Master's Thesis
Imputation of Panel Data

Submitted by
Tipu Sultan
Email: s4tisult@uni-trier.de
Matriculation Number: 1599320
Degree: M.Sc. in Applied Statistics
Address: Pommernweg 71, 89075, Ulm

Universität Trier, 1-Fach

Supervisors
Prof. Dr. Ralf Münnich
M.Sc. Jan Weymeirsch
Submission: January 06, 2025

Table of Contents

List of Figures

List of Tables

List of Symbols and Acronyms

Table 1: Symbols

Symbol	Explanation
α	Intercept
β_0	Intercept
β	Coefficients of explanatory variables
$\widehat{\beta}$	Pooled OLS estimated coefficient
$\tilde{\beta}$	Fixed Effects model estimated coefficient
$\hat{\beta}_{GLS}$	Random Effects model estimated coefficient using Generalized Least Squares (GLS)
u	Disturbance term ($u_{it} = \mu_i + \nu_{it}$)
ϵ	Error term
μ	Individual-specific effects not included in the model (Panel data model)
ν	Standard disturbance term (Panel data model)
μ	Probability distribution (Wasserstein distance)
ν	Probability distribution(Wasserstein distance)
Q	Transformation matrix
I_N	Identity matrix for individuals
I_T	Identity matrix for time periods
$\otimes$	Kronecker product
σ_ν^2	Variance of the idiosyncratic error term
σ_μ^2	Variance of individual-specific effects
$E(uu')$	Covariance matrix of the error term
$\text{cov}(u_{it}, u_{js})$	Covariance between error terms for different entities and time periods

Table 2: Acronyms

Acronym	Explanation
amelia	R package to Impute Missing Values
amelia II	R package to Impute Missing Values
CI	Confidence Interval

Continued on next page

Continued from previous page

Acronym	Explanation
cs	Cross-Sectional Identifier
DNR	Differential Non-Response
EM	Expectation-Maximization
E	Education (column name in the table of correlation)
EH	Employment Hours (column name in the table of correlation)
ET	Employment Types (column name in the table correlation)
FCS	Fully Conditional Specification
FD	First Difference
FE	Fixed Effects
FIML	Full Information Maximum Likelihood
GLS	Generalized Least Squares
GMM	Generalized Method of Moments
GIS	Geographic Information System
GDP	Gross Domestic Product
GRU	Gated Recurrent Unit
IC	Intraclass Correlations
LSTM	Long Short-Term Memory
LSDV	Least Squares Dummy Variables
MAR	Missing at Random
MCAR	Missing Completely at Random
MCSE	Multi-Layer-Security Architecture
MLE	Maximum Likelihood Estimator
MNAR	Missing Not at Random
mice	Multivariate Imputation by Chained Equations (R package)
MI	Multiple Imputation
ML	Maximum Likelihood
mitml	Multiple Imputation for Multilevel Modeling (R package)
NA	Not Available (used for missing values)
OLS	Ordinary Least Squares
pan	Panel Data Imputation (R package)

Continued on next page

Continued from previous page

Acronym	Explanation
PCA	Principal Component Analysis
plm	Panel Linear Models (R package)
pmm	Predictive Mean Matching
Pooled OLS	Method for Estimating Coefficients in Panel Data
R	Programming Language and Environment for Statistical Computing
RE	Random Effects
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SEM	Structural Equation Modeling
S-PLUS	Statistical Software
2SLS	Two-Stage Least Squares
ts	Time-Series Variable

1 Introduction

Panel data, also known as longitudinal data, plays an important role in many sectors, such as clinical trials, financial analysis, surveys and research. Panel data can capture the information of the same individuals over time, which efficiently controls individual-specific effects and dynamic changes. Data missingness is a challenging problem in panel data analysis due to unit-nonresponse, item-nonresponse, by design or subject dropout. These are significant hindrances in panel data study. Because of this loss of information, bias arises, and the model becomes more complex, which leads to invalid conclusions about any research or study. Addressing these issues requires robust imputation frameworks to predict and replace missing values while preserving the inherent characteristics of the dataset. Scholars have tried many different methods and tools to impute the missing values in panel data. However, the research is limited to different missingness scenarios and comparing the frameworks to find the most effective one. This study evaluates various imputation frameworks for balanced and unbalanced panel data, considering different types of missingness (MCAR, MAR, and MNAR) and varying levels (10%, 30%, and 50%) to identify the most effective one.

This study's four main areas of focus are: first, exploring the unique characteristics of panel data, including its correlation structures, benefits, drawbacks, and uses; second, the assessment of models for panel datasets, particularly in determining coefficients and intercepts; third, exploring the mechanisms of missingness, including its causes and methods for handling and detecting it; and fourth, exploring the efficiency and accuracy of different imputation frameworks by comparing Wasserstein-distance, Bias, RMSE, and computational time. The study also looks at the effects of imputation on the distribution, conditional distribution, and correlation of the simulated datasets.

This thesis study focuses on three R packages ("mice", "mitml", and "amelia") and one machine learning architecture ("LSTM-Network") to impute missing values from panel datasets. While the first three are well-established statistical packages for multiple imputations in panel data, the "LSTM-Network" represents an innovative approach. No research has examined "LSTM-Network" for panel data imputation and compared it with the other available packages designed to impute missing values from panel data. This thesis study connects the gap between statistical rigour and computational innovation by comparing traditional techniques with an emerging machine-learning approach and provides the groundwork for future advancements. The structure of the thesis includes a review of relevant literature in Section 2, a description of data in Section 3, a detailed explanation of methodology and simulation in Section 4, results and discussion in Section 5, and an ending conclusion in Section 6.

2 Literature Review

2.1 Panel Data

2.1.1 Theoretical Foundations of Panel Data

According to Baltagi (2008), panel data regression has a double subscript on its variables, different from a cross-sectional or regular time-series regression.

$$y_{it} = \alpha + X_{it}^{T}\beta + u_{it}, \quad i = 1,...,N; \quad t = 1,...,T \tag{1}$$

In equation (1), i represents various entities like households, individuals, firms, or countries, while t indicates time. α is a scalar. β is a matrix. X_{it} denotes observations on explanatory variables. The disturbance is,

$$u_{it} = \mu_i + \nu_{it} \tag{2}$$

In equation (2), μ_i captures unobservable individual-specific effects, and ν_{it} represents remaining disturbances. Let one example be considered. In labour economics, if y_{it} are household head earnings, while X_{it} could be variables like experience, education, etc. The μ_i term is constant over time. It also explains the individual-specific effects that are not included in the model, similar to unobserved abilities. On the other hand, ν_{it} varies across individuals and time, representing standard disturbances. Similarly, in a production function analyzing firms over time, y_{it} signifies output, X_{it} inputs, and μ_i captures unobservable firm-specific effects, such as managerial skills.

Selecting the appropriate model is a crucial stage in panel data analysis to produce accurate and insightful findings. Both the fixed [1] and random [2] effects observed in the data should be explained using panel data models. An analysis may be biased or inaccurate if these effects are not taken into consideration properly. It is not feasible for standard linear models to do so. One simple way to estimate the coefficients is using the pooled OLS (Ordinary Least Squared) method. To estimate the coefficients β in a pooled OLS model, one can minimize the sum of squared residuals. According to Abdullah et al. (2022), Greene (2008) and Gujarati (2003), the OLS estimator relies on five key assumptions. First, the model equation must have linear parameters. Second, the expected value of the error terms (disturbances) must be zero, meaning they

[1] Fixed effect is an unobserved, time-invariant characteristic specific to individuals or groups that can affect the dependent variable in regression models.

[2] Random effect is a special kind of fixed effect which is not correlated with the explanatory variables in regression models.

are not correlated with any of the regressors—a condition known as exogeneity. Third, the error terms should have a constant variance (homoscedasticity) and should not be correlated with each other (non-autocorrelation). Fourth, the observations on the independent variables must be non-stochastic. Lastly, there should be no multicollinearity among the independent variables. The pooled OLS model is a linear regression model applied to panel data without accounting for fixed or random effects. It assumes a constant intercept and identical slope coefficients across all groups and time periods. The pooled OLS equation from Abdullah et al. (2022) is given below,

$$y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_n X_{nit} + \epsilon_{it}, \quad \text{where } u_i = 0 \tag{3}$$

In equation (3), β_0 represents the intercept, $\beta_1\, \beta_2 \ldots \beta_n$ are the coefficients of the independent variables $X_{1it}, X_{2it} \ldots X_{nit}$, and ϵ_{it} is the error term. The assumption $u_i = 0$ implies that individual-specific effects, whether group- or time-specific, are not present.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{4}$$

OLS can produce efficient and consistent parameter estimates when the individual-specific effect (u_i) is truly absent ($u_i = 0$). However, if $u_i \neq 0$, the assumptions of exogeneity, homoscedasticity, and non-autocorrelation may be violated, especially in panel data settings. In such cases, the OLS estimator loses its status as the Best Linear Unbiased Estimator. The Fixed and Random Effects models can address difficulties related to individual-specific effects and take into consideration the structure of panel data. Additionally, advanced techniques like dynamic panel estimators, such as the Generalized Method of Moments, can be used for more complex settings where endogeneity [3] and time dynamics are present. It is possible to check whether panel effects exist or not using "Breusch-Pagan-Test" (Halunga (2017)). According to Baltagi (2008) and Bell and Jones (2015) The explanation of the Fixed Effects and Random Effects model is given below,

A Fixed Effect model allows each individual to have its own intercept in order to control for features unique to that individual that do not change over time. This method depends on the idea that these traits correlate with the explanatory variables. The Fixed Effect model focuses on within-individual changes over time while controlling for time-invariant individual differences by capturing them through individual-specific intercepts. The Fixed Effects model assumes that the parameters μ_i are fixed and to be estimated, while the remaining disturbances v_{it} are stochastic, independent,

[3] Endogeneity is a situation when the independent variable and the error term are correlated.

and identically distributed (IID$(0, \sigma_v^2)$). Additionally, X_{it} is assumed to be independent of v_{it} for all i and t.

In the Fixed Effects model, α represents the overall intercept or baseline effect common across all entities (e.g., firms, countries, individuals), while β is the slope coefficient indicating the relationship between the explanatory variable X. And the dependent variable is y. N denotes the number of entities in the dataset, and T refers to the number of time periods for each entity. The term μ captures the individual-specific, time-invariant effects that are unique to each entity but do not vary over time. ν is the idiosyncratic error term representing random noise. The matrix Z_μ consists of individual dummy variables used to capture the fixed effects, and the matrix Q is a transformation matrix that eliminates the individual fixed effects to facilitate the estimation of the model. The residuals u reflect the difference between the observed and predicted values, and the variance-covariance matrix var(β) shows the precision of the estimated coefficients. By substituting the disturbances into the model equation in a vector form the equation (Baltagi (2008)) is given below,

$$y = \alpha \iota_{NT} + X\beta + Z_\mu \mu + \nu = Z\delta + Z_\mu \mu + \nu, \tag{5}$$

In equation (5), Z is $NT \times (K+1)$ and Z_μ is $NT \times N$, OLS can be used to estimate α, β, and μ. For large N, this equation includes many individual dummies, leading to the inversion of a large $(N+K) \times (N+K)$ matrix. To simplify, the least squares dummy variables (LSDV) estimator can be used. By premultiplying the model by Q, individual effects are removed, leading to the transformed model,

$$Qy = QX\beta + Q\nu, \tag{6}$$

where $QZ_\mu = Q\iota_{NT} = 0$. This transformation regresses,

$$y = Qy \quad \text{(with elements } y_{it} - \bar{y}_i), \tag{7}$$

on

$$X = QX \quad \text{(with elements } X_{it,k} - \bar{X}_{i,k}), \tag{8}$$

where $k = 1, 2, \ldots, K$. The transformed regression requires inversion of a smaller $K \times K$ matrix, yielding the estimator:

$$\tilde{\beta} = (X'QX)^{-1} X'Qy, \tag{9}$$

with variance:

$$\text{var}(\tilde{\beta}) = \sigma_v^2 (X'QX)^{-1}. \tag{10}$$

Alternatively, generalized least squares (GLS) on the transformed model also yields β. The restriction $\sum_{i=1}^{N} \mu_i = 0$ prevents perfect multicollinearity, known as the "dummy variable trap." As a result, only β and $(\alpha + \mu_i)$ are identifiable, not α and μ_i separately. From the simple regression,

$$y_{it} = \alpha + \beta x_{it} + \mu_i + \nu_{it}, \tag{11}$$

averaging over time gives,

$$\bar{y}_i. = \alpha + \beta \bar{x}_i. + \mu_i + \bar{\nu}_i. \tag{12}$$

Subtracting equation (12) from (11) produces,

$$y_{it} - \bar{y}_{i.} = \beta(x_{it} - \bar{x}_{i.}) + (\nu_{it} - \bar{\nu}_{i.}) \tag{13}$$

Similarly, averaging across all observations, equation (11) provides,

$$\bar{y}_{..} = \alpha + \beta \bar{x}.. + \bar{\nu}.. \tag{14}$$

with $\sum_{i=1}^{N} \mu_i = 0$. Here, β is The estimates can be obtained from the transformed regression: $\tilde{\alpha} = \bar{y}_{..} - \tilde{\beta}\bar{x}_{..}$ and $\tilde{\mu}_i = \bar{y}_{i.} - \tilde{\alpha} - \tilde{\beta}\bar{x}_{i.}$. The Fixed Effects estimator loses degrees of freedom as it includes $N - 1$ extra parameters, leading to potential multicollinearity. Time-invariant variables (e.g., gender, race) are removed by the Q transformation, resulting in their effects being unestimable. If $T \to \infty$, the Fixed Effects estimator is consistent. However, for T fixed and $N \to \infty$, only the Fixed Effects estimator of β is consistent, as the individual effects $(\alpha + \mu_i)$ are not consistent as the total number of these parameters increase when N increases.

In a Random Effects model, individual-specific effects (μ_i) are considered random and assumed to be uncorrelated with the explanatory variables. This allows the model to account for variation within and between individuals. The individual differences are incorporated into the error term, enabling generalization across subjects, unlike the Fixed Effects model. The Random Effects model from Baltagi (2008) is given below,

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \cdots + \beta_n X_{n,it} + \mu_i + v_{it} \tag{15}$$

Y_{it} in equation (15) is the dependent variable for individual i at time t, and $X_{1,it}$, $X_{2,it}, \ldots X_{n,it}$ are the explanatory variables. The Random Effects model captures individual specific heterogeneity through μ_i, which varies across individuals but remains constant over time for each individual. The error term, v_{it}, varies across individuals and over time. The model assumes that μ_i is uncorrelated with the explanatory variables and that the errors are independently distributed across individuals and time.

This model is appropriate when the unobserved heterogeneity is randomly distributed, such as in studies where the sample is considered a random selection from a larger population (e.g., household panel studies). In such cases, assuming the individual effects are fixed would lead to a significant loss of degrees of freedom, particularly for large samples. The random effects model also incorporates the assumption that individual effects (μ_i) and idiosyncratic errors (v_{it}) are independently and identically distributed (IID) with $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$ and $v_{it} \sim \text{IID}(0, \sigma_v^2)$, and that the regressors are independent of both μ_i and v_{it}. According to Baltagi (2008), the variance-covariance matrix and $\hat{\beta}_{GLS}$ are,

$$E(uu') = \sigma_\mu^2 (I_N \otimes J_T) + \sigma_v^2 (I_N \otimes I_T) \tag{16}$$

where I_N and I_T are identity matrices for individuals and time periods, and $\otimes$ denotes the Kronecker product [4]. The covariance structure implies that,

$$\text{cov}(u_{it}, u_{js}) = \begin{cases} \sigma_\mu^2 + \sigma_v^2 & \text{if } i = j \text{ and } t = s \\ \sigma_\mu^2 & \text{if } i = j \text{ and } t \neq s \\ 0 & \text{if } i \neq j \end{cases} \tag{17}$$

To estimate the model coefficients, generalized least squares (GLS) is used, with the regression equation transformed to account for the error structure. The GLS estimator for the coefficients is,

$$\hat{\beta}_{GLS} = (X'QX)^{-1} X'Qy \tag{18}$$

where Q is the weighting matrix derived from the error structure. This estimator combines the within- and between-individual variations, providing a more efficient estimate than OLS, especially when random effects are present. Variance components, σ_μ^2 and σ_v^2, can be estimated using Within and Between regressions. The GLS estimator is asymptotically efficient, especially when N (the number of individuals) or T (the

[4] A matrix operation which multiplies each element of one matrix by another matrix and produces a block matrix.

number of time periods) is large. Simpler methods, like ANOVA-based GLS, are often preferred in smaller samples. However, even in small samples, GLS tends to be more efficient than LSDV.

After explaining Fixed Effects and Random Effects models and their assumptions, the question arises: which model should one choose? The debate between these models has been a major topic in fields like biometrics and statistics and has influenced panel data econometrics. Early supporters of the Fixed Effects model include Mundlak (1961) and Wallace and Hussain (1969), while Balestra and Nerlove (1966) backed the Random Effects model. Mundlak (1978) pointed out that the Random Effects model assumes all regressors are unrelated to individual effects (exogenous), whereas the Fixed Effects model allows for some relationship (endogeneity). This creates a strict "all-or-nothing" choice about exogeneity. Hausman and Taylor (1981) proposed a middle ground, allowing some regressors to be endogenous and others exogenous, with additional testable restrictions. Baltagi (2008) noted that for researchers using software like Stata or R, to run Fixed and Random Effects analyses, performing the Hausman test (Hausman and Taylor (1981)) is just the starting point. This thesis uses R to apply the Hausman test and decide which model (Fixed Effects or Random Effects) is more suitable for estimating the model's coefficients.

The R code in this study uses the "plm" package to implement the models mentioned above. During the model selection process, while finding the coefficients of the model of the simulated datasets, the "plm" package is being used extensively. Croissant and Millo (2008) first introduced the "plm" package for R. The primary function, "plm", extends the "lm" function by transforming data for panel analysis, supporting methods such as pooled OLS (model = "pooling"), Fixed Effects (model = "within"), Random Effects (model = "random"), first differences (model = "fd"), and between estimators (model = "between"). Additionally, the package can accommodate unbalanced panels and two-way effects for most methods. In panel data econometrics, pooled OLS is typically the most effective estimator for β when individual effects are absent, forming the basis of the pooling model. Fixed Effects GLS applies when individual-specific errors are correlated with the regressors. First-differencing, which eliminates time-invariant components, is particularly useful in highly persistent errors, reducing serial correlation and yielding the first-difference (FD) estimator. The between model, which averages data over time, is often used to estimate long-term relationships, particularly in cases of non-stationarity. Models with variable coefficients allow β to vary across individuals and over time, while fixed coefficient models restrict β to vary along one dimension. Random coefficient models assume random variations in β around a shared mean.

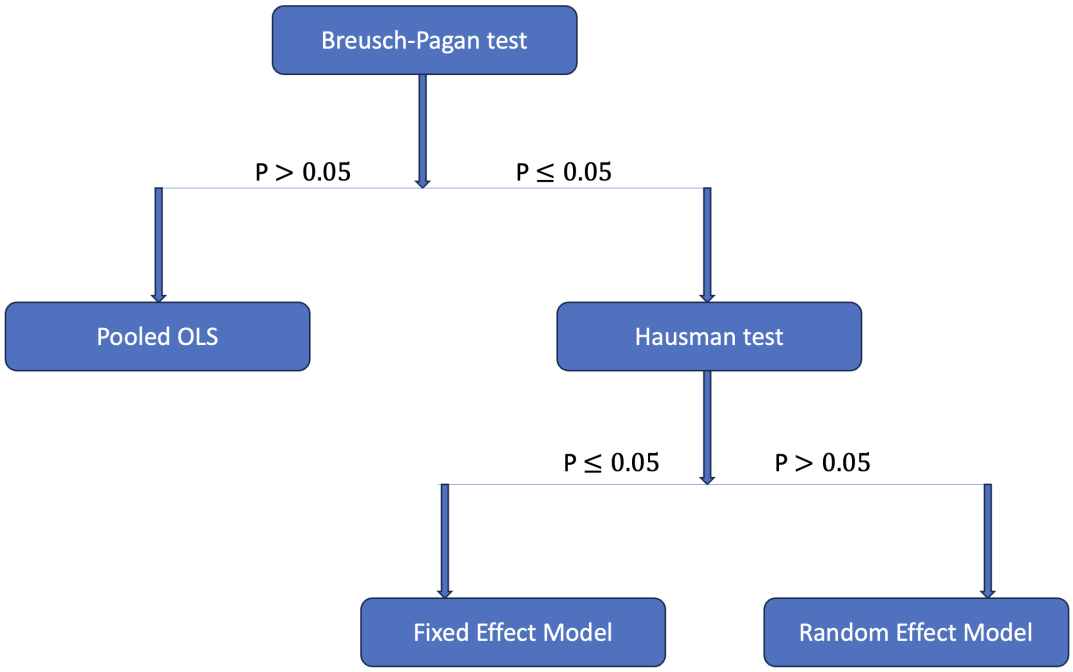


Figure 1: Model Selection for Panel Data

The flowchart in figure 1 explains how the models are selected to find the coefficients in this thesis study. First, The Breusch-Pagan test is employed, where H_0 : Panel effect does not exist. The p-value > 0.05 indicates no panel effect, suggesting pooled OLS is appropriate. Conversely, if the p-value ≤ 0.05, rejects null hypothesis. Then, the analysis proceeds with a Hausman test, where H_0 : Correlations do not exist between error and explanatory variables. The $p-value \leq 0.05$ indicates a significant correlation and rejects the null hypothesis. In contrast, the Random Effects model is suitable when a p-value is greater than 0.05. This theory is implemented to find the model for two original balanced and unbalanced panel data. According to the tests in this study, no random effects are present in the data. Because of that, the Fixed Effect model has been selected for all the simulated and original datasets, both in balanced and unbalanced panel scenarios.

2.1.2 Types and Properties

According to Buuren (2018), there are two formats for arranging longitudinal data: "long" and "wide." Every individual is represented by a single record in a wide format, with observations from various time periods kept in distinct columns. Time-invariant variables in the record stay constant, but time-varying variables are dispersed over these columns. The long format contains several records for each individual. Time-invariant variables don't change during these recordings, while time-independent vari-

ables do. This format also requires a "Time" variable to provide the time point for each record and an "ID" variable to identify entries that belong to the same individual. According to Baltagi (2008), Hsiao (2022) the variations of the panel data are,

Balanced Panel Data: In balanced panel data, the entire period has an equal and same number of observations.

Unbalanced Panel Data: In unbalanced panel data, an unequal number of observations are assigned for each level for the entire period.

Dynamic Panel Data: Dynamic Panel Data is panel data that contains lagged values of the response variable (Y_{t-i}) as a regressor.

Short or Micro Panel Data: Short or micro panel data are the panel data where the number of levels is more than the number of periods.

Long or Macro Panel Data: Long or macro panel data are the panel data where the number of periods is more than the number of levels.

While analysing all types of panel data is important, this thesis paper focuses only on balanced and unbalanced panel data because dynamic, long, and short panel data could be defined as the division of balanced and unbalanced panel data. Subject dropout in a survey or while recording data is extremely common. After a subject drops out for any period, the balance panel becomes unbalanced automatically. That is why comparing both balanced and unbalanced panels is significantly important. The following describes several attributes of panel data provided by Klevmarken (1989) and Hsiao (2003).

Panel data analysis addresses the issue of individual heterogeneity in studies involving individuals, firms, or countries. If the individual heterogeneity is not controlled, one might receive biased and inaccurate results. Panel data allows for controlling such state and time-invariant factors[5], unlike time-series or cross-section studies individually. Panel data can handle some issues better than time series data. such as degrees of freedom, multicollinearity, and increases parameter estimate efficiency. Panel data makes studying dynamic changes over time easier—for instance, unemployment and poverty estimation. Panel data tracks changes and can be used to determine whether experiences are transient or ongoing and the impact of policies. Which

[5]State and time-invariant factors refer to characteristics or variables that remain constant within a particular state or across different time periods.

cross-sectional data cannot. Comparing panel data models to cross-sectional or time-series data allows for greater flexibility in developing and testing more complicated behavioural models. Compared to time-series studies, panel data modelling provides fewer constraints when analyzing distributed lag models [6]. Panel data models are more flexible in finding hidden effects like fixed or mixed effects, which are difficult to find in cross-sectional or time-series data models. Micropanel data can provide more accurate measurements than similar variables measured at the macro level, such as individuals, firms, and households.

Kasprzyk et al. (1989) covered numerous issues with panel data survey, design and data collection. For example, data collection, nonresponse, coverage, interview frequency, spacing, and measurement errors. According to Kalton et al. (1989), the reasons behind the measurement errors are unclear questions, memory lapses, deliberate distortion, misrecording, interviewer effects, and overestimating work hours. Selectivity problems include self-selectivity, nonresponse, and attrition problems. While longer panel spans improve asymptotic arguments, they also raise the risk of attrition and complicate computation. In macro panels, ignoring cross-country dependence could lead to erroneous inferences. Tests for panel unit roots [7] should consider this dependence.

Robertson and Symons (1992) studied the biases that can happen when estimating panel data models, mainly when parameter variation [8] among individuals is not considered. The authors found these biases hard to identify and can be substantial even with slight changes. They examined the behaviour of the Anderson-Hsiao estimator under such mis-specifications using Monte Carlo simulations. They show that biases are likely to occur unless the regressor is random or uncorrelated [9]. The study suggests that even slight parameter variation can lead to serious estimation issues. Levin et al. (2002) utilized a technique to pool cross-section time-series data to test the unit root hypothesis. They suggested that pooling data can increase the power of unit root tests compared to separate tests for each individual time series and allow for varying de-

[6]Distributed lag models are a useful tool for understanding the long-term effects of one event on another.

[7]A statistical technique called a unit root test is employed to determine whether a time series data set contains a unit root. The value of a time series at any given point is heavily dependent on its value in the preceding time period; this is a feature of a stochastic process known as a unit root. When a time series has a unit root, its behaviour is dominated by shocks, making long-term predictions difficult.

[8]Parameter variation among individuals refers to differences in the effects being measured across different individuals in a model, meaning not every individual will experience the same influence from a given factor. For example, the impact of education on income might not be the same for everyone; it could vary depending on the person.

[9]When a model incorrectly considers that these effects are the same for every individual, it is known as misspecification, and it can cause biases or inaccuracies in the results.

grees of persistence across individuals. The study from Hansen (2007) examined the asymptotic properties of a covariance matrix estimator for panel data, focusing on scenarios where the cross-section dimension (n) grows large with a fixed time dimension. The estimator remains consistent despite arbitrary correlation within each individual under such asymptotics. According to Frees (1995), test statistics from Breusch and Pagan (1980) work well for calculating cross-sectional correlation when there are many time periods compared to units. Frees explained that it does not have the desired asymptotic characteristics when there are more units than time periods. Despite no cross-sectional correlations, this disparity indicates a reliance on the parent population. A distribution-free statistic that exhibits stable asymptotic properties under different conditions is presented by Frees (1995) to overcome this limitation. Frees emphasized comprehending cross-sectional correlation, especially in complex models, as Frees (1992; 1993) noted. These analyses compare the statistic's performance with alternative test statistics and evaluate the statistic's ability to detect cross-sectional correlation. Hayakawa (2007) explained why, despite using more instruments, the GMM (Generalized Method of Moments) estimator in panel data models is less biased than first differencing or level estimators. He stated that a weighted combination of biases from level and first differencing estimators makes up the bias of the system GMM estimator. The system GMM estimator has a slight bias when the variances of the individual effects and disturbances are similar. However, all GMM estimators are highly biased if the variance of individual impacts is substantially larger. He proposed bias-corrected GMM estimators as a solution to biases. On the other hand, the system estimator has a stronger downward bias than the level estimator if the variance of individual effects is smaller. Phillips and Moon (2007) examined developments in panel data analysis, emphasizing nonstationary[10] panels and provides a fresh perspective on individual effect models. It explored the theory of processes with two indexes that can grow indefinitely. The study clarified how individual effects in nonstationary panel data are perceived by investigating new limit theories [11] and their applications. The idea of a panel regression coefficient, which quantifies long-term relationships between panel sections and is comparable to coefficients in classical regression, is the main argument of his discussion. He also investigated the asymptotic properties of estimators in various nonstationary panel data models. Furthermore, his study discussed and reviewed recent developments in panel regression models and panel unit root tests.

[10] A time series whose mean, variance and other statistical properties change over time. This makes the time series unpredictable and unsuitable for further analysis.

[11] Limit theories study the estimators when the sample size approaches infinity.

2.2 Imputation

2.2.1 Foundations of Imputation

Missing data can be defined as the primary or permanent absence of a certain portion of values in a data set. Several reasons could be responsible for the missing values in a dataset. Such as unit-nonresponse[12], item-nonresponse[13] or by design[14]. According to Salgado et al. (2016), Missing data can occur due to (i) accidental omission, (ii) inapplicability to the instance, or (iii) irrelevance to the instance. For example, data may be missing in a medical setting due to technical errors, specific medical decisions, or lack of clinical relevance. They also explained the distinction between recoverable and non-recoverable missing data. If data is missing for identifiable reasons, imputation may introduce bias and make it non-recoverable. If the reasons are unidentifiable, then the missing data is considered recoverable and can be imputed. According to Scheffer (2002), data missing mechanisms are,

MCAR (Missing Completely at Random): Missing values has no relationship with any variables in the dataset, including the variable of interest. Data is collected randomly, and case deletion is valid.

$$f(r \mid y;\xi) = f(r \mid y_{\text{obs}}, y_{\text{mis}};\xi) = f(r;\xi) \quad \forall Y \tag{19}$$

The probability that data is missing r is independent of both the observed data y_{obs} and the unobserved (or missing) data y_{mis}. This means the missingness is unrelated to any of the data values, and the missing data mechanism $f(r)$ depends only on ξ, a parameter unrelated to the data.

MAR (Missing at Random): In a dataset, missing values depend on other observed variables but not the variable of interest.

$$f(r \mid y;\xi) = f(r \mid y_{\text{obs}}, y_{\text{mis}};\xi) = f(r \mid y_{\text{obs}};\xi) \quad \forall Y_{mis} \tag{20}$$

In this case, the probability of data being missing r depends on the observed data y_{obs} but not on the missing data y_{mis}. Once the observed data is accounted for, the miss-

[12] Unit-nonresponse usually happens when people refuse to participate or cannot be contacted in surveys. Furthermore, missing observations are not identified as missing here. They are just omitted from the dataset.

[13] Item-nonresponse usually happens when people participate in surveys but refuse to answer single or multiple questions.

[14] Fusioning multiple sets of data together, split questionnaire survey designs, or synthetic data could be examples of occurring missingness by design.

ingness does not depend on the unobserved data. Rather, it depends on the observed data and the parameter ξ.

NMAR (Not Missing at Random): Missing data depends on the actual value of the missing data itself, making it the hardest to model.

$$f(r \mid y; \xi) = f(r \mid y_{\text{obs}}, y_{\text{mis}}; \xi) \tag{21}$$

Here, the probability of data being missing r depends on both the observed data y_{obs} and the missing data y_{mis}. This implies that the missingness is directly related to the value of the missing data itself, which makes it more difficult to model or correct the parameters.

Scheffer (2002) also argued about what could be done if the missing data could not be sourced by other means. some of the approaches are,

Case Deletion: Remove cases with missing data. This can be done either listwise (removing the entire case) or pairwise (removing only the specific missing values for the analysis). This approach is not recommended for obvious reasons. Deleting single values or entire rows could remove important information from the data and make the inference biased or inaccurate.

Single Imputation: Replace missing values with a single estimate, such as the mean, median, or mode. The main issue with this imputation is that it overlooks nonresponse bias [15]. There could be specific reasons why some individuals choose not to respond to certain questions. As Rubin (1987) pointed out, people with very high or very low socioeconomic status frequently fail to report their income. Other methods include regression imputation, stochastic regression (adding random error), Expectation-Maximization (EM) algorithm, or hot deck imputation. Linear regression uses only complete cases in predictive equations to generate imputed missing values. According to Little and Rubin (1989), this approach can be effective when the predictors are strong, but it still suffers from insufficient variability, leading to underestimating standard errors. Stochastic regression offers an alternative approach. It adds a residual error term to the predicted score (Landerman et al. (1997)). However, both techniques can lead to biased estimates of parameters and standard errors. The hot deck approach involves matching individuals with missing data to others with similar values in a set of related variables, then imputing the known value into the missing data cell. According to Rubin (1987), the hot deck imputation treats the imputed data as certain, which can significantly underestimate variability. It could be a severe flaw.

[15] Nonresponse bias occurs when a single imputation overlooks systematic differences between missing and observed data, leading to biased estimates and distorted relationships.

Additionally, it assumes there is no difference between respondents and nonrespondents. These methods provide a single complete dataset. According to Schafer (1999), single imputation methods fail to account for variability between imputations since only a single value is imputed.

Multiple Imputation (MI): To fix the issues of Single Imputation, Rubin (1977; 1987) developed a new approach called Multiple Imputation. Donald B. Rubin is said to be the father of Multiple Imputation. According to Little and Rubin (1987) and Rubin (1996; 2018), Multiple Imputation is a method used to create several logical values based on observed variables and their relationships to fill in missing data. Multiple Imputation generates multiple complete datasets by imputing the missing data several times (e.g., using Bayesian or Frequentist methods). Each dataset is then analyzed separately. Finally, the results are combined to provide a single set of estimates that reflect both the variability within each imputation and the uncertainty across all imputations. This method is more accurate than Case Deletion or Single Imputation but can be complex for non-statisticians. Li et al. (2015) described that Multiple Imputation produces multiple plausible values for each missing value, in contrast to Single Imputation techniques that only fill in the missing data once. By quantifying uncertainty in missing value estimation, this method avoids false precision. It lowers the possibility of drawing false conclusions by producing precise estimates for various analyses. However, the process is complex and needs special statistical specialities. Allison (2000) and Schafer (1997; 2000) further elaborated on Rubin's concept of Multiple Imputation (MI). Freedman and Wolf (1995) and Schafer (1998) explained in their study how MI (Multiple Imputation) seeks to generate plausible imputations for missing values in a way that accurately reflects uncertainty while preserving key data relationships and maintaining the integrity of the data distribution.

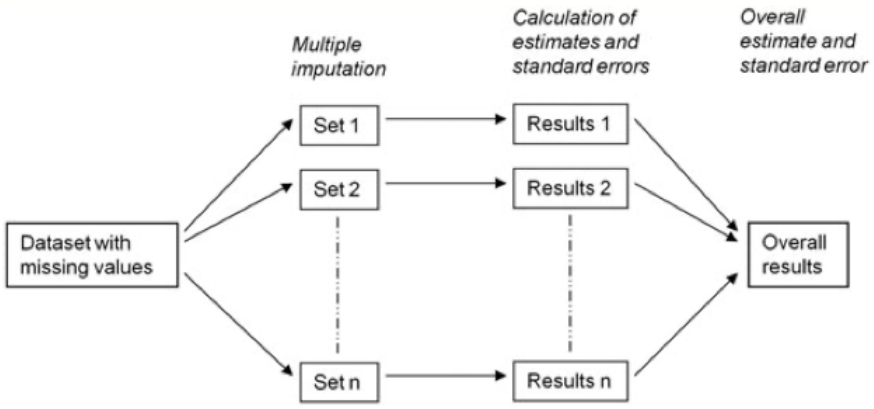


Figure 2: Graphic Representation of Multiple Imputation. Graphic Source: de Goeij et al. (2013)

According to Allison (2000), Some of the assumptions of Multiple Imputations are: missing data should be Missing at Random (MAR), and the imputation model must align with the analysis model[16]. Schafer (1999) emphasized that the model must preserve important variable associations, including interactions, and include the dependent variable. He also explained the algorithm for generating imputations must correctly account for necessary variables and their relationships. Allison (1998) showed that using all relevant information about missing cases leads to better imputation outcomes. It has some drawbacks as well. The main drawbacks are that it takes significant computing time and statistical knowledge. According to Patrician (2002), another disadvantage of multiple imputation (MI) is that it doesn't produce a unique answer. Since randomness is preserved in the MI process, each imputed dataset can yield slightly different estimates and standard errors, making it challenging to reproduce exact results. He also included in his study the advantages of MI, which allows the use of complete-data methods and incorporates the data collector's knowledge. It includes random error by requiring random variation in the imputation process, leading to more accurate estimates of standard errors. MI can be applied to any model or data without specialized software and improves efficiency by minimizing standard errors (Rubin (1987)). It also supports randomly drawn imputations under multiple models, simulating proper inferences from the data.

Maximum Likelihood(ML): Newman (2014), added Maximum Likelihood(ML) to this list. This outlines the direct calculation of the relevant parameters from an incomplete data matrix or calculates summary estimates (means, Standard Deviations, and correlations) using an algorithm like the EM algorithm, which then uses these estimates to guide the analysis. He also explained why missing data is crucial to understand and what happens when a researcher ignores it or uses the wrong method to tackle the problem. The guidelines are given below,

Listwise deletion goes against the core idea of missing data analysis: to use all available data by eliminating all cases with any missing data. This method produces biased parameter estimates under MAR and MNAR by needlessly lowering sample size, statistical power, and inference accuracy. Furthermore, it ignores the work of partial respondents, which raises ethical questions, and limits the target population to those with complete data—a rarely acceptable restriction. Listwise deletion is widely used in research, but it is both ethically and theoretically unacceptable because it frequently leads to extreme inferential errors, inflated standard errors, and less-than-ideal results. Robust methods like ML and MI, which handle missing data effectively, render listwise deletion unnecessary and irrelevant.

[16]Rubin (1987) called this model a "proper" imputation model.

Single Imputation techniques, such as mean imputation, hot deck imputation, and regression imputation, involve replacing missing data with "best guesses" but suffer significant drawbacks. These methods often introduce bias, even under MCAR, by distorting variances and correlations. Regression imputation, while slightly improved with stochastic adjustments, still falls short of more robust methods. A critical flaw of Single Imputation is its inability to produce accurate standard errors, often leading to Type I [17] errors by treating partially imputed data as complete. Given that multiple imputation eliminates these issues while retaining the benefits, Single Imputation is outdated and should be avoided for modern analyses.

The following guideline emphasizes the use of Maximum Likelihood (ML) or Multiple Imputation (MI) for addressing missing data when 10% or more of respondents provide incomplete answers to certain questions or variables. Researchers should report response rates (overall, full, and partial) and Construct-Level [18] respondent counts to understand missing data bias. If the Partial Response Rate exceeds 10%, ML or MI is recommended over pairwise deletion, as these methods reduce bias and provide accurate parameter estimates under MAR and MCAR conditions. MI generates multiple imputations to address uncertainty. ML uses likelihood functions to estimate parameters directly. Both methods combine theoretical and auxiliary variables in the imputation model to reduce bias. Auxiliary variables can help convert MNAR data to MAR. This improves accuracy and enables valid hypothesis testing.

For item-level analysis, especially in methods like Structural Equation Modeling, it is recommended to use Maximum Likelihood (e.g., Full Information Maximum Likelihood and Expectation-Maximization) or Multiple Imputation to handle missing data effectively. However, these methods face challenges when the number of variables exceeds 100. For construct-level analyses, it is recommended to use the mean of available item responses to compute scale scores, even if only one item is answered, as this retains more data and offers better statistical power compared to the listwise deletion cutoff method, which discards participants with insufficient item responses. While both approaches are biased under MAR or MNAR conditions, the mean method minimizes power loss and avoids unnecessary data deletion.

The final guideline is that person-level missingness (e.g., low response rates) is problematic because it cannot be treated as MAR, requiring missingness to be predictable from observed data. Current missing data techniques like Maximum Likelihood and Multiple Imputation are designed to be unbiased under MAR, but their ef-

[17] Type I error occurs when a true null hypothesis gets rejected in a statistical test, leading to a false positive result.

[18] Construct level refers to the level at which data are aggregated or analyzed based on specific theoretical concepts or variables (constructs) being measured.

fectiveness diminishes when applied to person-level missingness. Researchers facing high nonresponse rates should report the response rate and systematic nonresponse parameters and conduct sensitivity analyses to estimate the potential bias in their results. While systematic nonresponse parameters can indicate the likelihood of bias, missing data bias is typically small, and low response rates tend to lead to underestimation of relationships among constructs. However, guidelines for handling person-level missingness are still tentative, with the most practical approach for studies with less than a 30% response rate being to report and correct for potential bias in a transparent manner.

Acknowledging the complex nature of panel data structures, scholars have tried to create models and methods that can impute missing values from data on several levels. Chamberlain (1982; 1984) introduced a multivariate regression model designed for panel data analysis. It offers a reliable technique for looking at longitudinal data. Based on the Multiple Imputation framework, the "mice" package was developed by Buuren and Oudshoorn first in S-PLUS in 2000, then in R in 2001 according to Buuren and Oudshoorn (2011). Now, "mice" is considered the standard imputation tool for imputing missing data in numerous mechanisms. Schafer (2001) introduced the "pan" package in R. It makes imputing missing values in multilevel datasets easier. "pan" uses principal component analysis (PCA) for imputation, making it efficient with high-dimensional data and integrating cross-validation to enhance accuracy. In contrast, "mice" uses an iterative method where a model based on other variables imputes each variable with missing data. Which offers flexibility and detailed diagnostics but often requires more computational resources. "pan" is well-suited for large datasets with many variables compared to "mice". Later, Grund et al. (2016) criticized the "pan" for being overly complicated and dependent on specialized statistical knowledge. As a result, attempts have been made to improve usability and accessibility. The "pan" package was interfaced using user-friendly interfaces, and the "mitml" package was created. Croissant and Millo (2008) created a package called "plm" for R aims to simplify the process of estimating linear panel models. "plm" offers functions for making robust inferences and for estimating a wide range of models. First, "amelia" was written by Honaker et al. (2010) and then Honaker et al. (2011) upgraded "amelia II", is an R package for imputing missing data in surveys, time series, or panel datasets. It uses a fast, bootstrapping-based algorithm that is more efficient and user-friendly than traditional methods. It supports Bayesian priors and accounts for trends in data. The package provides graphical diagnostics and can be used via a command line or a user-friendly interface. Using factor structure, Cahan et al. (2023) presented a tall-project algorithm to handle missing data in large panel datasets. They discussed the impli-

cations for factor-augmented regressions and used estimated common components to impute missing values. They also offered three important outcomes: The first is a tall-project algorithm that, without iteration, can reliably estimate the whole low-rank matrix and gives a distribution theory for the estimates. The second finding clarifies the circumstances in which factor-augmented regressions can treat factors estimated from incomplete data as known. The estimation of covariances from partial data is covered in the third.

2.2.2 Previous Contributions

Panel data came into light in 1944 through the work of Haavelmo (1944). Since then, it has influenced many fields, such as biostatistics, epidemiology, sociology, economics, etc. This section first examines the history of panel data, following its development from the ground up through the influential works of Haavelmo (1944), Marschak (1950; 1953), Mundlak (1961), and Balestra and Nerlove (1966). Accordingly, this section thoroughly explains the scholarly literature on the "Imputation of Panel Data" scenario.

Haavelmo (1944) is credited with laying the groundwork for panel data by introducing the probability approach in econometrics. He mentioned how a general economic model could fail to capture individual economic behaviour. He pointed out the complex nature of the human decision-making process. And emphasize the importance of understanding these unknown economic behaviours by addressing their complex nature. Later, Marschak (1950; 1953) studied the work of Haavelmo and expanded it further by introducing the disturbances or shocks in economic models. In his work, he argued that residual errors in the economic model should be represented as a group using probability distributions rather than attributing causes to individual errors. This way of thinking has led other scholars to think that some factors are not present in the model as a variable but are enough to influence the results. Mundlak (1961) expanded the scope of panel data analysis by addressing the issue of management bias[19] in empirical production functions. He highlighted the difficulty of directly measuring managers' effectiveness and proposed a solution using covariance analysis. By assuming the stability of managerial effects over time, Mundlak demonstrated how panel data analysis could improve the accuracy of production efficiency calculations in factories. Balestra and Nerlove (1966) contributed to the advancement of panel data analysis by focusing on dynamic economic models, specifically examining the demand for natural gas. Their study considered the importance of unit changes over

[19]The error that could happen caused by the mismanagement of a manager.

time and proposed an alternative approach to estimating the model parameter. They were able to highlight the effectiveness of this approach by producing a reasonable conclusion. Balesta and Nerlove highlighted the dynamic relationship of the individuals in an economic model. Due to the massive economic growth in America since 1960, panel data became very important. Scholars and researchers have extensively researched the panel data, and numerous papers have been published to analyse and understand its intricate nature since then. They discover and overcome many limitations by studying and understanding these behaviours. The development of the economic model through studying panel data is still incomplete and ongoing. Panel data analysis, which allowed researchers to take into account individual heterogeneity, dynamic interactions, and unobserved factors, has revolutionised the field since its invention by Haavelmo and its subsequent improvement by Marschak, Mundlak, Balestra, and Nerlove. The growing literature on panel data confirms its fundamental significance in forming economic theory and policy.

Nordholt (1998) explored nonresponse in surveys. Mostly, he focused on unit and item nonresponse. Hot deck imputation was discussed for handling missing data, showing superior performance in simulation experiments compared to available case methods. Despite imputation's lower acceptance than weighing, Nordholt highlighted its necessity. For transparency, he recommended supplying both imputed and non-imputed datasets. The paper draws attention to current research efforts in imputation techniques because of inadequate databases.

The article from Hirano et al. (1998) addressed the issue of missing data in panel datasets caused by units dropping out [20] over time, an event referred to as attrition. To counteract attrition effects and preserve the advantages of panel data, researchers frequently include fresh refreshment samples from randomly selected units from the initial population. The research shows how to estimate models for attrition, including those assuming missing data are lost at random, by employing refreshing samples. The paper demonstrates that refreshment samples can increase model estimation and result precision by applying imputation techniques to travel behaviour data from the Netherlands. This provides important new statistical and economic modelling information in panel data analysis.

The study by Phillips and Chen (2011) examined regional growth in China using multiple imputation techniques on panel data from 1978 to 1999 across 30 provinces, autonomous regions, and cities. They used imputation by chained equations to ensure that data gaps did not bias the results. The process involves estimating a model using

[20]Unit drop out refers to the event where individuals or entities participating in a study or dataset fail to provide data for one or more subsequent observations or waves of data collection.

the complete sample. Then, missing values are predicted based on that model. Monte Carlo methods generated these missing observations by randomly drawing residuals from the complete sample. The study employs OLS regression analysis and the annual growth of GDP per capita is the dependent variable. The regressions include a set of potential regressors (49 in total) and control for fixed effects using province and year dummy variables. The study notes that multiple imputation techniques allowed the researchers to utilize a more complete dataset, which would otherwise be significantly reduced due to missing observations. They found that four out of 49 variables (regressors) they tested showed statistically significant effects on GDP growth in the original dataset (which had missing data) and the dataset where missing values had been filled in using imputation. And 23 out of 49 variables were significant in the imputed dataset. The analysis suggests multiple imputations helped preserve useful information and provided more reliable results. Because it reduced the potential bias caused by missing data very well. Although the exact influence of multiple imputations on the overall results has not been well examined, it has become clear that multiple imputations is a crucial tool for reducing the biases caused by missing data.

Christelis (2011) discussed imputing missing data in the first two waves of the SHARE survey, which faced significant item non-response. Buuren et al. (2006) utilized the fully conditional specification in their paper. They addressed convergence issues and adapted the theory for SHARE's multi-country, multi-section format, ensuring consistency with the original framework. Future improvements entail using information from subsequent survey waves and exploring methods to reduce missing values, including efforts to minimize non-response cases and utilize panel data from SHARELIFE [21]. His ultimate goal was to refine imputation techniques while simultaneously tackling the underlying problem of missing data prevalence.

The imputation of missing data in hierarchical structures, typical in educational research, is addressed by Drechsler (2015). Although multilevel imputation models are theoretically consistent with multilevel analysis, practitioners tend to favour Fixed Effects imputation models because they are easier to apply. However, Drechsler argues that Fixed Effects models can introduce bias, especially for cluster effects [22], impacting educational research outcomes. Unless missing rates are very high or intraclass correlations are very low, empirical evidence indicates that the bias introduced by fixed effects models is negligible for most real datasets. Multilevel imputation models are preferred for unbiased findings, even though fixed effects models are more

[21] SHARELIFE: is an extension of the SHARE project that collects life history information from respondents, including detailed data on childhood, education, occupation, partnerships, and fertility.

[22] Cluster effects capture the variation in outcomes between different clusters, beyond what can be explained by individual-level variables.

convenient, particularly where random effects are important. However, implementing multilevel models into practice can be computationally demanding and require sophisticated tools. Researchers should weigh the trade-offs between simplicity and bias when choosing imputation models. They should pay particular attention to the applicability of conclusions based on random effects. Most of the discussion concentrates on random intercept models. There may also be issues with random coefficient models. These issues require more research.

Westermeier and Grabka (2016) evaluated several imputation techniques for filling in missing information in longitudinal wealth panel datasets. They compared six imputation technique combinations using simulation data derived from German wealth panel data. Three non-response mechanisms were considered: two types of differential non-response (DNR) [23] and MAR. The methods tested include multiple imputations using chained equations (mice), the row-and-column method for panel data, and regression prediction with correction for sample selection. The evaluation criteria were trend estimates, a measure of inequality, and wealth mobility over three survey waves. According to the results, the row-and-column approach performed well, especially for skewed variables like wealth. According to the study, using "mice" and the row-and-column method typically yielded better findings for estimations of trend and inequality. However, the decision regarding wealth mobility depended on data usage; "mice" frequently replicated observed mobility structures more accurately.

Young and Johnson (2015) highlighted the complex characteristics unique to panel data. For example, repeated measures, non-random study dropout, and missingness within and across the level. They also explained why traditional approaches to impute missing values from panel data could lead to bias and inaccurate results. They criticized the work of Little and Rubin (1987) and Schafer (1997) and argued that the multiple imputations they used overlooked some key aspects of panel data. They claimed the method from Little and Rubin (1987) ignored the correlation between overall missingness and within-level missingness, which could undermine the reliability of the research.

Building on the work of Basmann (1957; 1959) and Basmann et al. (1971) on two-stage least squares (2SLS) [24] estimation, McDonough and Millimet (2017) study imputation techniques for panel data with endogeneity [25]. They draw attention to the

[23] DNR refers to situations where the probability of missing data varies systematically across different groups or variables.

[24] Two-stage least squares (2SLS) is a method that fixes endogeneity in regression by using instruments to get accurate estimates.

[25] Endogeneity is a term when variables are correlated with the variables in a statistical model. This correlation violates the assumption of exogeneity, which states that independent variables are unrelated to the error term.

difficulty in choosing the best imputation strategies, pointing out that various factors can cause imputation accuracy to not always minimize bias in the 2SLS estimate. Their paper examines several imputation techniques, including nearest neighbour matching and regression imputation, and emphasizes the significance of addressing biases and increasing measurement error brought about by imputed data, especially in 2SLS estimation. They found that adding more variables during imputation strengthens the tools and reduces bias in 2SLS estimation. He showed it through simulations and research using the data on birth weight and cognitive development in low-income households. It highlights the significance of thoroughly assessing the impact of imputation techniques, mainly when dealing with endogeneity and structural equation modelling.

Dang et al. (2019) addressed the crucial questions arising when household consumption data are either missing or inadequate. They thoroughly reviewed poverty measurement methods applicable in such data-scarce scenarios, focusing on both snapshot [26] and dynamic [27] poverty estimates. Their study provided insights into approaches ranging from handling fully missing cross-sectional data to partially missing panel data by bringing different imputation techniques under a single framework. They simplified these techniques, making them accessible and practical. They used Vietnamese household survey data and provided examples and guidance for implementation, mainly in resource-limited settings. They also highlighted the potential of imputation methods to offer cost-effective solutions for poverty estimation and emphasized the importance of further validation and scaling up adoption in low-capacity settings [28]. They also proposed future research directions, emphasizing the potential of using subjective quality data to provide deeper insights into poverty analysis.

This literature review section explains panel data, its properties and applications. It explains why fixed and random effects are crucial before selecting a model to analyze panel data. Further, this chapter explored the types of missingness (MCAR, MAR, MNAR) of the panel data that could occur during panel data collection. Then, this section explains what happens when a researcher considers data missingness incorrectly and the possible solution to encountering the messiness of panel data. And why, out of numerous solutions, Multiple Imputation is selected to impute the missing values. This is the groundwork for the imputation of panel data.

[26] Snapshot poverty estimates refer to a single point in time assessment of poverty levels within a population, providing a snapshot of the poverty situation at that moment.

[27] To understand how people or households move in and out of poverty over time, dynamic poverty estimates monitor changes in poverty levels for different periods.

[28] Low-capacity settings are environments where resources, expertise, infrastructure, or institutional support are insufficient to carry out activities or implement methodologies effectively.

3 Data

3.1 Source and Variables Selection

The primary data source for this thesis is the MikroSim model, a dynamic microsimulation tool developed for Germany. The development and structure of the MikroSim model are detailed in Münnich et al. (2020). This model evaluates the potential impacts of various policies on social indicators while capturing local variations across more than 10,000 communities in the country. For this thesis, the semi-public "David" dataset of the MikroSim project has been implemented. It is the outermost layer in the "Multi-Layer-Security Architecture" and is deemed secure enough for research and education. The data is divided into three layers. The first two layers of the dataset have less noise and significantly reflect real-world scenarios. For example, the information about the individual is more similar to that of the people of Trier. In contrast, the final layer, known as the "David" dataset, contains a significantly higher level of noise, which diminishes its close reflection of real-world individual information. As a result, the "David" dataset is more suited for educational purposes and can be made available to academic researchers and individuals due to the research and academic purpose.

As described in Münnich et al. (2021), traditional microsimulation models typically operate at a national level due to the limited availability of detailed regional data and the high computational complexity. However, the demand for more granular modelling that captures regional and subgroup variations has led to the creation of the MikroSim model. The MikroSim model uses datasets from various sources, including surveys, administrative records, and other data types. These datasets allow the model to simulate dynamic processes (eg., births, deaths, education, and marriages). MikroSim model employs various methods of data integration and small-area estimation to construct a dataset that covers the entire population at the municipal level. The base dataset is derived from an anonymized national register of residents, which has been utilized for methodological research for the 2011 census in Germany. This register models the German population across all 11,339 municipalities. It provides a foundational structure for the dataset. Then, a synthetic dataset is created to supplement the anonymized national register. The purpose of the synthetic dataset is to reflect the characteristics of the real-world population. The synthetic dataset's characteristics, such as distributional parameters, correlations, cluster effects, and totals, are aligned with the real population data. The "David" dataset includes structural missing values, which will be handled during the process.

In summary, this thesis study uses a modified subset of the MikroSim population dataset, which reflects the population of Trier, Germany. This dataset is accurately suitable for simulated missingness and imputing the missing values from the simulated dataset. The data used in this study originates from a meticulously constructed and comprehensive synthetic dataset, augmented with additional information to facilitate dynamic microsimulation at a highly detailed regional level.

The "David" dataset has 92 variables. Out of 92 variables from the dataset, nine variables are selected for this thesis study. The nine variables—ID, Year, Age, Employment Types, Income, Marital Status, Employment Hours, Education, and Sex—are chosen for their relevance in simulating socioeconomic outcomes over time, particularly within the panel data context. There are several reasons for selecting these variables. These variables are significantly correlated with "Income", and one big part of this study is to simulate missingness artificially in the "Income" column. For instance, education and age often influence income, while year captures temporal trends. These correlations and trends are crucial for accurately generating the models to create missing scenarios and imputing missing data. These variables are practical and relevant to the simulations because they are commonly found in real-world panel surveys. Furthermore, they provide a solid foundation for studying socioeconomic trends and outcomes. While additional variables could enhance the analysis, these nine variables are sufficient to simulate missingness, formulate stochastic models and cover all possible scenarios. They offer a good balance of accuracy and precision in simulating the missingness and imputing the missing values in this thesis study. "ID" and "Year" are responsible for capturing the panel effects, "Income" will be the target variable, "Employment Hours" and "Education" are exogenous variables to check the external effect after the imputation, and the rest are predictors.

1. ID: "ID" is very important to track changes in the same individual over time. It uniquely identifies the individuals within the dataset. Before renaming it to "ID", it was named "PID" in the "David" dataset.

2. Year: The variable "Year" demonstrates how the data is collected for the same individuals over multiple periods of time. The initial dataset starts in 2011 and ends in 2024. The data was preliminary and inconsistent from 2011 to 2013. Furthermore, 2024 (During the writing of this thesis) is not yet complete. That is why the simulation has been simplified and made more consistent by filtering the year data from 2013 to 2023. Before renaming it to "Year", it was named "year" in the "David" dataset. After filtering the column from 2013 to 2023, the frequency for each year is given below,

Year	Frequency
2013	107343
2014	108312
2015	114651
2016	110284
2017	109885
2018	110593
2019	111211
2020	110500
2021	111193
2022	111519
2023	111966

Table 3: Frequency Table of "Year"

3. Age: The dataset includes individuals aged between 0 and 106 years. The median age is 47, and the mean age is 44.05, suggesting a fairly balanced distribution of ages around middle adulthood. Since Trier is a student city, there is a notable peak of 25 years, an anomaly not typically seen in other German regions except the student cities. The Age density curve also demonstrates a peak of 75 years old, which shows the retired ageing population. Before renaming it to "Age", it was named "EF44" in the "David" dataset. The density curve and summary statistics are given below,

Statistic	Data
Min.	0.00
1st Qu.	28.00
Median	47.00
Mean	44.05
3rd Qu.	59.00
Max.	106.00

Figure 3: Age Summary

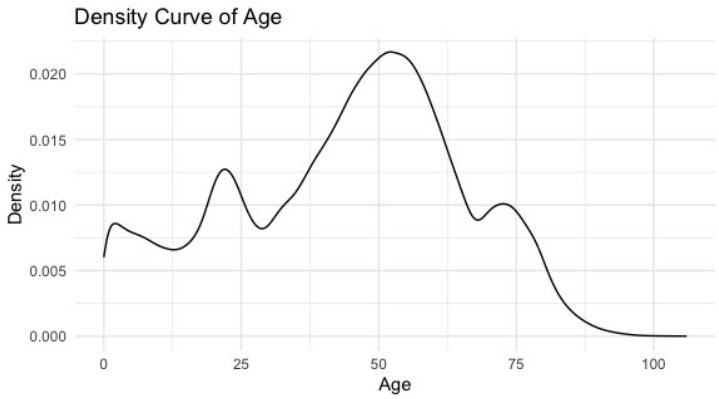


Figure 4: Density Curve of Age

4. Employment Types: "Employment Types" is a categorical variable that provides the employment situation of the individuals. It has three types of categories: Employed, Unemployed and Not in the workforce. Before renaming it to "Employment Types", it was named "erwerbstyp" in the "David" dataset. The frequency table is given below,

Value	Frequency	Label
0	724,241	Employed
1	25,058	Unemployed
2	468,158	Not in workforce

Table 4: Frequency Table of "Employment Types"

5. Income: The variable "Income" defines the average monthly money a person earns. The "Income" variable has 631,874 structural missing values. The missing values are for the people who do not earn any money. For example, individuals who are underage, retired, unemployed, students, or housewives. All the missing values are replaced with 0. Before renaming it to "Income", it was named as "inc.ind" in the "David" dataset

Statistic	Data
Min.	4.7
1st Qu.	796.6
Median	1650.6
Mean	2919.4
3rd Qu.	3374.5
Max.	311207.7
NA's	631874

Table 5: "Income" Before Replacing the Structural Missing Values

Statistic	Data
Min.	0.0
1st Qu.	0.0
Median	544.8
Mean	1723.1
3rd Qu.	2029.1
Max.	311207.7

Table 6: "Income" After Replacing the Structural Missing Values

The distribution of the variable "Income" in Figure 5 has four different scenarios. The first graph represents the variable without any transformation, which is highly skewed, with most points clustered near the lower end of the income scale. This indicates that most individuals have low incomes. The histogram shows significantly higher bars on the far left and a long tail extending to the right, showing a few individuals with much higher incomes, though these are rare. The x-axis spans from near zero to over 300,000, with income values sharply dropping off as they increase, while the y-axis reflects the frequency of income ranges, with nearly 600,000 individuals earning very little or nothing [29]. This suggests a significant income disparity, likely due to a large number of students and retired people who are not earning [30]. Other possible reasons for low income include being underage, retired, unemployed, or a housewife. To tackle this high disparity, the following three plots are generated from log-transformed "Income". Log transformation is also important to scale the data, reduce heteroskedasticity, improve the coefficients' interpretability, and remove outliers.

[29] For the entire time duration (2013-2023).

[30] There is a notable peak at 25 and 75 years old in figure 4, indicating a high number of retirees and students who do not earn.

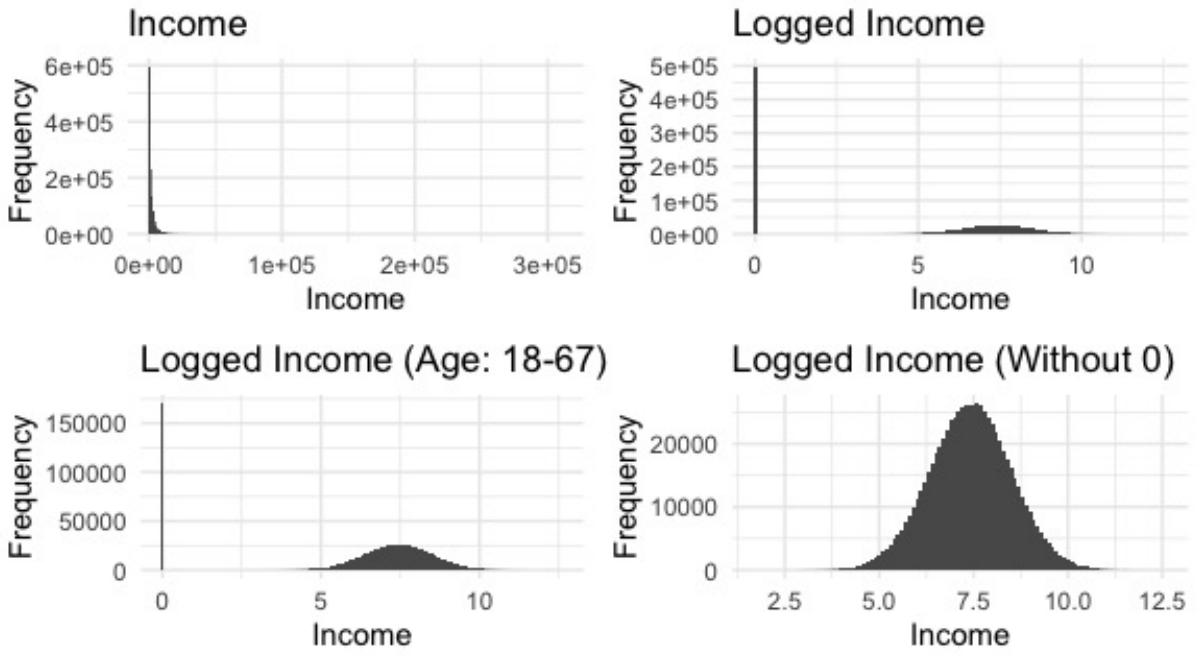


Figure 5: Histograms of "Income" in Different Scenarios

A log transformation is considered to better visualize the distribution, especially among higher incomes. The second plot compresses large values and makes the distribution more interpretable. Applying a logarithm normalizes the typically skewed income data, where a few individuals earn significantly more. On the log scale, there is a peak at 0, indicating that many individuals have zero or very low income. After this, there is a rise between Log values of around 5 to 10, showing a more balanced, bell-shaped distribution for those with non-zero income. The x-axis now indicates the logarithm of income values (adjusted by adding 1), and the y-axis depicts the frequency of individuals within each log-transformed income bin. Key takeaways include that most people either have no or very low income, and the log transformation reveals a more symmetric distribution for those with non-zero income, reducing the apparent income disparity.

The third plot shows "Income" for the filtered "Age"(18-67), and the fourth plot shows the histogram without zero income. The third plot shows the list of incomes only for the ages of 18 to 67. The range is considered the earning age for average individuals. After the income filtered by age, zero income is still significantly high. This shows the high number of working age people either not working or not in the workforce. The fourth plot generated the values without zero income, showing nearly a normal distribution with a mean of 7.5. The transformed variable "Income" of the second plot has been taken to improve the further imputation. The logged "Income"

makes the interpretation convenient, and, on the other hand, it has zero values and other variability of all income levels and characteristics of the original data.

6. Marital Status: "Marital Status" is a categorical variable that defines an individual's marital situation. Before renaming it to "Marital Status", the variable was named "EF49" in the "David" dataset. The frequency table for "Marital Status" is given below,

Value	Frequency	Label
1	524,632	Married
2	461,721	Single
3	101,066	Widowed
4	130,038	Divorced

Table 7: Frequency Table of "Marital Status"

7. Education: "Education" is a categorical variable. It indicates the highest general school degree attained by an individual in Trier. Before renaming it to "Education", the variable was named "EF310" in the "David" dataset. The variable "Education" initially consisted of seven education levels. Values 1-6 are assigned for their respective labels described in table 8. The last level, "Polytechnic secondary school of the GDR with completion of the 10th grade," initially had no assigned value. Although the value was missing, it was structurally missing[31]. The total number of structural missing values was 315756. Then, the value "7" is assigned for the label "Polytechnic secondary school of the GDR with completion of the 10th grade" using R so that it does not count as missing values. After the replacement, the complete frequency table is given below,

Label	Value	Frequency
Secondary (elementary) school diploma	1	483654
Polytechnic secondary school of the GDR with completion of the 8th or 9th grade	2	7088
Realschulabschluss (Mittlere Reife) or equivalent qualification	3	318535
Technical college entrance qualification	4	81481
General or subject-linked higher education entrance qualification (Abitur)	5	330684
Graduation after a maximum of 7 years of school attendance	6	4764
Polytechnic secondary school of the GDR with completion of the 10th grade	7	315756

Table 8: Frequency Table of "Education"

8. Employment Hours: "Employment Hours" is a categorical variable responsible for defining whether an individual is working part-time, full time or not working. Initially, the variable had 493216 structural missing values. The missing values are

[31] Structural missingness occurs when certain data points are systematically absent due to specific reasons rather than at random.

replaced with two because the initial categories were 0 and 1. Before renaming it to "Employment Hours", the variable was named "vollzeit" in the "David" dataset. The frequency table of "Employment Hours" is given below,

Value	Frequency	Label
0	179,798	Part-time
1	544,443	Full-time
2	493,216	Not-Working

Table 9: Frequency Table of "Employment Hours"

9. Sex: "Sex" is the last variable in the list. It is a categorical variable. Also, It is a time-invariant[32] in general. It will be interesting to see how a time-invariant variable reacts to imputation. Before renaming it to "Sex", the variable was named "EF46" in the "David" dataset. The frequency table is given below,

Value	Frequency	Label
1	607130	Male
2	610327	Female

Table 10: Frequency Table of "Sex"

3.2 Data Preprocessing

Balanced and Unbalanced panels are considered for this thesis described in section 2.1.2. Panel data could be either balanced or unbalanced. The symmetric relation that individuals have with the time in a balanced panel that is not present is unbalanced. On the other hand, during any study, balanced panel data could be unbalanced after some time due to subject dropout. That is why studying and comparing both balanced and unbalanced panels is essential. The original data set is called the David dataset. This thesis study breaks the original dataset into balanced and unbalanced panels and prepare for further transformation.

Balanced Panel: All the IDs must be present for every year in a balanced panel. Several steps have been implemented in the R code to ensure all individuals are present throughout the years. The code converts the " Year " column to numeric format. This is important for accurate comparison and filtering based on year values. Then, the rows

[32]In a time-invariant variable, the values do not change over time.

for unique IDs are filtered. These unique IDs are present for every year from 2013 to 2023. If the dataset is truly balanced, each ID must appear precisely 11 times for every year from 2013 to 2023. The occurrences of each ID are counted, and IDs that do not meet this criterion are filtered out, resulting in a subset of IDs that appear precisely 11 times every year. The dataset then includes only the rows corresponding to these IDs, resulting in a balanced panel where each ID has observations for every year in the specified range. To confirm that the panel is balanced, count the number of distinct years in which each ID appears. Every ID will appear every year from 2013 to 2023 on a balanced panel. The dataset can be confirmed as a balanced panel by verifying this condition. IDs that are not present or present more or less than 11 times in years from 2013-2023 are also identified and displayed. This step is unnecessary, but it helps understand the data better. The dataset is significantly large. Sampling was initially necessary for computational efficiency. Each year, there are 50483 observations after filtering and cleaning the dataset. This makes further sampling unnecessary for a balanced panel dataset. The dataset is named *balanced_panel_data* in the R code. This dataset is further used to simulate the missing values for balanced scenarios and compare the characteristics of the imputed balanced datasets as a benchmark.

Unbalanced Panel: The code in R samples 50483 distinct ID-Year combinations for each year from the original large dataset, resulting in an unbalanced panel dataset. The reason for the sample of 50483 observations per year is that 50483 observations per year are received from the balanced panel. To implement this, an empty data frame is first created to store the sampled data. The data is then filtered so that only records from a specific year are included by iterating through each distinct year. The method ensures the sampled data has 50483 unique ID-Year pairings yearly. These samples are appended to the main empty data frame, forming the unbalanced panel. Following the sampling process, the code determines if the panel is balanced by calculating the number of distinct years each ID appears in and comparing that number to the total years in the 2013–2023 range. It then prints whether the panel is balanced, lists any IDs that do not appear in all years, and displays a summary and count of observations per year. The dataset is called *unbalanced_panel_data* in the R code. Likewise, *balanced_panel_data*, this dataset is further used to simulate the missing values for unbalanced scenarios and compare the characteristics of the imputed unbalanced datasets as a benchmark.

To assess imputation techniques, missing data needs to be artificially generated. Since the true missing values in real-world datasets are always missing, it is impossible to evaluate the precision and accuracy of imputation methods directly. Researchers

can compare imputed datasets to the original dataset and determine the most efficient techniques by creating missingness in a controlled way. This thesis employs visualization tools in R to detect and analyze missing data patterns, complemented by theoretical approaches discussed earlier. The R code simulates missing data in both balanced and unbalanced panel datasets using three distinct missing data mechanisms: MNAR, MCAR, and MAR. The probabilistic model is implemented for MAR and MNAR missingness. The code is part of a more extensive process to simulate and study the imputation effects of different types of missingness on panel datasets, which is crucial to measure the effects of the imputation of panel data. In missing data mechanisms, linear combinations describe how missing values are introduced based on observed variables. For MCAR, missing values are assigned randomly and independent of other dataset variables. For MAR, the probability of missingness is modelled by the following equation,

$$L_{MAR} = \beta_0 + \beta_1 \text{Employment Types} + \beta_2 \text{Age} + \epsilon \tag{22}$$

In equation 22, The parameter β_0 is intercept, β_1 and β_2 are coefficients. ϵ represents random noise. This model reflects how the likelihood of missing data is influenced by observed variables such as "Employment Types" and "Age", assuming that missingness depends on these factors but not on the missing values themselves. For Missing Not at Random (MNAR), the model includes the variable itself and is given by,

$$L_{MNAR} = \beta_0 + \beta_1 \text{Employment Types} + \beta_2 \text{Age} + \beta_3 \text{Income} + \epsilon \tag{23}$$

In equation 23, β_3 is added as the effect of the variable Income itself on the probability of missingness. This approach indicates that the missingness is related to the actual values of the variable in question and other observed variables.

Coefficient	Unbalanced Panel	Balanced Panel
β_0	0.5	0.5
β_1	0.002	0.004
β_2	-0.07	-0.07
β_3	.1	.1

Table 11: Regression Coefficients for Balanced and Unbalanced Panels

The chosen intercepts and coefficients of the models are given in table 11. The choice of the intercepts and the coefficients of the models could not be made entirely data-driven. The pooled OlS model could be used to find the coefficients from the original data. However, pooled OlS ignores any fixed or random effects in the data. So,

the result could be erroneous. After running the "Hausman Test," no random effect was found in the data. So, the next choice would be the Fixed Effect model. The "within" method for fixed effects from the "plm" package absorves the intercept and adds the effects in the coefficients. So, the results could also be misleading without intercept. In that case, the intercepts and the coefficients of the models are chosen based on the guess from the correlation table. The correlation between "Income" and "Employment Types" is negative and high. On the other hand, the correlation of "Income" and "Age" age is positive and low. As the correlation of the "Income" with "Income" is 1, a high coefficient is assigned for "Income" as β_3. The random noise follows a standard normal distribution with $\mathcal{N}(0,1)$

In conclusion, the original dataset is first divided into balanced and unbalanced panels. Each panel type generates datasets for the three missing data mechanisms (MCAR, MAR, and MNAR). For each mechanism, missingness levels of 10%, 30%, and 50% are applied, resulting in 18 datasets. A summary of all the generated datasets is given below,

- **Balanced Panel with MCAR**: `balanced_panel_data_mcar_50`, `balanced_panel_data_mcar_30`, `balanced_panel_data_mcar_10`
- **Balanced Panel with MAR**: `balanced_panel_data_mar_50`, `balanced_panel_data_mar_30`, `balanced_panel_data_mar_10`
- **Balanced Panel with MNAR**: `balanced_panel_data_mnar_50`, `balanced_panel_data_mnar_30`, `balanced_panel_data_mnar_10`
- **Unbalanced Panel with MCAR**: `unbalanced_panel_data_mcar_50`, `unbalanced_panel_data_mcar_30`, `unbalanced_panel_data_mcar_10`
- **Unbalanced Panel with MAR**: `unbalanced_panel_data_mar_50`, `unbalanced_panel_data_mar_30`, `unbalanced_panel_data_mar_10`
- **Unbalanced Panel with MNAR**: `unbalanced_panel_data_mnar_50`, `unbalanced_panel_data_mnar_30`, `unbalanced_panel_data_mnar_10`

Table 12 below provides all the above-explained information together. First, the "David" dataset is broken into a balanced panel and an unbalanced panel. Then, for each balanced and unbalanced panel, MCAR, MAR, and MNAR types of missingness are generated. Then, for each type, different levels of missingness(10%, 30%, and 50%) are generated.

Panel Type	Mechanism	50% Missingness	30% Missingness	10% Missingness
Balanced	**MCAR**	Randomly selects 50% of the rows and sets the "Income" column to NA.	Randomly selects 30% of the rows and sets the "Income" column to NA.	Randomly selects 10% of the rows and sets the "Income" column to NA.
	MAR	Creates missing values based on a linear combination of "Employment Types" and "Age" with added random noise, setting those below the 50th percentile to NA in the "Income" column.	Similar to the 50% case of MAR but sets those below the 30th percentile to NA in the "Income" column.	Similar to the 50% case of MAR but sets those below the 10th percentile to NA in the "Income" column.
	MNAR	Creates missing values based on a linear combination of "Employment Types", "Age", and "Income" with added random noise, setting those below the 50th percentile to NA in the "Income" column.	Similar to the 50% case of MNAR but sets those below the 30th percentile to NA in the "Income column".	Similar to the 50% case of MNAR but sets those below the 10th percentile to NA in the "Income" column.

Panel Type	Mechanism	50% Missingness	30% Missingness	10% Missingness
Unbalanced	**MCAR**	Randomly selects 50% of the rows and sets the "Income" column to NA.	Randomly selects 30% of the rows and sets the "Income" column to NA.	Randomly selects 10% of the rows and sets the "Income" column to NA.
	MAR	Same procedure as for the balanced panel, using "Employment Types" and "Age" to determine missing values in unbalanced panel data.	Same procedure as for the balanced panel, using 30% missingness.	Same procedure as for the balanced panel, using 10% missingness.
	MNAR	Same procedure as for the balanced panel, using "Employment Types", Age, and "Income" to determine missing values in unbalanced panel data.	Same procedure as for the balanced panel, using 30% missingness.	Same procedure as for the balanced panel, using 10% missingness.

Table 12: Data Preprocessing Before Simulation

By the end of this chapter, it is clear what data missingness is and why a real-world dataset with missing values cannot be used for research. To overcome this issue, artificial missing values have to be generated so the original dataset can be compared with the imputed dataset to analyse the performance of the imputation.

4 Methodology and Simulation

This section outlines the imputation frameworks, the algorithms underlying the packages, the R implementation of the chosen frameworks, and the simulation process for generating the imputed datasets. Three R packages called "mice", "milml", "amelia", and the development of a machine learning architecture called "LSTM Network" is being used in this thesis study to impute the missing values. Together, this thesis paper defined them as imputation frameworks. "mice" uses the "pmm" method, "mitml" uses "Gibs sampler", "amelia" uses "bootstrap", and "LSTM-Network" is based on a machine learning architecture called "Recurrent Neural Network". "mice", "milml", and "amelia" use multiple imputations and "LSTM-Network" uses single imputation mechanism. Selecting these four frameworks ensures comparing the above-mentioned methods, single and multiple imputation scenarios and traditional vs. an emerging machine-learning approach. R is used to implement the theories, get the results, and compare them with the original balanced and unbalanced datasets. After implementing the frameworks ("mice", "mitml", "amelia", and "LSTM-Network"), 18 datasets are simulated from each framework. Nine balanced and nine unbalanced simulated datasets from each framework. 36 datasets for the balanced panel and 36 datasets for the unbalanced panel in total. This provides a total of 72 datasets at the end of the simulation. These 72 datasets are then compared to the original two datasets called *balanced_panel_data* and *unbalanced_panel_data*.

4.1 mice

Buuren and Oudshoorn (1999) first published the mice package for S-PLUS, later developed for R in 2001 (Buuren et al. (2015)). "mice" is also known as "Multivariate Imputation by Chained Equations". The primary goal of the "mice" package was to handle missing values from a dataset. The methodology was first used in Buuren et al. (1999) to analyze the missing blood pressure data. The package generates multiple imputations (with replacement) for multivariate missing data. The technique is based on "Fully Conditional Specification", in which a different model is imputed to each incomplete variable [33]. Initially introduced as the fully conditional specification in Buuren et al. (2006), this general class of methods specifies the imputations model for multivariate data as a series of conditional distributions. "mice" can impute mixtures of continuous, binary, unordered, and ordered categorical data. Moreover, "mice" can

[33] Fully Conditional Specification (FCS) is a technique for imputing missing data that uses models based on the other variables in the dataset to iteratively fill in the missing values for each variable, one at a time. This procedure is repeated multiple times until the imputed values become consistent.

use passive imputation[34] to maintain consistency between imputations and impute continuous two-level data. The pseudo algorithm of "mice" from Buuren (2018) is given below,

1. Specify an imputation model $P(Y_j^{\text{mis}} \mid Y_j^{\text{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, T$:
   (a) Repeat for $j = 1, \ldots, p$:
      i. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ as the currently complete data except $\dot{Y}_j$.
      ii. Draw $\dot{\theta}_j^t \sim P(\theta_j \mid Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
      iii. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}} \mid Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\theta}_j^t)$.
   (b) End repeat j.
4. End repeat t.

"mice.impute.2l.pan", "mice.impute.panImpute", or "mice.impute.jomoImpute" are some methods in "mice" that can incorporate panel data. Nonetheless, these methods are connected to the "pan" and "mitml" packages either directly or indirectly. Since "mitml" is built on the "pan" package and used in this thesis, simulating the datasets using the "pmm" method could be interesting. Little (1988) proposed the term predictive mean matching. Later, Buuren and Oudshoorn (2000; 2012) developed mice.impute.pmm. The algorithm behind "mice.impute.pmm" is given below,

1. Calculate the cross-product matrix $S = X'_{\text{obs}} X_{\text{obs}}$.
2. Calculate $V = (S + \text{diag}(S)\kappa)^{-1}$, with some small ridge parameter κ.
3. Calculate regression weights $\hat{\beta} = V X'_{\text{obs}} y_{\text{obs}}$.
4. Draw q independent $\mathcal{N}(0,1)$ variates in vector $\dot{z}_1$.
5. Calculate $V^{1/2}$ by Cholesky decomposition.
6. Calculate $\dot{\beta} = \hat{\beta} + \dot{\sigma} \dot{z}_1 V^{1/2}$.
7. Calculate $\dot{\eta}(i,j) = \left| X_{\text{obs},[i]} \hat{\beta} - X_{\text{mis},[j]} \dot{\beta} \right|$ with $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_0$.

[34] Passive imputation fills missing values of a variable by deriving them based on its known relationships with other variables, rather than imputing it independently. It ensures consistency in the dataset by preserving predefined constraints or dependencies.

8. Construct n_0 sets Z_j, each containing d candidate donors, from Y_{obs} such that $\sum_d \dot{\eta}(i,j)$ is minimum for all $j = 1,\ldots,n_0$. Break ties randomly.

9. Draw one donor i_j from Z_j randomly for $j = 1,\ldots,n_0$.

10. Calculate imputations $\dot{y}_j = y_{i_j}$ for $j = 1,\ldots,n_0$.

The "mice" package provides functions for inspecting, imputing, diagnosing, analyzing, pooling, storing, and exporting incomplete data. The "mice" function simulates multiple imputed data set objects using Gibbs sampling, where missing values are imputed based on other columns in the data. After applying the analysis model—such as regression—the "with()" function yields a multiply imputed repeated analysis object. Finally, Rubin's rules are applied to the results as they are combined by the "pool()" function to create a pooled result object that resembles typical regression output. Default imputation methods depend on the measurement level of the target column. Custom imputation methods can also be incorporated. Categorical variables are handled by creating dummy variables and imputing them accordingly.

A function is written in R code called *Data_Imputation_mice*, which simulates new datasets to impute the missing values (datasets containing missing values are described in 3.2). The function imputes missing values in the "Income" column, incorporating lagged variables for improved prediction accuracy. The predictor variables are "Age", "Employment Types", "Marital Status", and "Sex". "Employment Hours" and "Education" are exogenous variables. The function first converts the "ID" column to a factor and the Year column to a numeric to enable lagging. It creates lagged versions of specified variables—"Employment Types", "Marital Status", "Sex", and "Income"—representing their values from the previous year for each individual. Then, a predictor matrix is generated to incorporate the predictors with "mice". The predictor matrix excludes "ID". Then, it includes the lagged variables. "Age" is then added as a predictor inside the predictor matrix. The squared "Age" could provide a good prediction, but the dataset is very sensitive. Instead of a curve shown in Figure 5 very unusual curve were generated for squared "Age". As a result, only "Age" provided a significantly good prediction compared to squared "Age". The function simulates five datasets for each imputation by providing m = 5. "maxit" = 100 ensures 100 iterations, and the method used is "pmm". The values of "m" and "maxit" could be given more to generate more datasets over longer iterations. However, it would take a significant amount of system memory and would not be computationally efficient. Then, the exogenous variables are added to the simulated datasets and stored in another dataframes.

4.2 mitml

The "pan" package in R was first introduced in Schafer (2001), which makes imputing missing values in multilevel datasets easier. It employs the Gibbs sampler for the multivariate linear mixed model outlined by Schafer (1997). This package employs multiple imputations for missing data in a multivariate panel or clustered data. The underlying model is:

$$y_i = X_i\beta + Z_i b_i + e_i, \quad i = 1, \ldots, m, \tag{24}$$

where

$y_i = (n_i \times r)$ matrix of incomplete multivariate data for subject or cluster i,

$X_i = (n_i \times p)$ matrix of covariates,

$Z_i = (n_i \times q)$ matrix of covariates,

$\beta = (p \times r)$ matrix of coefficients common to the population (Fixed Effects),

$b_i = (q \times r)$ matrix of coefficients specific to subject or cluster i (Random Effects),

$e_i = (n_i \times r)$ matrix of residual errors.

The rows of e_i are assumed to be independently normally distributed. The mean is zero and has an unstructured covariance matrix σ [35]. When stacked into a single column, the matrix b_i is also normally distributed. It has a mean of zero and an unstructured covariance matrix ψ. y_i may contain missing values in any pattern. The first column of X_i and Z_i will often be constant (one) in these applications, and Z_i will contain a subset of X_i's columns. Where p is the number of fixed-effect covariates (the number of columns in X_i), and q is the number of random-effect covariates (the number of columns in Z_i).

The "pan" package was later criticized by Grund et al. (2016). According to him, "pan" is not user-friendly and depends on specialized statistical knowledge. Consequently, initiatives to enhance accessibility and usability have been undertaken. To interface the "pan" package with user-friendly interfaces, the "mitml" package was created (Grund et al. (2019)). "mitml" offers tools for multiple missing data imputations in multilevel modelling. Multiple functions for visualizing, managing, and analyzing multiple imputed data sets are included in this package, along with an intuitive interface to the algorithms implemented in the R packages. The function "panImpute,"

[35] For an unstructured covariance matrix, no specific assumptions are made about the correlation among the variables.

which is the primary interface to "pan," enables the specification of imputation models for continuous variables with level 1 missing data[36] (Schafer and Yucel (2002)). The command "type" inside the function "PanImpute" specifies variable roles numerically. The following numerical codes denote distinct roles within the imputation model for the "type" argument: -2 for cluster indicator variables, -1 for grouping variables where imputation is carried out separately, 0 for variables not included in the model, 1 for target variables with missing data, 2 for predictors with fixed effects on all targets (fully observed) and predictors with random effects on all targets (fully observed) denotes by 3.

The R code first estimates Fixed Effects and Random Effects models for both balanced and unbalanced panel datasets to analyze the relationship between Income and predictors. Before setting up the imputation process, one must determine whether fixed or random effects are present in the predictor variables. The Hausman test is used for both balanced and unbalanced panels, whether fixed or random effects are present. Due to significant p-values (< 0.05), the results found fixed effects in the predictor variables. The predictor variables are "Age", "Employment Types", "Marital Status", and "Sex". "Employment Hours" and "Education" are kept as exogenous variables. Missing data is imputed using the panImpute function, with separate functions called *Data_Imputation_mitml_Bal* and *Data_Imputation_mitml_Unbal* tailored for balanced and unbalanced panels. A single function could do the simulation, but for the balanced and unbalanced panels, the fixed and random effects are measured separately; therefore, each function is written for both balanced and unbalanced panels. The functions select key variables ("ID", "Year", "Age", "Employment Types", "Income", "Marital Status", "Employment Hours", "Education", "Sex") and define a type vector to specify variable roles during imputation. The parameter "type" is assigned with -2, -1, 2, 2, 1, 2, 0, 0, 2 for the columns "ID", "Year", "Age", "Employment Types", "Income", "Marital Status", "Employment Hours", "Education", "Sex", respectively. "m" is equal to 5 to simulate five imputed datasets. "n.iter" is 100 to iterate 100 times. "n.burn" is 1000[37]. Bigger "m" or more iterations may provide a more accurate simulation; however, this is not computationally efficient and could slow down the function. This process is applied to the datasets described in 3.2. The simulated datasets that provide multiple completed datasets are stored separately for further analysis.

[36] Level 1 in multilevel modelling refers to the individual data within the bigger units (level 2). When data points are absent from these unique observations, such as student test results at specific times, it's known as level 1 missing data. Where test results for students are recorded at several intervals (level 1) at various schools (level 2). Level 1 missing data is when parts of test results for particular students at particular times are absent.

[37] n.burn refers to the burn-in period. The number of burn-in iterations before any imputations are drawn allows the Markov Chain to reach a stable distribution.

4.3 amelia

"amelia" is an R package written by Honaker et al. (2010) for the multiple imputation of incomplete multivariate data. It utilizes the bootstrap EM algorithm to simulate imputed datasets from incomplete data. Two versions of "amelia" mostly share the same underlying code and algorithms but have different benefits and limitations. They are "amelia" (Honaker et al. (2010)) and "amelia II" (Honaker et al. (2011)). This thesis paper only focuses on "amelia".

Bootstrap was first introduced by Efron (1979) as a statistical method of resampling [38]. According to Davison and Hinkley (1997) and Singh and Xie (2008),

$$\hat{P} \rightarrow Z^* = (z_1^*, z_2^*, \ldots, z_n^*) \rightarrow f(Z^*) \tag{25}$$

Where Z^* is getting through random sampling from estimated population $\hat{P}$. $Z^* = (z_1^*, z_2^*, \ldots, z_n^*)$ is the bootstrap dataset and $f(Z^*)$ is the bootstrap estimate. They also explain how the bootstrap method is categorized under the broader framework of resampling methods. To understand the bootstrap method conceptually, let a scenario be considered where it is possible to repeatedly collect samples of the same size from the population of interest. If this process could be performed many times, it would provide a robust approximation of the sampling distribution for a given statistic. However, such an approach is impractical due to the associated costs and the purpose of a sample study, which is to obtain information efficiently and at a lower cost. The bootstrap addresses this issue by treating the available sample data as a "surrogate population" from which repeated resampling (with replacement) can be performed. These resamples are known as bootstrap samples, which estimate the sampling distribution of the statistic. By simulating many bootstrap samples (typically thousands), the desired statistic is computed for each, resulting in a distribution of these values. This distribution is referred to as the bootstrap distribution of the statistic. One of the fundamental applications of the bootstrap is to produce numerous replications of the sample statistic through these bootstrap samples. From this collection of computed statistics, a small percentage, such as $100(\alpha/2)\%$ (commonly with $\alpha = 0.05$), is removed from both the lower and upper ends of the distribution. The remaining $100(1-\alpha)\%$ values define the confidence interval for the unknown population parameter, with the associated confidence level being approximately $100(1-\alpha)\%$. This procedure is known as the bootstrap percentile method. The next concept to understand the algorithm behind "amelia" is "Expectation-Maximization".

[38] Resampling method repeatedly draws samples from the observed data and estimates the sampling distribution of any statistic without depending on its parametric properties.

According to Ng et al. (2012), the Expectation-Maximization (EM) approach makes parameter estimation simpler in complex models where it isn't easy to maximise the likelihood function directly. They claim that it is beneficial for addressing missing or incomplete data while producing maximum likelihood estimations iteratively.

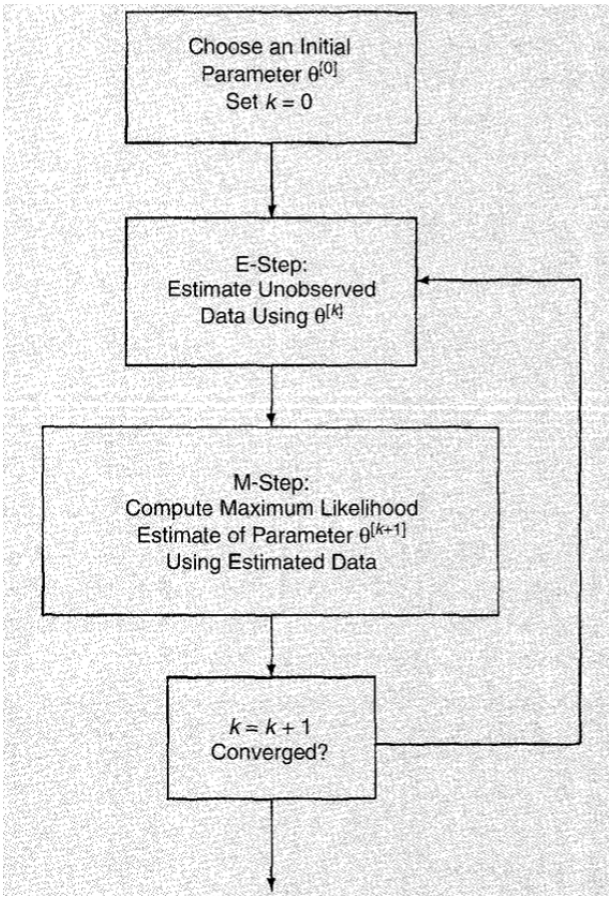


Figure 6: Simple Overview of EM Algorithm. Graphic Source: Moon (1996)

Let $g(y;\theta)$ be considered the probability density function (p.d.f.) of the observed data y, parameterized by θ. Finding the parameter vector θ that maximises the likelihood function $L(\theta) = g(y;\theta)$ is the objective of maximum likelihood estimation or MLE. This involves solving the score equations $\frac{\partial L(\theta)}{\partial \theta} = 0$ or, equivalently, $\frac{\partial \log L(\theta)}{\partial \theta} = 0$. However, most often, the likelihood function often exhibits nonlinearity concerning the parameters. This nonlinearity may arise from models with nonlinear parameter relationships, non-normal error distributions, missing data, or dependencies among variables. Traditional methods like the Newton-Raphson could be used here. However, they can be demanding analytically and computationally. This is especially true in complex scenarios. The EM algorithm provides a systematic and iterative approach. It helps address these challenges. It is particularly advantageous when dealing with incomplete or missing data. The algorithm alternates between two main steps: the Expectation (E) step and the Maximization (M) step. The algorithm then calculates

the expected value of the log-likelihood function in the E-Step. This is done using the current parameter estimates. It also takes the incomplete data into account. This is formally represented as $Q(\theta|\theta^{(k)}) = \mathbb{E}[\log L(\theta; y)|y^{(k)}]$, where $\theta^{(k)}$ denotes the parameter estimates from the k-th iteration. The M-Step then involves maximizing this expected log-likelihood function to update the parameter estimates, which is expressed as $\theta^{(k+1)} = \arg\max_\theta Q(\theta|\theta^{(k)})$. The EM algorithm iterates between these E-Step and M-Step until convergence when the parameter estimates stabilize and the likelihood function reaches a local maximum. It is highly versatile. This makes it effective for models with missing data or latent variables. It is also useful for complex structures where direct likelihood maximization is impractical. The EM algorithm decomposes the optimization problem into smaller steps. This makes the problem more manageable. The algorithm provides a computationally efficient method for parameter estimation. It is an important tool in modern statistical analysis. The visual representation of how "amelia" works is given below,

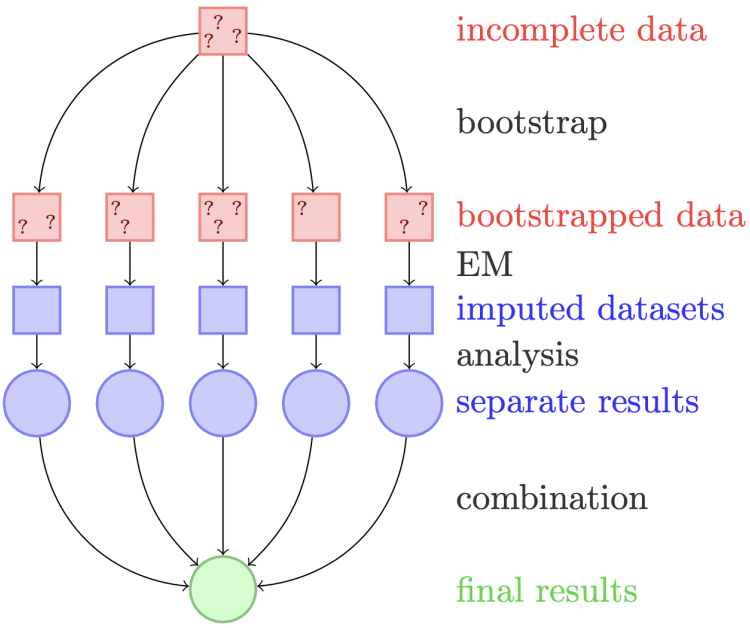


Figure 7: Framework of "amelia". Graphic Source: Honaker et al. (2011)

According to Honaker et al. (2011), the process of "amelia" could be,

1. **Initialize:**
    - Observed data D_{obs}
    - Missing data indicator matrix M
    - Initial parameters $\theta_0 = (\mu_0, \Sigma_0)$

2. **For each bootstrap iteration** $b = 1$ to B:

   (a) **Bootstrap the data:**

   - Draw a bootstrap sample D_b from the observed data D_{obs}.

   (b) **EM Algorithm:**

   - Set initial guess $\theta_b = \theta_0$.
   - **Repeat until convergence:**

     i. **E-step:** Estimate missing data D_{mis} from the current parameters θ_b.

     ii. **M-step:** Update the parameters $\theta_b = (\mu_b, \Sigma_b)$ using both D_{obs} and D_{mis}.

     (iii) Recalculate the mean μ_b and covariance matrix Σ_b.

   (c) **Draw parameter sample** θ_b:

   - After the EM algorithm converges, draw posterior samples of the complete-data parameters $\theta_b = (\mu_b, \Sigma_b)$.

3. **Impute missing data:**

   - For each bootstrap sample θ_b:

     (i) For each imputation $i = 1$ to m:

     i. Draw missing data D_{mis} from the conditional distribution $p(D_{\text{mis}}|D_{\text{obs}}, \theta_b)$.

     ii. Create a complete dataset by combining D_{obs} and the imputed D_{mis}.

4. **Repeat for** m **imputations:**

   - Generate m imputed datasets, each with imputed values drawn from the posterior of θ_b.

Output: A set of m imputed datasets.

The R code imputes missing data using the "amelia" package for balanced and unbalanced panel datasets. The function *Data_Imputation_Amelia* selects relevant variables ("ID", "Year", "Age", "Employment Types", "Income", "Marital Status", "Sex") and converts the Year column to numeric for time-series handling. The "amelia" algorithm imputes missing values with panel-specific settings, treating "Year" as the time variable (ts) and "ID" as the cross-sectional identifier (cs) while specifying "Employment Types", "Marital Status", and "Sex" as categorical variables (noms). After simulation, exogenous variables ("Employment Hours" and "Education") are added back to each simulated dataset, and the column order is restored. The simulated datasets are stored separately, providing multiple completed datasets for further analysis while utilizing the panel structure of the data.

4.4 LSTM-Network

As described in Pascanu et al. (2013), a recurrent neural network (RNN) is designed to simulate a discrete-time dynamical sequential data that has an input x_t, an output y_t, and a hidden state h_t. In the given notation, the subscript t represents time. The dynamical system is defined by

$$h_t = f_h(x_t, h_{t-1}) \tag{26}$$

$$y_t = f_o(h_t) \tag{27}$$

where f_h and f_o are a state transition function and an output function. Each function is parameterized by a set of parameters: θ_h and θ_o.

A set of N training sequences $D = ((x_1^{(n)}, y_1^{(n)}), ..., (x_{T_{(n)}}^{(n)}, y_{T_{(n)}}^{(n)}))_{n=1}^{N}$, the parameters of the RNN can be estimated by minimizing the following cost function:

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{T=1}^{T_n} d(y_t^{(n)}, f_o(h_t^{(n)})) \tag{28}$$

where $h_t^{(n)} = f_h(x_t^{(n)}, h_{t-1}^{(n)})$ and $h_o^{(n)} = 0$. d(a, b) is a predefined divergence measure between a and b. For example, Euclidean distance or cross-entropy.

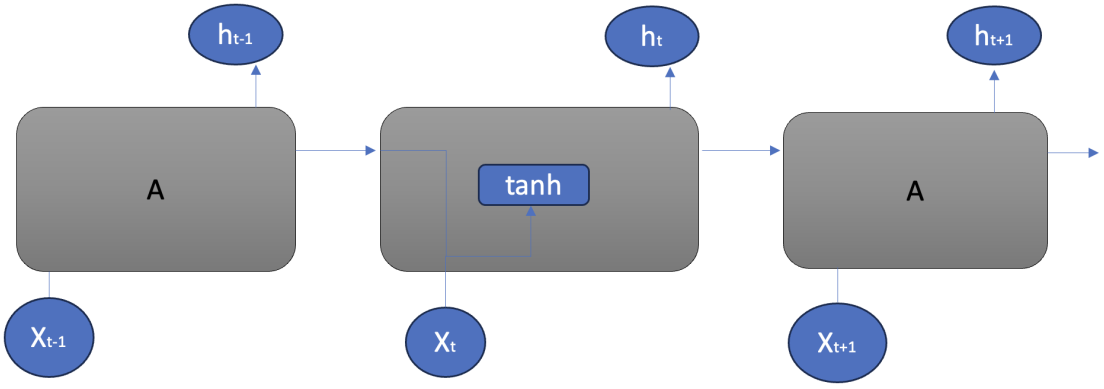


Figure 8: The Repeating Module in a Standard RNN for a Single Layer

"A" represents a segment of the neural network in Figure 8. Where X_{t-1} as the input and h_{t-1} as the output. Recurrent neural networks include a sequence of repeating neural network modules. This module usually consists of a single tanh layer in a regular RNN. However, the number of layers could be higher. RNN is critical for tasks like time series data analysis based on earlier data because they can connect earlier data to the present when relevant information is tightly tied to time. For example, utiliz-

ing earlier stock data may help a researcher to comprehend the current stock price. However, RNN may be unable to handle relationships, including long-term dependencies. RNN struggles to learn correlations as the gap between relevant information widens, as shown in Staudemeyer and Morris (2019). Chang et al. (2017) improved computational efficiency with dilated RNN by reducing the number of parameters. Although theoretically capable of handling long-term dependencies, RNN often fail to learn them in practice, necessitating careful feature selection.

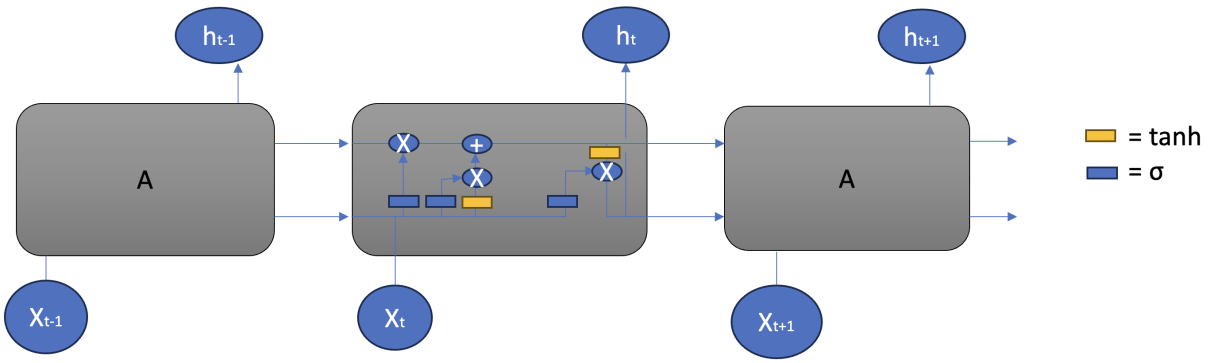


Figure 9: The Repeating Module in a Standard RNN for Multiple Layers

Figure 9 describes how the four layers connect one to another to capture long-term dependencies. It portrays how the multiple layers look like, where each line represents an entire vector from one node's output to another's input. Instead of a single neural network layer, four layers are interacting in a unique manner. "LSTM-Network" also has a chain-like structure[39], but its repeating module is different. Fortunately, "LSTM-network"—a particular kind of recurrent neural network—do many jobs more effectively than regular RNN because they do not suffer from this limitation. Long-term dependencies can be learned using "LSTM-Network" discussed in Muzaffar and Afshari (2019). "LSTM-Network", or Long Short Term Memory Network, is a special kind of RNN that can identify long-term dependencies. "LSTM-Network" was first presented by Hochreiter and Schmidhuber (1997), and it has since been improved upon and made more widely known.

In Figure 10, orange circles indicate pointwise operations, such as vector addition, while blue boxes signify learned neural network layers. Merging lines indicate concatenation, whereas forking lines show content copied and sent to different locations. The horizontal line at the top represents the cell state, crucial to the "LSTM-Network", functioning like a conveyor belt with minimal linear interactions, allowing information to flow unaltered.

[39]The "LSTM-Network" concept is understood through the graphical and equational representation from Li et al. (2020) and Olah (2015).

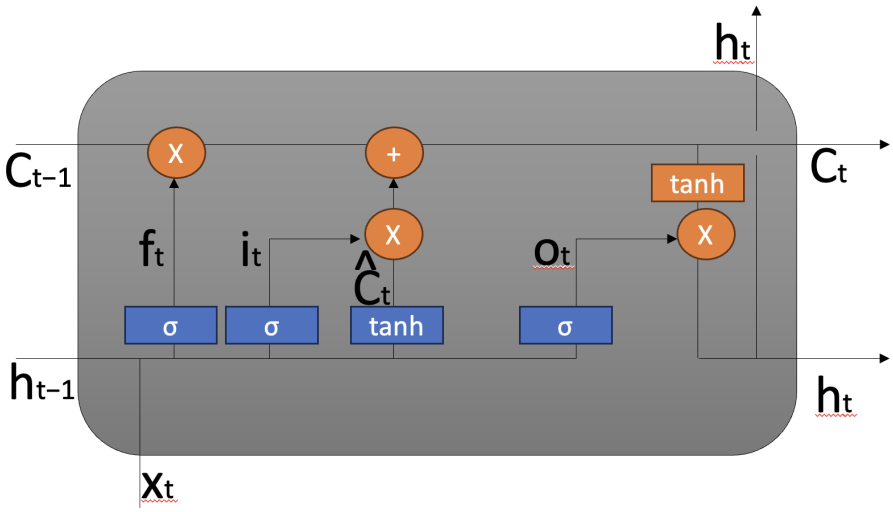


Figure 10: Moving Information Via the Cell State of "LSTM Network"

The "LSTM-Network" uses gates to control the addition or removal of information from the cell state. Gates uses a sigmoid layer and pointwise multiplication to regulate this process. The sigmoid layer generates values between zero and one to indicate how much each component should be allowed through. When a value is zero, it means "let nothing through". When a one means "let everything through". These three gates safeguard and regulate the cell state in an "LSTM-Network". The whole process of "LSTM-Network" could be categorized into four steps. The first step is to choose which information about the cell state must be ignored by the "LSTM-Network". To put this concept into practice, a sigmoid layer called the "forget layer" is responsible. It examines each number of h_{t-1} and x_t in the cell state C_{t-1}. The cell state C_{t-1} produces a number between 0 and 1. Completely keeping something is represented by 1, whereas completely getting rid of something is represented by 0.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{29}$$

$$\widehat{C_t} = tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{30}$$

The next stage is choosing the new information to be stored in the cell state. There are two components to this. First, the values to be updated are selected via a sigmoid layer known as the "input gate layer." After that, a tanh layer generates a vector, $\widehat{C_t}$, which includes potential new candidate values that could be included in the state. Combining these two to produce a state update is the next step.

$$C_t = f_t * C_{t-1} + i_t * \widehat{C_t} \tag{31}$$

In the following steps, the previous cell state C_{t-1} has to be changed to the current cell state C_t. Implementing the suitable actions that the previous steps have already determined is necessary. The old state is multiplied by f_t in order to remove the information that the preceding step had decided to remove. Next, add $i_t * \widehat{C_t}$. These are the updated candidate values, scaled by each state value's estimated update.

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{32}$$

$$h_t = o_t * tanh.(C_t) \tag{33}$$

The output, a filtered representation of the cell state, must be determined at the last stage. Initially, a sigmoid layer selects which cell state components to output. After that, a tanh function is applied to the cell state to limit its values to -1 and 1. This result is multiplied by the sigmoid gate's output to ensure that just the chosen parts are output.

Some versions of "LSTM-Network" are as follows: Gers and Schmidhuber (2016) includes peephole connections allowing gate layers to access cell states. The Gated Recurrent Unit (GRU), introduced by Cho et al. (2014), simplifies the model by combining forget and input gates into a single "update gate" and merging cell and hidden states. Other notable variants include Depth Gated RNN by Yao et al. (2015) and Clockwork RNNs by Koutnik et al. (2014). Comparative studies, such as Greff et al. (2015), indicate that these variants perform similarly, although Jozefowicz et al. (2015) found some RNN architectures that outperform "LSTM-Network" on specific tasks.

The function *Data_Imputation_LSTM* is designed to perform missing data imputation using an "LSTM-Network" model. R package "keras" is used to implement the model (see Arnold (2017), and Gulli and Pal (2017)). Initially, the function excludes the exogenous ("Employment Hours" and "Education") variables. It then removes rows with missing values in the target variable ("Income"), ensuring that only complete data is used to train the model. Next, the continuous variables "Age" and "Income" are normalized by subtracting their means. Then, dividing by their standard deviations ensures the features are on a similar scale, which is important for neural network performance. The categorical variables, including "Year", "Employment Types", "Mari-

tal Status", and "Sex", are converted to numeric values using factor encoding, making them compatible with the "LSTM-Network" model. The prepared data is then formatted into a 3D array. This is the required input format for LSTM-Network models. The dimensions correspond to samples, timesteps and features. Timesteps are set to 1, and features set to 5. A dense layer that outputs a single value after an "LSTM-Network" layer with 50 *units* [40], is trained using the rearranged data. The model is trained by the dataset for 50 *epochs* and 32 *batch_size* [41] using the mean squared error loss function and the adam optimizer for compilation. The iterations of the model through the datasets are determined by epochs during training. The model has more learning opportunities for a large number of epochs. However, it can lead to overfitting, where the model memorizes the training data and cannot perform on the unseen data. After the model is trained on the complete data, it is used to predict the missing "Income" values for the rows that initially had missing data. These rows are preprocessed in the same way as the training data, including encoding categorical variables and normalising "Age". The predictions made by the "LSTM-Network" model are then rescaled to their original scale by multiplying by the standard deviation of "Income" and adding the mean. This ensures that the imputed values are comparable to the original "Income" data. Finally, the imputed values are inserted back into the dataset, and the exogenous variables "Education" and "Employment Hours" are added to each simulated dataset. The dataset is then reordered to match the original structure, and the simulated dataset is returned. This process is then repeated for all the datasets described in 3.2 and stored the simulated datasets in different dataframes for further analysis.

This is the end of the "Methodology and Simulation" section. This is also the end of explaining the theories behind the application part. From the first three sections, it is clear why studying the imputation of panel data is important. It is also clear what the researchers have explored in this field, what gaps need to be explored, and what needs to be considered before studying the imputation of panel data. This chapter explains the frameworks and the algorithm behind their mechanism. Furthermore, the simulation process of imputing the missing values. The next sections will explain the application of the imputation frameworks to impute the missing values. Finally, based on the comparison of the results, this thesis study will find suitable frameworks to impute missing values in panel data for any specific condition.

[40] The unit parameters define the number of cells in the hidden layers. More units ensure the learning capacity of distinct features. However, it brings more computational complexity to the equation. Fifty units in this equation mean this model can learn 50 different relationships of the data.

[41] *batch_size* is responsible for controlling the number of samples the model can process before updating its weight.

5 Results and Discussion

All the calculations are implemented using macOS (12.1.1(24.B91)), 8 GB Memory, and Apple M1 Chip. Each package ("mice", "mitml", and "amelia") simulated five sets of imputed datasets. Then, five datasets are stacked in one data set and used for further performance analysis. "LSTM-Network" provides a single simulated dataset. The computational time of each framework is counted to compare the computational efficiency. Distribution and conditional distribution curves are generated for all simulated datasets and compared to the original datasets to analyse the distribution of the simulated datasets. Correlation tables compare the relationship among the variables before and after the imputation. Wasserstein distance, Bias, and RMSE are used to understand the distributional distance among the simulated and original datasets. Various resources are consulted to address coding challenges encountered during the thesis work, including "R documentation", "stack overflow discussions", "video tutorials from multiple creators", and generative AI tools like "ChatGPT". These resources provided essential guidance for troubleshooting, refining and optimizing the R codes.

5.1 Computational Efficiency

The imputation frameworks are different and the parameters used in the frameworks are also different and not comparable. That is why deciding which framework is the fastest is not wise based only on computational time. The table below provides an overview of the computational time and the parameters used. The parameters not present in the table are used as default values from the installed packages.

Frameworks	Version	Computational Time	Parameter Values
mice	3.16.0	13.11498 mins	m = 5, maxit = 100, method = 'pmm'
mitml	0.4.5	25.23591 mins	m = 5, n.iter = 100, n.burn = 1000
LSTM-Network	Keras (2.15.0)	2.409152 hours	units = 50, epochs = 50, batch_size = 32
amelia	1.8.2	4.240052 hours	m = 5, ts = "Year", cs = "ID"

Table 13: Computational Efficiency of the Frameworks

From the table 13, "mice" is simply the fastest package to impute the missing values. "mitml" comes in second place with acceptable speed. However, the "LSTM-Network" and "amelia" are significantly slow due to their imputing algorithms inside their respective frameworks.

5.2 Distributions Comparison

In data analysis, density curves are responsible for showing the distribution and giving a smooth, continuous representation of the underlying structures of datasets. Density curves provide a clearer representation of properties like kurtosis[42], skewness[43], and tails than histograms, since they are not affected by bin size like a histogram. Density curves are especially helpful in determining how closely imputed datasets resemble the distribution of the original data. Density curves are crucial for validating imputation techniques because they show whether the data's general structure and important characteristics have been maintained. Differences in density curves highlight limitations in the imputation process, which is crucial information for enhancing the frameworks. The original balanced and unbalanced panel data is the benchmark for imputation performance. Providing the performance rank is difficult based on the density curves because the performance of the imputation frameworks is close and could be subjective. It might depend on the view of the observer. That is why no rank of the performance has been given.

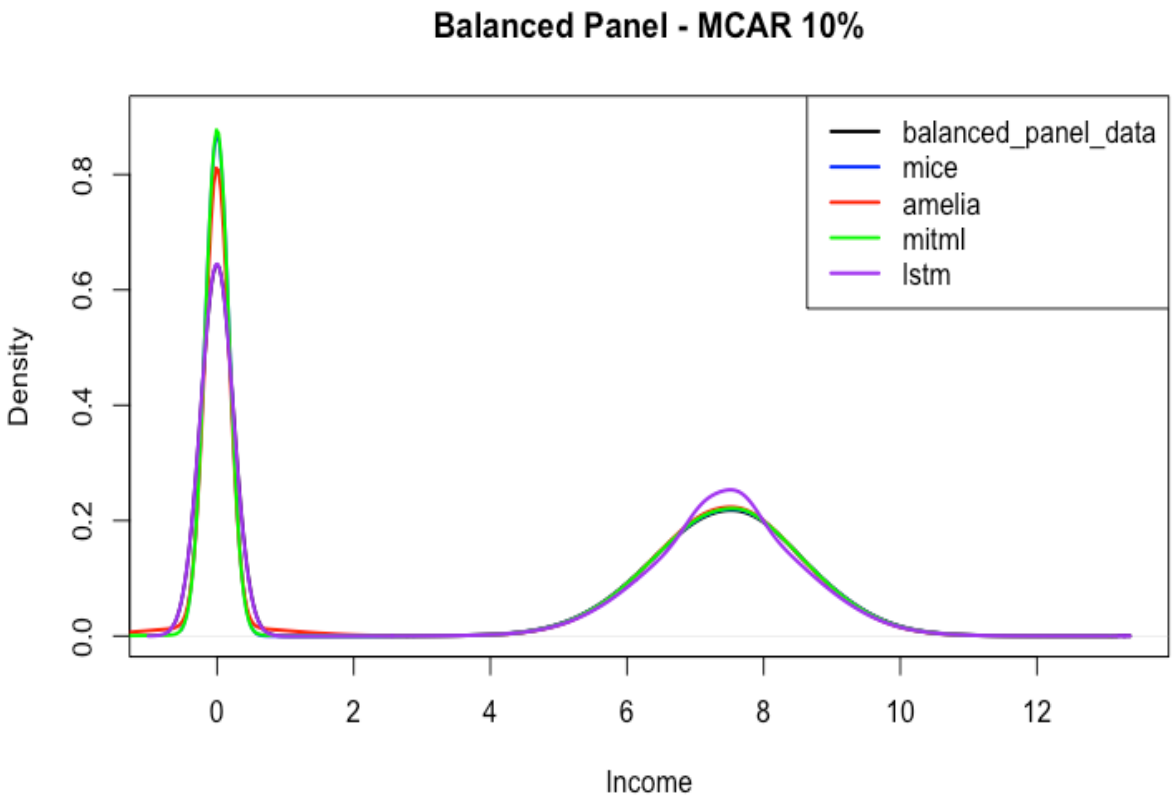
Figures 11-19 illustrate the distribution of "Income" without any condition over it. Overall, All of them struggled more or less while performing MAR and MNAR conditions. "mice" performed better in overall imputation but not at its best in MNAR settings. "amelia" and "mitml" are mostly the same, but "amelia" is slightly better. The performance of "LSTM-Network" is not commendable except for 10% missingness.

The comparison of distributions highlights how imputation frameworks respond to overall "Income." However, it is not clear how these frameworks perform across specific categories within variables. To address this, the conditional distribution is examined. Analyzing plots for every scenario (MCAR, MAR, MNAR at 10%, 30%, and 50% missingness) would be overwhelming. Therefore, this discussion focuses on the 30% missingness conditional distribution scenarios for MCAR, MAR, and MNAR. The remaining plots are provided in the appendix.

Figures 20-25 illustrate the conditional distribution of the "Income" variable over "Age" and "Marital Status." "Marital Status" includes four categories: Single, Married, Divorced, and Widow. "Age" is filtered from 15 to 65, as this range encompasses most working individuals. The legend follows the distribution of "Income": black lines represent the original data, blue corresponds to "mice," red to "amelia," green to "mitml," and purple to "LSTM-Network." A detailed explanation of the individual performance of the frameworks is given below,

[42] Kurtosis measures the peak of a distribution.

[43] Skewness measures how symmetric the distribution is.



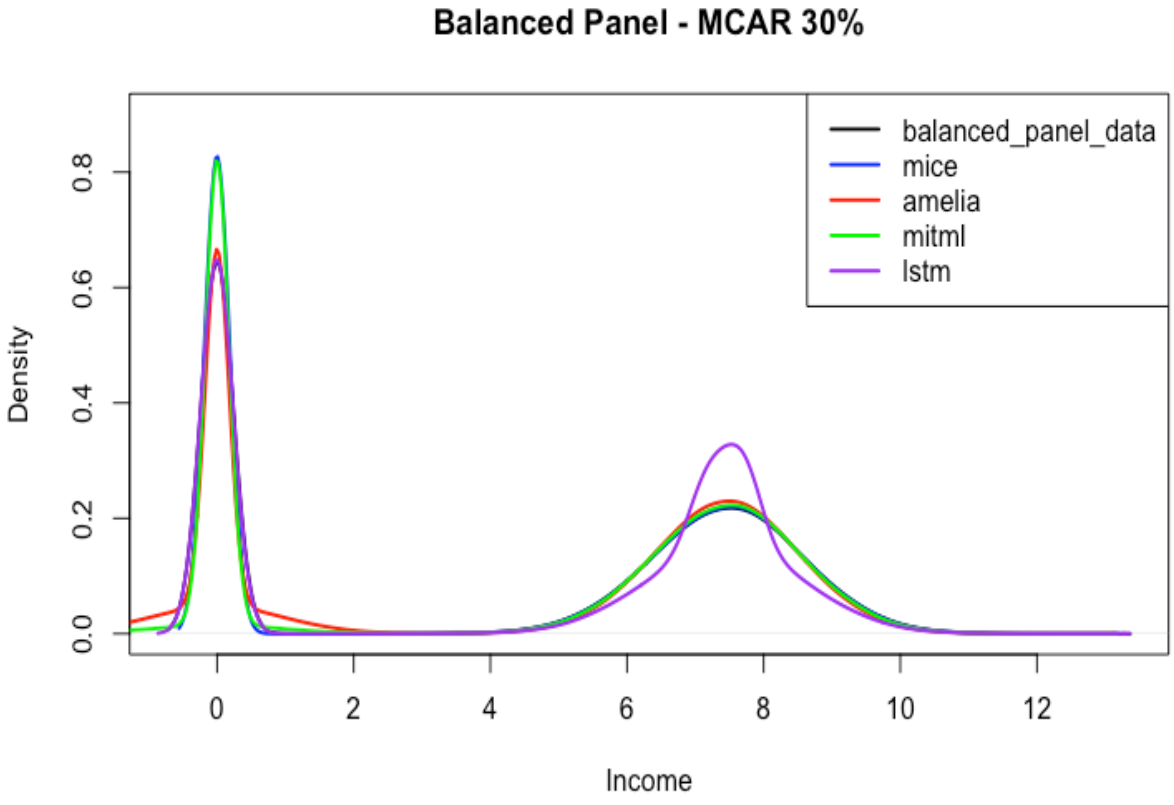
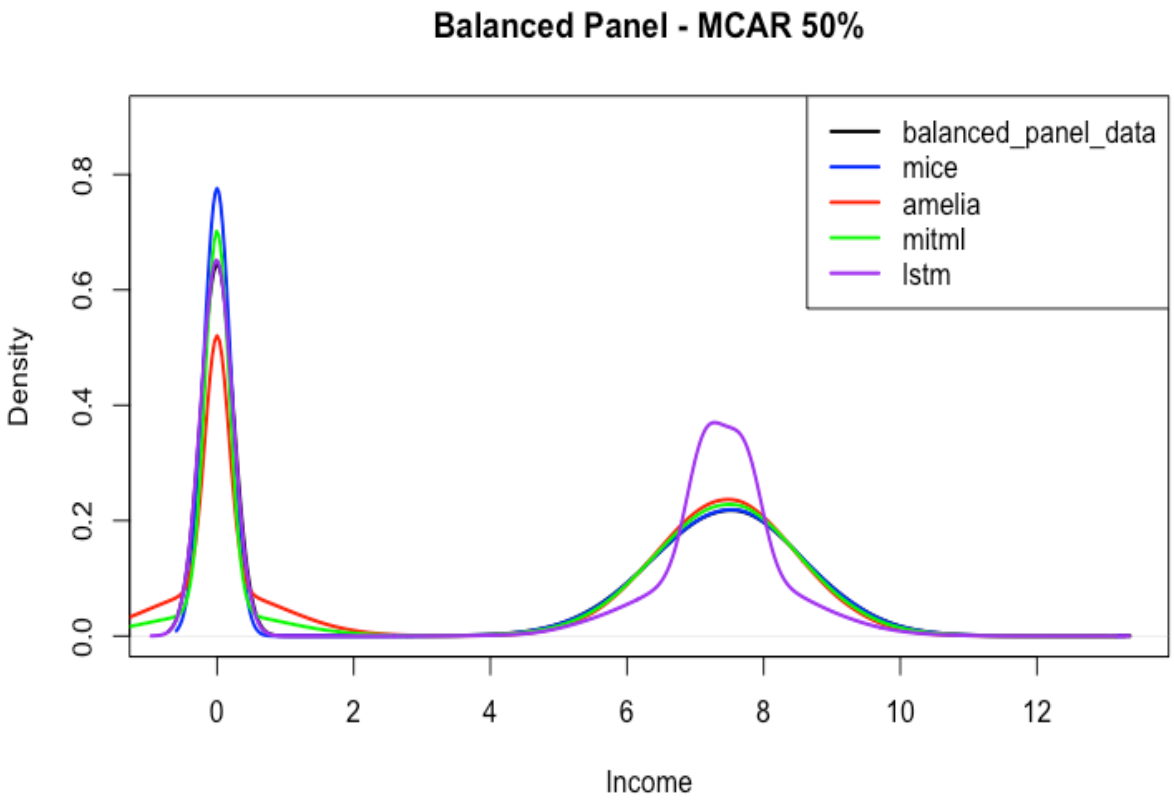


Figure 11: Distributions Comparison of "Income" for Different Frameworks Part-01



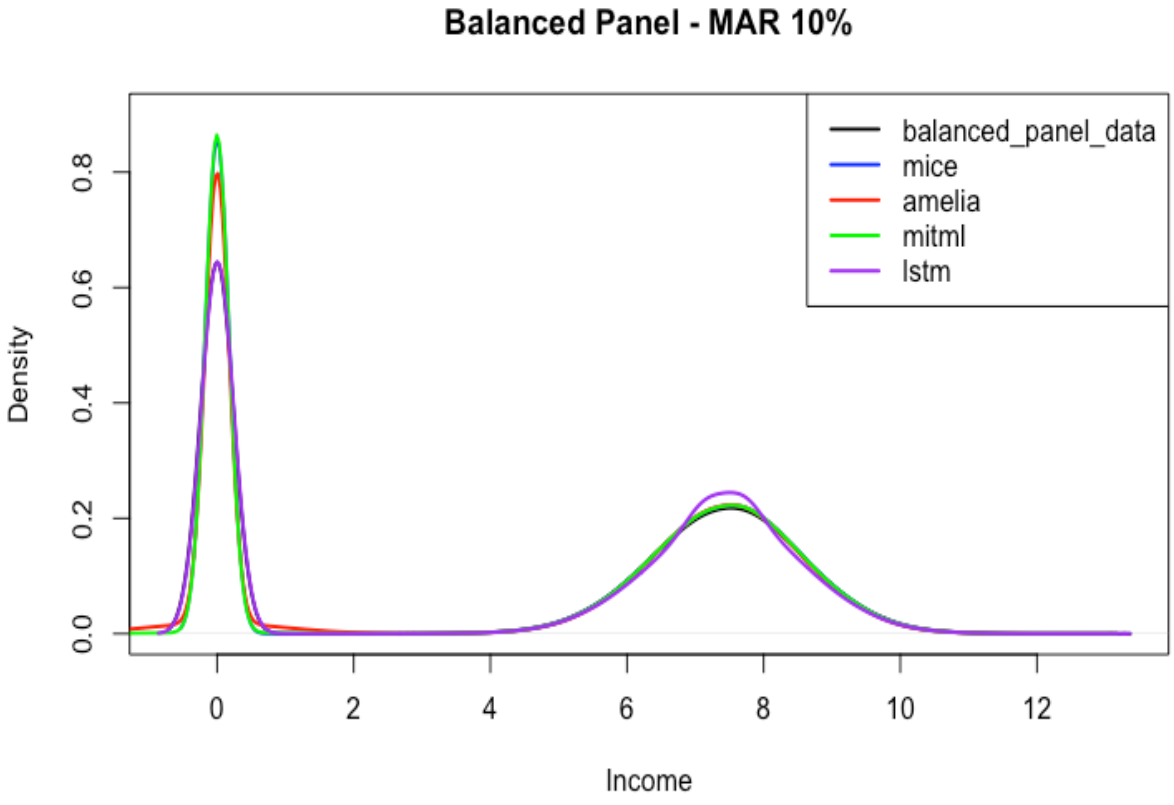
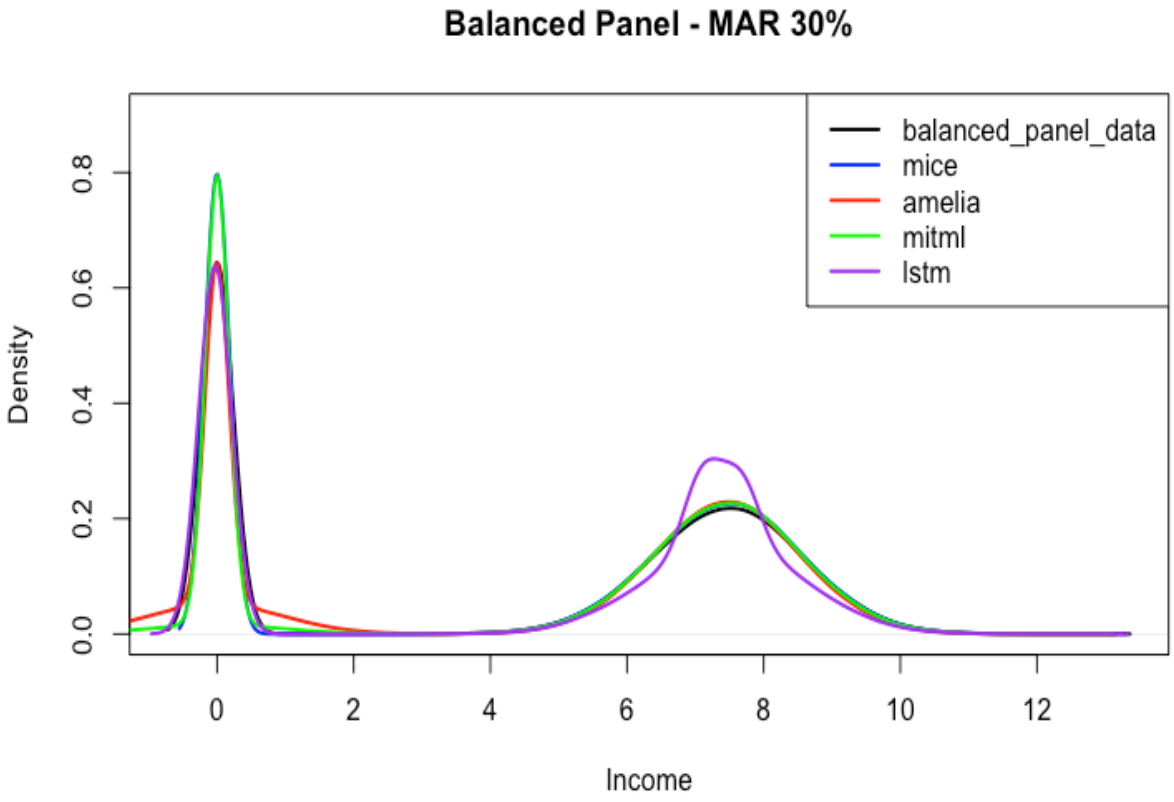


Figure 12: Distributions Comparison of "Income" for Different Frameworks Part-02



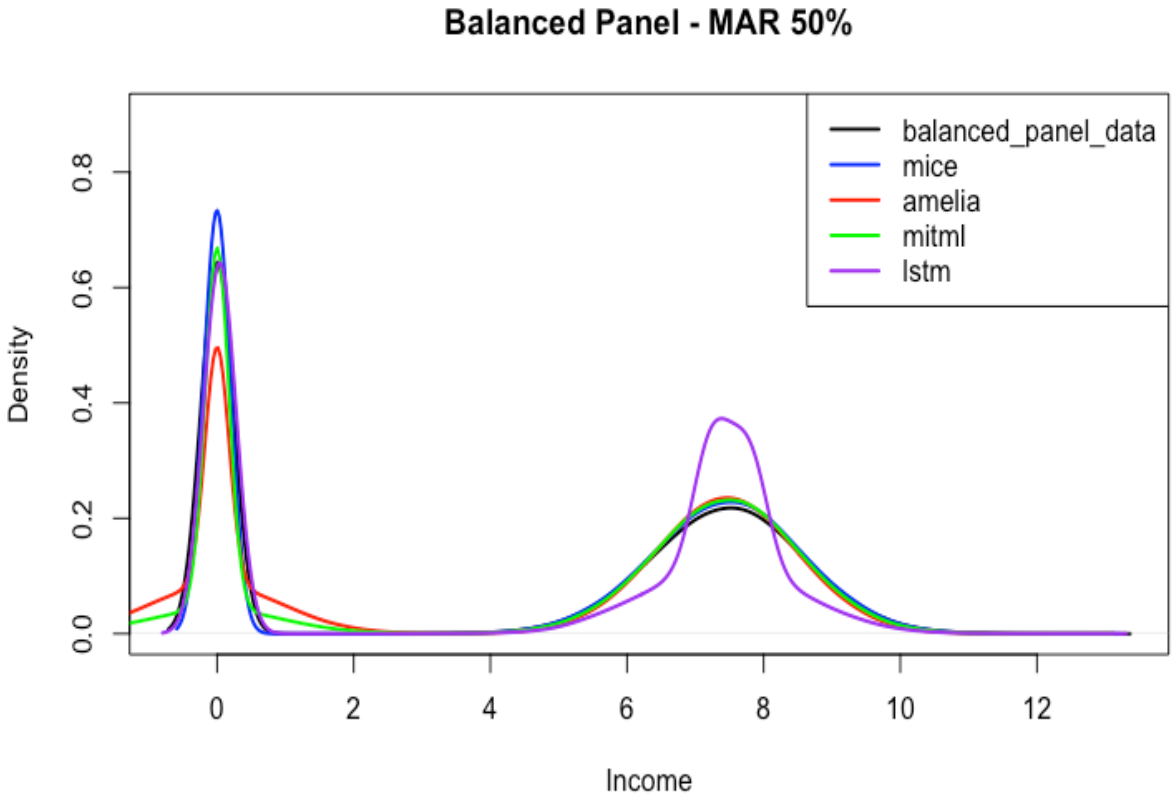
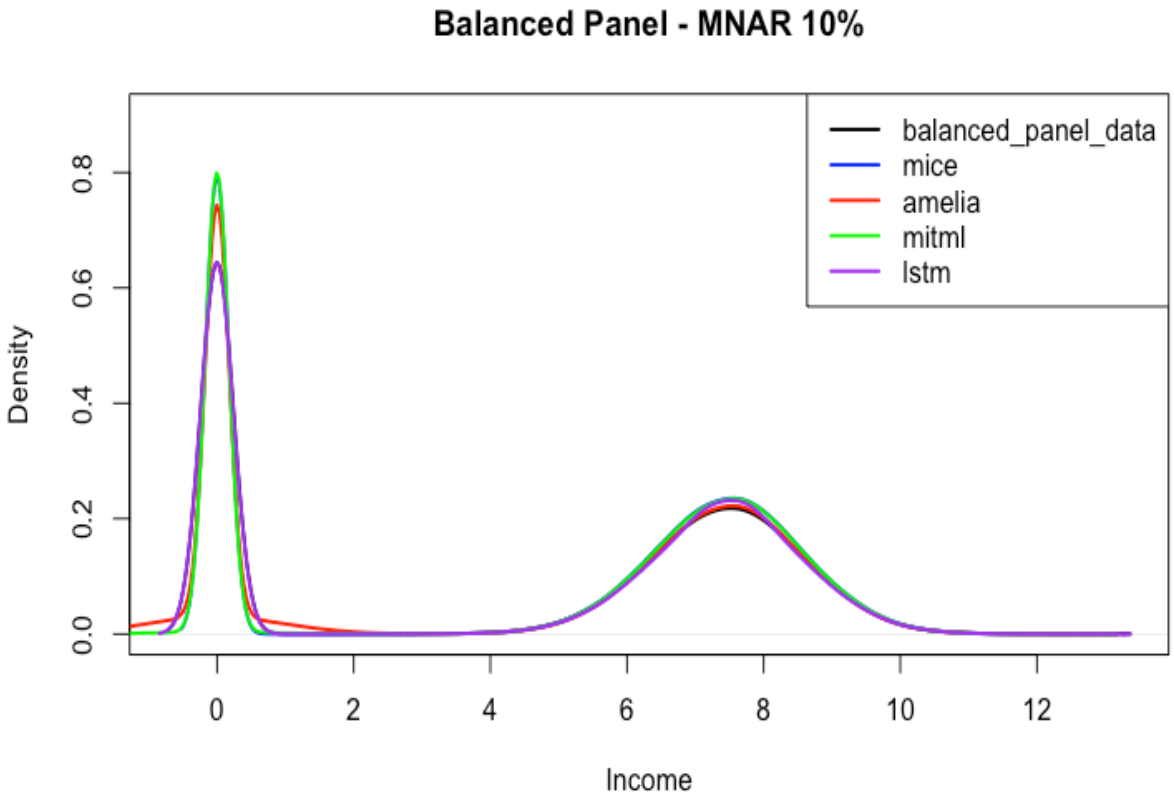


Figure 13: Distributions Comparison of "Income" for Different Frameworks Part-03



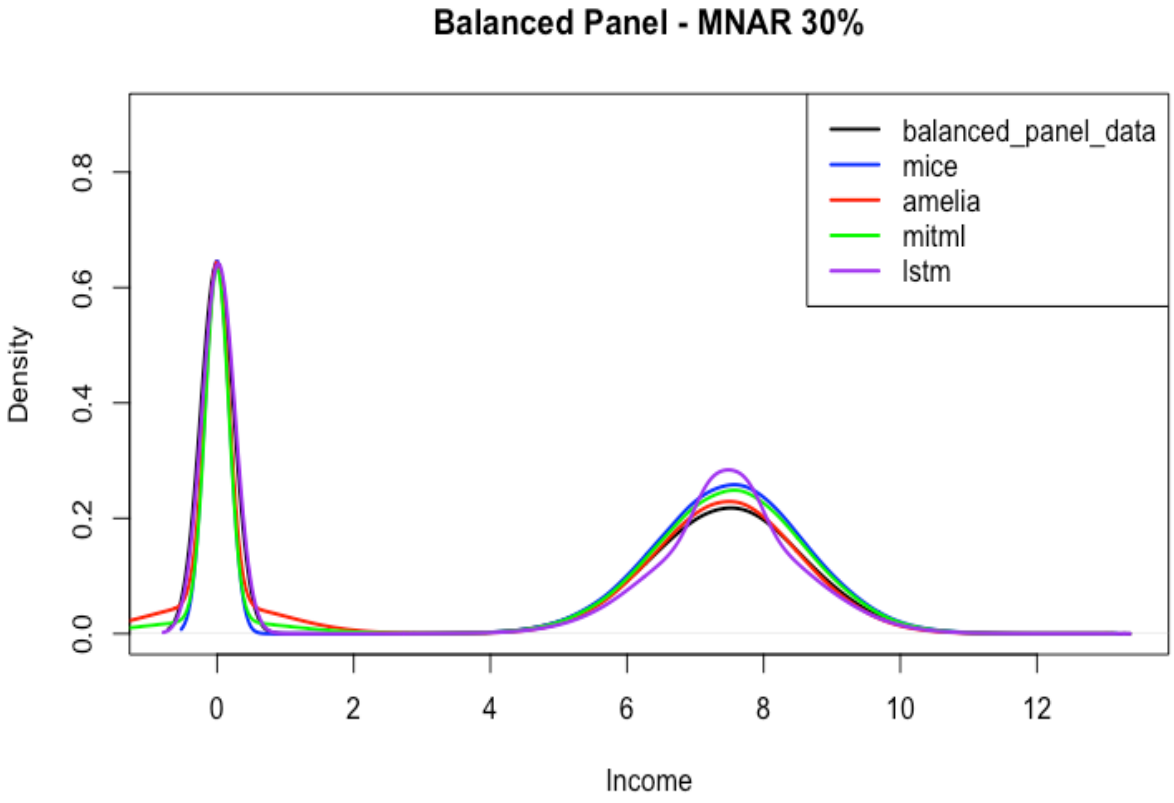
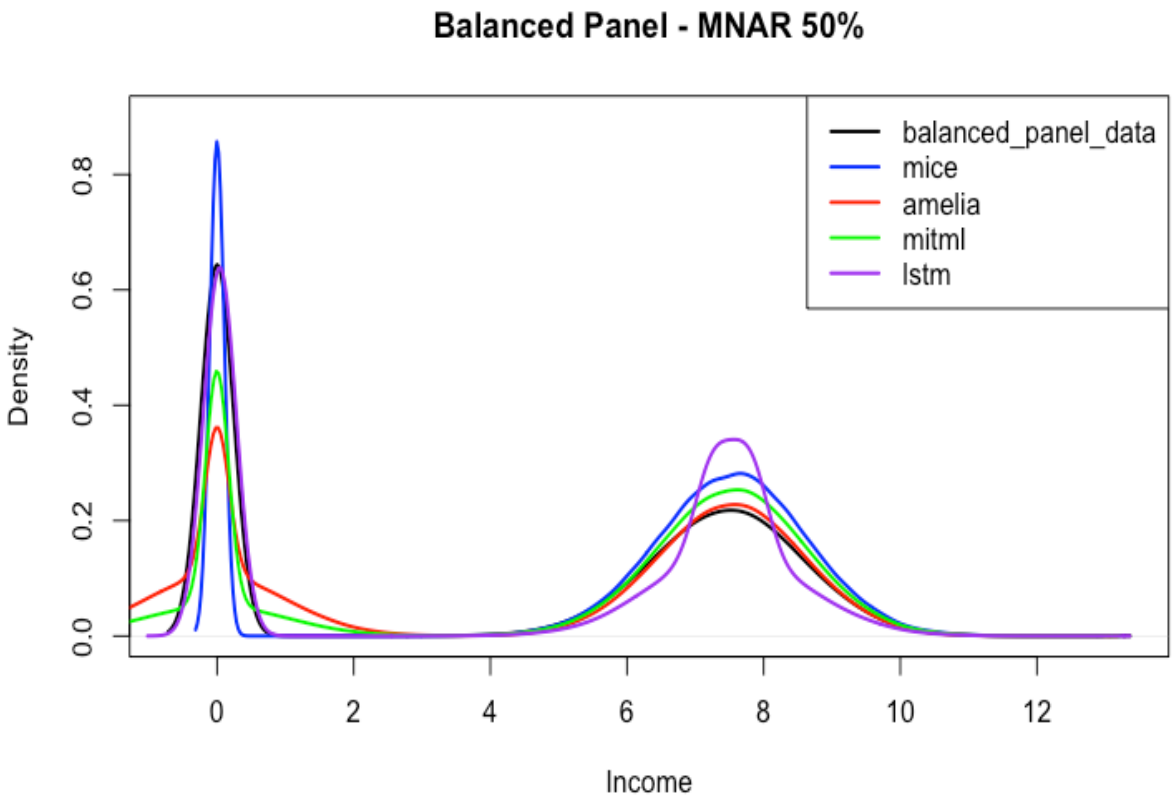


Figure 14: Distributions Comparison of "Income" for Different Frameworks Part-04



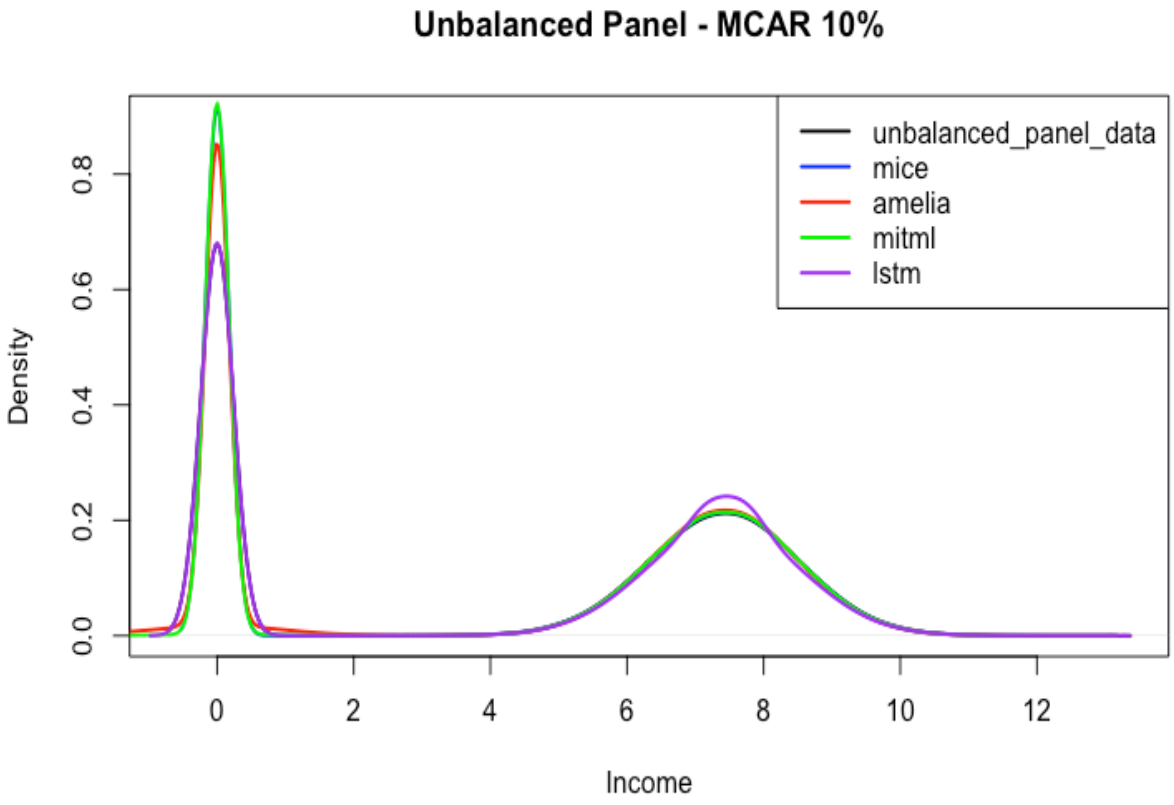
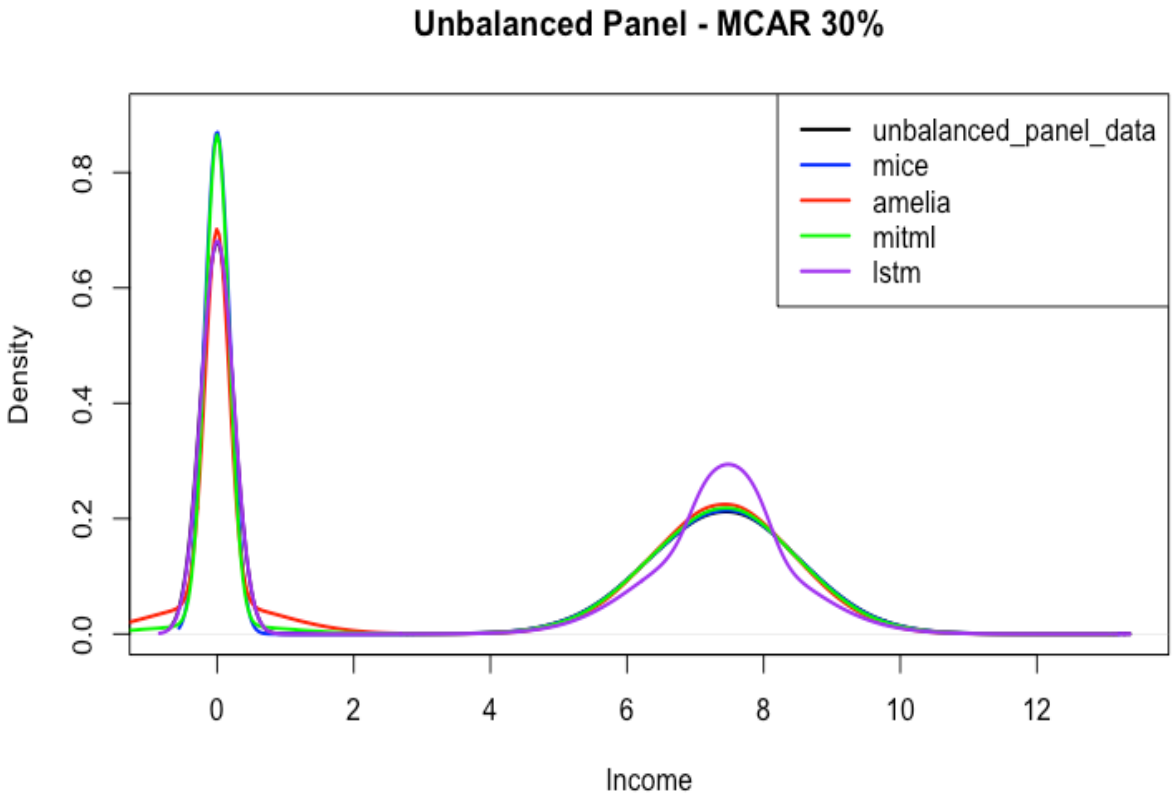


Figure 15: Distributions Comparison of "Income" for Different Frameworks Part-05



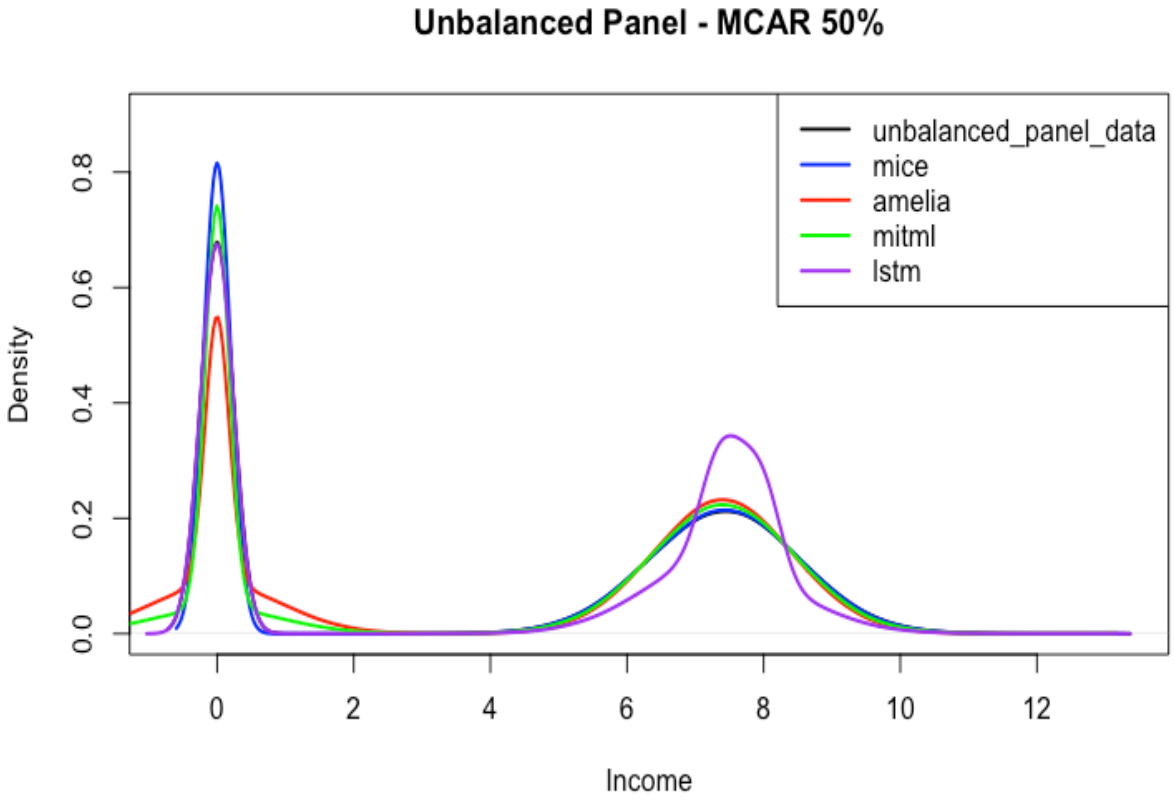
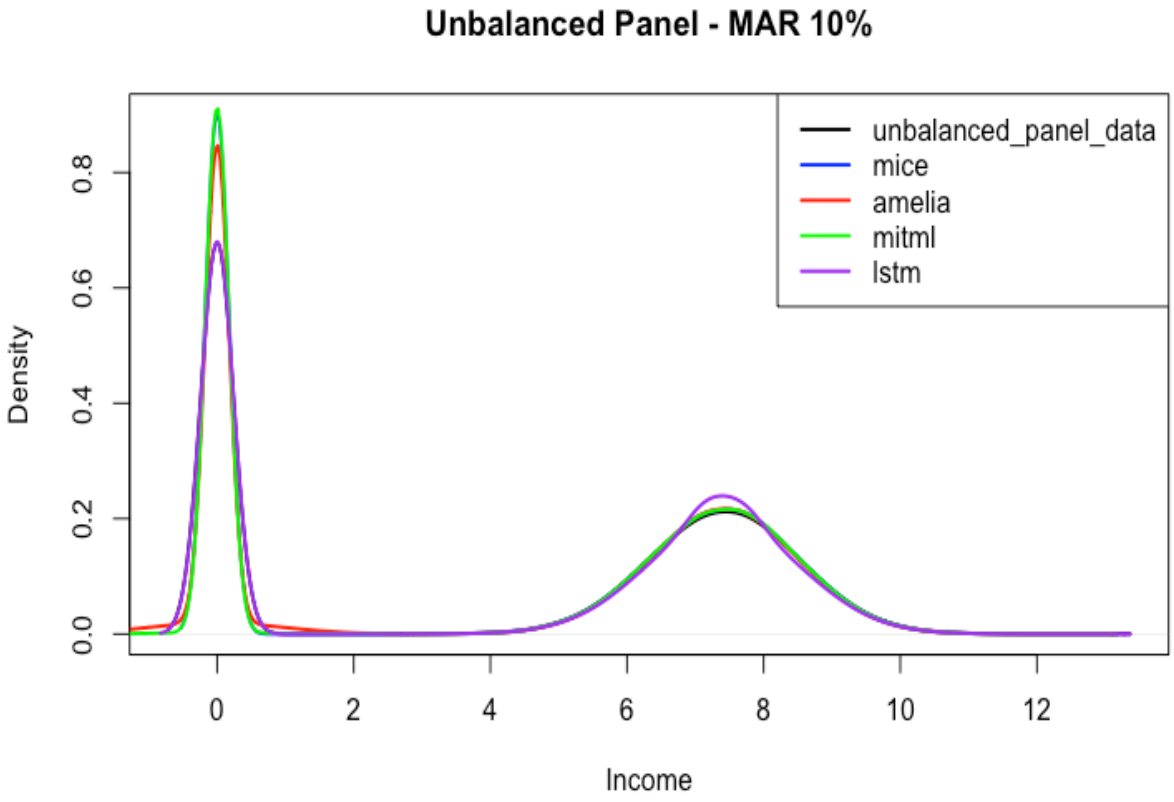


Figure 16: Distributions Comparison of "Income" for Different Frameworks Part-06



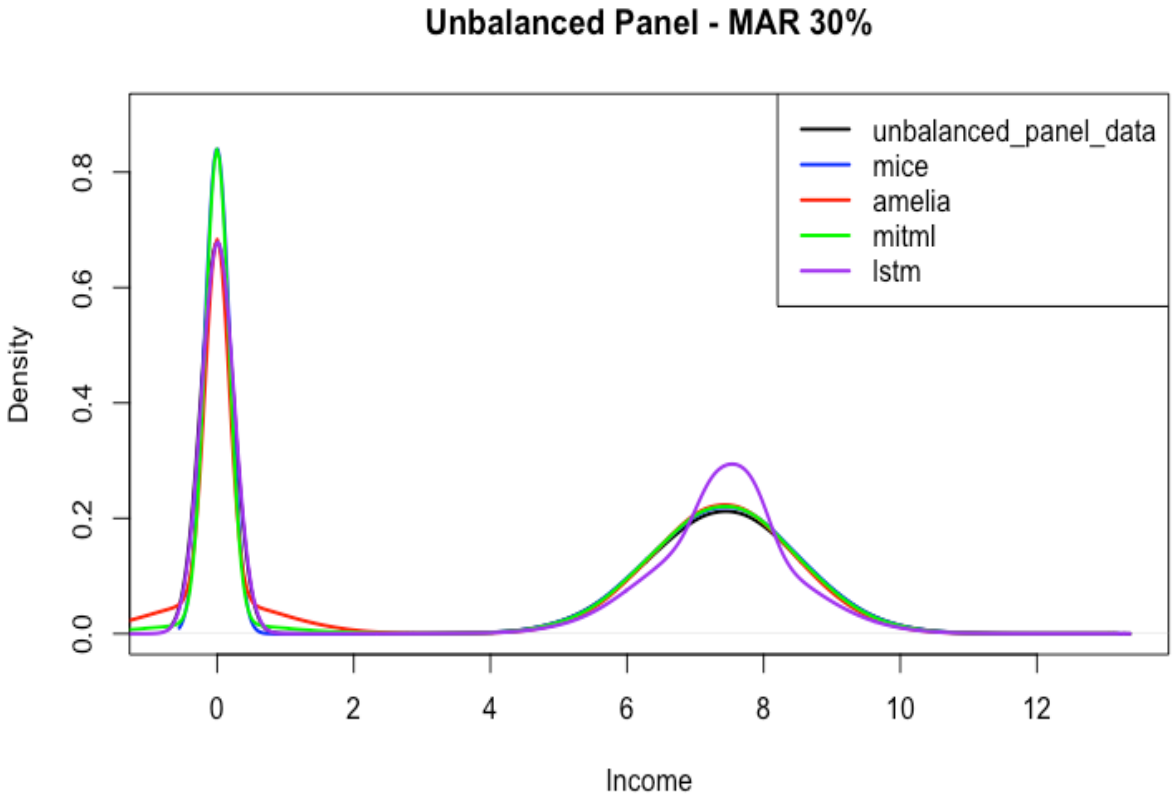
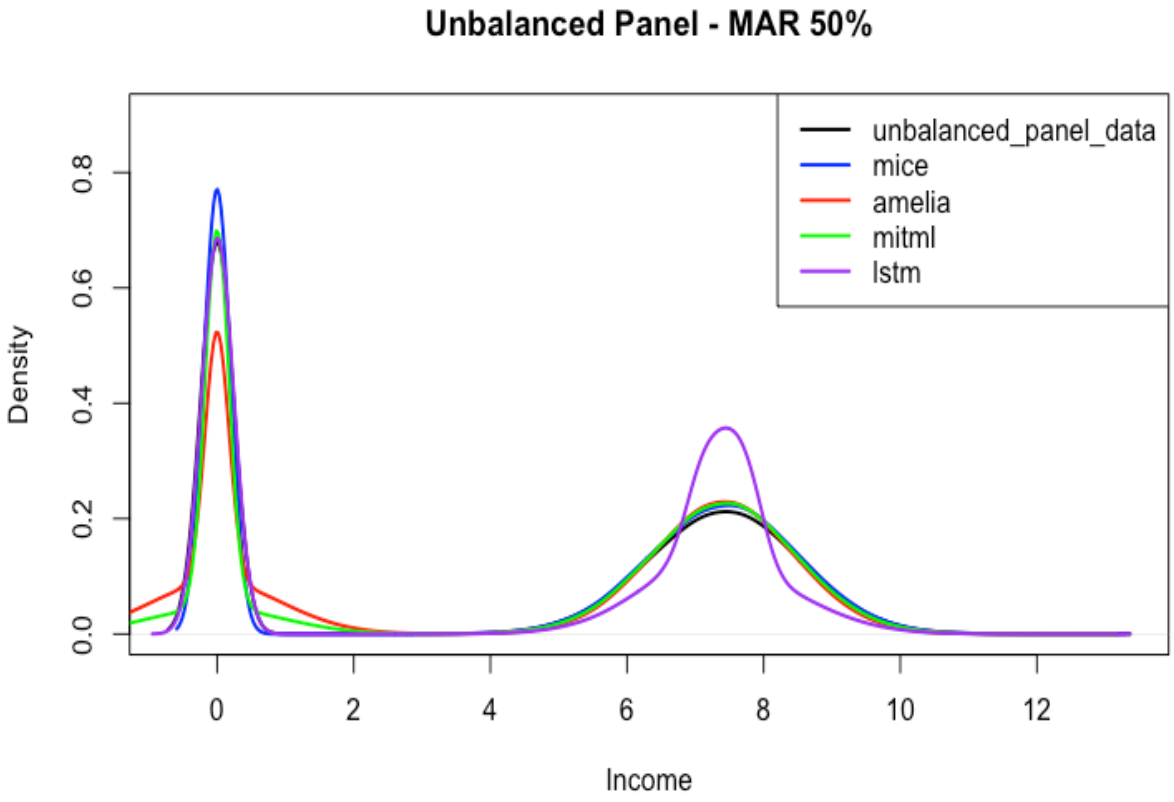


Figure 17: Distributions Comparison of "Income" for Different Frameworks Part-07



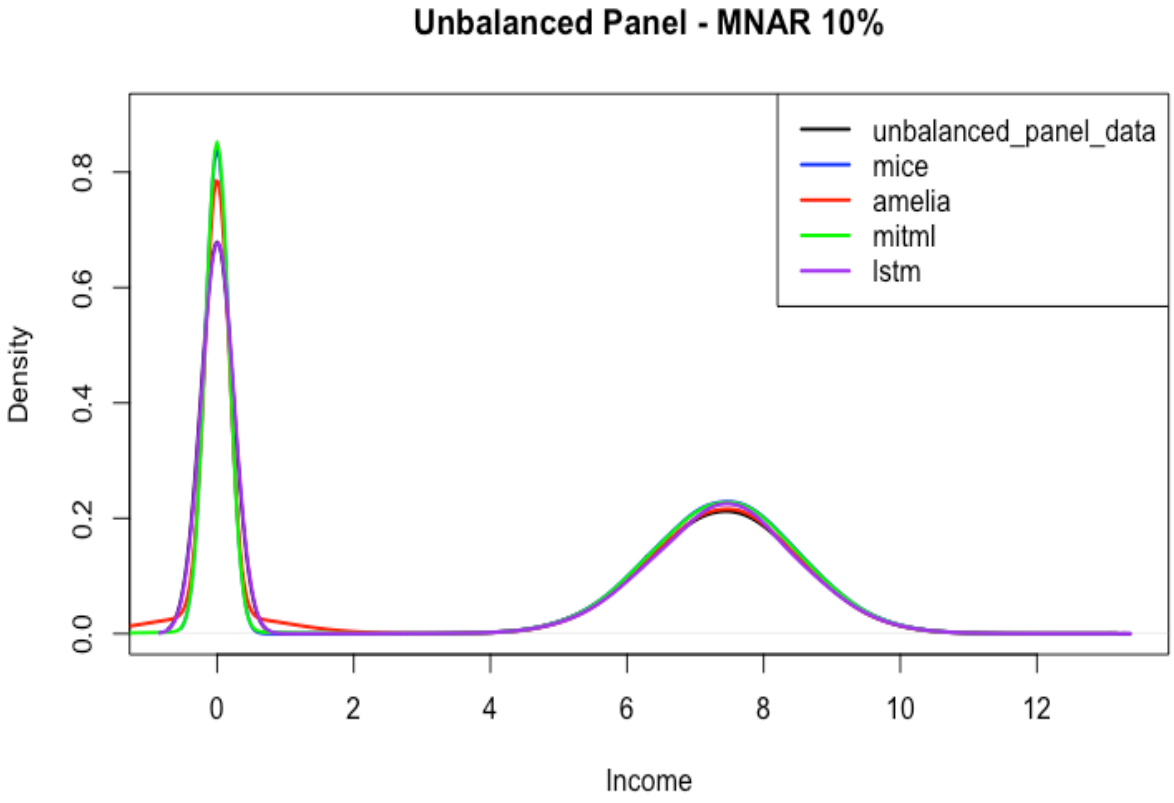
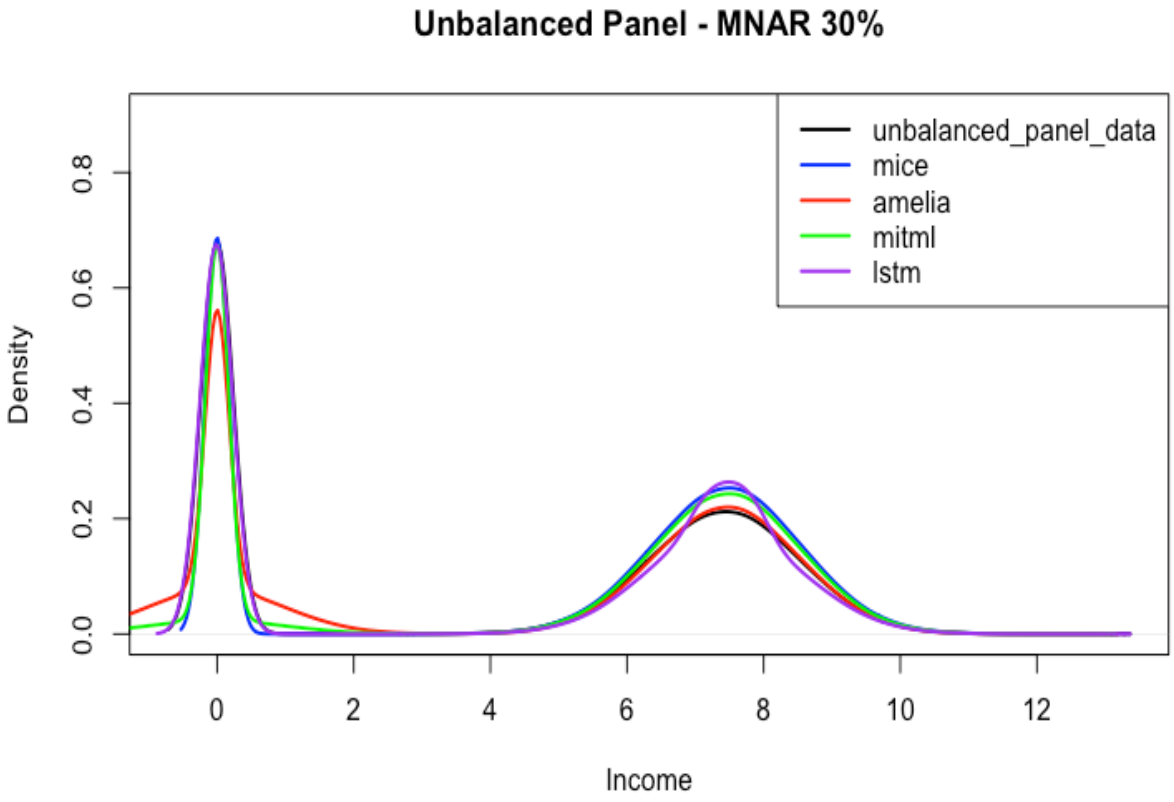


Figure 18: Distributions Comparison of "Income" for Different Frameworks Part-08



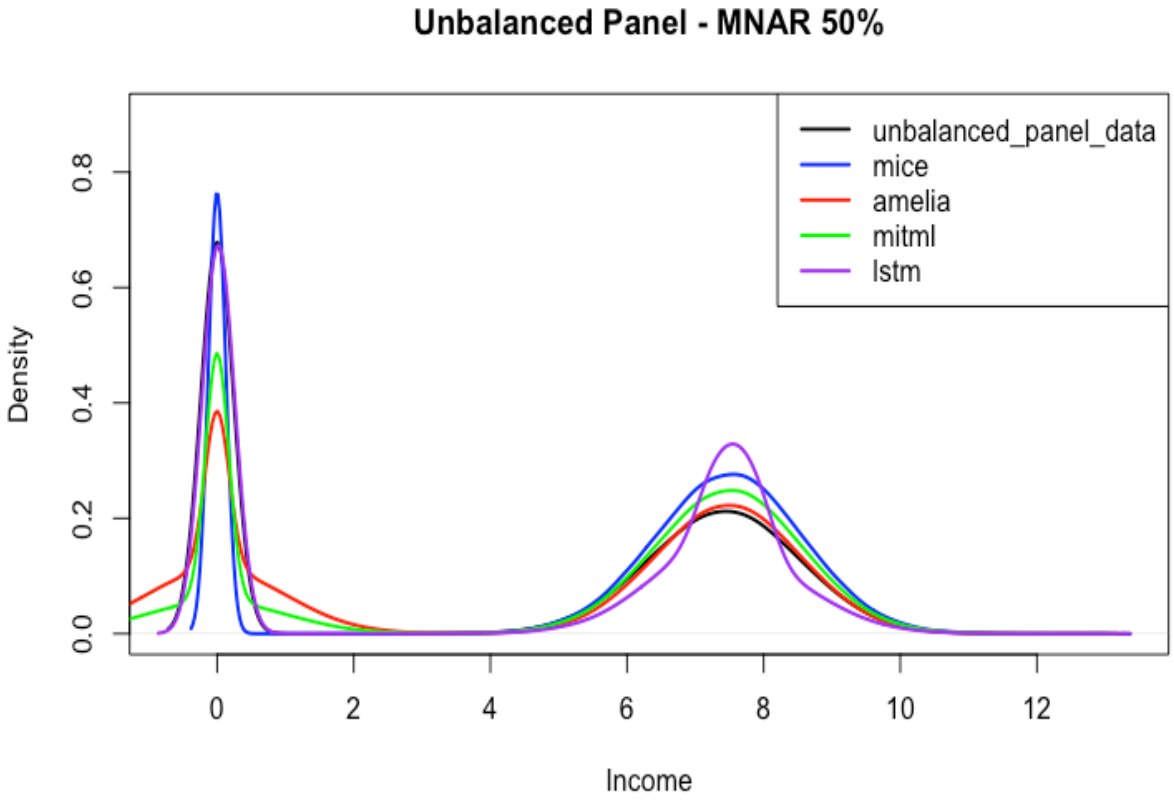


Figure 19: Distributions Comparison of "Income" for Different Frameworks Part-09

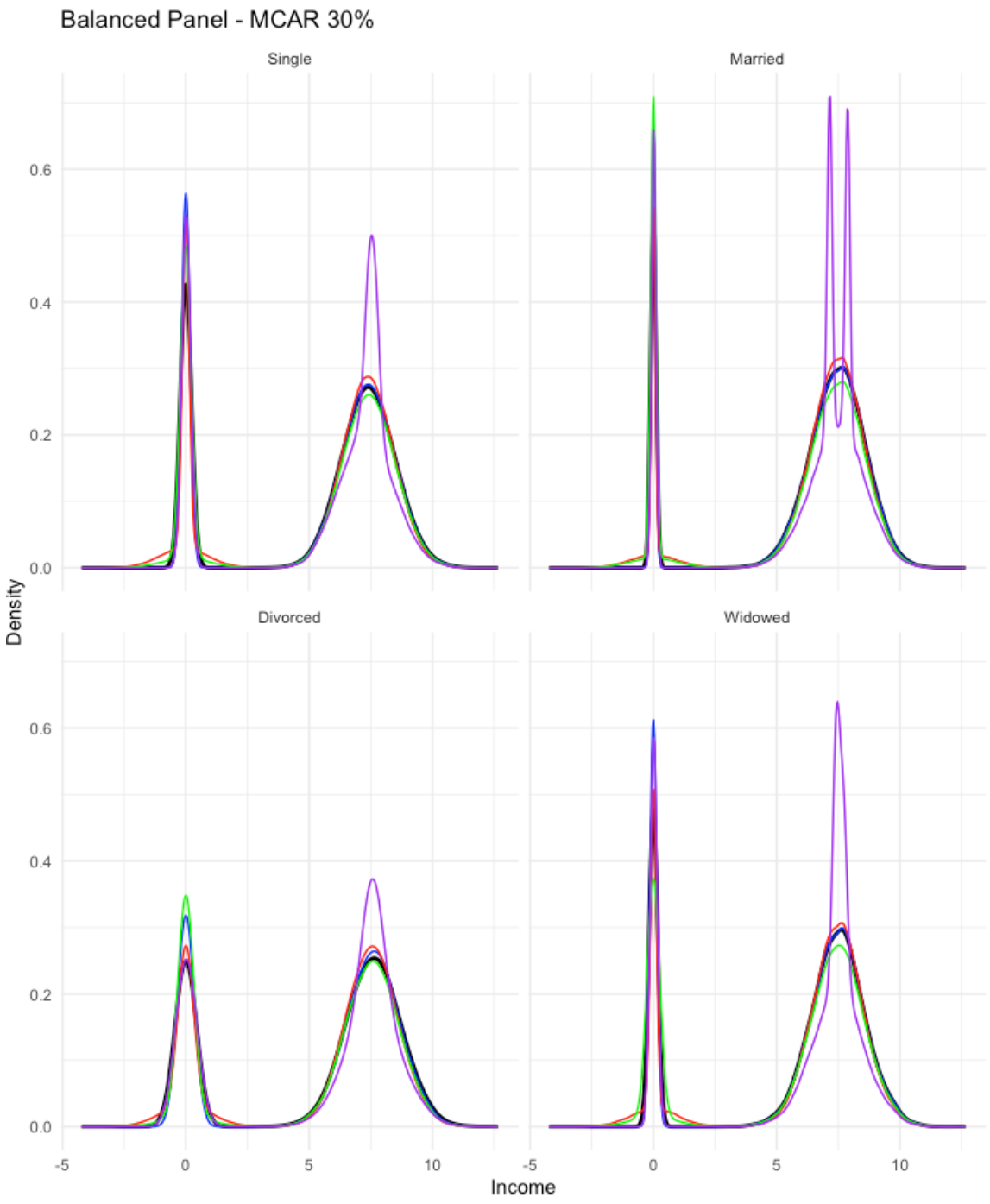


Figure 20: Conditional Distributions of "Income" for Balanced Panel MCAR at 30%

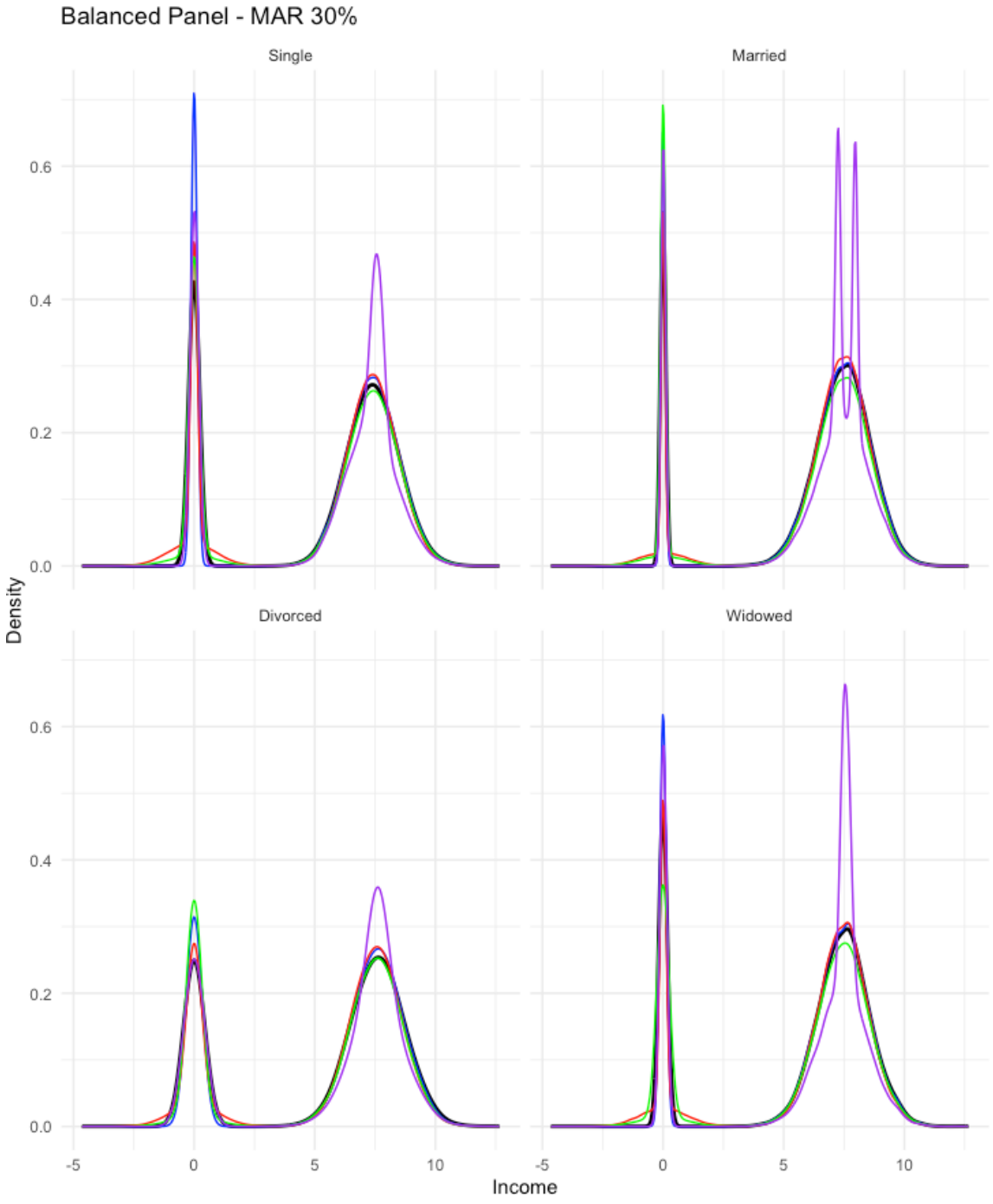


Figure 21: Conditional Distributions of "Income" for Balanced Panel MAR at 30%

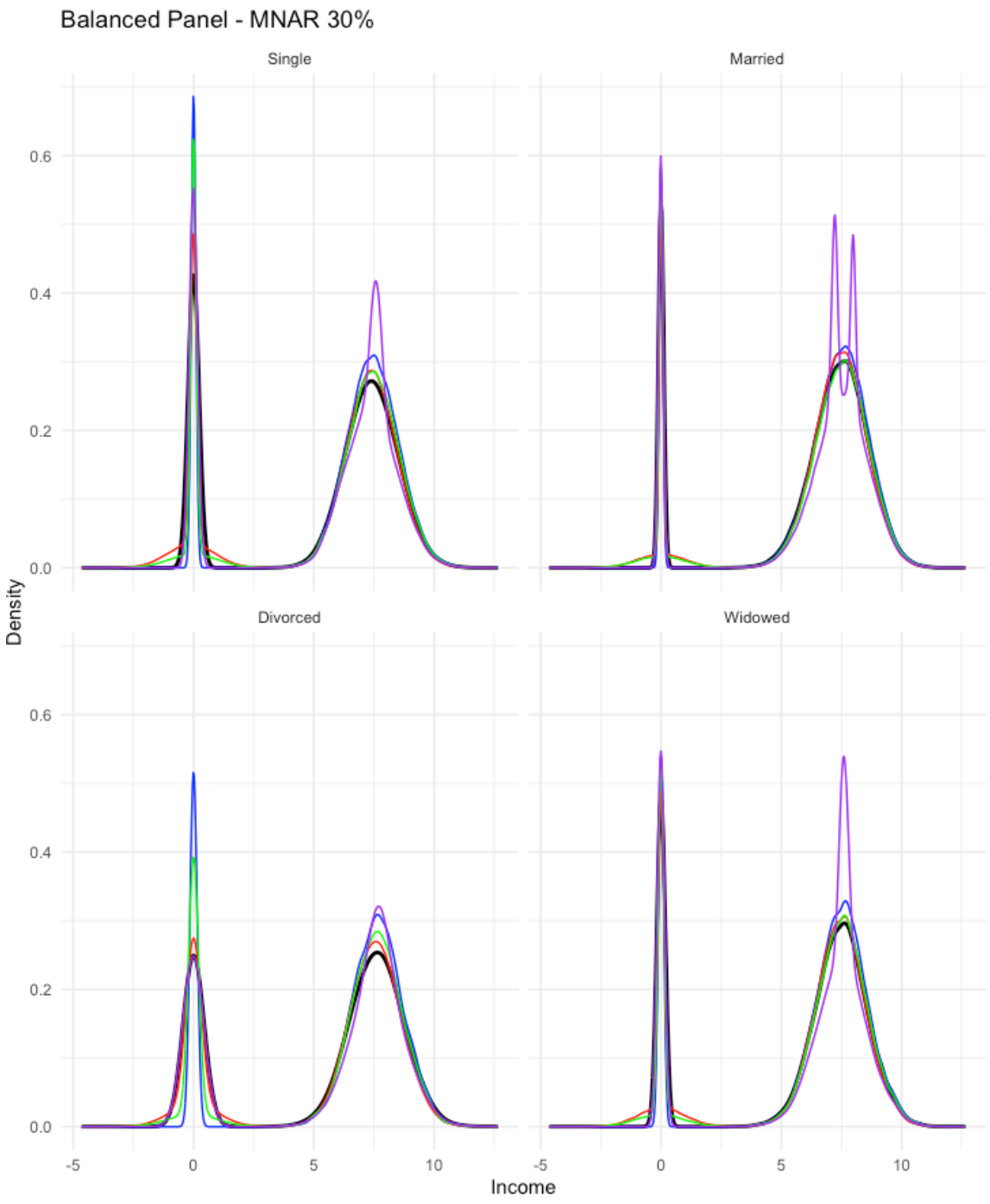


Figure 22: Conditional Distributions of "Income" for Balanced Panel MNAR at 30%

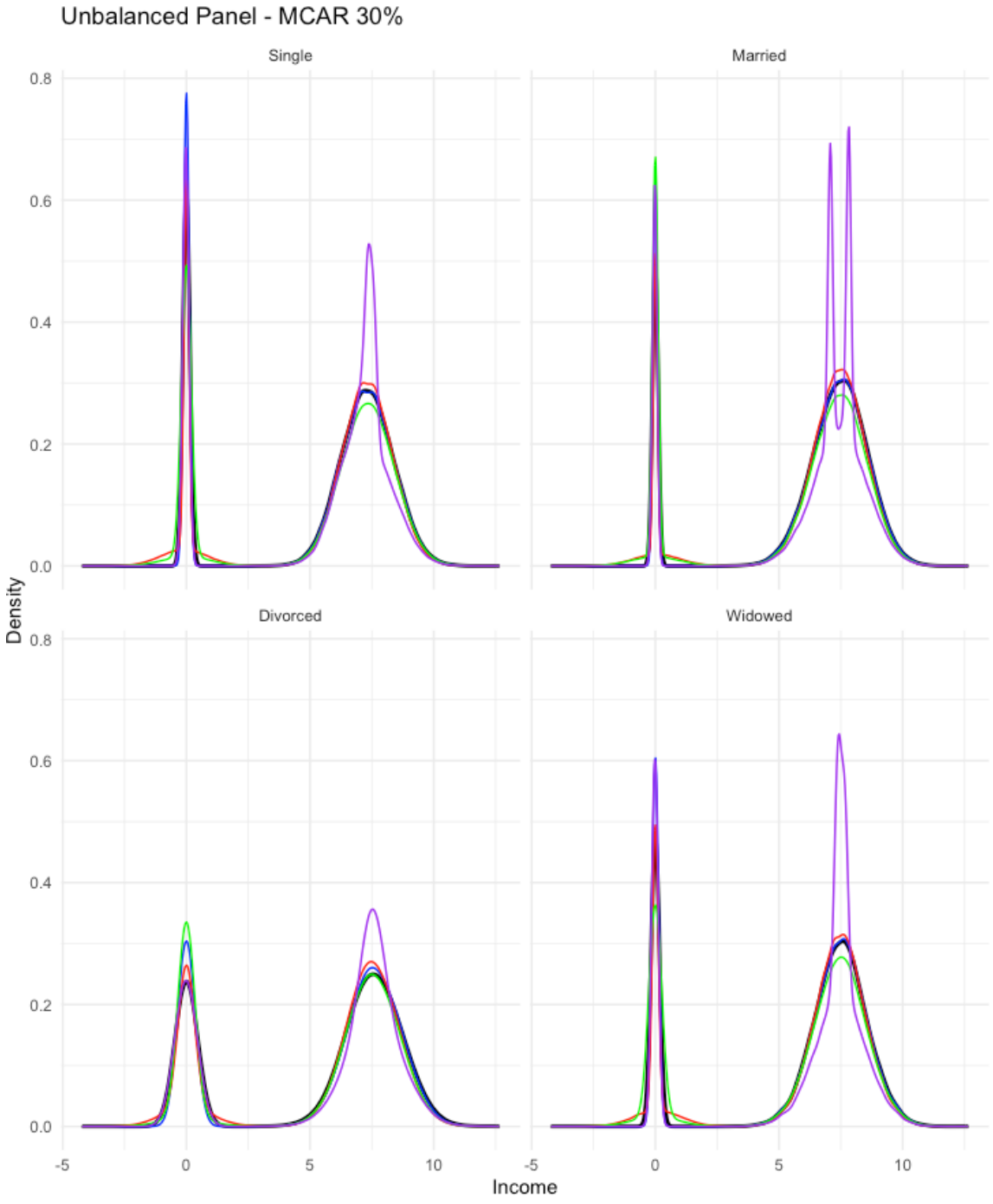


Figure 23: Conditional Distributions of "Income" for Unbalanced Panel MCAR at 30%

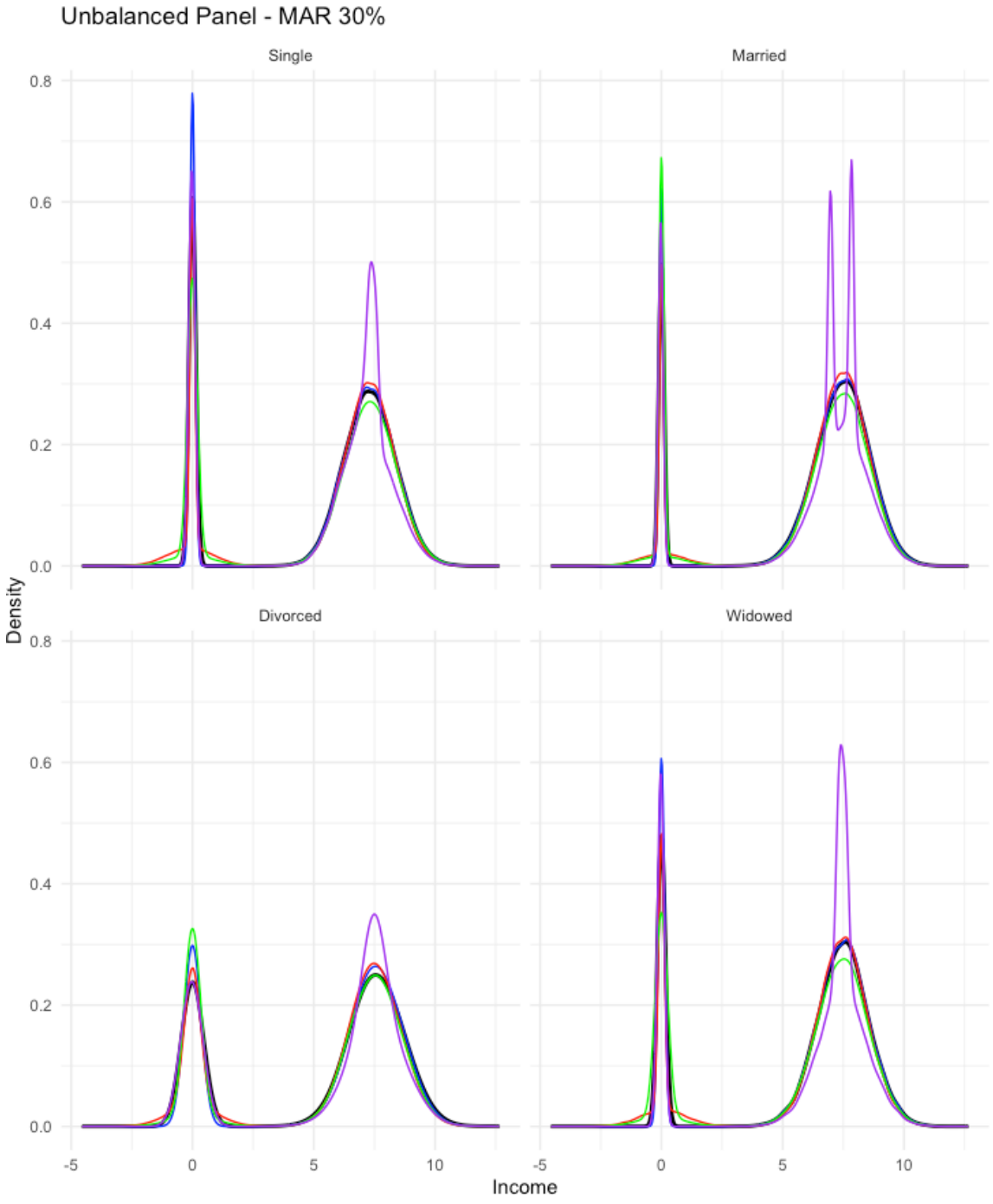


Figure 24: Conditional Distributions of "Income" for Unbalanced Panel MAR at 30%

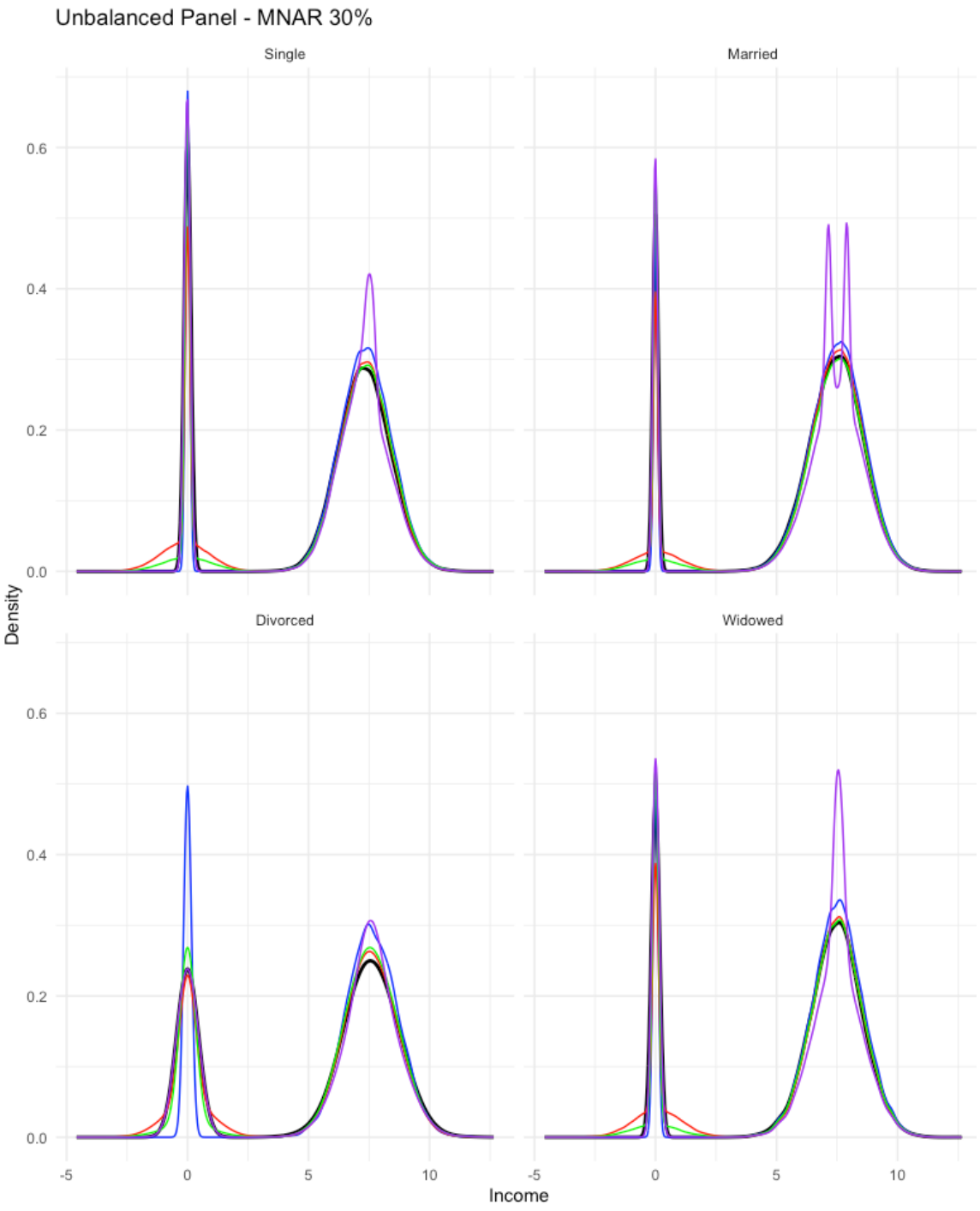


Figure 25: Conditional Distributions of "Income" for Unbalanced Panel MNAR at 30%

For the distribution without any condition, the performance of "mice" is closely aligned with the real dataset. But shows difficulties in handling MNAR conditions. However, compared to "mitml" and "amelia", "mice" does not fail to predict zero values effectively at the lower end of the density curve. Compared to "mitml" and "amelia", it struggled more with imputing MNAR condition at a higher level of missingness (50%). Unlike "mitml" and "amelia", "mice" struggled a little more in unbalanced than balanced panels.

"amelia" performs well when the missingness level is 10%, but its effectiveness declines with 30% and 50% missingness, particularly in predicting zero values. Interestingly, the lower end of the density curve for zero values in "amelia" is similar to that of "mitml". It exhibits less difficulty with MNAR conditions compared to "mice". In fact, "amelia" looks the best at predicting missingness for the MNAR conditions at a higher level of missingness. Like "mitml", the performance in unbalanced and balanced scenarios is not very different for "amelia".

"mitml" demonstrates similarities to "amelia", but it shows more difficulties in handling MNAR at a higher level of missingness compared to "amelia". The skewness and kurtosis are mostly similar to the real dataset for lower levels of missingness. However, the tail near the zero value strays a little for all MCAR, MAR, and MNAR cases. "amelia" has this similar trait, but "mitml" looks slightly better than "amelia". It outperformed "amelia" in predicting missing zero values. For MNAR, "mitml" performs better than "mice" for 30% and 50% missingness. No significant differences can be seen between the balanced and unbalanced cases.

The "LSTM-Network" performs poorly overall compared to the other frameworks. For 30% and 50% missing values for all MCAR, MAR, and MNAR, the skewness and kurtosis got completely distorted. Nonetheless, it excels at imputing zero values and performed significantly better for 10% missingness across all MCAR, MAR, and MNAR scenarios. This superior performance in imputing zero values could be an overfitting[44] issue, as the dataset contained a substantial number of zero values. The "LSTM-Network" architecture may have captured these specific characteristics, enabling it to predict zero values more effectively than the other packages. Like the "amelia" and "mice", it performs almost the same for balanced and unbalanced scenarios.

For the conditional distribution, "mice" performs well across most categories, with a few exceptions. The performance for balanced panels under MCAR is nearly consistent across all categories. However, under MNAR conditions, mice struggle to predict zero values for the Single and Divorced categories compared to other frameworks.

[44]In a machine learning architecture, overfitting occurs when a model learns from the underlying properties of the data, noise and any random fluctuations in the data.

This issue becomes more pronounced in unbalanced panels. Overall performance in MNAR, "mice" struggles more than the other packages.

"amelia", while not as robust as "mice", demonstrates consistent performance. Unlike mice, its performance is generally symmetric across categories and consistent for balanced and unbalanced panels. However, "amelia" distorts the lower tail near the zero values across all categories, showing a similar pattern regardless of the panel types.

The "mitml" package handles zero-value predictions better than "mice" but struggles to predict values for the Married and Divorced categories under MCAR and MAR. "mitml" also distorts at the lower tail near zero values like "amelia", but to a lesser extent. It's performance is stable across balanced and unbalanced panels like "amelia", which is not very likely in "mice".

The "LSTM-Network" produces very strange triple curves for the Married category and generally performs poorly overall. Despite the limitations, it excels at predicting zero values, outperforming other frameworks. This strength likely stems from its ability to learn effectively from the high proportion of zero values in the data. Its performance is comparatively better for the Divorced category and remains consistent across balanced and unbalanced panels.

5.3 Correlation Comparison

The correlations in table 14 and 15 reflect the relationships between different variables after missing data have been imputed using different frameworks. The tables highlight how well each framework preserves the relationships among variables. The tables compare the correlations of various imputation frameworks with the original balanced and unbalanced panel datasets across the other variables. The names of the columns "ET" means "Employment Types", "MT" means "Marital Status", "EH" means "Employment Hours", and "E" means "Education", respectively.

Dataset	Age	ET	MT	EH	E	Sex
balanced_panel_data	-0.05	-0.96	-0.00	-0.82	-0.11	-0.05
amelia_imp_bal_mcar_50	-0.05	-0.96	-0.00	-0.83	-0.12	-0.05
amelia_imp_bal_mcar_30	-0.05	-0.96	-0.00	-0.82	-0.12	-0.05
amelia_imp_bal_mcar_10	-0.05	-0.96	-0.00	-0.82	-0.11	-0.05
amelia_imp_bal_mar_50	-0.05	-0.96	-0.00	-0.83	-0.12	-0.06
amelia_imp_bal_mar_30	-0.05	-0.96	-0.00	-0.82	-0.12	-0.06
amelia_imp_bal_mar_10	-0.05	-0.96	-0.00	-0.82	-0.11	-0.05

Dataset	Age	ET	MS	EH	E	Sex
amelia_imp_bal_mnar_50	-0.04	-0.95	-0.00	-0.82	-0.12	-0.06
amelia_imp_bal_mnar_30	-0.05	-0.95	-0.00	-0.82	-0.12	-0.06
amelia_imp_bal_mnar_10	-0.05	-0.96	-0.00	-0.82	-0.11	-0.06
mice_imp_bal_mcar_50	-0.05	-0.96	-0.00	-0.82	-0.12	-0.06
mice_imp_bal_mcar_30	-0.05	-0.96	-0.00	-0.82	-0.12	-0.06
mice_imp_bal_mcar_10	-0.05	-0.96	-0.00	-0.82	-0.11	-0.05
mice_imp_bal_mar_50	-0.10	-0.95	-0.03	-0.81	-0.08	-0.06
mice_imp_bal_mar_30	-0.09	-0.96	-0.02	-0.81	-0.09	-0.06
mice_imp_bal_mar_10	-0.07	-0.96	-0.01	-0.82	-0.10	-0.06
mice_imp_bal_mnar_50	-0.12	-0.94	-0.03	-0.73	-0.05	-0.08
mice_imp_bal_mnar_30	-0.10	-0.95	-0.02	-0.77	-0.07	-0.07
mice_imp_bal_mnar_10	-0.07	-0.95	-0.00	-0.80	-0.10	-0.06
mitml_imp_bal_mcar_50	-0.02	-0.48	0.00	-0.41	-0.06	-0.03
mitml_imp_bal_mcar_30	-0.03	-0.67	-0.00	-0.57	-0.08	-0.04
mitml_imp_bal_mcar_10	-0.04	-0.86	-0.00	-0.74	-0.10	-0.05
mitml_imp_bal_mar_50	-0.04	-0.47	-0.01	-0.40	-0.04	-0.03
mitml_imp_bal_mar_30	-0.06	-0.66	-0.01	-0.57	-0.06	-0.04
mitml_imp_bal_mar_10	-0.06	-0.86	-0.01	-0.73	-0.09	-0.05
mitml_imp_bal_mnar_50	-0.03	-0.43	-0.01	-0.36	-0.04	-0.04
mitml_imp_bal_mnar_30	-0.05	-0.63	-0.01	-0.53	-0.06	-0.05
mitml_imp_bal_mnar_10	-0.06	-0.84	-0.01	-0.73	-0.09	-0.05
lstm_bal_mcar_50	-0.05	-0.97	-0.00	-0.84	-0.12	-0.06
lstm_bal_mcar_30	-0.05	-0.96	-0.00	-0.83	-0.12	-0.05
lstm_bal_mcar_10	-0.05	-0.96	-0.00	-0.82	-0.12	-0.05
lstm_bal_mar_50	-0.05	-0.97	-0.00	-0.84	-0.12	-0.05
lstm_bal_mar_30	-0.05	-0.96	-0.00	-0.83	-0.12	-0.06
lstm_bal_mar_10	-0.05	-0.96	-0.00	-0.82	-0.12	-0.05
lstm_bal_mnar_50	-0.05	-0.97	-0.00	-0.83	-0.12	-0.05
lstm_bal_mnar_30	-0.05	-0.96	-0.00	-0.83	-0.11	-0.05
lstm_bal_mnar_10	-0.05	-0.96	-0.00	-0.82	-0.11	-0.05

Table 14: Correlation Table for Balanced Panel (Rounded to 2 Decimal Places)

Table 14 represents the correlations between "Income" and other variables ("Age", "Employment Types", "Marital Status", "Employment Hours", "Education", and "Sex") for the balanced panel. Each imputation framework shows varying effects on the cor-

relations, especially in the presence of different missing data patterns (MCAR, MAR, and MNAR).

"amelia" appears to be the most stable imputation package here, with minimal changes in the correlation values across different scenarios. The correlations remain close to those in the original dataset, especially for variables like "Employment Types" and "Education", which show a strong negative correlation with Income (-0.96) and (-0.11), respectively. For the rest of the variables, MCAR and MAR are acceptable, but the correlation fluctuates a little for the MNAR conditions.

"mice" shows more variability in the correlations than "amelia", especially in the MNAR settings. For instance, in the *mice_imp_bal_mnar*_50 dataset, the correlations with "Income" decrease noticeably, particularly for "Employment Types" and "Employment Hours". The reason could be how "mice" is set to impute the values using the lag values of the predictors. This indicates that "mice" may struggle more with maintaining accurate correlations when the data is missing in a non-random manner, especially with higher levels of missing data (50%).

"mitml" also demonstrates significant variability in the correlation values, particularly under the MNAR scenarios. For example, the correlations between "Income" and "Employment Types" drop significantly in the *mitml_imp_bal_mnar*_50 dataset, showing a decline to -0.42. This suggests that "mitml" has more difficulty in preserving the correlation structure when the missing data is non-random, especially with higher levels of missingness. However, "mitml" provides somewhat consistent results in MCAR and MAR settings, though it still lags in preserving the original data's correlation patterns.

While "LSTM-Network" shows notably good results than "mitml", it cannot be considered because overfitting is observed in the density curves. However, it shows a significantly good match for MAR and MNAR conditions compared to the other packages at 10% missingness. This shows the future potential for "LSTM-Network."

Dataset	**Age**	**ET**	**MT**	**EH**	**E**	**Sex**
unbalanced_panel_data	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
amelia_imp_unbal_mcar_50	-0.00	-0.96	0.00	-0.84	-0.17	-0.06
amelia_imp_unbal_mcar_30	-0.00	-0.96	0.00	-0.84	-0.17	-0.06
amelia_imp_unbal_mcar_10	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
amelia_imp_unbal_mar_50	-0.00	-0.96	0.00	-0.84	-0.17	-0.06
amelia_imp_unbal_mar_30	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
amelia_imp_unbal_mar_10	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
amelia_imp_unbal_mnar_50	0.01	-0.95	0.01	-0.83	-0.17	-0.07

Dataset	Age	ET	MS	EH	E	Sex
amelia_imp_unbal_mnar_30	0.00	-0.96	0.01	-0.83	-0.17	-0.07
amelia_imp_unbal_mnar_10	0.00	-0.96	0.00	-0.83	-0.16	-0.06
mice_imp_unbal_mcar_50	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
mice_imp_unbal_mcar_30	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
mice_imp_unbal_mcar_10	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
mice_imp_unbal_mar_50	-0.03	-0.96	-0.01	-0.82	-0.15	-0.06
mice_imp_unbal_mar_30	-0.02	-0.96	-0.00	-0.82	-0.15	-0.06
mice_imp_unbal_mar_10	-0.01	-0.96	0.00	-0.83	-0.16	-0.06
mice_imp_unbal_mnar_50	-0.03	-0.95	-0.01	-0.75	-0.12	-0.08
mice_imp_unbal_mnar_30	-0.02	-0.95	-0.00	-0.79	-0.14	-0.07
mice_imp_unbal_mnar_10	-0.07	-0.95	-0.01	-0.80	-0.10	-0.06
mitml_imp_unbal_mcar_50	0.00	-0.48	0.00	-0.42	-0.08	-0.03
mitml_imp_unbal_mcar_30	0.00	-0.67	0.00	-0.58	-0.11	-0.04
mitml_imp_unbal_mcar_10	0.00	-0.86	0.00	-0.75	-0.15	-0.06
mitml_imp_unbal_mar_50	-0.01	-0.47	0.00	-0.41	-0.07	-0.03
mitml_imp_unbal_mar_30	-0.01	-0.66	0.00	-0.58	-0.11	-0.05
mitml_imp_unbal_mar_10	-0.01	-0.86	0.00	-0.75	-0.14	-0.06
mitml_imp_unbal_mnar_50	-0.00	-0.44	0.00	-0.37	-0.07	-0.04
mitml_imp_unbal_mnar_30	-0.01	-0.64	0.00	-0.55	-0.10	-0.05
mitml_imp_unbal_mnar_10	-0.01	-0.85	0.00	-0.73	-0.14	-0.06
lstm_unbal_mcar_50	0.00	-0.97	0.00	-0.85	-0.17	-0.06
lstm_unbal_mcar_30	-0.00	-0.97	0.00	-0.84	-0.17	-0.06
lstm_unbal_mcar_10	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
lstm_unbal_mar_50	-0.00	-0.97	0.00	-0.85	-0.17	-0.07
lstm_unbal_mar_30	-0.00	-0.97	0.00	-0.84	-0.17	-0.06
lstm_unbal_mar_10	-0.00	-0.96	0.00	-0.83	-0.16	-0.06
lstm_unbal_mnar_50	-0.01	-0.97	0.00	-0.85	-0.16	-0.06
lstm_unbal_mnar_30	-0.00	-0.96	0.00	-0.84	-0.16	-0.06
lstm_unbal_mnar_10	-0.00	-0.96	0.00	-0.83	-0.16	-0.06

Table 15: Correlation Table for Unbalanced Panel (Rounded to 2 Decimal Places)

Table 15 provides the correlation table of the simulated and original unbalanced panels. Here, "amelia" demonstrates good performance for MCAR and MAR conditions, producing correlations that closely align with the original data, especially at lower levels of missingness. As the percentage of missing data increases, deviations from the original data become more apparent, particularly under MNAR conditions.

Variables like "Employment Types" and "Sex" are better retained in terms of correlation, but "Employment Hours" exhibits larger deviations under MNAR scenarios. "amelia" overall could be a reliable package for MCAR and MAR conditions, but its performance declines with higher percentages of missingness in MNAR cases.

"mice" performs well for MCAR and MAR conditions, particularly at lower levels of missing data, where its correlations closely match the original data. However, its effectiveness diminishes under MNAR conditions, with significant deviations observed at 50% missingness. Features like "Employment Hours" and "Employment Types" show marked inconsistencies in MNAR scenarios, while correlations for "Sex" remain relatively stable. Overall, "mice" is a reliable choice for handling MCAR and MAR data at lower levels of missingness but struggle when data is missing, not at random or in large proportions.

"mitml" consistently exhibits the largest deviations from the original data across all missingness mechanisms and percentages. Even for MCAR and MAR conditions, mitml struggles to maintain correlations, particularly for variables like "Employment Types" and "Employment Hours". While variables like "Age" and "Marital Status" show some stability, most others deviate significantly from the original correlations. "mitml" performance is weaker than "mice" and "amelia", making it less suitable for datasets with complex missing data patterns and high levels of missingness.

"LSTM-Network" outperforms other packages in maintaining correlations close to the original data for MCAR and MAR scenarios. It demonstrates robustness even at higher missingness levels, particularly for variables like "Employment Types" and "Sex". This could be due to the overfitting effects mentioned in the section 5.2. Despite this, LSTM consistently shows smaller deviations of 10% in MNAR conditions compared to the rest of the frameworks. Its ability to preserve relationships in the data under MCAR and MAR conditions at 10% missingness shows its future potentiality.

In conclusion, "mice" and "amelia" performed well in these scenarios, with "amelia" being more consistent at moderate levels of missingness, preserving the original correlations with minimal changes. "mice" struggled more in the Unbalanced panel than the balanced panel. Deviation of the correlation in the exogenous variables ("Employment Hours" and "Education") is a little higher compared to the predictor variables. "mice" and "mitml", while effective, are more sensitive to the amount and type of missing data, especially in MNAR settings, leading to more significant deviations from the original correlation structure. "LSTM-Network" excels in MNAR for 10% missingness conditions, whereas the other packages struggled.

5.4 Wasserstein Distances

According to Villani (2009), the Wasserstein distance is described as the measure of the optimal transport cost required to move one probability distribution (μ) to another (ν) over a given space X while considering a given metric d. The Wasserstein distance of order p between two probability measures μ and ν is given below,

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_X d(x,y)^p \, d\pi(x,y) \right)^{1/p} \tag{34}$$

$W_p(\mu,\nu)$ is the cost required to transform μ into ν. $\Pi(\mu,\nu)$ is the set of all joint probability measures. $d(x,y)$ is the cost or distance metric between points x and y in the space X. A higher Wasserstein distance indicates less similarities between two probability distributions.

R package "transport" is used to find the Wasserstein distance for this thesis study. Three colours are defined for different levels of missingness in figures 26 and 27. Red is for 10%, blue is for 30%, green is for 50%. The plots are set in a way that each group consists of four imputation frameworks ("mice", "amelia", "mitml", and "lstm"). Four frameworks are compared for 10%, 30%, and 50% missingness for MCAR, then MAR, and then MNAR. All the frameworks are performing better in MCAR conditions than MAR and MNAR. "mice" has lower Wasserstein distances in MCAR and MAR compared to the other frameworks. It struggles significantly in MNAR at higher levels of missingness. "amelia" and "mitml" both have lower scores than "mice" in MNAR settings for all levels of missingness. "amelia" is showing another interesting trend here. It is more comfortable to impute a lower level of missingness (10%) than focusing on the types (MCAR, MAR, and MNAR) of missingness. For example, "amelia" has lower scores in MNAR at 10% than MCAR at 50%. Following this, "amelia" could be an option for lower levels of missingness in MNAR settings. The performance of "mitml" is not good in general. But it is better in MNAR conditions for all levels of missingness than "mice". "LSTM-Network" provides acceptable results, which are misleading explained in section 5.2. However, it excelled, whereas the other packages struggled. In MNAR, the condition for all levels of missingness is that "LSTM-Network" has the lowest Wasserstein distances. Especially in MNAR at 10% missingness. Which shows the future possibility to improve. The difference between the balanced and unbalanced panels is not high, except the frameworks struggle a little more in unbalanced than balanced panels, particularly "mice".

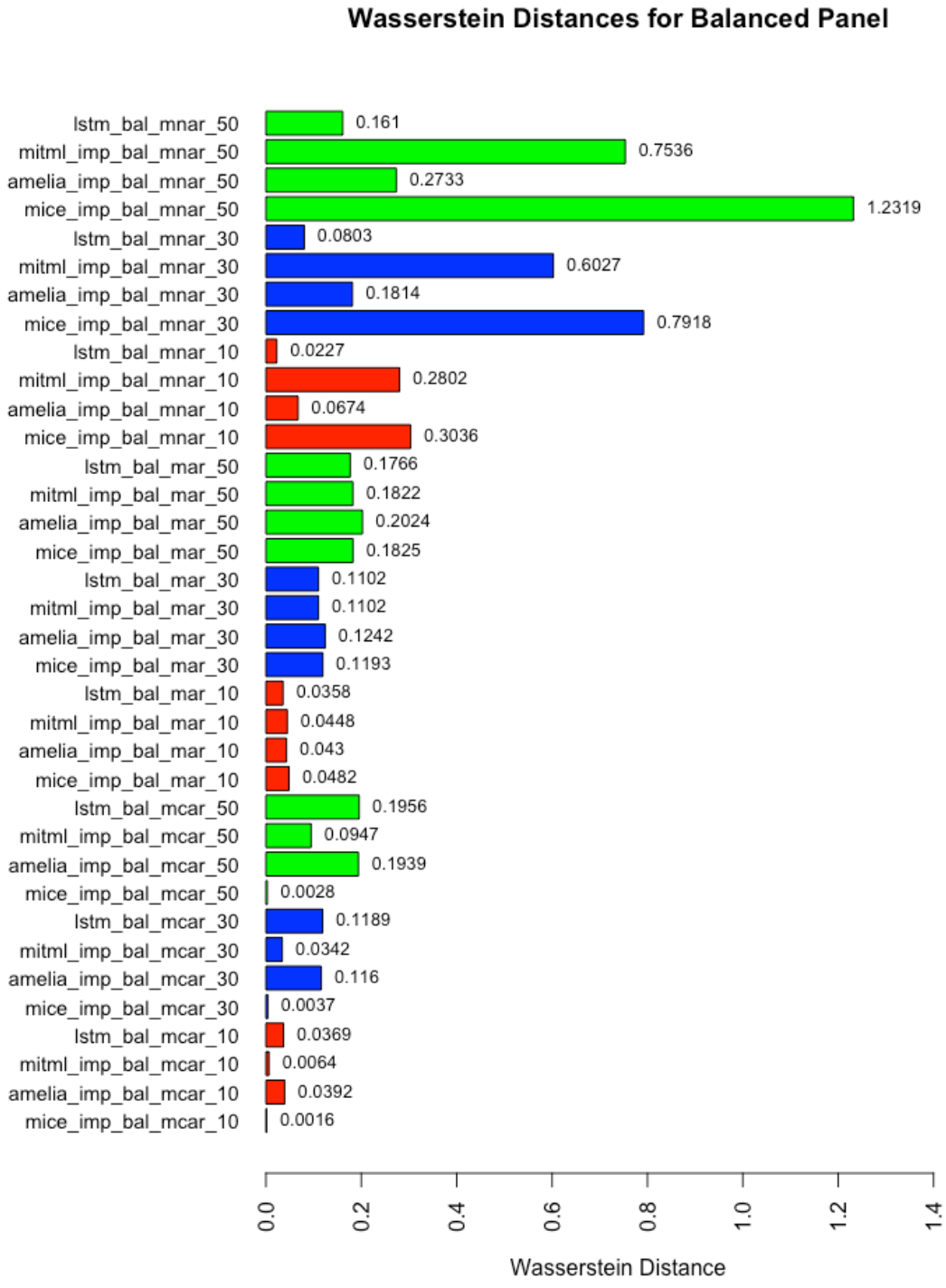


Figure 26: Wasserstein Distances for Balanced Panel Across the Simulated Datasets

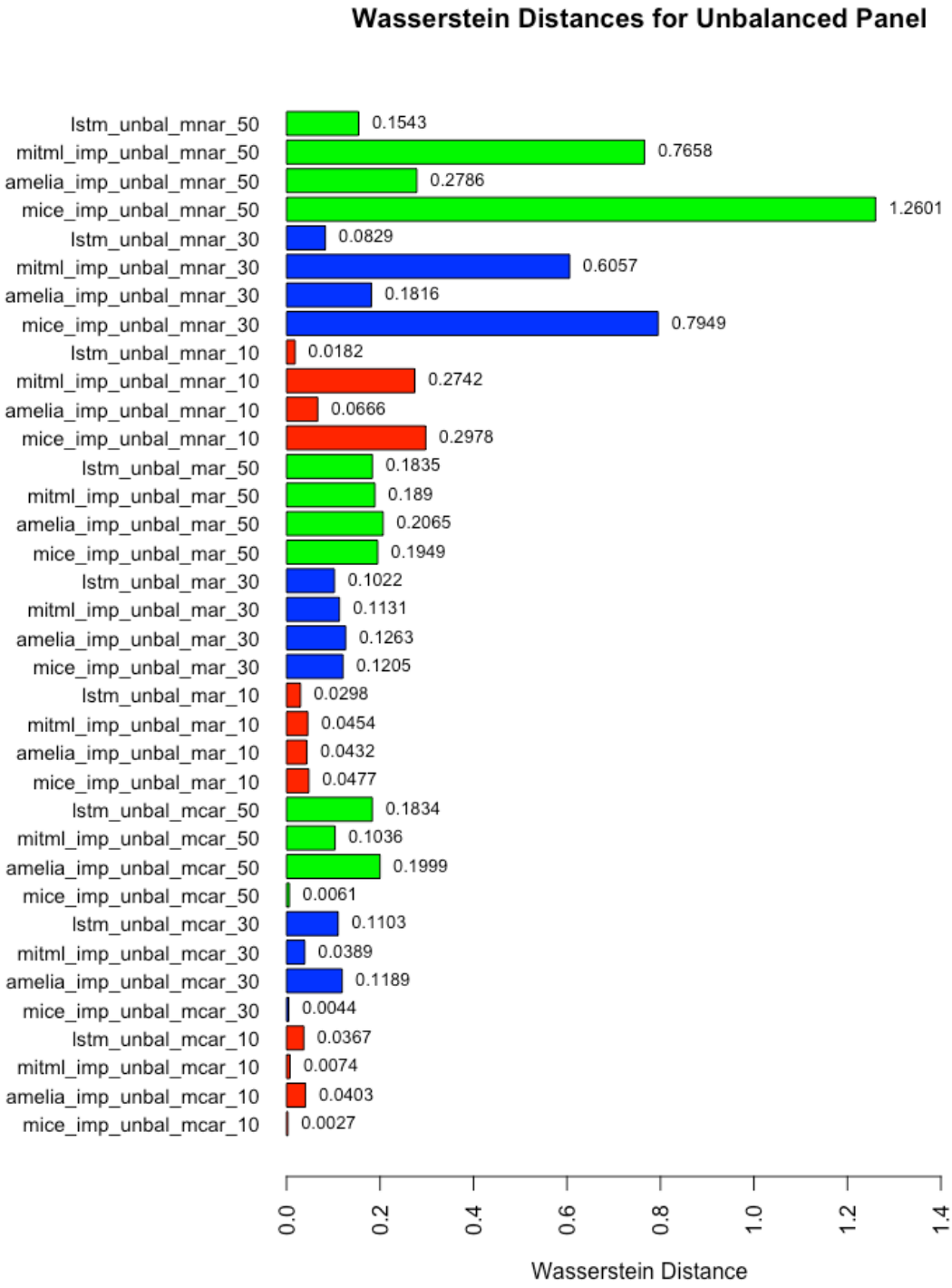


Figure 27: Wasserstein Distances for Unbalanced Panel Across the Simulated Datasets

5.5 Bias and RMSE

RMSE and Bias are two important criteria in this study for assessing how well the imputation frameworks perform. Bias can be used to determine if it tends to overestimate or underestimate by measuring the systematic departure of the estimated coefficients from the true values. More accurate estimates have a reduced Bias. Contrarily, RMSE calculates the total error magnitude and offers information on how closely the model's predictions match the actual values across all datasets. Overall prediction performance is better when the RMSE values are lower. The accuracy and dependability of the imputation framework are evaluated using both measures, respectively.

Before calculating Bias and RMSE, the datasets underwent preliminary preprocessing. A Fixed Effects model estimates the coefficients, as no random effects are identified in the original balanced or unbalanced panel datasets. Likewise, in correlation assessment, the exogenous variables "Employment Hours" and "Education" are included to assess the model's overall performance. However, no coefficients are found for "Education" and "Sex" because these variables are time-invariant. They are automatically dropped out from the model as they do not change over time. A Fixed Effects model is fitted for each imputed dataset, and the coefficients are extracted. The model used in R code is,

$$\begin{aligned} \text{Model} \leftarrow \text{plm}(&\text{Income} \sim \text{Age} + \text{Employment Types} \\ &+ \text{Marital Status} + \text{Employment Hours} \\ &+ \text{Education} + \text{Sex}, \\ &\text{data} = \text{data}, \text{model} = \text{"within"}) \end{aligned} \tag{35}$$

In equation (35), the model is derived using "plm" package in R for the original balanced and unbalanced data. The goodness of fit is checked using "Breusch-Pagan-Test" and "Hausman-Test" (see detail in figure 1). The results reveal one coefficient for "Employment Hours", two coefficients for "Employment Types", and three coefficients for "Marital Status", attributable to dummy variable effects. Additionally, the intercepts are absorbed by other coefficients due to the mechanism of the Fixed Effects model. The same process is used for "mice", "mitml", and "amelia". Rubin's rule is implemented to average the coefficients from multiple datasets. As the original datasets and the simulated datasets from "LSTM-Network" have only one dataset inside the data frame, using Rubin's rule to pool the coefficients is unnecessary. These coefficients from all the imputed datasets are then combined into a single data frame for further analysis. The coefficients table for all the simulated datasets is in the appendix section, and the coefficients of the original datasets are given below,

Variable	Balanced Panel	Unbalanced Panel
Age	-0.00	0.00
EmploymentTypes1	-7.20	-7.18
EmploymentTypes2	-7.03	-7.01
MaritalStatus2	0.07	0.09
MaritalStatus3	0.16	0.15
MaritalStatus4	0.12	0.12
EmploymentHours1	0.28	0.27

Table 16: Coefficients for Balanced and Unbalanced Panel Data

Table 16 provides the estimated coefficients for various variables in balanced and unbalanced datasets. These coefficients show the relationship between each independent variable and "Income" while holding other factors constant. "Employment hours" and "Income" have a positive correlation in both the balanced and unbalanced panels, while "Income" and "Employment Types" have a strong negative correlation. The table also compares the balanced and unbalanced panels. While most coefficients are similar, minor variations exist, such as for "Marital Status". These differences highlight how the balance of panel data can slightly influence model estimates, though the overall trends remain consistent. Such comparisons help assess the stability of the model when applied to different types of datasets.

Statistic	Balanced Panel	Unbalanced Panel
Minimum	-7.20	-7.18
1st Quartile	-3.51	-3.50
Median	0.07	0.09
Mean	-1.94	-1.94
3rd Quartile	0.14	0.13
Maximum	0.28	0.27

Table 17: Coefficients Summary Statistics for Balanced and Unbalanced Panel Data

Table 17 provides the summary statistics of both datasets to compare. The minimum coefficient in both datasets is highly negative, driven by the variable "Employment Types", which shows a strong negative effect on "Income" (because of the ranking of the categories). On the other hand, the maximum coefficient, contributed by "Employment Hours", reflects its consistently positive impact on "Income" (also because of the ranking of the categories). The median and mean provide further context about the central tendency of the coefficients. The median values indicate that most coefficients are either positive or near zero, reflecting moderate to negligible effects for many variables. The mean coefficient (-1.94 for both datasets) suggests a slight overall negative trend, likely influenced by the strongly negative coefficients like "Employment Types". Additionally, the 1st and 3rd quartiles reveal a small interquartile range. This

indicates that most variables have effects that do not deviate substantially from the median, suggesting consistency in the model's outputs.

The Bias shows the deviation of the estimated coefficients $\widehat{\beta_i}$ from the true coefficients β_i in the original dataset. The Bias for each coefficient in each dataset is calculated as follows:

$$\text{Bias}_i = \hat{\beta}_i - \beta_i \tag{36}$$

$$\text{Mean Bias} = \frac{1}{p}\sum_{i=1}^{p} \text{Bias}_i \tag{37}$$

i indicates the coefficients of the variables inside each dataset, and p is the total number of variables. Then, the mean of the Mean Bias for each dataset is computed by averaging the Mean Bias across all variables.

RMSE (root mean square error) combines Bias and variance to measure overall error. For each dataset, the RMSE and Mean RMSE are calculated as:

$$\text{RMSE}_i = \sqrt{\frac{1}{p}\sum_{i=1}^{p} (\text{Bias}_i)^2} \tag{38}$$

$$\text{Mean RMSE} = \frac{1}{p}\sum_{i=1}^{p} \text{RMSE}_i \tag{39}$$

For a balanced panel in Figure 28, "mice" is performing better in MCAR and MAR conditions for lower levels of missingness; most values are close to zero. Compared to "mice", "amelia" does not perform very well. "amelia" has Biases chronologically close to zero (10%, then 30%, and then 50%). Like "mice", "amelia" also suffers while imputing MNAR settings. "mitml" is not performing well compared to "mice" and "amelia". "mitml" provides positive and very high Biases, while most frameworks provide negative biases and close to zero. "LSTM-Network" cannot be considered even if the Biases look better than "mitml" due to overfitting. But for the MNAR condition and 10% missingness, the Bias is the same as "amelia" (-0.013), even performing better in MAR and 10% missingness, showing its future potentiality.

Figure 29 shows the Bias of the simulated unbalanced panel datasets. The performance, in general, is not very distinctive from the balanced panel, except the values are a little more dispersed than balanced panels. "mice" mostly have a Bias close to zero compared to "amelia". "mice" mostly performed better for MCAR conditions and struggled for MAR and MNAR. "amelia" also struggled for MAR and MNAR. Interestingly,

under the MNAR condition, "mice" exhibits lower Bias than "amelia" at higher levels of missingness in both balanced and unbalanced panels. The performance of "amelia" is chronological (10% than 30% and 50%)."mitml" is showing high biases compared to both "mice" and "amelia". "LSTM-Network" in an unbalanced panel performs even better than the balanced panel for MNAR, with 10% missingness and is better than all the other packages.

Figure 30 shows the RMSE scores in the balanced panel for the frameworks. "mice" achieves lower RMSE values under MCAR conditions and at lower levels of missingness. However, RMSE values increase as the level of missingness grows and under MNAR conditions. While "amelia" does not perform as well as "mice" regarding RMSE, its scores remain acceptable. Notably, "amelia" struggles more with higher levels of missingness than with the types of missingness (MCAR, MAR, MNAR). It chronologically provides good results (10%, then 30%, then 50%). For example, "amelia" is better in MNAR at 10% missingness compared to MCAR at 50% missingness. The higher RMSE values show that it is not as reliable as "mice" and "amelia." It does not give a better result in any scenario. Like Bias, the performance of "LSTM-Network" is affected by overfitting, resulting in misleading scores. However, its capability to handle MAR and MNAR conditions at lower levels of missingness is notable. For example, it outperforms "amelia" in MAR and MNAR conditions at 10% missingness, highlighting its potential for improvement in these areas.

For an unbalanced panel, as shown in Figure 31, the results differ slightly. "mice" struggle more in this case, with outcomes that are not as good as for the balanced panel. However, it still generally outperforms the other frameworks. For MNAR at 10% missingness, "amelia" achieves RMSE values comparable to "mice." As observed in the balanced panel, "amelia" again demonstrates better performance at lower levels of missingness. The performance of "mitml" is not commendable in any scenario compared to "mice" and "amelia." Meanwhile, "LSTM-Network" scores better than "amelia" in MAR and MNAR conditions at lower levels of missingness.

The key takeaways from both Bias and RMSE are as follows: "mice" struggle more with unbalanced panels than with balanced panels. It performs well under MCAR and MAR conditions at lower levels of missingness. "amelia" performs better at lower levels of missingness and does not differentiate as strongly between MAR and MNAR conditions as "mice" does. Moreover, "amelia" maintains relatively consistent performance across balanced and unbalanced panels. In contrast, "mitml" is less reliable than "mice" and "amelia." Finally, "LSTM-Network" shows potential in MAR and MNAR conditions at lower levels of missingness, indicating areas for further improvement.

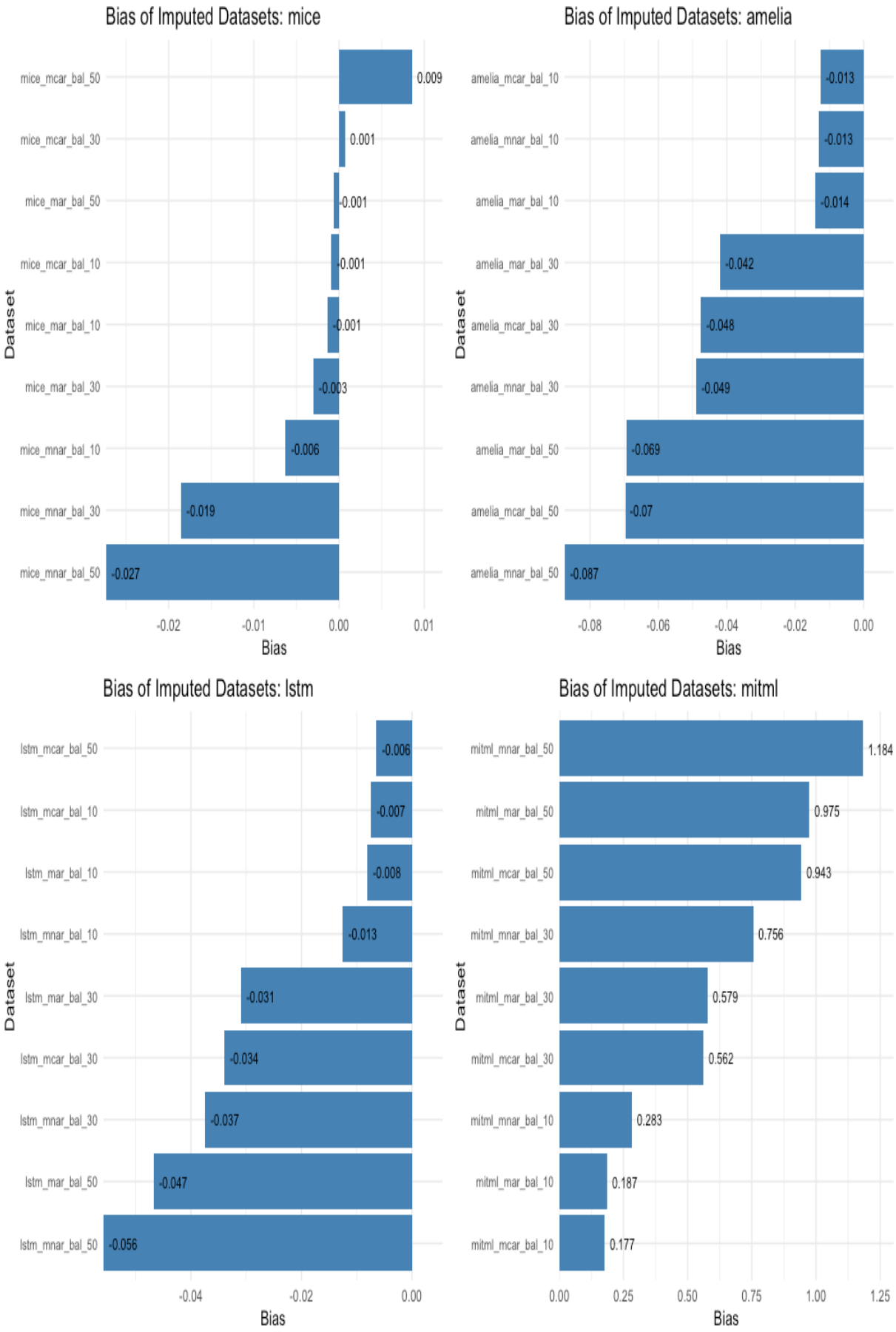


Figure 28: Bias Histogram for Balanced Panel

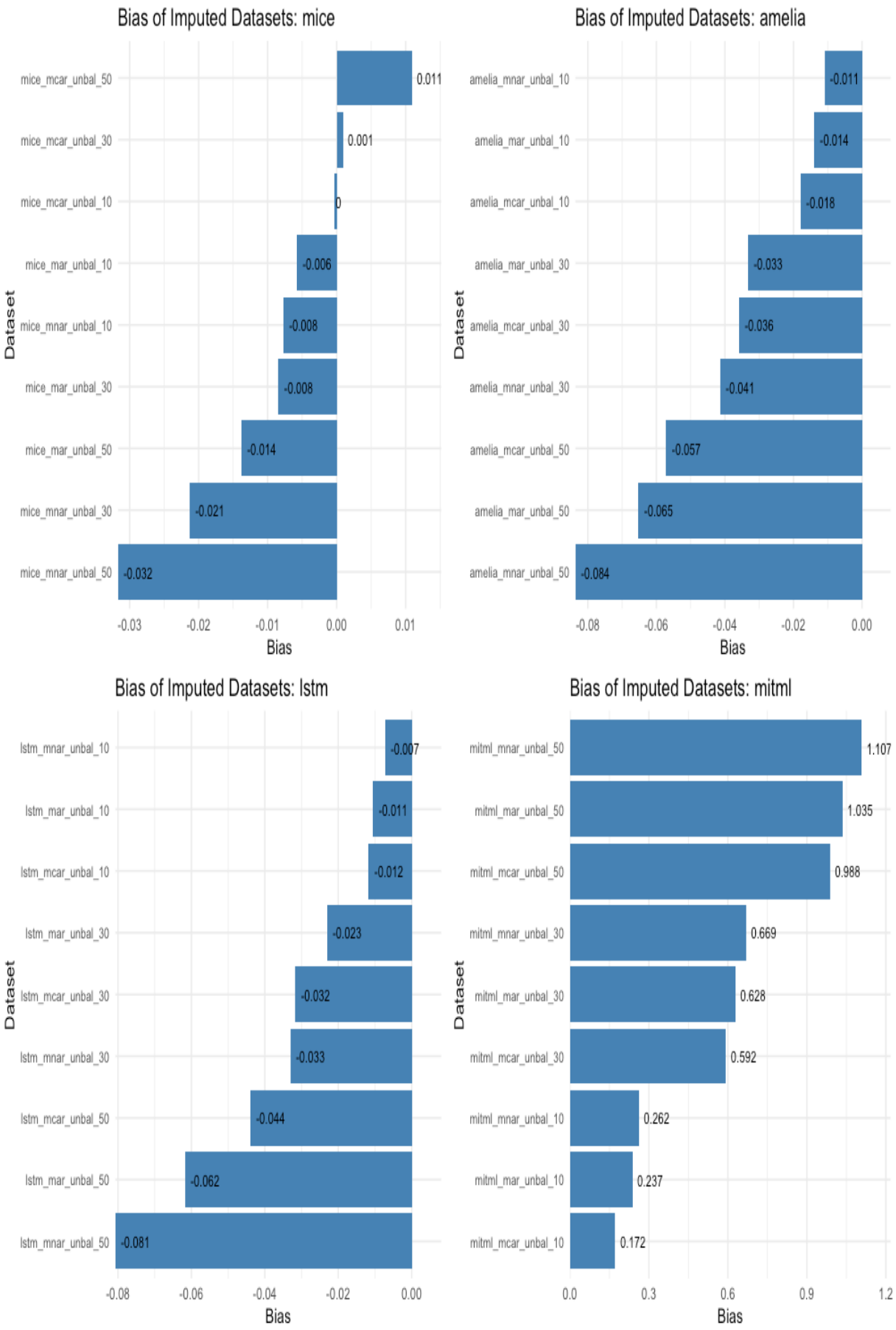


Figure 29: Bias Histogram for Unbalanced Panel

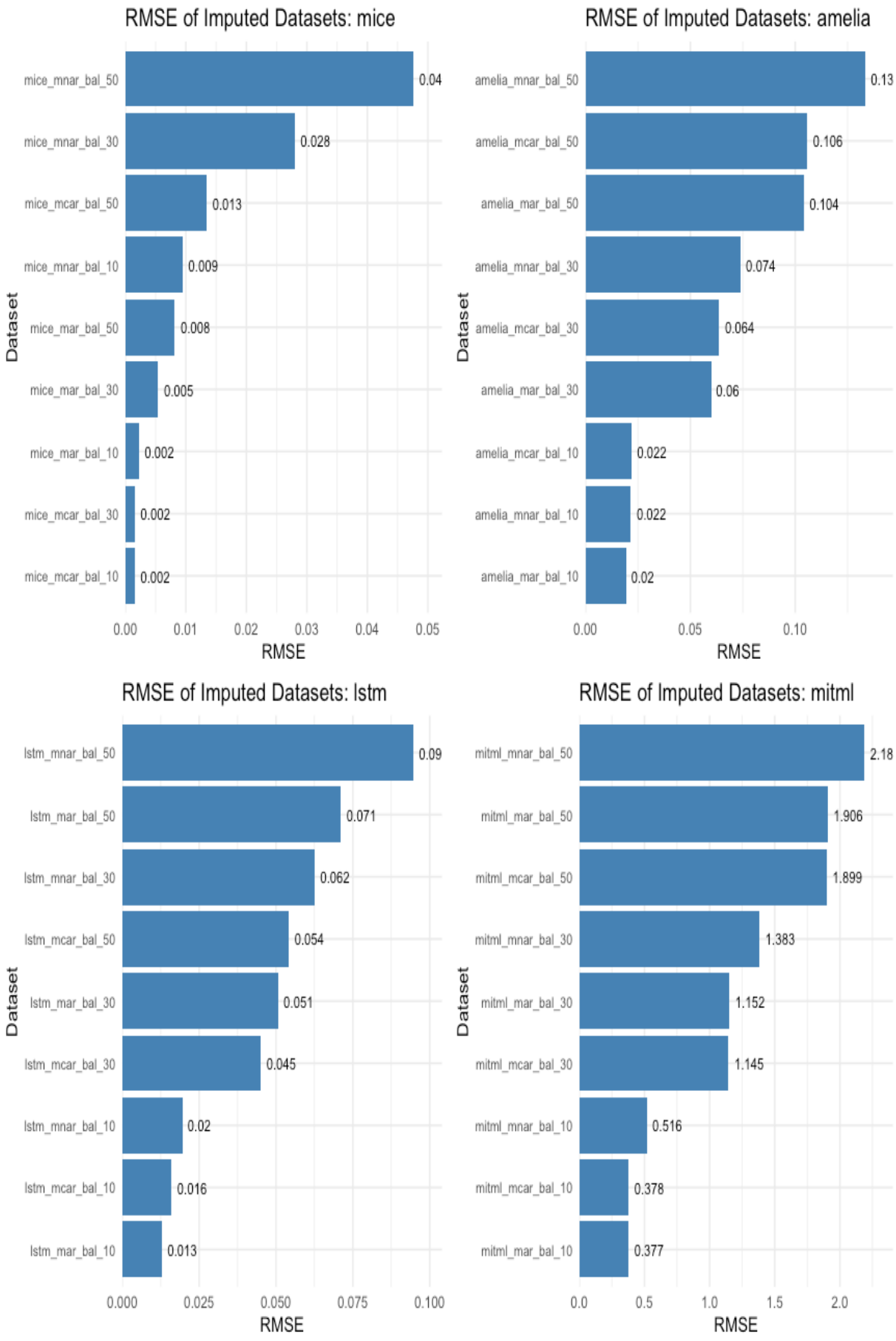


Figure 30: RMSE Histogram for Balanced Panel

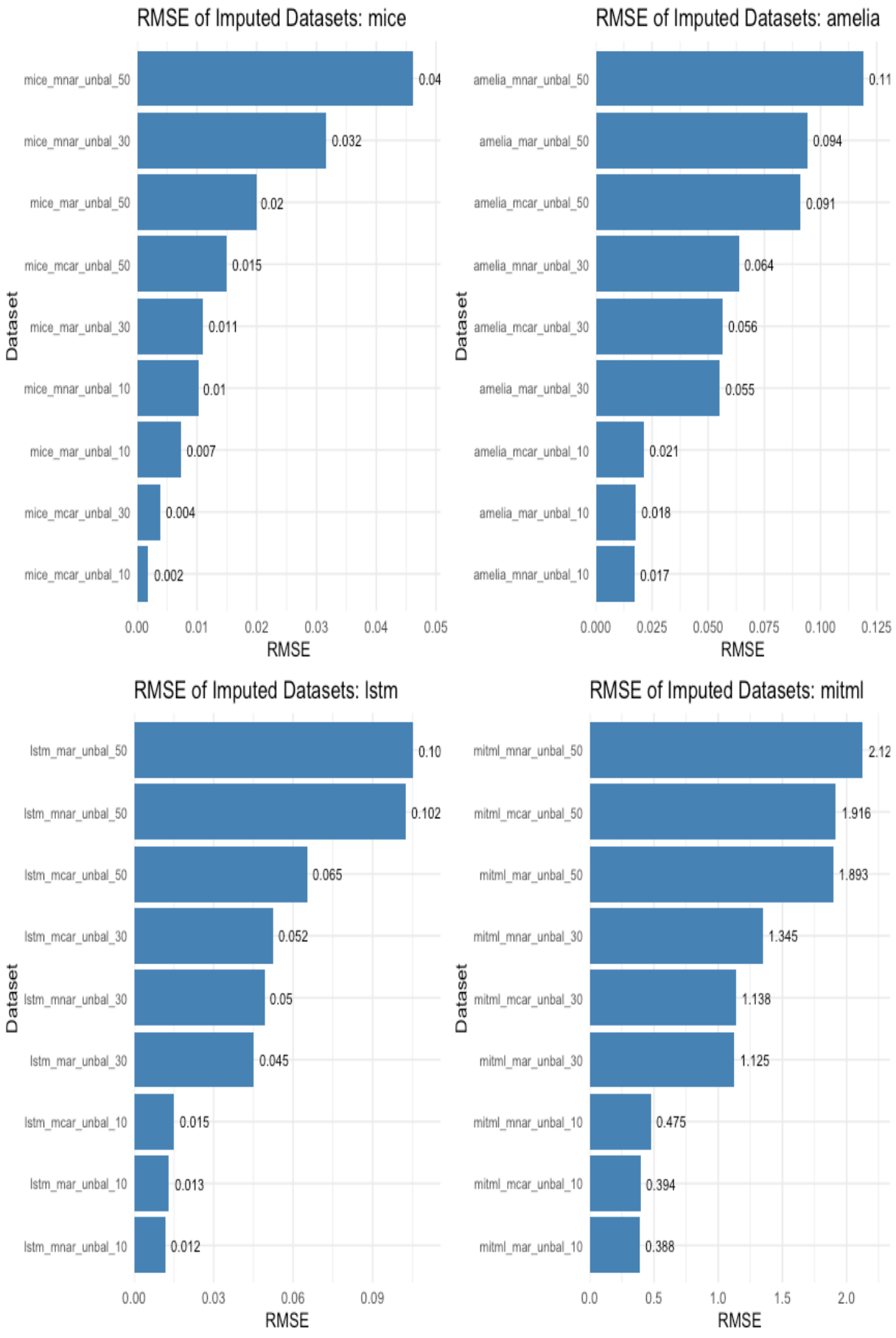


Figure 31: RMSE Histogram for Unbalanced Panel

Initially, the Bias and RMSE are calculated for each variable within every dataset. Subsequently, the mean Bias and mean RMSE are computed for each dataset. These dataset-level means are then aggregated to calculate the overall mean Bias and mean RMSE for each framework.

Algorithm	Mean Bias (Balanced)	Mean RMSE (Balanced)	Mean Bias (Unbalanced)	Mean RMSE (Unbalanced)
amelia	-0.0450	0.0671	-0.0399	0.0595
LSTM-Network	-0.0266	0.0474	-0.0337	0.0511
mice	-0.00543	0.0130	-0.00858	0.0163
mitml	0.627	1.22	0.632	1.20

Table 18: Average Bias and RMSE for Balanced and Unbalanced Panels

In table 18, the comparison of overall mean Bias and RMSE values across balanced and unbalanced panel datasets highlight that "mice" consistently outperform other frameworks, with the lowest Bias and RMSE, demonstrating its robustness and accuracy in both settings. "amelia" perform moderately well, showing slightly higher Bias and RMSE, with improving its RMSE in the unbalanced dataset. Conversely, "mitml" struggles significantly in both datasets, with higher Bias and RMSE than "amelia" and "mice", indicating consistent overestimation and significant errors in imputing missing values. Even if "LSTM-Network" provides acceptable results, it cannot be considered because the results here are misleading due to the overfitting. In overall conditions, "mice" is a more reliable imputation package, maintaining low error and Bias across balanced and unbalanced scenarios. Then, "amelia" provides acceptable results after "mice". After that, "mitml" shows poor performance, suggesting it is less suitable for accurate imputations.

5.6 Critical Analysis

This thesis study highlighted as many scenarios as possible. However, some conditions could not be highlighted due to time and resource scarcity. Simulating missingness rather than only in "income" would give the study more insight into the imputation. With density curves, more measures would provide more information about the simulated datasets and their respective frameworks, such as box plots. Information like extreme values, outliers, median and quantiles could be analysed thoroughly through box plots. The "David" dataset has 92 variables to select. This study uses nine variables. Adding more variables would provide more edge to the study. By adding more variables, more missingness could be simulated. More predictors could be added to the imputation model, providing more extra information to explore. This thesis study uses four frameworks to explore. More frameworks (mentioned in 2.2.1) could be implemented to explore more insights about the nature of the panel data and the missingness. Five-panel data variants are mentioned in section 2.1.2. This thesis study

explored balanced and unbalanced panel data. Even if the rest of the variants are the parts of the main two branches, exploring the rest could also provide more information to study. While simulating the imputed datasets, five datasets are simulated for multiple imputations. More datasets could provide more information on the simulation and imputation. No framework provides reasonable results for MNAR conditions at higher levels of missingness. Exploring more frameworks for the MNAR conditions at a higher level of missingness could also be a future endeavour.

"mice" may has performed well in MCAR and MAR, but it has some major drawbacks. For example, a significant amount of statistical knowledge is required to impute the missing values. The lagged values of the predictors to impute "Income" worked better because "Income" is highly dependent on its previous values. Other variables might not work as expected. Lagged and squared values of "Age" are also implemented before "Age" is used as a predictor. Only "Age" gave better results. Using a predictor matrix also requires a significant amount of coding experience and statistical knowledge, which restricts "mice" from becoming available to non-statistician users. "amelia" does not have any complications like "mice", and the user interface is much simpler; nonetheless, it is significantly slower than "mitml" and "mice". Compared to the 13.11 and 25.23 minutes in "mice" and "mitml", it took 4.24 hours to accomplish the same task, which is a significant drawback. "mitml" does not have any complications like "mice" and is also not as slow as "amelia". However, it could not perform like the other frameworks. It came in almost third place among all the measurements. The "LSTM-Network" has some drawbacks that need to be discussed. The first drawback is its complexity; it is a challenging framework to understand. The initial implementation requires significant statistical knowledge and coding experience. Second, it is slow. In terms of computational efficiency, it came third with 2.409152 hours. It also needs a lot of system memory and computational power. Third, it shares the same problem as other machine learning models: overfitting. It provides better results in correlation, Bias, and RMSE. However, within the density curves, it is evident that this is possible only because the data contains a lot of zero values, and it significantly outperformed in predicting these zero values. The model is trained primarily on this data and struggles to impute values that are not zero. This is why, even if it appears to have performed better than "mitml" on paper, in real-life scenarios, it does not. Nevertheless, it demonstrated some positive aspects as well. The main point is that it performed very well with 10% missing values. It even better predicted the MAR and MNAR conditions for a 10% missingness, where all the other frameworks struggled, showing immense potential for future research and innovation with these frameworks.

6 Conclusion

"mice" shows better performance in general except for MNAR conditions. "amelia" ranks second overall in terms of performance. Even first in MNAR at a lower level of missingness. It also reserves minimal changes in the correlation values across different scenarios. However, it shows a notable slowdown in the simulation process. "mitml" is notable for the simulation process faster than "amelia." Nonetheless, the measurements do not show a better performance than "amelia" and "mice". "LSTM-Network" ranked fourth in terms of performance. What distinguishes "LSTM-Network" is its ability to impute 10% missingness effectively compared to "amelia" and "mitml". And nearly on par with mice, even outperforming them in the MNAR conditions. This highlights its future potential. In general, the frameworks can impute missing values with less struggle in MCAR, a little in MAR and a lot in MNAR conditions. Generally, they suffer more to impute in unbalanced panels than in balanced panels, particularly "mice". Many surveys encounter difficulties because of missing data. This problem with data missingness is not limited to surveys. It also affects other data collection fields. Therefore, understanding panel data, data missingness and using suitable frameworks for imputing missing values are essential. Conventional frameworks for panel data imputing are adequate, but they have drawbacks. Investigating new frameworks that can successfully handle the difficulties posed by traditional frameworks is essential. Emerging techniques like machine learning provide the best example of this progress. It has been demonstrated that the "LSTM Network" can effectively overcome the challenges associated with imputing MAR and MNAR situations. Notwithstanding its drawbacks, this framework has a lot of room for development and may even outperform traditional imputation frameworks. Future research could enhance the potential of this study. Research such as adding more simulated missing values across different variables, including more variables as predictors, introducing additional exogenous variables, using various panel datasets, and employing more frameworks of missingness are some examples. The drawbacks mentioned in the 5.6 section could be addressed in future studies. For example, explore a framework that can impute higher levels of missingness (more than 30%) in MNAR condition, to make "mice" less complicated and less dependent on statistical knowledge, make "amelia" faster, improve the efficiency of "mitml" in imputing, and eliminate the overfitting issue of the "LSTM-Network", etc. This is how the future of panel data imputation could lie in combining the powerful potential of machine learning with the reliability of traditional frameworks. As adopting new ideas while improving current data imputation frameworks are not just essential but a responsibility.

References

Abdullah, H., Sahudin, Z., & Bahrudin, N. Z. (2022): Short Guides to Static Panel Data Regression Model Estimator. Asian Journal of Accounting and Finance, 4(4), pp. 1-6.
https://doi.org/10.55057/ajafin.2022.4.4.1

Allison, P.D. (1998): Multiple imputation for missing data: A cautionary tale. Retrieved January 10, 1999, University of Pennsylvania.
http://www.ssc.upenn.edu/~allison

Allison, P.D. (2000): Missing data. Thousand Oaks, CA: Sage Publications.
https://doi.org/10.1177/0049124100028003003

Arnold, T. B. (2017): kerasR: R interface to the keras deep learning library. J. Open Source Softw., 2(14), 296.
10.21105/joss.00296

Balestra, P., & M. Nerlove (1966): "Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Economic Model: The Demand for Natural Gas," Econometrica, 34, pp. 585–612.
https://www.jstor.org/stable/1909771

Baltagi, B. H. (2008): Econometric analysis of panel data (Vol. 4, pp. 135-145). Chichester: Wiley.
https://doi.org/10.1007/978-3-030-53953-5

Basmann, R. L. (1957): A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica: Journal of the Econometric Society, pp. 77-83
https://doi.org/10.2307/1907743

Basmann, R. L. (1959): Basmann, R. L. (1959). The computation of generalized classical estimates of coefficients in a structural equation. Econometrica: Journal of the Econometric Society, pp. 72-81

Basmann, R. L., Brown, F. L., Dawes, W. S., & Schoepfle, G. K. (1971): Exact finite sample density functions of GCL estimators of structural coefficients in a leading, exactly identifiable case. Journal of the American Statistical Association, 66(333), pp. 122-126
https://doi.org/10.1080/01621459.1971.10482231

Bell, A., & Jones, K. (2015): Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. Political Science Research and Methods, 3(1), pp. 133-153.
https://doi.org/10.1017/psrm.2014.7

Breusch, T. S., & Pagan, A. R. (1980): The Lagrange multiplier test and its applications to model specification in econometrics. The review of economic studies, 47(1), pp. 239-253.
https://doi.org/10.2307/2297111

van Buuren, S., Boshuizen, H.C., Knook, & D.L. (1999): Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18, pp. 681–694.
https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6%3C681::AID-SIM71%3E3.0.CO;2-R

Van Buuren, S., & Oudshoorn, K. (1999): Flexible multivariate imputation by MICE (pp. 1-20). Leiden: TNO

Van Buuren, S., & Oudshoorn, C. G. (2000): Multivariate imputation by chained equations.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006): Fully conditional specification in multivariate imputation. Journal of statistical computation and simulation, 76(12), pp. 1049-1064.
https://doi.org/10.1080/10629360600810434

Van Buuren, S. (2018): Flexible imputation of missing data. Second Edition, CRC press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011): mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45, pp. 1-67.
https://doi.org/10.18637/jss.v045.i03

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015): Package 'mice'. Computer software, 20.

Cahan, E., Bai, J., & Ng, S. (2023): Factor-based imputation of missing values and covariances in panel data of large dimensions. Journal of Econometrics, 233(1), pp. 113-131
https://doi.org/10.1016/j.jeconom.2022.01.006

Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., ... & Huang, T. S. (2017): Dilated recurrent neural networks. Advances in neural information processing systems, P. 30.

Chamberlain, G. (1982): Multivariate regression models for panel data. Journal of econometrics, 18(1), pp. 5-46.
https://doi.org/10.1016/0304-4076(82)90094-X

Chamberlain, G. (1984): Panel data. Handbook of econometrics, 2, pp. 1247-1318.
https://doi.org/10.1016/S1573-4412(84)02014-6

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014): Learning phrase representations using RNN encoder-decoder for statistical machine translation.
https://arxiv.org/pdf/1406.1078

Christelis, D. (2011): Imputation of Missing Data in Waves 1 and 2 of SHARE. Available at SSRN 1788248.
https://dx.doi.org/10.2139/ssrn.1788248

Christopher Olah (2015): Understanding LSTM Networks. Google research
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Croissant, Y., & Millo, G. (2008): Panel data econometrics in R: The plm package. Journal of statistical software, 27(2), pp. 1-43.
https://hal.univ-reunion.fr/hal-01245304

Davison, A. C., & Hinkley, D. V. (1997): Bootstrap methods and their application (No. 1). Cambridge university press.

Dang, H. A., Jolliffe, D., & Carletto, C. (2019): Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments. Journal of Economic Surveys, 33(3), pp. 757-797
https://doi.org/10.1111/joes.12307

de Goeij, M. C., Van Diepen, M., Jager, K. J., Tripepi, G., Zoccali, C., & Dekker, F. W. (2013): Multiple imputation: dealing with missing data. Nephrology Dialysis Transplantation, 28(10), pp. 2415-2420.
https://doi.org/10.1093/ndt/gft221

Drechsler, J. (2015): Multiple imputation of multilevel missing data—Rigor versus simplicity. Journal of Educational and Behavioral Statistics, 40(1), pp. 69-95.
https://doi.org/10.3102/1076998614563393

Efron, B. (1979): Bootstrap methods: Another look at jackknife. Ann. Stat. 7, pp. 1-26.
https://link.springer.com/chapter/10.1007/978-1-4612-4380-9_41

Freedman, V.A., & Wolf, D.A. (1995): A case study on the use of multiple imputation. Demography, 32, pp. 459-470.
https://doi.org/10.2307/2061691

Frees, E. W. (1992): Forecasting state-to-state migration rates. Journal of Business & Economic Statistics, 10(2), pp. 153-167.
https://doi.org/10.1080/07350015.1992.10509895

Frees, E. W. (1993): Short-term forecasting of internal migration. Environment and Planning A, 25(11), pp. 1593-1606
https://doi.org/10.1068/a251593

Frees, E. W. (1995): Assessing cross-sectional correlation in panel data. Journal of econometrics, 69(2), 393-414.
https://doi.org/10.1016/0304-4076(94)01658-M

Gers, F. A., & Schmidhuber, J. (2000, July): Recurrent nets that time and count. In Proceedings of the IEEE INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium (Vol. 3, pp. 189-194). IEEE.
https://ieeexplore.ieee.org/abstract/document/861302

Greene, W. H. (2008) Econometric Analysis. Upper Saddle River, NJ: Pearson Prentice Hall.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016): LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), pp. 2222-2232.
https://arxiv.org/pdf/1503.04069

Grund, S., Lüdtke, O., & Robitzsch, A. (2016): Multiple imputation of multilevel missing data: an introduction to the R package pan. Sage Open, 6(4), 2158244016668220.
https://doi.org/10.1177/2158244016668220

Grund, S., Robitzsch, A., Luedtke, O., & Grund, M. S. (2019): Multiple imputation of multilevel missing data: an introduction to the R package pan. Sage Open, 6(4), 2158244016668220.

Gujarati, D. N. (2003): Basic Econometrics, International Edition - 4th Edition: McGraw-Hill Higher Education.
https://doi.org/10.1177/2158244016668220

Gulli, A., & Pal, S. (2017): Deep learning with Keras. Packt Publishing Ltd.

Trygve Haavelmo (1994): "The Probability Approach in Econometrics," Econometrica, pp. 50-56
https://www.jstor.org/stable/1906935

Halunga, A. G., Orme, C. D., & Yamagata, T. (2017): A heteroskedasticity robust Breusch–Pagan test for Contemporaneous correlation in dynamic panel data models. Journal of econometrics, 198(2), pp. 209-230.
https://doi.org/10.1016/j.jeconom.2016.12.005

Hansen, C. B. (2007): Asymptotic properties of a robust variance matrix estimator for panel data when T is large. Journal of Econometrics, 141(2), pp. 597-620.
https://doi.org/10.1016/j.jeconom.2006.10.009

Hausman, J.A. & W.E. Taylor (1981): Panel data and unobservable individual effects, Econometrica 49, pp. 1377–1398.

Hayakawa, K. (2007): Small sample bias properties of the system GMM estimator in dynamic panel data models. Economics letters, 95(1), pp. 32-38.
https://doi.org/10.1016/j.econlet.2006.09.011

Hirano, K., Imbens, G., Ridder, G., & Rebin, D. B. (1998): Combining panel data sets with attrition and refreshment samples. National Bureau of Economic Research
http://dx.doi.org/10.3386/t0230

Hochreiter, S., & Schmidhuber, J. (1997): Long short-term memory. Neural computation, 9(8), pp. 1735-1780.
https://ieeexplore.ieee.org/abstract/document/6795963

Honaker, J., King, G., & Blackwell, M. (2011): Amelia II: A program for missing data. Journal of statistical software, 45, pp. 1-47.
https://doi.org/10.18637/jss.v045.i07

Honaker, J., King, G., Blackwell, M., & Blackwell, M. M. (2010): Package 'Amelia'. Version. View Article
http://gking.harvard.edu/amelia

Hsiao, C. (2003): Analysis of Panel Data. ISBN 0521818559. Cambridge, Cambridge University Press, pp. 382.

Hsiao, C. (2022): Analysis of panel data (No. 64). Cambridge university press.
https://doi.org/10.1017/9781009057745

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015): An empirical exploration of recurrent network architectures. In International conference on machine learning (pp. 2342-2350). PMLR.

Kalton, G., D. Kasprzyk & D. McMillen (1989): Nonsampling errors in panel surveys, in D. Kasprzyk, G.J. Duncan, G. Kalton and M.P. Singh, eds., Panel Surveys. John Wiley, New York, pp. 249– 270

Kasprzyk, D., G.J. Duncan, G. Kalton & M.P. Singh (1989): Panel Surveys. John Wiley, New York.

Kennickell, A. B. (2017): Look again: Editing and imputation of SCF panel data. Statistical Journal of the IAOS, 33(1), pp. 195-202
http://dx.doi.org/10.3233/SJI-160268

Klevmarken, N. A. (1989): Panel Studies: what can we Learn from them?. European Economic Review, 33, pp. 523-529.
https://doi.org/10.1016/0014-2921(89)90131-1

Koutnik, J., Greff, K., Gomez, F., & Schmidhuber, J. (2014): A clockwork rnn. In International conference on machine learning (pp. 1863-1871). PMLR.
https://arxiv.org/pdf/1402.3511v1

Levin, A., Lin, C. F., & Chu, C. S. J. (2002): Unit root tests in panel data: asymptotic and finite-sample properties. Journal of econometrics, 108(1), pp. 1-24.
https://doi.org/10.1016/S0304-4076(01)00098-7

Landerman, L.R., Land, K.C., & Pieper, C.F. (1997): An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods & Research, 26, pp. 3-33.
https://doi.org/10.1177/0049124197026001001

Li, P., Stuart, E. A., & Allison, D. B. (2015): Multiple imputation: a flexible tool for handling missing data. Jama, 314(18), pp. 1966-1967.
https://doi.org/10.1001%2Fjama.2015.15281

Li, T., Li, G., Xue, T., & Zhang, J. (2020): Analyzing brain connectivity in the mutual regulation of emotion–movement using bidirectional granger causality. Frontiers in Neuroscience, 14, pp. 369.
https://doi.org/10.3389/fnins.2020.00369

Little, R. J. A., & Rubin, D. B. (1987): Multiple imputation for nonresponse in surveys. Wiley Series in Probability and Statistics, pp. 244–250.
https://doi.org/10.1002/9780470316696.refs

Little, R.J.A. (1988): Missing data adjustments in large surveys (with discussion), Journal of Business Economics and Statistics, 6, pp. 287–301
https://doi.org/10.1080/07350015.1988.10509663

Little, R. J. A., & Rubin, D. B. (1989): The analysis of social science data with missing values. Sociological Methods & Research, 18, pp. 292-326.
https://doi.org/10.1177/0049124189018002004

Marschak, J. (1950): "Statistical Inference in Economics: An Introduction," in Statistical Inference in Dynamic Economic Models, T. C. Koopmans Ed., New York: John Wiley, 12, pp. 1–50.

Marschak, J. (1953): "Economic Measurements for Policy and Prediction," in Studies in Econometric Method, W. C. Hood and T. P. Koopmans, Eds., New York: John

Wiley, pp. 1–26.

McDonough, I. K., & Millimet, D. L. (2017): Missing data, imputation, and endogeneity. Journal of Econometrics, 199(2), pp. 141-155. https://doi.org/10.1016/j.jeconom.2017.05.006

Moon, T. K. (1996): The expectation-maximization algorithm. IEEE Signal processing magazine, 13(6), pp. 47-60. https://doi.org/10.1109/79.543975

Mundlak, Y. (1961): On the pooling of time series and cross-section data, Econometrica 46, pp. 69–85. https://www.jstor.org/stable/1235460

Mundlak, Y. (1978): "Empirical Production Functions Free of Management Bias," Journal of Farm Economics, 43, pp. 44–56.

Münnich, R., Schnell, R., Kopp, J., Stein, P., Zwick, M., Dräger, S., ... & Schmaus, S. (2020): Zur Entwicklung eines kleinräumigen und sektorenübergreifenden Mikrosimulationsmodells für Deutschland. Mikrosimulationen: Methodische Grundlagen und ausgewählte Anwendungsfelder, pp. 109-138.

Münnich, R., Schnell, R., Brenzel, H., Dieckmann, H., Dräger, S., Emmenegger, J., ... & Stein, P. (2021): A population based regional Dynamic microsimulation of Germany: the MikroSim model. methods, data, analyses, 15(2), p. 24. https://doi.org/10.12758/mda.2021.03

Muzaffar, S., & Afshari, A. (2019): Short-term load forecasts using LSTM networks. Energy Procedia, 158, 2922-2927 https://doi.org/10.1016/j.egypro.2019.01.952

Newman, D. A. (2014): Missing data: Five practical guidelines. Organizational research methods, 17(4), pp. 372-411. https://doi.org/10.1177/1094428114548590

Nordholt, E. S. (1998): Imputation: methods, simulation experiments, and practical examples. International Statistical Review, 66(2), pp. 157-180. https://doi.org/10.1111/j.1751-5823.1998.tb00412.x

Ng, S. K., Krishnan, T., & McLachlan, G. J. (2012).: The EM algorithm. Handbook of computational statistics: concepts and methods, pp. 139-172.
https://doi.org/10.1007/978-3-642-21551-3_6

Patrician, P. A. (2002): Multiple imputation for missing data. Research in nursing & health, 25(1), pp. 76-84.
https://doi.org/10.1002/nur.10015

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013): How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026.
https://doi.org/10.48550/arXiv.1312.6026

Phillips, K. L., & Chen, B. (2011): Regional growth in China: An empirical investigation using multiple imputation and province-level panel data. Research in Economics, 65(3), pp. 243-253.
https://doi.org/10.1016/j.rie.2010.11.001

Phillips, P. C., & Moon, H. R. (2000): Nonstationary panel data analysis: An overview of some recent developments. Econometric reviews, 19(3), pp. 263-286.
https://doi.org/10.1080/07474930008800473

Robertson, D., & Symons, J. (1992): Some strange properties of panel data estimators. Journal of Applied econometrics, 7(2), pp. 175-189.
https://www.jstor.org/stable/2285027

Rubin, D. B. (2018): Multiple imputation. In Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC. pp. 29-62.
https://doi.org/10.1201/9780429492259

Rubin, D.B. (1987): Multiple imputation for nonresponse in surveys. New York: Wiley.

Rubin, D. B. (1996): Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434), pp. 473-489.
https://doi.org/10.1080/01621459.1996.10476908

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016): Missing data. Secondary analysis of electronic health records, pp. 143-162.

Schafer, J. L. (1997): Imputation of missing covariates under a multivariate linear mixed model. Technical report 97-04, Dept. of Statistics, The Pennsylvania State University

Schafer, J. L. (1998): The practice of multiple imputation. Paper presented at the meeting of the Methodology Center, Pennsylvania State University, University Park, PA.

Schafer, J.L. (1999): Multiple imputation: A primer. Statistical Methods in Medical Research, 8, pp. 3-15.
https://doi.org/10.1177/096228029900800102

Schafer, J.L. (2000): Analysis of incomplete multivariate data. Boca Raton, FL: Chapman & Hall/ CRC. (Original work published 1997)

Schafer, J. L. (2001): Multiple imputation with PAN. In L. M. Collins & A. G. Sayer (Eds.), New methods for the analysis of change (pp. 357–377). American Psychological Association.
https://doi.org/10.1037/10409-012

Scheffer, J. (2002): Dealing with missing data.

Schafer, J. L., & Yucel, R. M. (2002): Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics, 11, pp. 437-457.
https://doi.org/10.1198/106186002760180608

Singh, K., & Xie, M. (2008): Bootstrap: a statistical method. Unpublished manuscript, Rutgers University, USA, pp. 1-14.

Staudemeyer, R. C., & Morris, E. R. (2019): Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586.
https://doi.org/10.48550/arXiv.1909.09586

Villani, C. (2009): The Wasserstein distances. In: Optimal Transport. Grundlehren der mathematischen Wissenschaften, vol 338. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/978-3-540-71050-9_6

Wallace, T.D. & A. Hussain, (1969): The use of error components models in combining cross-section and time-series data, Econometrica 37, pp. 55–72.

Westermeier, C., & Grabka, M. M. (2016): Longitudinal wealth data and multiple imputation: An evaluation study. In Survey Research Methods (Vol. 10, No. 3, pp. 237-252). Konstanz: Universität Konstanz.
https://doi.org/10.18148/srm/2016.v10i3.6387

Wooldridge, J. M. (2010): Econometric analysis of cross section and panel data. MIT press.

Yao, K., Cohn, T., Vylomova, K., Duh, K., & Dyer, C. (2015) Depth-gated recurrent neural networks. arXiv preprint
https://arxiv.org/pdf/1508.03790v2

Young, Rebekah & Johnson, David R. (2015): Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. Journal of Marriage and Family, 77(1), pp. 277-294.
http://dx.doi.org/10.1111/jomf.12144

Appendix

A.1 Distribution Comparison

The density curves below compare the simulated datasets to the original data, focusing on individual datasets rather than frameworks as in section 5.2. While the purpose remains the same—to evaluate alignment with the original data—this visualization provides an alternative perspective by grouping the analysis by datasets instead of frameworks.

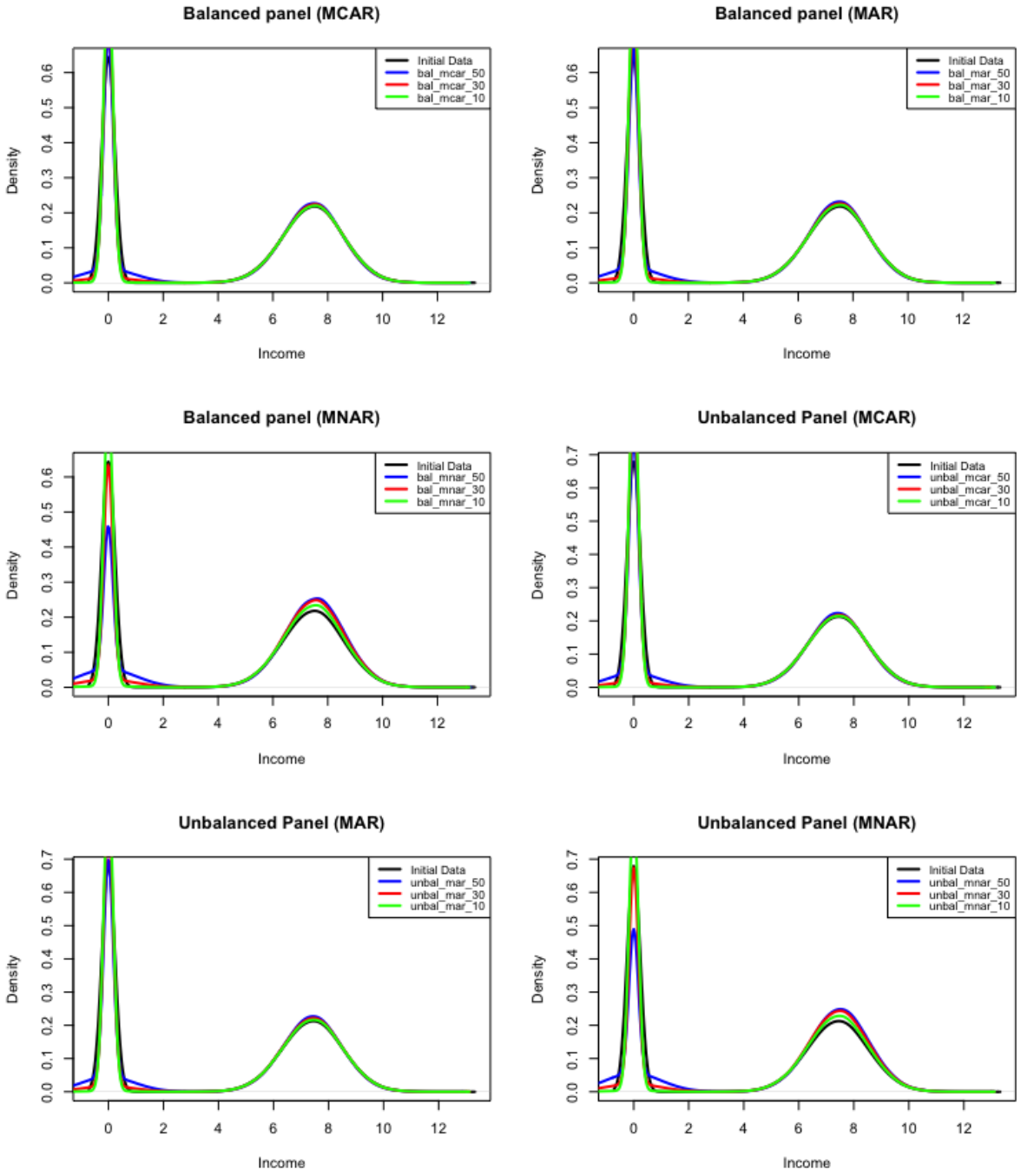


Figure A1: Income Distribution from "mitml"

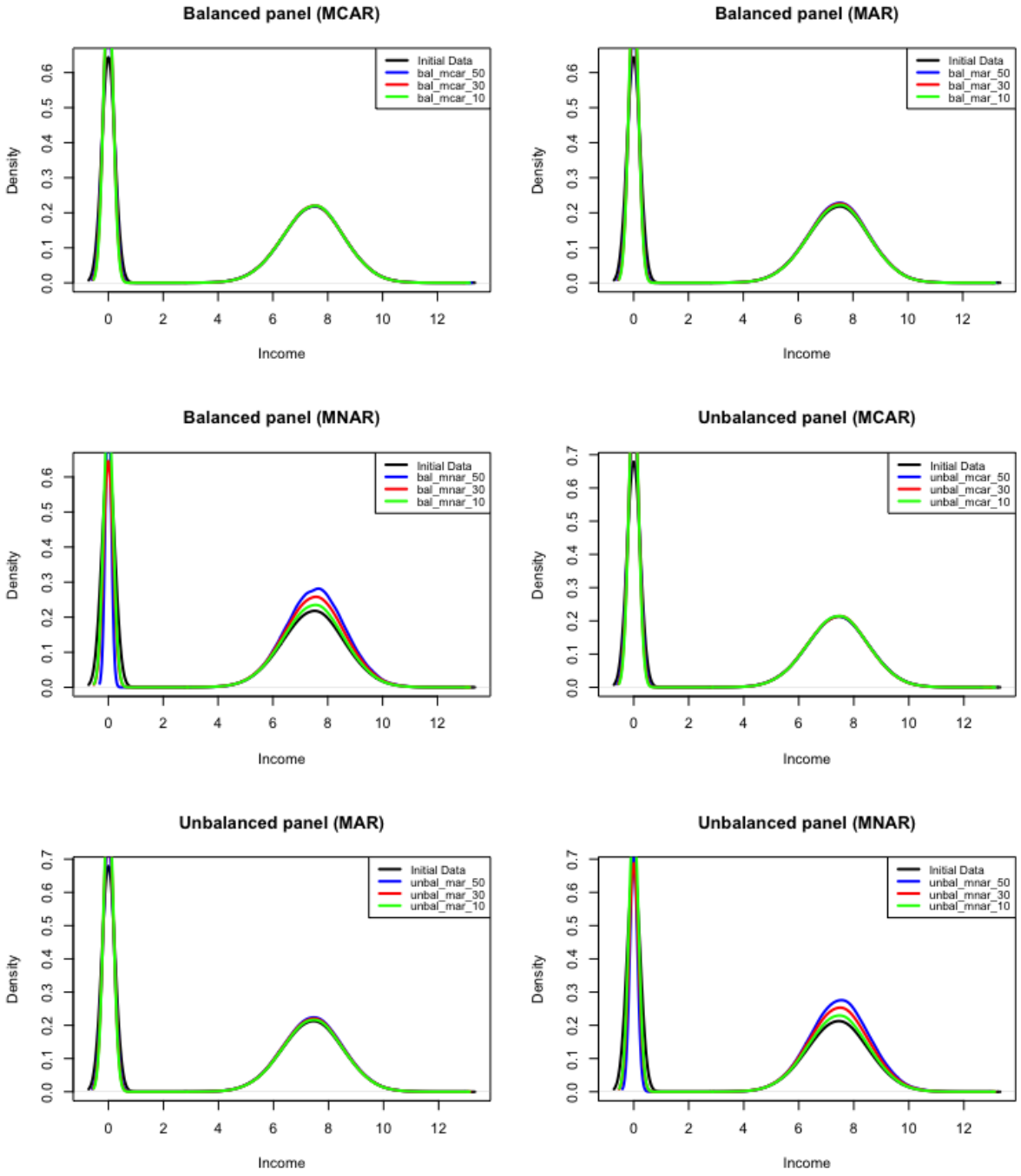


Figure A2: Income Distribution from "mice"

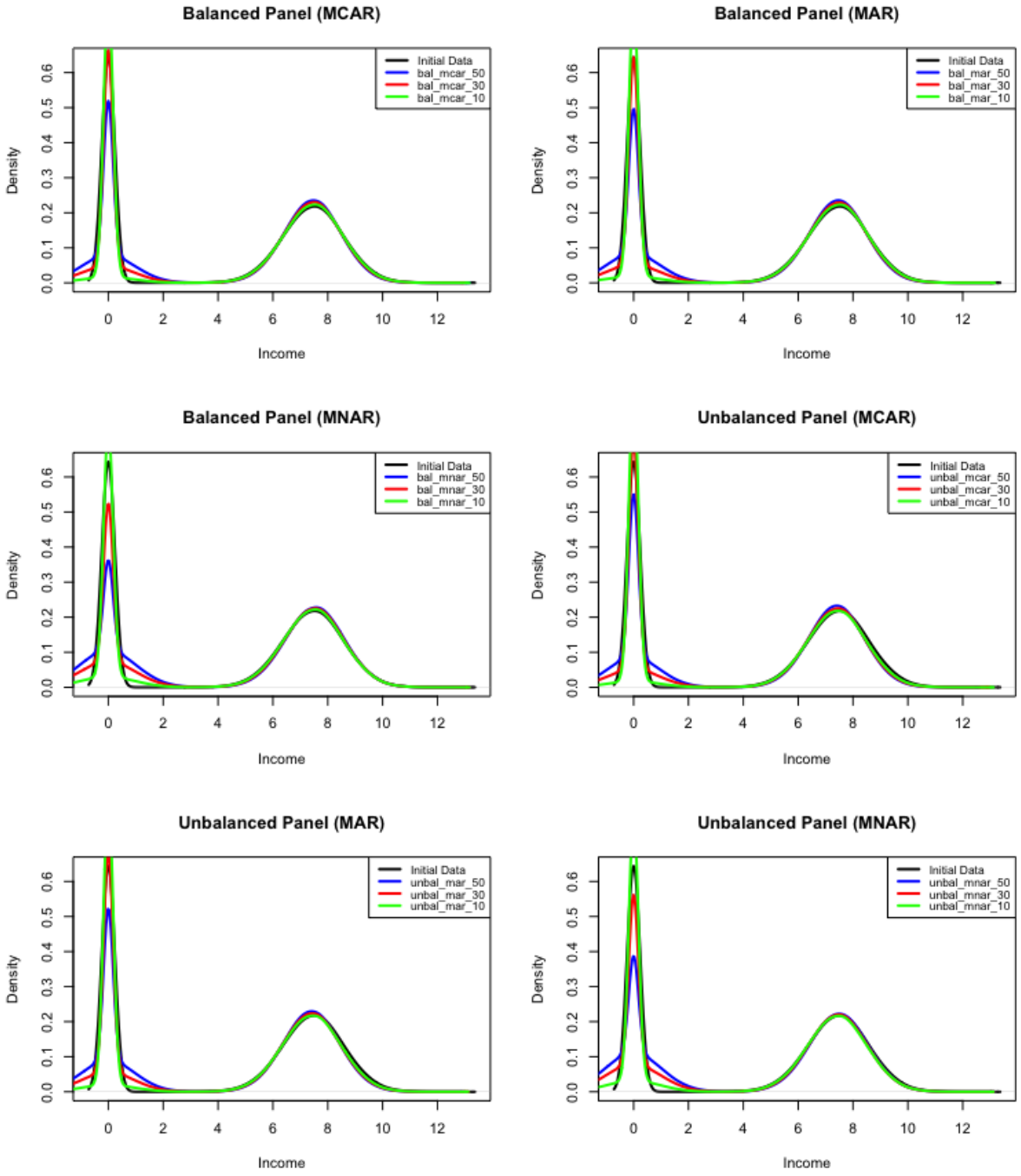


Figure A3: Income Distribution from "amelia"

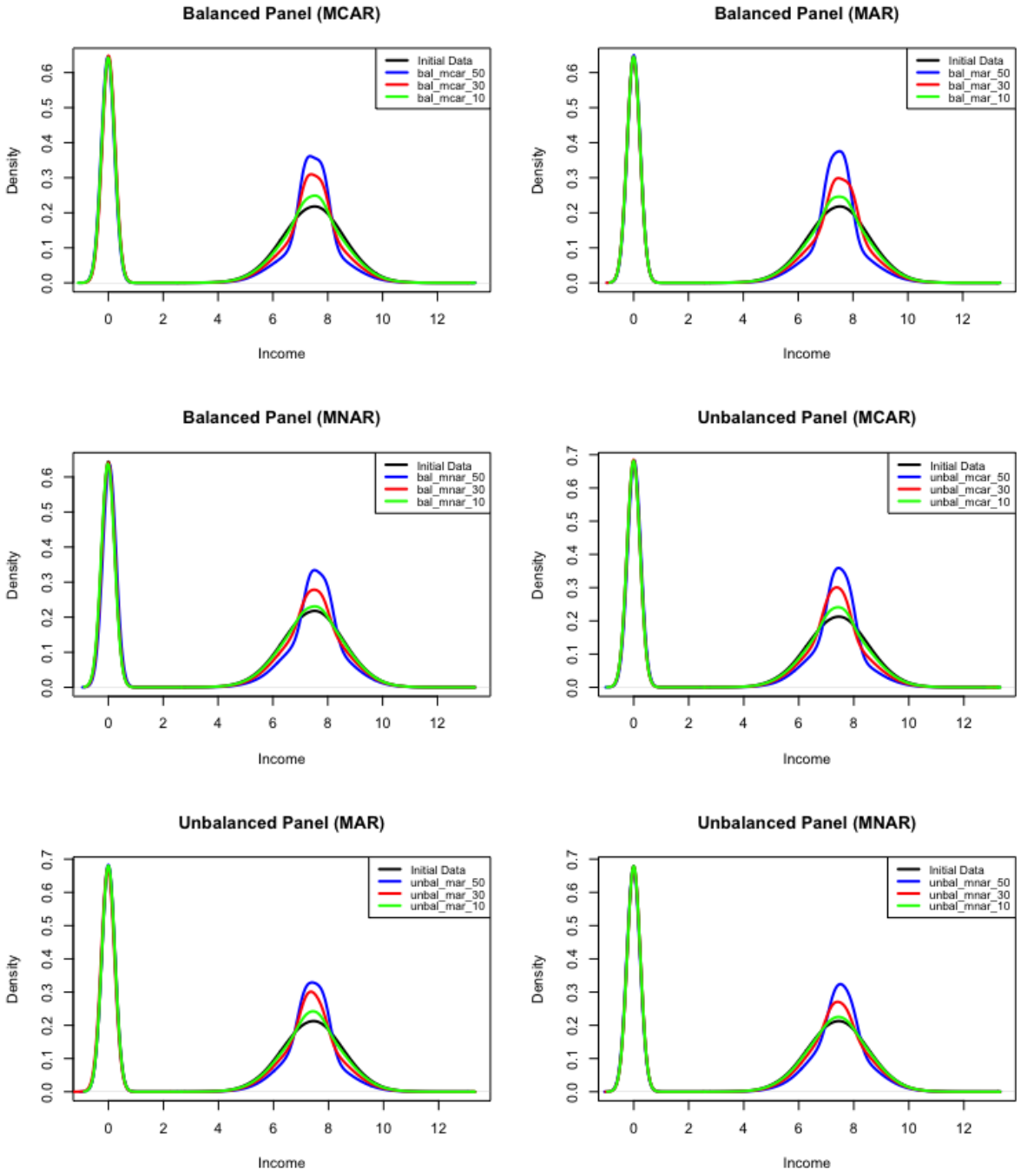


Figure A4: Income Distribution from "LSTM-Network"

A.2 Conditional Distribution Comparison

The conditional distribution below compares the simulated datasets with the original datasets but at 10% and 50% missingness. The general conclusion is not very different for the explained 30% missingness datasets in section 5.2.

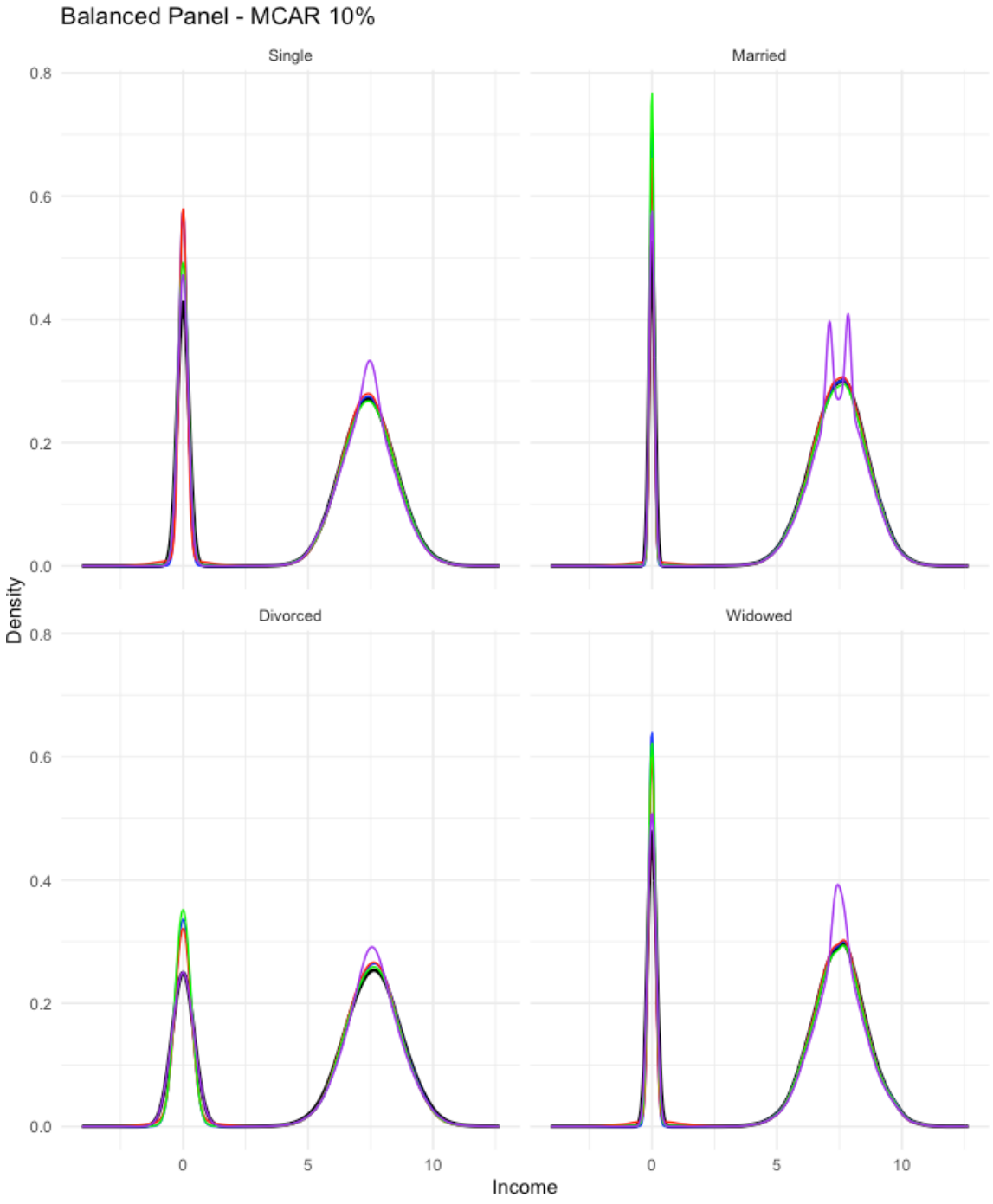


Figure A5: Conditional Distribution of Income for Balanced Panel MCAR at 10%

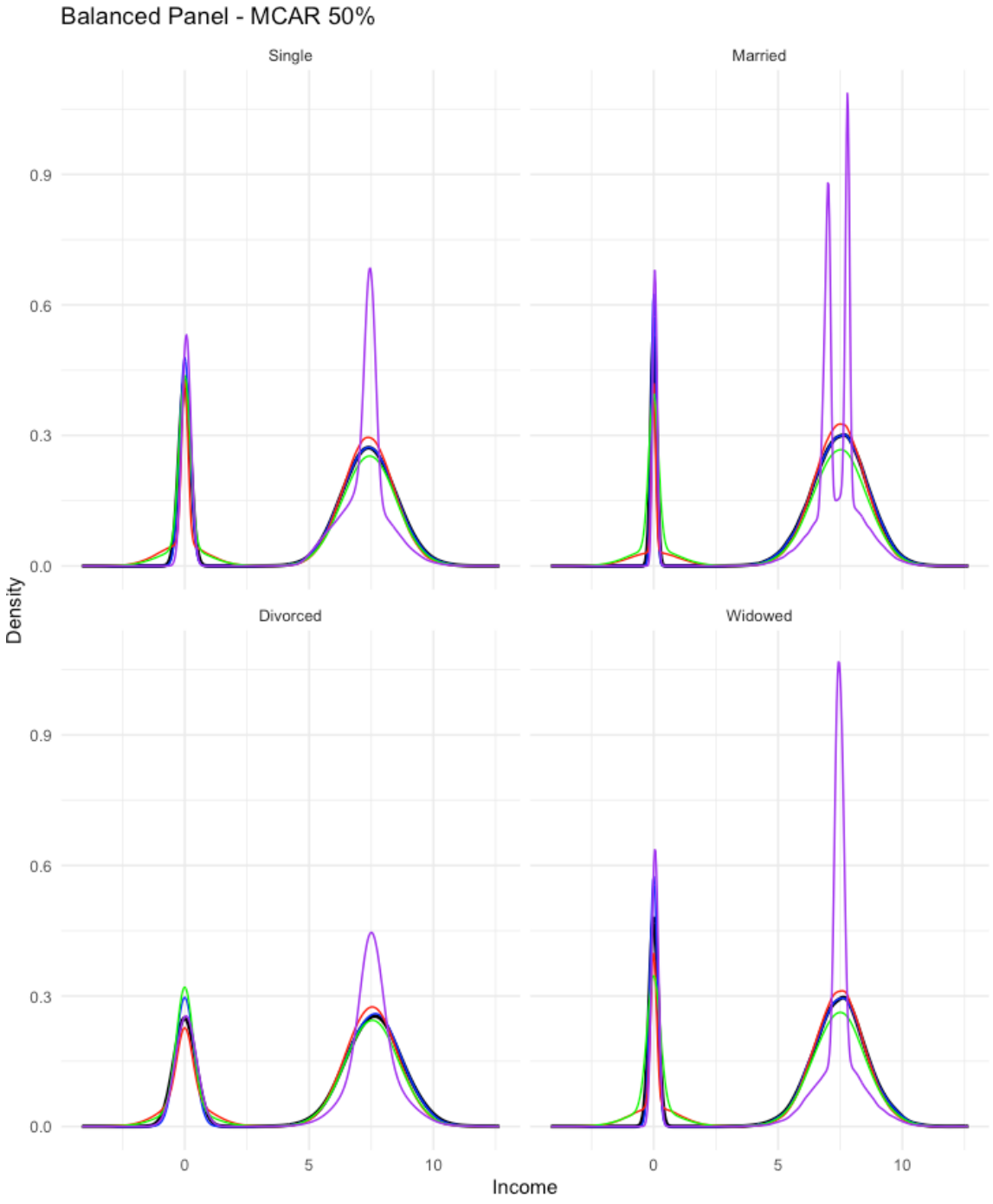


Figure A6: Conditional Distribution of Income for Balanced Panel MCAR at 50%

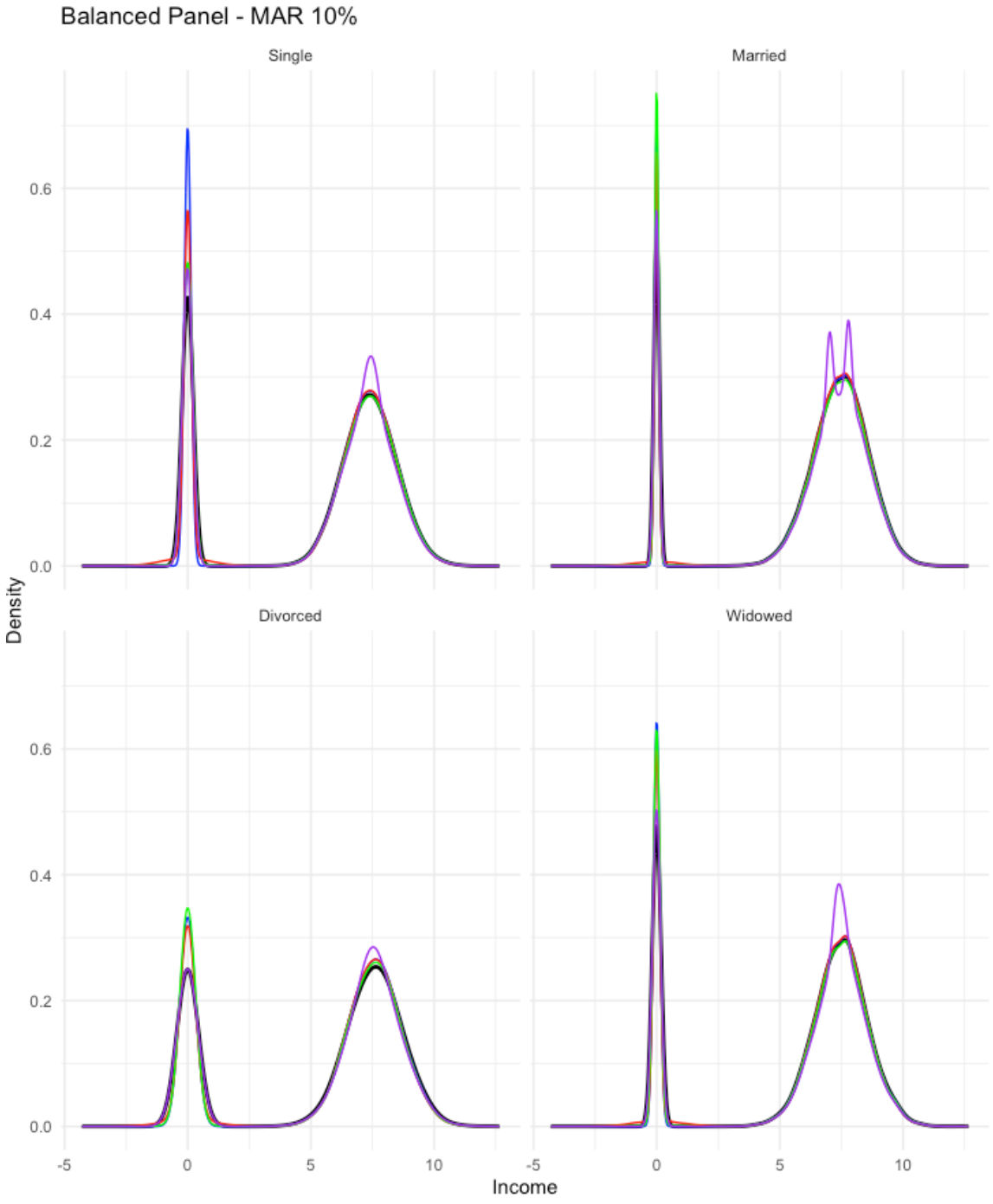


Figure A7: Conditional Distribution of Income for Balanced Panel MAR at 10%

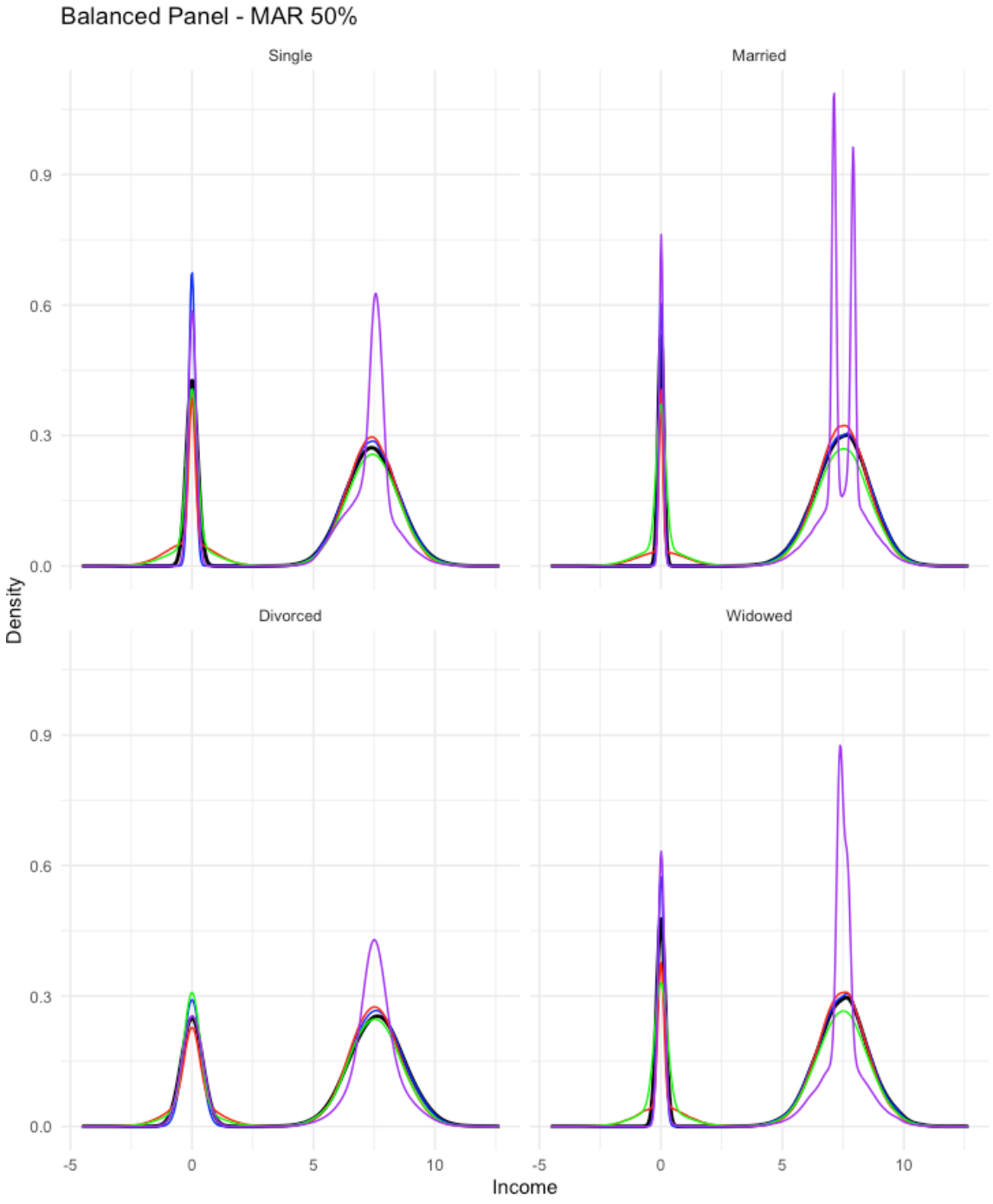


Figure A8: Conditional Distribution of Income for Balanced Panel MAR at 50%

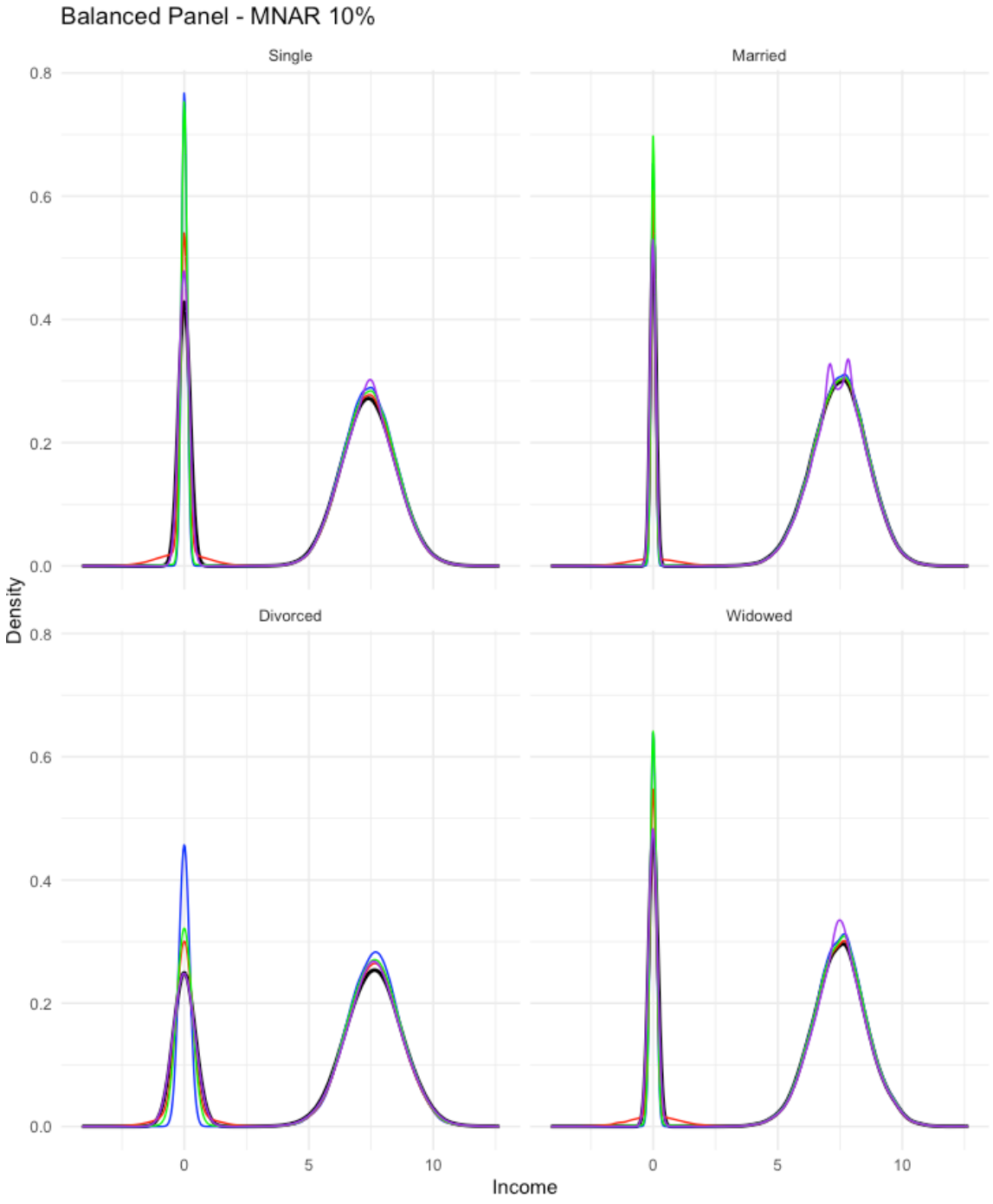


Figure A9: Conditional Distribution of Income for Balanced Panel MNAR at 10%

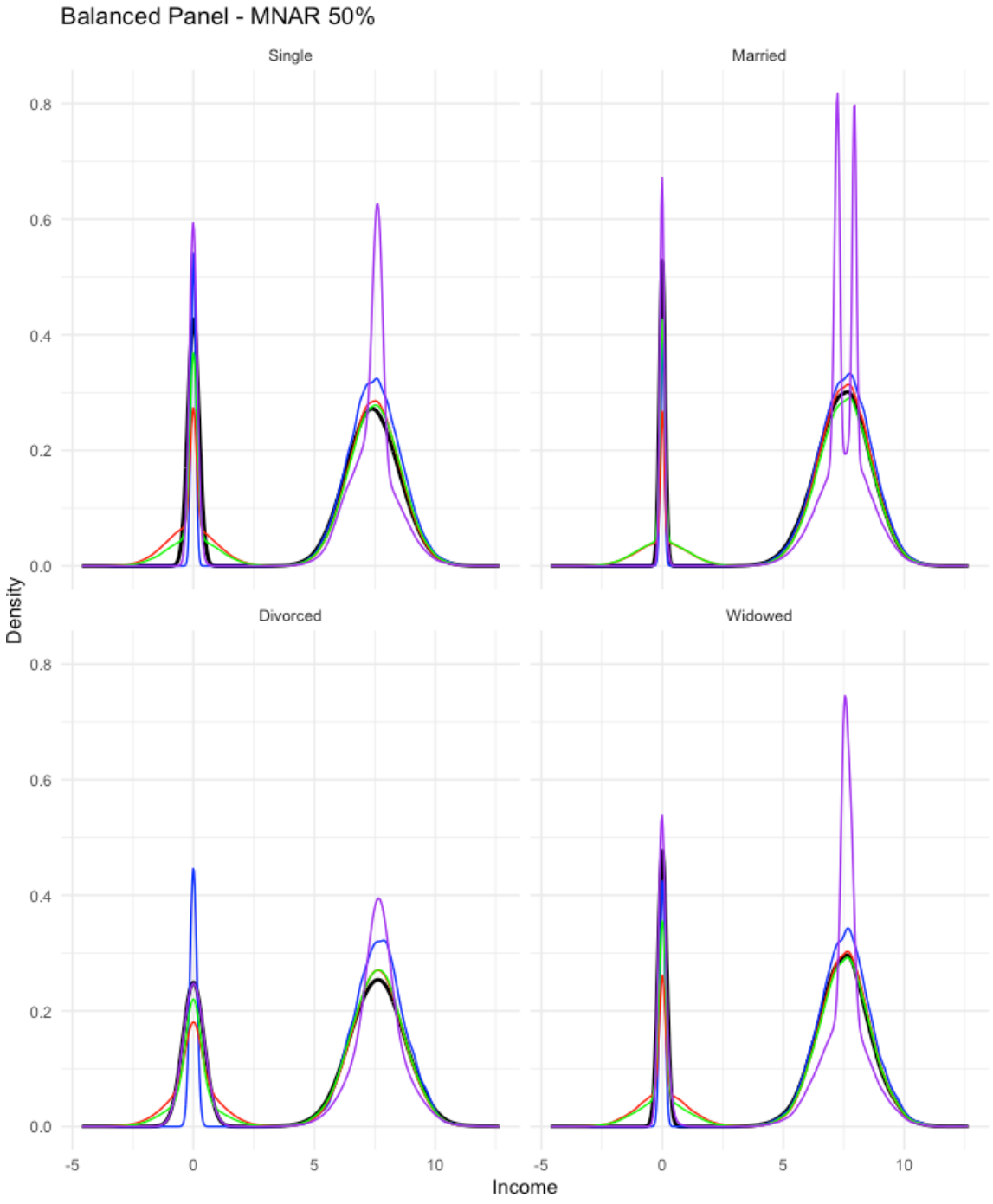


Figure A10: Conditional Distribution of Income for Balanced Panel MNAR at 50%

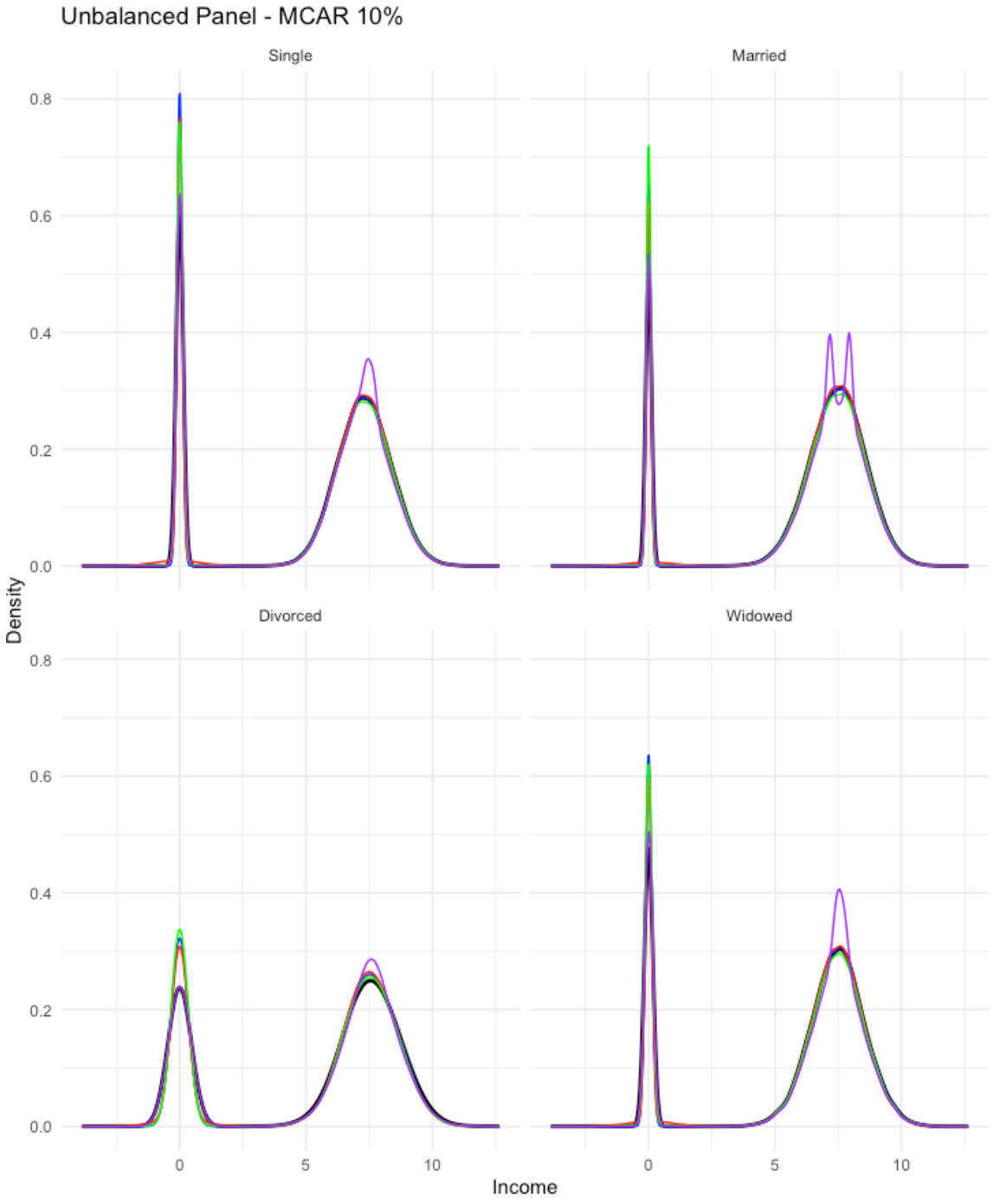


Figure A11: Conditional Distribution of Income for Unbalanced Panel MCAR at 10%

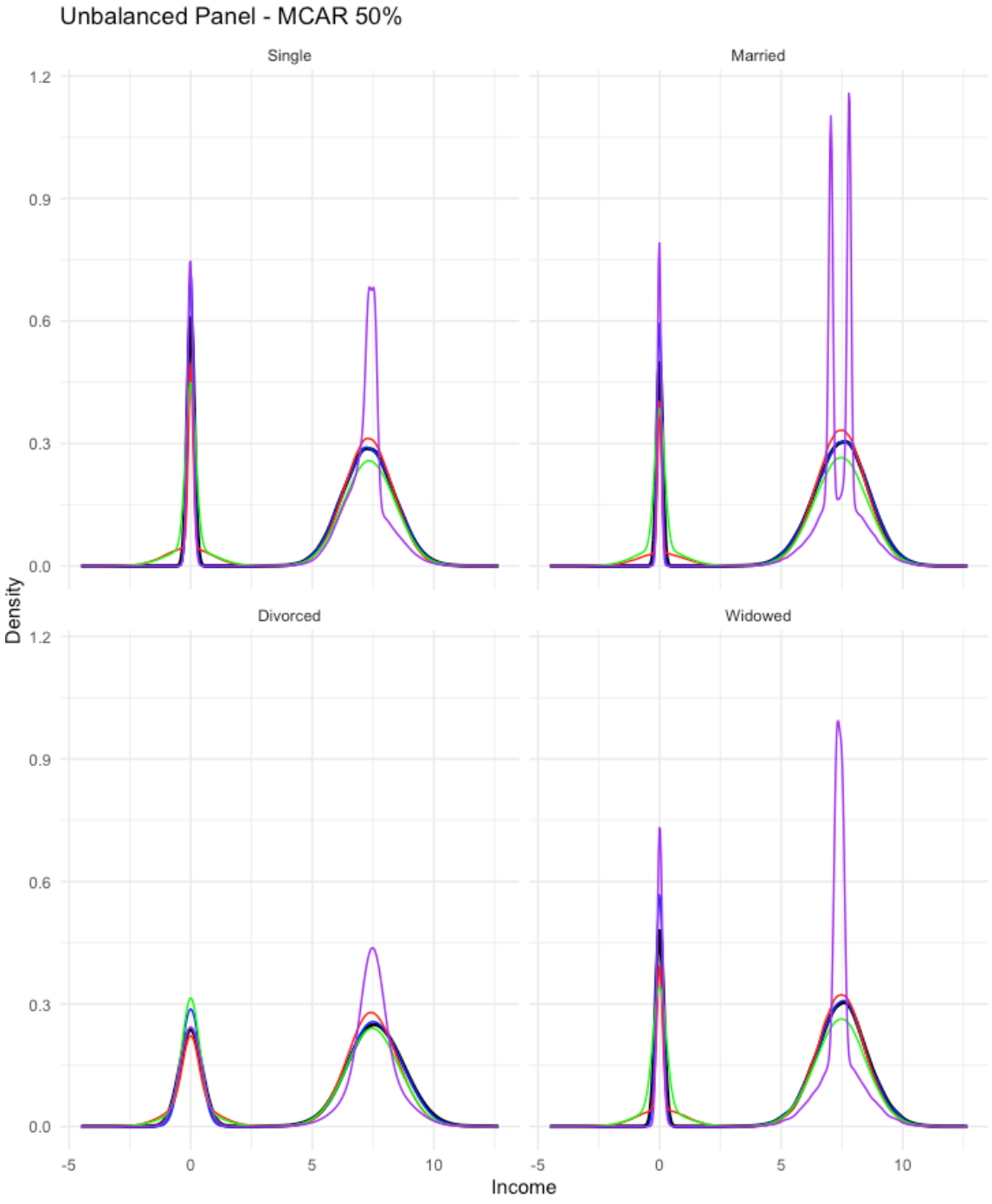


Figure A12: Conditional Distribution of Income for Unbalanced Panel MCAR at 50%

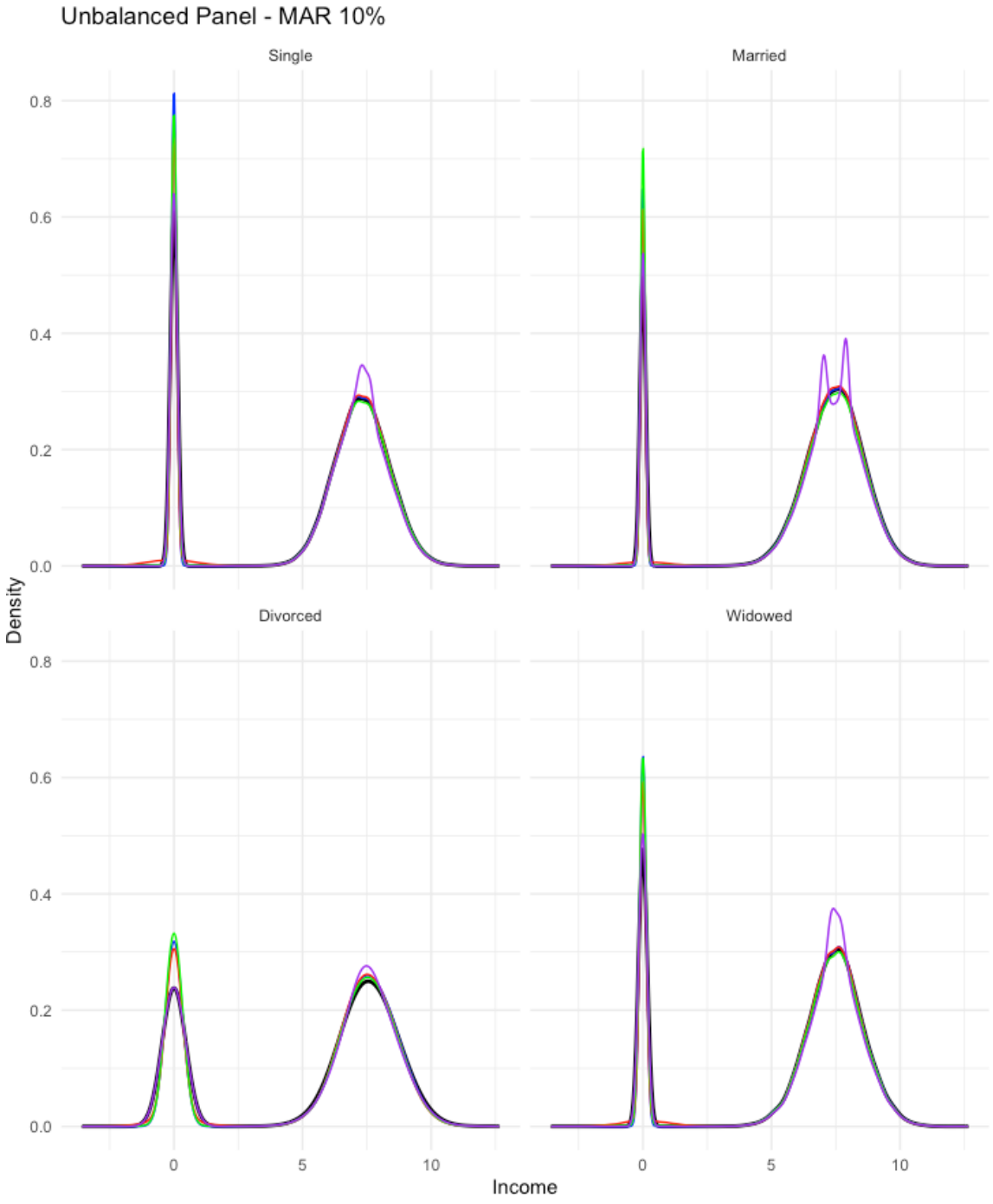


Figure A13: Conditional Distribution of Income for Unbalanced Panel MAR at 10%

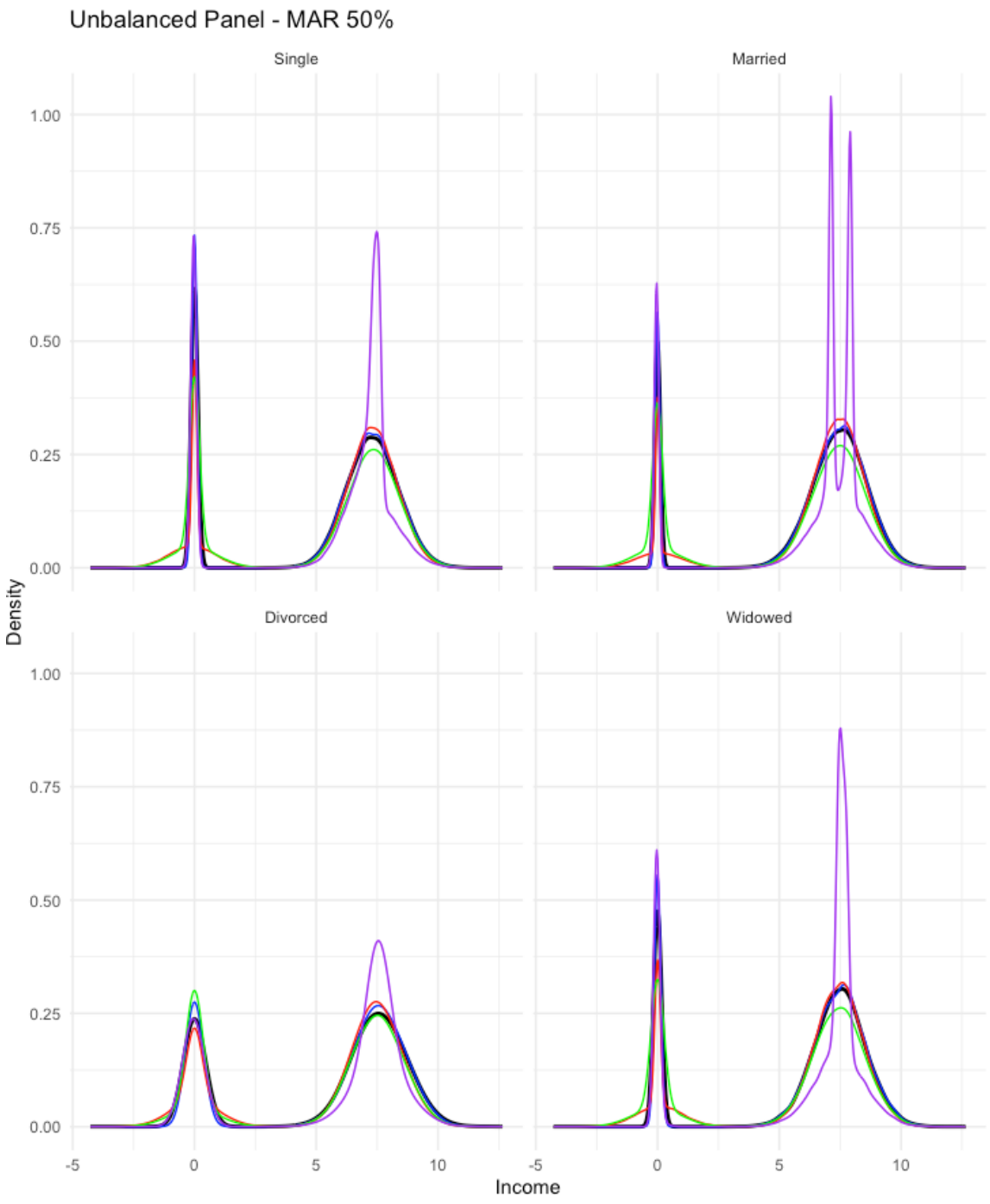


Figure A14: Conditional Distribution of Income for Unbalanced Panel MAR at 50%

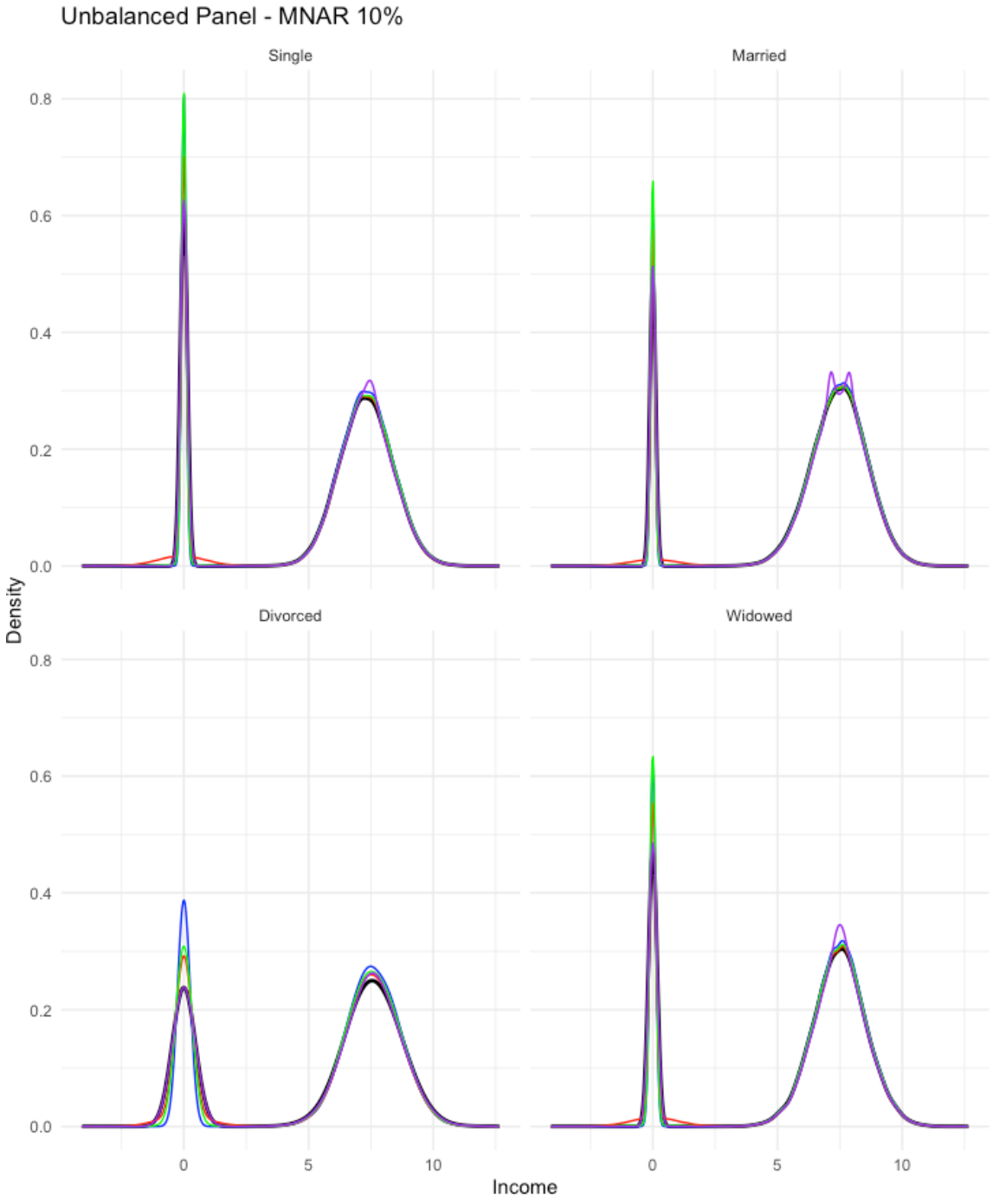


Figure A15: Conditional Distribution of Income for Unbalanced Panel MNAR at 10%

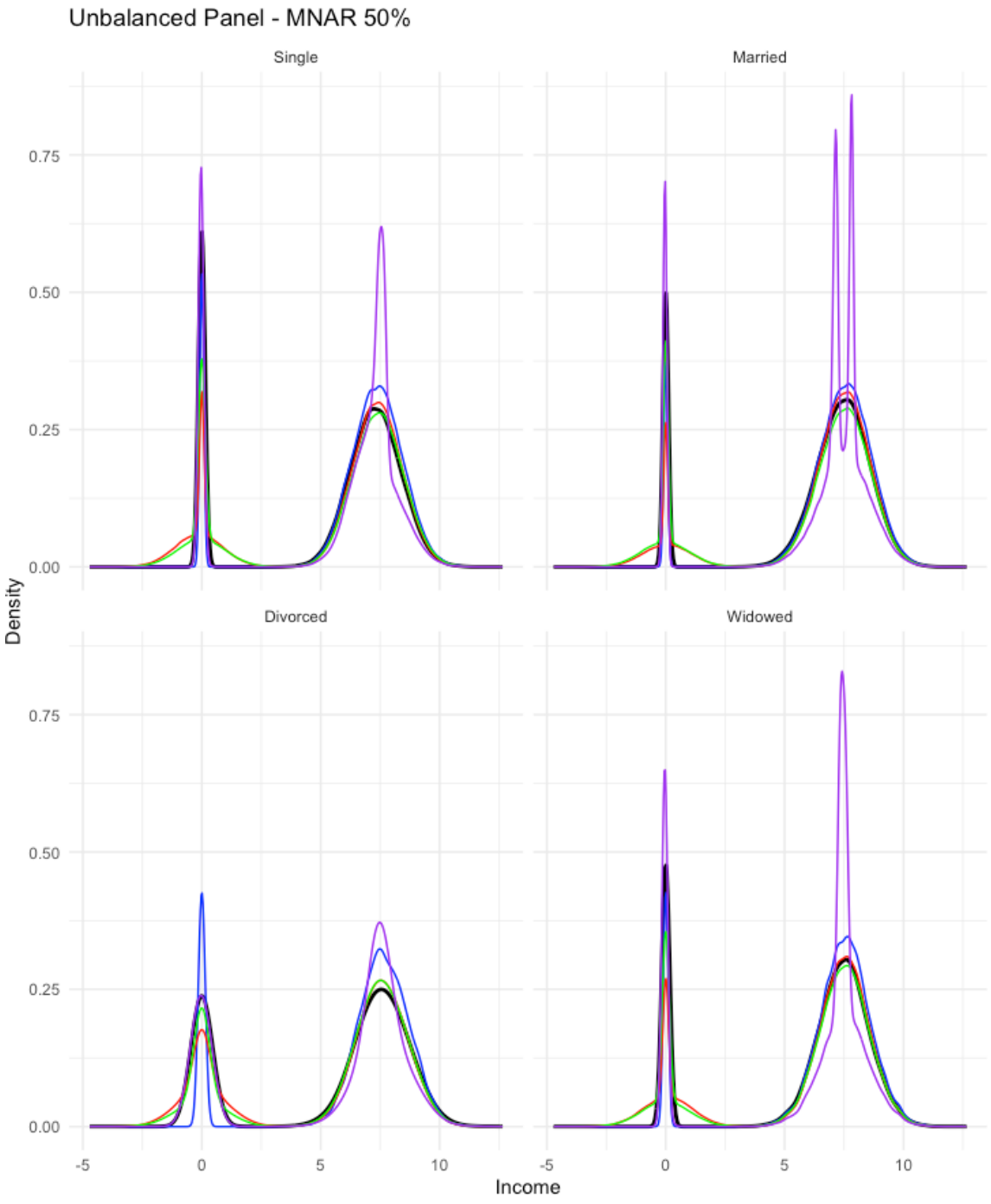


Figure A16: Conditional Distribution of Income for Unbalanced Panel MNAR at 50%

A.3 Coefficients

The names of the columns are as follows: ET1 = Employment Types1, ET2 = Employment Types2, MS2 = Marital Status2, MS3 = Marital Status3, MS4 = Marital Status4, EH1 = Employment Hours1

Table A1: Coefficients for Balanced Panel

Datasets	Age	ET1	ET2	MS2	MS3	MS4	EH1
mice_mcar_bal_50	-0.00098	-7.204	-7.026	0.0890	0.1830	0.1360	0.2793
mice_mcar_bal_30	-0.00089	-7.204	-7.025	0.0745	0.1595	0.1180	0.2792
mice_mcar_bal_10	-0.00089	-7.207	-7.026	0.0704	0.1586	0.1156	0.2795
mice_mar_bal_50	-0.00009	-7.202	-7.045	0.0744	0.1662	0.1135	0.2856
mice_mar_bal_30	-0.00036	-7.202	-7.033	0.0671	0.1548	0.1066	0.2830
mice_mar_bal_10	-0.00056	-7.207	-7.030	0.0691	0.1578	0.1151	0.2816
mice_mnar_bal_50	0.00463	-7.275	-7.127	0.0611	0.1640	0.0972	0.2802
mice_mnar_bal_30	0.00258	-7.240	-7.085	0.0545	0.1541	0.0961	0.2838
mice_mnar_bal_10	0.00034	-7.217	-7.045	0.0649	0.1565	0.1108	0.2819
mitml_mcar_bal_50	-0.02728	-3.581	-3.552	-0.0637	0.0634	0.0245	0.1362
mitml_mcar_bal_30	-0.01734	-5.022	-4.930	-0.0266	0.0990	0.0482	0.1817
mitml_mcar_bal_10	-0.00592	-6.492	-6.331	0.0287	0.0779	0.1093	0.2487
mitml_mar_bal_50	-0.03264	-3.650	-3.453	0.0349	0.0806	0.1166	0.1247
mitml_mar_bal_30	-0.02149	-5.079	-4.845	0.0310	0.0871	0.0899	0.1842
mitml_mar_bal_10	-0.00882	-6.531	-6.292	0.0506	0.1408	0.1058	0.2423
mitml_mnar_bal_50	-0.03436	-3.099	-2.945	0.0920	0.2000	0.2760	0.1929
mitml_mnar_bal_30	-0.02379	-4.645	-4.414	0.0912	0.2109	0.2438	0.2265
mitml_mnar_bal_10	-0.01084	-6.256	-6.044	0.0803	0.1704	0.1672	0.2713
amelia_mcar_bal_50	0.00106	-7.330	-7.232	0.0690	0.1492	0.1130	0.1392
amelia_mcar_bal_30	0.00010	-7.280	-7.146	0.0564	0.1339	0.1040	0.1942
amelia_mcar_bal_10	-0.00047	-7.231	-7.067	0.0742	0.1615	0.1210	0.2505
amelia_mar_bal_50	0.00136	-7.330	-7.233	0.0762	0.1491	0.0955	0.1511
amelia_mar_bal_30	0.00067	-7.271	-7.147	0.0696	0.1509	0.0946	0.2040
amelia_mar_bal_10	-0.00031	-7.225	-7.066	0.0686	0.1517	0.1130	0.2563
amelia_mnar_bal_50	0.00295	-7.392	-7.303	0.0507	0.1583	0.0949	0.1722
amelia_mnar_bal_30	0.00198	-7.306	-7.181	0.0586	0.1606	0.0915	0.2281
amelia_mnar_bal_10	0.00008	-7.235	-7.073	0.0685	0.1625	0.1114	0.2695
lstm_mcar_bal_50	-0.00106	-7.246	-7.104	0.0852	0.1826	0.1060	0.1472
lstm_mcar_bal_30	-0.00159	-7.261	-7.096	0.0434	0.1351	0.0863	0.2008
lstm_mcar_bal_10	-0.00068	-7.231	-7.050	0.0701	0.1630	0.1215	0.2533

Datasets	Age	ET1	ET2	MS2	MS3	MS4	EH1
lstm_mar_bal_50	-0.00174	-7.290	-7.107	0.0829	0.1767	0.1011	0.1556
lstm_mar_bal_30	0.00003	-7.269	-7.079	0.0888	0.1747	0.1080	0.2068
lstm_mar_bal_10	-0.00081	-7.212	-7.034	0.0712	0.1596	0.1164	0.2577
lstm_mnar_bal_50	0.00022	-7.326	-7.175	0.0385	0.1249	0.0524	0.1855
lstm_mnar_bal_30	0.00042	-7.247	-7.110	0.0612	0.1458	0.0911	0.2354
lstm_mnar_bal_10	-0.00060	-7.227	-7.061	0.0684	0.1598	0.1163	0.2725

Table A2: Coefficients for Unbalanced Panel

Datasets	Age	ET1	ET2	MS2	MS3	MS4	EH1
mice_mcar_bal_50	-0.00098	-7.204	-7.026	0.0890	0.1830	0.1360	0.2793
mice_mcar_bal_30	-0.00089	-7.204	-7.025	0.0745	0.1595	0.1180	0.2792
mice_mcar_bal_10	-0.00089	-7.207	-7.026	0.0704	0.1586	0.1156	0.2795
mice_mar_bal_50	-0.00009	-7.202	-7.045	0.0744	0.1662	0.1135	0.2856
mice_mar_bal_30	-0.00036	-7.202	-7.033	0.0671	0.1548	0.1066	0.2830
mice_mar_bal_10	-0.00056	-7.207	-7.030	0.0691	0.1578	0.1151	0.2816
mice_mnar_bal_50	0.00463	-7.275	-7.127	0.0611	0.1640	0.0972	0.2802
mice_mnar_bal_30	0.00258	-7.240	-7.085	0.0545	0.1541	0.0961	0.2838
mice_mnar_bal_10	0.00034	-7.217	-7.045	0.0649	0.1565	0.1108	0.2819
mitml_mcar_bal_50	-0.02728	-3.581	-3.552	-0.0637	0.0634	0.0245	0.1362
mitml_mcar_bal_30	-0.01734	-5.022	-4.930	-0.0266	0.0990	0.0482	0.1817
mitml_mcar_bal_10	-0.00592	-6.492	-6.331	0.0287	0.0779	0.1093	0.2487
mitml_mar_bal_50	-0.03264	-3.650	-3.453	0.0349	0.0806	0.1166	0.1247
mitml_mar_bal_30	-0.02149	-5.079	-4.845	0.0310	0.0871	0.0899	0.1842
mitml_mar_bal_10	-0.00882	-6.531	-6.292	0.0506	0.1408	0.1058	0.2423
mitml_mnar_bal_50	-0.03436	-3.099	-2.945	0.0920	0.2000	0.2760	0.1929
mitml_mnar_bal_30	-0.02379	-4.645	-4.414	0.0912	0.2109	0.2438	0.2265
mitml_mnar_bal_10	-0.01084	-6.256	-6.044	0.0803	0.1704	0.1672	0.2713
amelia_mcar_bal_50	0.00106	-7.330	-7.232	0.0690	0.1492	0.1130	0.1392
amelia_mcar_bal_30	0.00010	-7.280	-7.146	0.0564	0.1339	0.1040	0.1942
amelia_mcar_bal_10	-0.00047	-7.231	-7.067	0.0742	0.1615	0.1210	0.2505
amelia_mar_bal_50	0.00136	-7.330	-7.233	0.0762	0.1491	0.0955	0.1511
amelia_mar_bal_30	0.00067	-7.271	-7.147	0.0696	0.1509	0.0946	0.2040
amelia_mar_bal_10	-0.00031	-7.225	-7.066	0.0686	0.1517	0.1130	0.2563
amelia_mnar_bal_50	0.00295	-7.392	-7.303	0.0507	0.1583	0.0949	0.1722
amelia_mnar_bal_30	0.00198	-7.306	-7.181	0.0586	0.1606	0.0915	0.2281

Datasets	Age	ET1	ET2	MS2	MS3	MS4	EH1
amelia_mnar_bal_10	0.00008	-7.235	-7.073	0.0685	0.1625	0.1114	0.2695
lstm_mcar_bal_50	-0.00106	-7.246	-7.104	0.0852	0.1826	0.1060	0.1472
lstm_mcar_bal_30	-0.00159	-7.261	-7.096	0.0434	0.1351	0.0863	0.2008
lstm_mcar_bal_10	-0.00068	-7.231	-7.050	0.0701	0.1630	0.1215	0.2533
lstm_mar_bal_50	-0.00174	-7.290	-7.107	0.0829	0.1767	0.1011	0.1556
lstm_mar_bal_30	0.00003	-7.269	-7.079	0.0888	0.1747	0.1080	0.2068
lstm_mar_bal_10	-0.00081	-7.212	-7.034	0.0712	0.1596	0.1164	0.2577
lstm_mnar_bal_50	0.00022	-7.326	-7.175	0.0385	0.1249	0.0524	0.1855
lstm_mnar_bal_30	0.00042	-7.247	-7.110	0.0612	0.1458	0.0911	0.2354
lstm_mnar_bal_10	-0.00060	-7.227	-7.061	0.0684	0.1598	0.1163	0.2725

Ehrenwörtliche Erklärung

Name:	Tipu Sultan
Geburtstag:	October 11, 1996
Matrikelnummer:	1599320
Studiengang:	M.Sc. Applied Statistics
Titel der besuchten Veranstaltung:	Master Thesis
Titel der Arbeit:	Imputation of Panel Data

Hiermit erkläre ich, dass die oben genannte Arbeit meine eigene ist und ich sie selbstständig verfasst habe. Des Weiteren bestätige ich folgendes:

- Ich habe diese Arbeit ohne fremde Hilfe verfasst;
- ich habe alle Quellen (Bücher, Zeitschriftenartikel, Internetquellen, ...) in Übereinstimmung mit den Lehrstuhlanforderungen sowohl im Text als auch im Literaturverzeichnis gekennzeichnet;
- alle Daten und Erkenntnisse in der Arbeit wurden weder gefälscht noch verschönert;
- die vorliegende Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form zum Erwerb eines Leistungsnachweises vorgelegt;
- die vorliegende Arbeit wurde bisher nicht veröffentlicht.

Ich verstehe, dass jegliche falsche Angabe in der Arbeit disziplinarische universitäre Konsequenzen nach sich ziehen wird. Ich bin damit einverstanden, dass meine Arbeit elektronisch mit Hilfe einer Software zur Aufdeckung von Plagiaten überprüft wird. Ich bin auch damit einverstanden, dass meine Arbeit für zukünftige Überprüfungen gespeichert wird.

Datum: 06.01.2025

Unterschrift: