap length $k = 1, \ldots, 6$, we randomly select **10%** of non-missing observations.
  - For each selected firm, the corresponding BS (or BE) values are masked to NA for k **consecutive periods**.
- **Repetition**
  - Each simulation setting is repeated **30 times** to ensure stable and robust evaluation.

Evaluation Metrics

- Compute Accuracy, Cohen's κ, and Spearman's ρ for each of 30 imputed datasets

Calibration

- Select optimal imputation methods for BS and BE based on evaluation metrics
- Check if predicted class probabilities match observed frequencies

BS/BE Balance

- Use optimal methods to impute actual missing BS and BE values
- Aggregate original and imputed values to compute balances:
  - Balance of BS = (BS = "good" - BS = "bad") / Total
  - Balance of BE = (BE = "favourable" - BE = "unfavourable") / Total
- Compare balances to assess impact of imputation

Business Situation Imputation

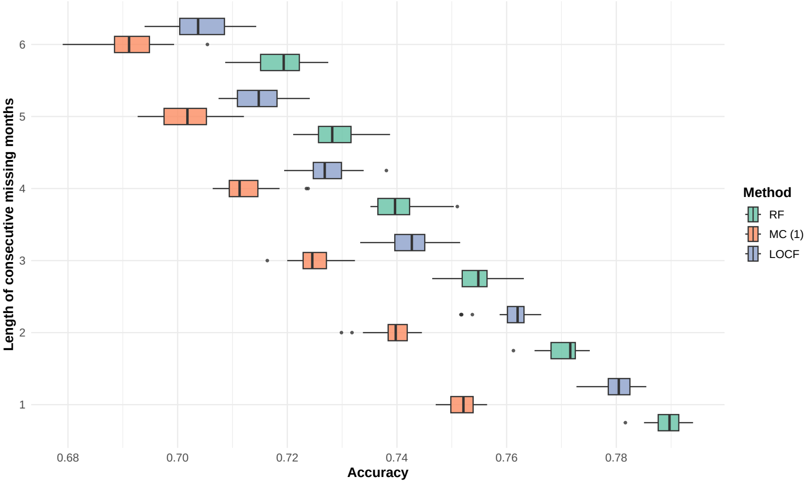
Evaluation Metric – Accuracy

- Accuracy measures the proportion of correctly classified observations.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\hat{y}_i = y_i),$$

where $\hat{y}_i$ and y_i denote the predicted and true class labels for observation i.

- Random Forest achieves the highest accuracy across all k.



Evaluation Metric – Kappa

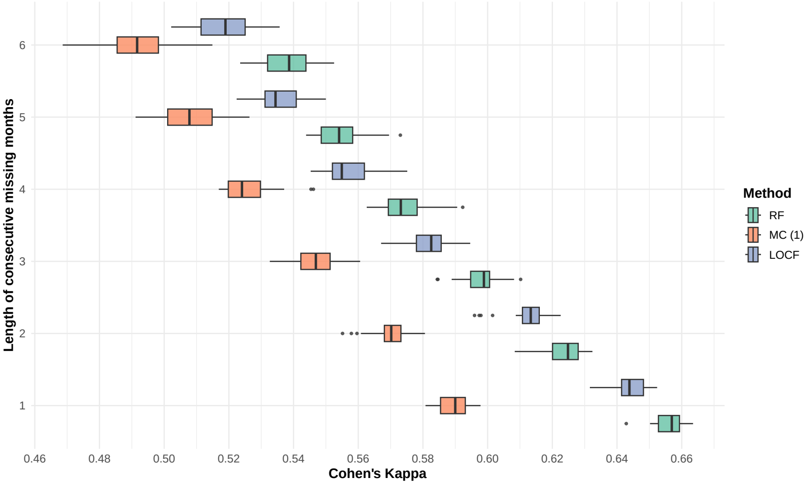
- Cohen's κ measures agreement beyond chance.
- Higher κ indicates better predictive agreement.

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement and p_e is the expected agreement under random guessing.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Random Forest achieves the highest κ values across all k.



Evaluation Metric – Rho

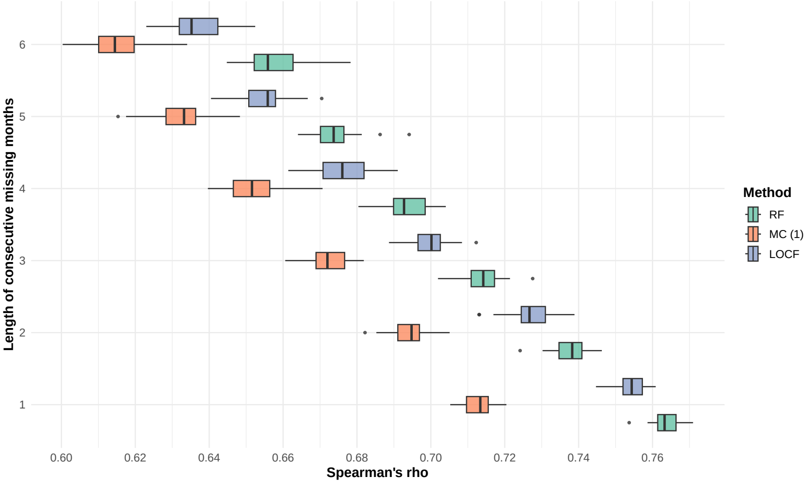
- Spearman's ρ measures rank correlation between predicted and true categories.
- Higher ρ indicates stronger monotonic agreement.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where $d_i = \mathrm{rank}(\hat{y}_i) - \mathrm{rank}(y_i)$ and n is the sample size.

Evaluation Metric – Rho

- Random Forest achieves the highest ρ values across all k.



What is Calibration?

- Calibration assesses whether predicted probabilities reflect true outcome frequencies.
- A model is well calibrated if, among all observations with predicted probability p, the outcome occurs approximately p proportion of the time.
- Assesses whether the predicted probabilities produced by the imputation model are trustworthy

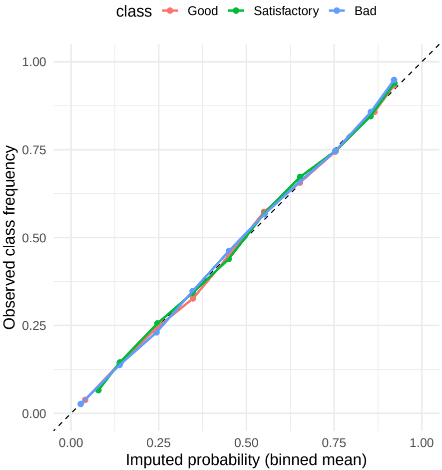
Formally, for class c, calibration evaluates whether

$$\mathbb{E}[\mathbf{1}(y = c) \mid \hat{p}_c = p] \approx p,$$

i.e., whether observed class frequencies match the predicted probabilities.

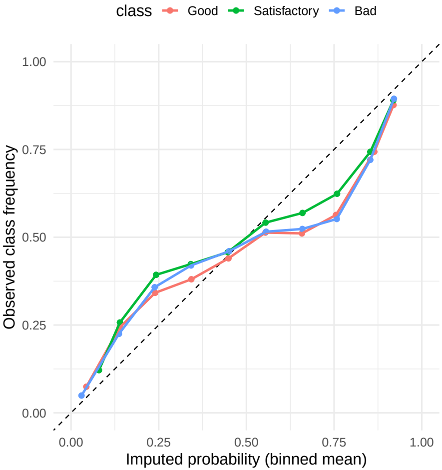
Calibration (Random Forest) k = 1

- Well calibrated
- No preference for any particular class



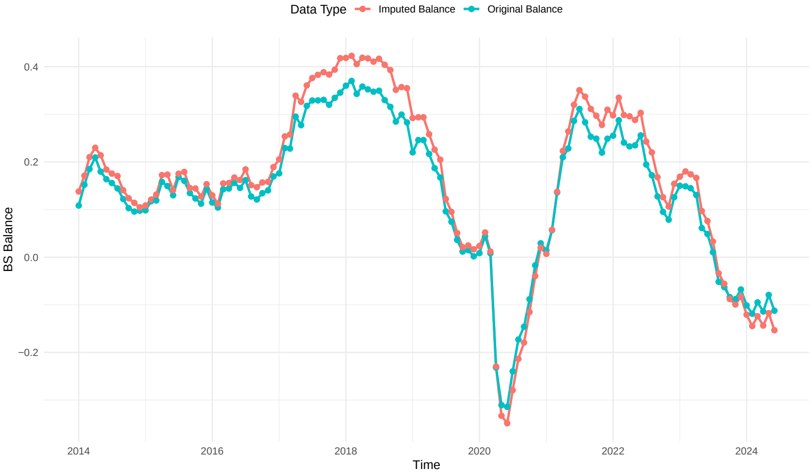
Calibration (Random Forest) k = 6

- Underestimation at low predictions; overestimation at high predictions.
- Class Satisfactory is best calibrated in the high-prediction region.



Balance (Random Forest)

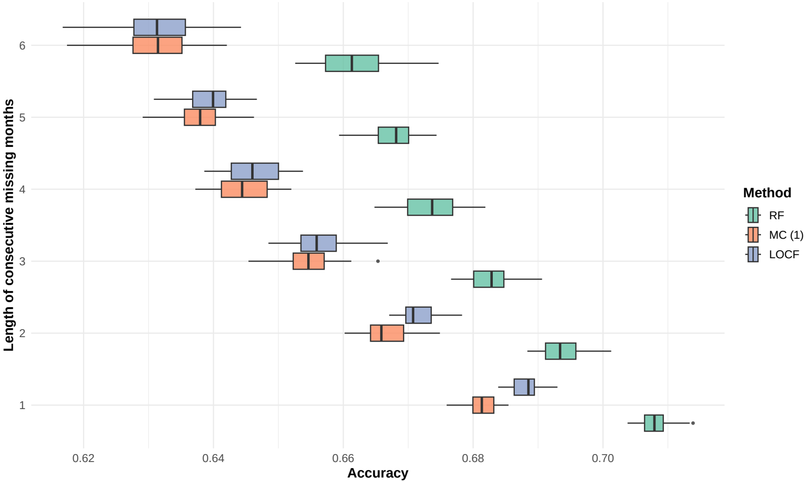
- Imputed balances tend to be higher than balances computed from observed data for much of the sample period, with more pronounced differences in 2017–2019 and mid-2021 to mid-2022.



Business Expectation Imputation

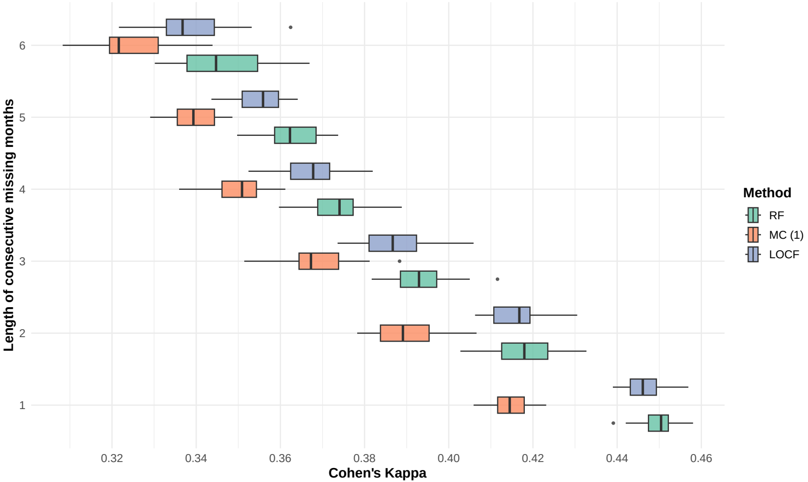
Evaluation Metric - Accuracy

- Random Forest achieves the highest accuracy across all k.

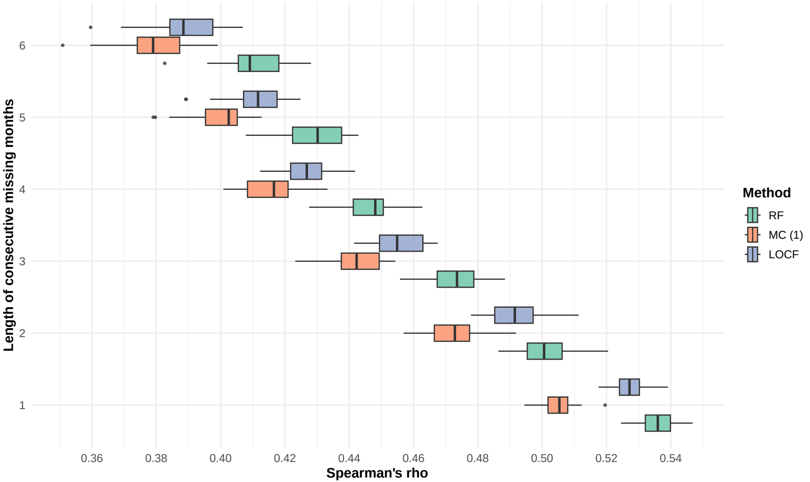


Evaluation Metric - Kappa

- Random Forest achieves the highest κ values across all k.

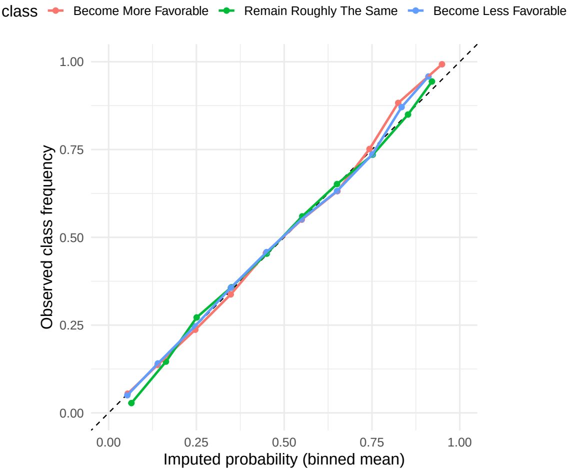


- Random Forest achieves the highest ρ values across all k.



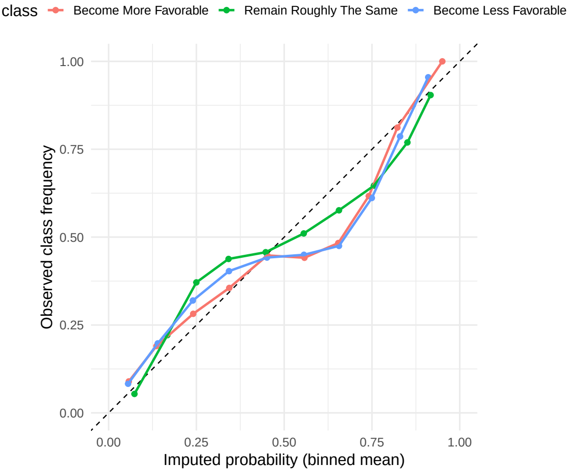
Calibration (Random Forest) k = 1

- Well calibrated
- No preference for any particular class



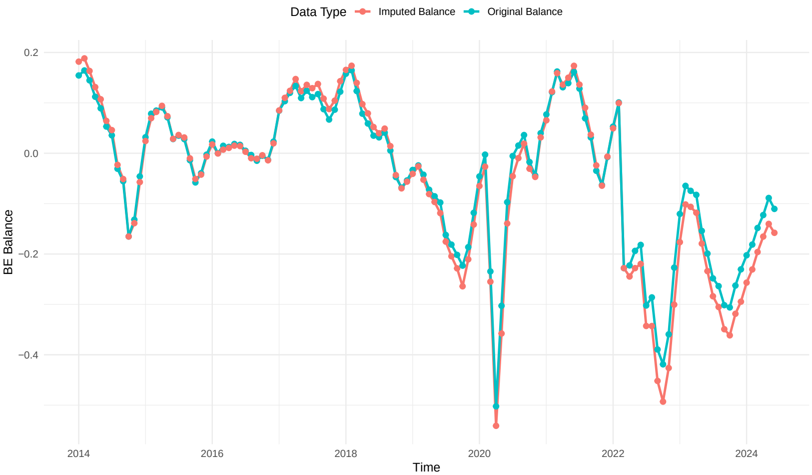
Calibration (Random Forest) $k = 6$

- Low predictions: underestimation; high predictions: overestimation.
- Best calibrated: class Become More Favourable (low-prediction region), class Remain Roughly The Same (high-prediction region).



Balance (Random Forest)

- After 2022, imputed BE balances are lower than balances computed from observed data.



Conclusion

Performance Trends

- Imputation performance deteriorates with longer consecutive missingness

BS vs. BE Comparison

- BE imputations perform worse than BS in simulation
- Expectations are harder to predict than current conditions

Method Comparison

- Markov Chain performs worst among all methods
- Random Forest outperforms LOCF

Outline

1. Introduction
2. Descriptive Analysis
3. Quantitative Analysis
4. Imputation
5. Summary
6. Extension

- Unit nonresponse shows systematic temporal, regional, and behavioral patterns.
- All continuous covariates exhibit nonlinear effects.
- Random Forest consistently outperforms LOCF and homogeneous Markov Chain across all evaluation metrics.
- Imputation performance deteriorates with longer consecutive missingness.

Limitations

- Neither the quantitative analysis nor the Random Forest imputation incorporates external macroeconomic factors.
- A homogeneous Markov Chain model is used, whereas Random Forest allows heterogeneous, covariate-dependent transitions.
- The simulation missingness mechanism is MCAR-like, whereas the quantitative analysis reflects MAR tendencies. This inconsistency may affect how well the simulation results generalize to the real missing-data mechanism.

Outlook

- Incorporate external macroeconomic indicators into both the nonresponse modeling and the imputation framework to improve interpretability and imputation performance.
- Extend the Markov Chain approach to heterogeneous or covariate-dependent transitions, making it more comparable to flexible machine-learning models such as Random Forest.
- Introduce simulation studies under MAR mechanisms to better reflect the missing-data patterns observed in real survey participation behavior.

Outline

1. Introduction
2. Descriptive Analysis
3. Quantitative Analysis
4. Imputation
5. Summary
6. Extension

Notation

$R_{it} = 1$ if outcome is observed, $R_{it} = 0$ if missing.
X_{it}: observed covariates; Y_{it}: outcome; $Y_{\text{obs}}/Y_{\text{mis}}$: observed/missing parts of Y.

- **MCAR (Missing Completely at Random):**

$$P(R \mid X, Y) = P(R)$$

- **MAR (Missing at Random):**

$$P(R \mid X, Y_{\text{obs}}, Y_{\text{mis}}) = P(R \mid X, Y_{\text{obs}})$$

- **MNAR (Missing Not at Random):**

$$P(R \mid X, Y_{\text{obs}}, Y_{\text{mis}}) \neq P(R \mid X, Y_{\text{obs}})$$

We model the response indicator R_{it} using observable covariates and past outcomes. Let H_{it} denote observable history (e.g., $Y_{i,t-1}, \ldots, Y_{i,t-L}$).

$$\operatorname{logit} P(R_{it} = 1 \mid X_{it}, H_{it}) = \eta(X_{it}, H_{it})$$

MAR restriction: the current outcome Y_{it} is not included.

To match an overall missing rate of 10%, we add an intercept shift:

$$\operatorname{logit} P(R_{it} = 1 \mid X_{it}, H_{it}) = \alpha + \eta(X_{it}, H_{it}).$$

Missingness is generated by $R_{it} \sim \text{Bernoulli}(P(R_{it} = 1))$.

The logit link is used for convenience; $f(\cdot)$ may be nonparametric. An intercept shift α allows simple calibration of the missing rate.

Limitations of MAR-Based Simulation

(1) Class imbalance in the missingness model
The target variable R_{it} (missing vs. observed) is highly unbalanced (approximately 16% missing). As a result, it is difficult to estimate a response model with strong and reliable predictive performance, which may weaken the realism of the simulated MAR mechanism.

(2) No direct control of continuous missingness
The MAR mechanism operates at the observation level and does not directly generate consecutive missing periods.

Takeaway: Introducing MAR improves upon MCAR, but additional modeling is needed to capture missingness patterns.

Thanks for your attention!