BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Generating and validating synthetic unbalanced panel data: evidence from Italian firm microdata

by Marco Langiulli, Alessandro Moro, Mattia Orzincolo and Daniele Piras

September 2026

Number 1056

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Generating and validating synthetic unbalanced panel data: evidence from Italian firm microdata

by Marco Langiulli, Alessandro Moro, Mattia Orzincolo and Daniele Piras

Number 1056 – September 2026

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of Banca d'Italia and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within Banca d'Italia, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of Banca d'Italia

GENERATING AND VALIDATING SYNTHETIC UNBALANCED PANEL DATA: EVIDENCE FROM ITALIAN FIRM MICRODATA

by Marco Langiulli*, Alessandro Moro*, Mattia Orzincolo** and Daniele Piras*

Abstract

This study develops and evaluates a framework for generating synthetic microdata for the Survey of Industrial and Service Firms (INVIND), a large dataset of Italian firms observed between 1993 and 2024. We extend synthetic-data methodologies, traditionally designed for cross-sectional settings, to panel data and compare a sequential modelling approach based on regression and classification trees with a joint modelling strategy employing Gaussian copulas. The procedure differentiates between short and long firm histories to accommodate heterogeneous time spans. We synthesize some key variables, including geographic area, sector, employment, and turnover, and assess the resulting dataset along multiple dimensions: reproduction of univariate and conditional distributions, stability of structural relationships through fixed-effects regressions, and disclosure risk. While both modelling approaches ensure strong fidelity to the original data distributions, re-identification risk is also sizeable. To address this trade-off, we introduce controlled measurement errors, which significantly reduce risk while preserving essential statistical properties.

JEL Classification: C15, C18, C23, C81, C82.

Keywords: synthetic data, re-identification risk, disclosure control, unbalanced panel, INVIND.

DOI: 10.32057/0.QEF.2026.1056

* Banca d'Italia - Directorate General for Economics, Statistics and Research.

** Former trainee at Banca d'Italia - Directorate General for Economics, Statistics and Research.

1 Introduction[1]

In recent years, synthetic data have become a promising tool for statistical disclosure control of confidential data. Several Research Data Centers (RDCs)—including those of national statistical institutes and central banks—are increasingly experimenting the use of synthetic microdata as an alternative or complementary practice for data dissemination. The appeal of synthetic data lies in their ability to mimic the statistical properties of original microdata while avoiding the legal and operational constraints aimed at ensuring confidentiality of granular data.

From the perspective of a RDC, virtual laboratories populated with high-quality synthetic datasets offer a promising approach to reconciling the goal of rich data dissemination with strict confidentiality requirements. In such environments, researchers are able to explore the data structure, conduct exploratory analysis, develop and debug code, and specify empirical models without interacting directly with confidential microdata. Once the analysis is finalized, the corresponding code is submitted to the RDC and executed on the original confidential data by RDC staff. The resulting outputs are released to the researcher only after they have passed standard confidentiality checks. Importantly, disclosure control is concentrated in a single final validation step, rather than being required throughout the research process, as in remote execution procedures. This workflow preserves confidentiality while substantially reducing the burden typically associated with restricted-access environments. In particular, it alleviates constraints such as physical access to on-site laboratories, scheduling bottlenecks, repeated iterations of remote script submissions, long turnaround times for output validation, and the high coordination costs between researchers and RDC staff. By shifting most of the empirical work to a synthetic-data setting, both researchers and RDCs can operate more efficiently.

The use of synthetic data, however, entails a fundamental tension. On the one hand, analytical validity requires synthetic microdata to reproduce key features of the original dataset, including marginal distributions, cross-variable dependencies, and temporal dynamics. On the other hand, excessive similarity may increase the risk that synthetic records lie too close to real ones, undermining confidentiality protections. This intrinsic similarity–risk trade-off is the central methodological challenge in synthetic-data generation: models that more accurately capture the joint distribution of the original dataset tend to raise disclosure risk, whereas methods that inject additional variability offer stronger privacy at the expense of statistical similarity.

A further challenge arises from the fact that the synthetic-data literature and

[1]The authors wish to thank Laura Bartiloro, Marco Bottone, Raffaela Giordano, Michele Loberto, and Alfonso Rosolia for their insightful comments and suggestions. The opinions expressed in the paper are those of the authors and do not necessarily reflect the views of Banca d'Italia. Contact emails: Marco.Langiulli@bancaditalia.it, Alessandro.Moro2@bancaditalia.it and Daniele.Piras@bancaditalia.it

accompanying tools were primarily developed for cross-sectional data. Extending these methods to panel data is non-trivial: a credible synthetic panel must preserve time dependence, firm-specific heterogeneity, and the unbalanced nature of the underlying microdata.

Against this background, our work contributes to the growing effort to build high-quality synthetic data by focusing on a large dataset of Italian firms drawn from the Survey of Industrial and Service Firms (INVIND). INVIND is a repeated cross-sectional survey in which a fraction of firms is followed over multiple waves. As a result, although the survey is not designed as a balanced panel, it naturally gives rise to a rich unbalanced panel structure that can be exploited for longitudinal analysis. In this paper, we first develop and evaluate a synthetic version of INVIND, focusing on a relatively large set of key quantitative and qualitative variables—around 20 in total, including potentially indirect identifiers—substantially more than is typically considered in comparable studies. Second, we adapt and extend synthetic-data methodologies originally designed for cross-sectional databases to an unbalanced panel setting with heterogeneous firm-level histories. Third, we compare the two most widely used strategies for synthetic-data generation—the sequential and the joint modelling approaches—in terms of fitting accuracy and disclosure risk. Fourth, we show how the introduction of controlled measurement errors can substantially improve the accuracy–privacy trade-off, reducing disclosure risk while preserving the essential statistical properties of the synthetic dataset.

The literature on synthetic data is rapidly expanding but remains largely practice-driven, with limited theoretical foundations. The concept of synthetic data is not new and dates back more than three decades, having been first introduced by Rubin (1987, 1993). He suggested using multiple imputation techniques to generate synthetic microdata that could be released without disclosing confidential information about any individual unit.

As formalized in Drechsler (2011), a central aspect concerns the modelling strategy adopted to generate, from a set of real observations Y_{obs}, synthetic values Y_{new} drawn from the posterior predictive distribution $P(Y_{\text{new}} \mid Y_{\text{obs}})$. Within this framework, two main approaches can be distinguished. The *joint modelling approach* aims to directly estimate the full joint distribution $P(Y)$, often relying on Gaussian copulas or their extensions (Patki et al., 2016; Gambs et al., 2021; Meyer et al., 2021; Hofert et al., 2026). The *sequential modelling approach*, instead, orders variables according to a predetermined criterion (e.g., from more exogenous to more endogenous) and estimates a sequence of conditional distributions $P(Y^{(j)} \mid Y^{(j-1)}, \ldots, Y^{(1)})$, where the distribution of the jth variable $Y^{(j)}$ is conditioned on all preceding variables $Y^{(1)}, \ldots, Y^{(j-1)}$ (Reiter, 2005; Kinney et al., 2011; Nowok et al., 2016).

Focusing on firm-level data, one distinctive feature is the substantial heterogeneity they typically exhibit. In particular, many firms display unique characteristics that make them readily identifiable, for instance due to the highly skewed distribution of key - and sometimes publicly available - variables like turnover

or employment. The use of synthetic micro-data as an approach to release secure firm level data remains relatively limited (see, e.g. Drechsler, 2012) and is even more uncommon for panel data on firms. The longitudinal structure poses additional challenges, as it entails higher disclosure risks due to the increased potential for re-identification of individual units over time. Kinney et al. (2011) provided the first application of synthetic data to firm-level longitudinal data in the United States, through the release of the Synthetic Longitudinal Business Database (SynLBD), which covers more than 20 millions firms, albeit with a very limited set of variables. In a subsequent work, Kinney et al. (2014) improve their original framework introducing greater flexibility in the synthesis methods using Classification and Regression Trees (CART) models and including additional years. Jahangir Alam et al. (2020) replicate the SynLBD methodology by using the same code on firm-level longitudinal data from Germany and Canada to assess the extent to which this approach can serve as a cost-effective solution for synthetic data generation. Drawing on this line of work, we aim to generate a synthetic dataset testing alternative methods and including a richer set of qualitative and quantitative variables than those considered in previous applications.

In addition, our paper relates to the literature on the use of controlled noise as a tool to reduce the risk of re-identification (Brand, 2002; Kim and Winkler, 2003; Domingo-Ferrer et al., 2004; Shlomo, 2010). Adding noise or measurement errors is a well–established disclosure–control technique, and it can be integrated into synthetic–data workflows to attenuate excessive similarity to the original micro–data.

The rest of the paper is organized as follows. Section 2 describes the dataset used in the empirical application. Section 3 outlines the different methodologies employed to generate the synthetic data. Section 4 evaluates the quality of the data generated using the different methodologies and presents the corresponding metrics used to assess disclosure risk. Finally, Section 5 offers some concluding remarks.

2 Data

The *Survey of Industrial and Service Firms* (INVIND), annually conducted by Banca d'Italia, covers non-financial firms operating in industry and private services and collects detailed information on production, employment, investment, and expectations.

The survey is designed as a repeated cross-sectional survey. While each annual wave provides a cross-sectional snapshot of the firm population, survey waves partially overlap, so that a subset of firms is observed for multiple years. Although the survey is not conceived as a balanced longitudinal dataset, this overlap naturally gives rise to a large unbalanced panel that researchers often exploit for longitudinal analysis. The dataset used in this study spans the period from 1993 to 2024 and includes 13,837 distinct firms, with substantial heterogeneity in the

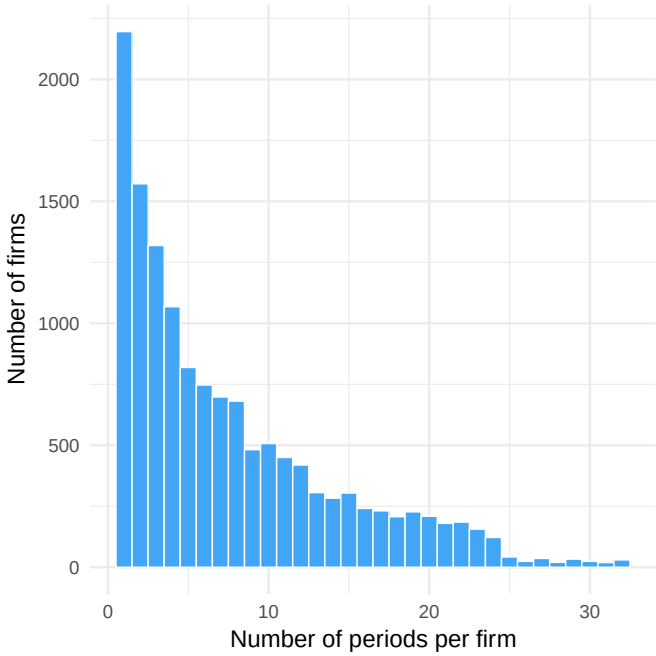

Figure 1: Distribution of firms by number of observed time periods.

number of observed periods T_i across firms (Fig. 1).

In this paper, we work with a subset of variables extracted from INVIND. This choice reflects the *proof-of- concept* nature of our contribution: our aim is not to replicate the full structure of the survey, but rather to develop and validate a methodology for synthetic-data generation tailored to unbalanced firm-level panels. The selected variables span the key dimensions of firm activity and expectations, and they are sufficiently rich to stress-test the methodological components of our pipeline. Moreover, this reduced set includes some of the most challenging variables for synthetization, as they may act as potential indirect identifiers of firms participating in the survey.[2]

Each firm i is identified by the tax code ι_i, which will be replaced by a pseudo-identifier in the disseminated synthetic dataset. For each firm we also have the following time-invariant characteristics: the founding year F_i; the geographic macro-area A_i of residence (taking four possible values, i.e. North-West, North-East, Centre, and South); and the industry classification S_i based on the Italian NACE system (aggregated in 11 sectors, matching the survey's stratification structure).

The time-varying variables used in the analysis cover firm size, activity, and short-term dynamics: employment is denoted by $L_{i,t}$; hours worked by $H_{i,t}$; turnover by $R_{i,t}$; export turnover by $X_{i,t}$; investment by $I_{i,t}$; and the percent-

[2]Additional, less sensitive INVIND variables may be added to the synthetic dataset using nearest-neighbor matching, following the same logic described in Subsection 4.3.

age change in the price set by firm i between $t-1$ and t by $\Delta P_{i,t}$. At each time t, INVIND also collects one-period lags for several of these variables: in particular, we observe $L_{i,t-1}$, $H_{i,t-1}$, $R_{i,t-1}$, $X_{i,t-1}$, and $I_{i,t-1}$.[3] In addition, the survey includes forward-looking questions, which elicit expectations formed at time t about next-period outcomes, for turnover $R^e_{i,t+1}$, export turnover $X^e_{i,t+1}$, investment $I^e_{i,t+1}$, and price percentage variation $\Delta P^e_{i,t+1}$. These expectation variables play an important role in the modelling of cross-sectional heterogeneity and are synthesised together with the corresponding level and lagged quantities. Table 1 below lists all the variables considered in our exercise.

Table 1: List of INVIND variables used in the analysis.

Description	**Symbol**
Firm identifier	ι_i
Calendar year	t
Geographic area (4 levels)	A_i
Industry sector (11 levels)	S_i
Founding year	F_i
Employment (cur., lag)	$L_{i,t}$, $L_{i,t-1}$
Hours worked (cur., lag)	$H_{i,t}$, $H_{i,t-1}$
Turnover (cur., lag, exp.) in k€	$R_{i,t}$, $R_{i,t-1}$, $R^e_{i,t+1}$
Export turnover (cur., lag, exp.) in k€	$X_{i,t}$, $X_{i,t-1}$, $X^e_{i,t+1}$
Investment (cur., lag, exp.) in k€	$I_{i,t}$, $I_{i,t-1}$, $I^e_{i,t+1}$
Price percentage variation (cur., exp.)	$\Delta P_{i,t}$, $\Delta P^e_{i,t+1}$

3 Methodology

This section outlines the methodological framework underlying the construction of the synthetic dataset. Given the heterogeneous temporal structure of the data, we distinguish between cross-sectional and panel components, allowing the synthesis to flexibly account for their distinct informational content. Within this framework, we develop and compare three alternative data-generating strategies: (i) a sequential fully conditional specification based on regression and classification trees; (ii) a joint multivariate model based on Gaussian copulas; and (iii) a naïve generator based on lognormal and normal approximations, used as a benchmark.The framework is further complemented by procedures aimed at balancing analytical validity and disclosure protection, including the introduction of controlled measurement error and post-processing steps designed to restore internal coherence.

[3]For panel firms, these lags correspond to the contemporaneous values reported at time $t-1$.

3.1 Cross-sectional and panel partition

We begin by partitioning the dataset into a cross-sectional component $\mathcal{C}$ and a panel component $\mathcal{P}$. The purpose of this partition is methodological: the two subsets are characterized by markedly different time structures and therefore call for synthesis strategies that are better tailored to their respective features. Synthetic data are generated separately for the two components using methods appropriate to each case and are subsequently recombined into a single, coherent synthetic dataset.

The classification is based on a threshold value θ for the number of survey appearances. Let T_i denote the number of years in which firm i appears in the survey. Firms observed for fewer than θ periods are treated as cross-sectional, whereas firms with $T_i \geq \theta$ are treated as panel units:

$$i \in \mathcal{C} \iff T_i < \theta, \qquad i \in \mathcal{P} \iff T_i \geq \theta.$$

Although the natural cutoff for distinguishing between cross-sectional and longitudinal observations would be $\theta = 2$, generating credible synthetic trajectories requires richer time information. After a robustness analysis over $\theta \in \{2, \ldots, 20\}$, balancing the need to preserve the panel structure of the original data against the requirement of sufficiently long time spans, we set $\theta = 10$. With this threshold, 31% of firms (accounting for 65% of total observations) are in the panel component. Across the three methodologies used to synthesize the dataset, outlined below, cross-sectional and panel components are treated differently.

3.2 Sequential modelling approach via trees

The sequential method is based on a fully conditional specification and is implemented using the synthpop package in R (Nowok et al., 2016). Following Reiter (2005), each conditional distribution is estimated using flexible nonparametric methods such as classification and regression trees (CART; see Breiman et al., 1984). Given an ordering of the vector $\mathbf{Y}_{i,t}$ of variables, at step j we estimate the following conditional model:

$$Y_{i,t}^{(j)} \;=\; f^{(j)}\big(Y_{i,t}^{(1)}, \ldots, Y_{i,t}^{(j-1)}, \mathbf{Z}_{i,t}\big) \;+\; \varepsilon_{i,t}^{(j)}, \tag{1}$$

where $f^{(j)}(\cdot)$ denotes the predictive function estimated using CART for variable j, and $\varepsilon_{i,t}^{(j)}$ is the model residual. The vector $\mathbf{Z}_{i,t}$ may include variables that are not synthesised but used in prediction (e.g., a time trend for panel data).

Once $\widehat{f}(.)^{(j)}$ is estimated, synthetic values are then generated recursively from the conditional distribution:

$$\tilde{Y}_{i,t}^{(j)} \;\sim\; \widehat{F}^{(j)}\Big(\cdot \,\Big|\, \tilde{Y}_{i,t}^{(1)}, \ldots, \tilde{Y}_{i,t}^{(j-1)}, \mathbf{Z}_{i,t}\Big), \tag{2}$$

where $\widehat{F}^{(j)}(\cdot \mid \cdot)$ is the estimated conditional distribution used for sampling (i.e., the stochastic completion around $\widehat{f}^{(j)}$). In this setting, the stochastic component

$\varepsilon_{i,t}^{(j)}$ does not represent an additive error term drawn from a parametric distribution. Rather, it captures the randomness arising from the resampling mechanism used in CART, whereby synthetic values are generated by drawing from the empirical distribution of observed outcomes within each terminal node of the fitted tree.

We adopt an ordering of variables that follows a progression from more exogenous to more endogenous outcomes. Specifically, we begin with time-invariant individual characteristics, followed by lagged variables, contemporaneous realizations, and finally expectations about future outcomes formed conditional on information available at time t. This ordering reflects a natural causal and informational sequence and is consistent with the fully conditional specification framework. Importantly, our results are robust to alternative variable orderings, yielding no material differences in the main distributional features of the synthetic data. The method is implemented differently for the cross-sectional and panel components, as detailed below.

Cross–section. We estimate fully conditional models separately for each calendar year t, with no conditioning variables $\mathbf{Z}_{i,t}$:

$$\mathbf{Z}_{i,t}^{\mathcal{C}} \equiv 0, \qquad (i,t) \in \mathcal{C}_t.$$

The order of variables in the cross–section prioritizes time-invariant variables, lags, contemporaneous levels, then expectations:

$$\begin{aligned}\text{order}_{\mathcal{C}} = \big(&A_i,\ S_i,\ F_i;\ L_{i,t-1},\ H_{i,t-1},\ R_{i,t-1},\ X_{i,t-1},\ I_{i,t-1},\\ &L_{i,t},\ H_{i,t},\ R_{i,t},\ X_{i,t},\ I_{i,t},\ \Delta P_{i,t},\\ &L^e_{i,t+1},\ R^e_{i,t+1},\ X^e_{i,t+1},\ I^e_{i,t+1},\ \Delta P^e_{i,t+1}\big),\end{aligned}$$

The first element A_i (geographic area, see Table 1) is generated by re–sampling from its empirical distribution in year t (not modeled). For each subsequent step j in the order, we fit

$$Y_{i,t}^{(j)} \;=\; f_{\mathcal{C},t}^{(j)}\big(Y_{i,t}^{(1)}, \ldots, Y_{i,t}^{(j-1)}\big) \;+\; \varepsilon_{i,t}^{(j)}, \tag{3}$$

and draw

$$\tilde{Y}_{i,t}^{(j)} \;\sim\; \widehat{F}_{\mathcal{C},t}^{(j)}\Big(\cdot \,\Big|\, \tilde{Y}_{i,t}^{(1)}, \ldots, \tilde{Y}_{i,t}^{(j-1)}\Big), \tag{4}$$

where $f_{\mathcal{C},t}^{(j)}(\cdot)$ and $\widehat{F}_{\mathcal{C},t}^{(j)}(\cdot \mid \cdot)$ are year-specific.

Panel. For observations in $\mathcal{P}$, we pool all years and explicitly condition on time and firm identifiers:

$$\mathbf{Z}_{i,t}^{\mathcal{P}} = (t, \iota_i), \qquad (i,t) \in \mathcal{P}.$$

The firm identifier ι_i is treated as a categorical variable and acts as a firm fixed effect, capturing all time-invariant firm-specific characteristics. In contrast to

the cross–sectional component, geographic area and sector of activity are not synthesized in the panel. These dimensions are instead absorbed by the fixed effects structure induced by conditioning on the firm identifier itself.[4]

The order of variables in the panel follows a progression from lagged outcomes to contemporaneous levels and, finally, expectations about future outcomes:

$$\begin{aligned}\text{order}_{\mathcal{P}} = \big(&L_{i,t-1},\, H_{i,t-1},\, R_{i,t-1},\, X_{i,t-1},\, I_{i,t-1},\\ &L_{i,t},\, H_{i,t},\, R_{i,t},\, X_{i,t},\, I_{i,t},\, \Delta P_{i,t},\\ &L^{e}_{i,t+1},\, R^{e}_{i,t+1},\, X^{e}_{i,t+1},\, I^{e}_{i,t+1},\, \Delta P^{e}_{i,t+1}\big).\end{aligned}$$

At step j, the conditional model and the synthetic draw are given by

$$Y^{(j)}_{i,t} \;=\; f^{(j)}_{\mathcal{P}}\big(Y^{(1)}_{i,t}, \ldots, Y^{(j-1)}_{i,t}, \mathbf{Z}^{\mathcal{P}}_{i,t}\big) \;+\; \varepsilon^{(j)}_{i,t}, \tag{5}$$

$$\tilde{Y}^{(j)}_{i,t} \;\sim\; \widehat{F}^{(j)}_{\mathcal{P}}\Big(\cdot \,\Big|\, \tilde{Y}^{(1)}_{i,t}, \ldots, \tilde{Y}^{(j-1)}_{i,t}, \mathbf{Z}^{\mathcal{P}}_{i,t}\Big), \tag{6}$$

with $f^{(j)}_{\mathcal{P}}(\cdot)$ denoting the tree-based predictor and $\widehat{F}^{(j)}_{\mathcal{P}}(\cdot \mid \cdot)$ the associated conditional distribution used for sampling.

3.3 Joint modelling via Gaussian copulas

For the copula-based method, we model the joint distribution of the vector of variables to be synthesized, $\mathbf{Y}_{i,t}$. By Sklar's theorem (Sklar, 1959), copulas allow the separation of the dependence structure from the marginal distributions. Accordingly, each component of $\mathbf{Y}_{i,t}$ is transformed into a uniform random variable on the unit interval using its empirical cumulative distribution (CDF) function. The dependence among the resulting uniform variables is then modeled through a multivariate Gaussian copula. The entire procedure is applied within pre-defined strata, which are identified by different sets of variables for the cross-sectional and panel components, as detailed below.

In detail, the synthesis method proceeds as follows. Within each stratum c, pseudo-observations are obtained by applying the empirical CDF of each variable to the observed data:

$$U^{(j)}_{i,t} \;=\; \hat{F}_{c,j}\big(Y^{(j)}_{i,t}\big), \qquad j = 1, \ldots, D, \tag{7}$$

where $\hat{F}_{c,j}$ denotes the empirical CDF of the j-th variable computed using only the observations in stratum c. We then model the joint distribution of the uniform vector $\mathbf{U}_{i,t}$ conditional on the stratum identifier c. The vector $\mathbf{Z}_{i,t}$ denotes the variables that define the stratum c and are not synthesised in the panel; conditioning on $\mathbf{Z}_{i,t}$ therefore corresponds to estimating a copula and the associated marginal distributions within the corresponding cell $c(\mathbf{Z}_{i,t})$.

[4]Because panel firms are economically important units, synthesis is performed conditional on observed firm identifiers while preserving their original geographic location and sector, ensuring realistic firm trajectories and a credible synthetic dataset.

Conditional on a stratum c, the dependence structure is a Gaussian copula

$$C_{\Sigma_c}\left(u_{i,t}^{(1)},\ldots,u_{i,t}^{(D)}\right)=\Phi_{\Sigma_c}\Big(\Phi^{-1}\left(u_{i,t}^{(1)}\right),\ldots,\Phi^{-1}\left(u_{i,t}^{(D)}\right)\Big), \tag{8}$$

where Σ_c is estimated by maximum likelihood from $\{\Phi^{-1}(U_{i,t}^{(j)}) : \forall j,(i,t)\in c\}$. Synthetic uniforms are drawn and mapped back to data space via empirical inversion:

$$\tilde{\mathbf{U}}_{i,t}\sim C_{\widehat{\Sigma}_c},\qquad \tilde{Y}_{i,t}^{(j)}=\hat{F}_{c,j}^{-1}\big(\tilde{U}_{i,t}^{(j)}\big), \tag{9}$$

with $\hat{F}_{c,j}^{-1}$ the empirical quantile function of coordinate j computed within c. The Gaussian-copula based synthetic data are generated in R using the copula package (Hofert et al., 2026)[5]. The method is implemented differently for the cross-sectional and panel components, as detailed below.

Cross–section. In the cross-section we estimate one copula per stratum $c(A_i,S_i,t)$ —identified by Geographic Area, Industry Sector and year—, so that both dependence and marginals are area-, sector-, and year-specific. Given c, we compute the elements of $\mathbf{U}_{i,t}$ using (7), estimate $\widehat{\Sigma}_c$ via (8), draw $\tilde{\mathbf{U}}_{i,t}$ as in (9), and invert using the cell-specific $\hat{F}_{c,j}^{-1}$.

Panel. In the panel component we adopt broader strata $c(A_i,S_i)$, pooling information across years to obtain more stable estimates of the dependence matrix Σ_c. In contrast to the cross-sectional setting, the calendar year t is included among the variables to be synthesized and therefore enters the copula as an additional coordinate. This allows the Gaussian copula to capture how the joint distribution of $\mathbf{Z}_{i,t}$ co-moves with time within each area–sector stratum.

For each firm i and for each observed year $t\in T_i$, we first draw a synthetic uniform vector $\tilde{\mathbf{U}}_{i,t}$ from the copula associated with the firm's stratum. The coordinate corresponding to the year variable is then mapped through the empirical quantile function of the firm's observed years, producing a synthetic year $\tilde{t}_{i,t}=\hat{F}_{i,\text{year}}^{-1}(\tilde{U}_{i,t}^{(\text{year})})$. The values $\tilde{t}_{i,t}$ act as a latent time index and are used to order the T_i synthetic observations of the firm. Once this ordering is determined, the synthetic years are discarded and replaced by the firm's actual years, sorted in increasing order. All remaining coordinates of $\tilde{\mathbf{U}}_{i,t}$ are then mapped back to the data space using firm-specific empirical quantile functions: $\tilde{Y}_{i,t}^{(j)}=\hat{F}_{i,j}^{-1}\big(\tilde{U}_{i,t}^{(j)}\big)$.

3.4 Naïve generative model

The naïve model synthesises each component of $\mathbf{Y}_{i,t}$ independently, without modelling cross–variable or temporal dependence. This method serves as a benchmark for assessing the extent to which the sequential and joint modelling approaches improve upon it.

[5]For additional details, see also Hofert and Mächler, 2011, Kojadinovic and Yan, 2010, and Yan, 2007.

For strictly positive variables (e.g. employment, turnover, investment and their lags or expectations), synthetic values are generated from a lognormal distribution,

$$\tilde{Y}_{i,t}^{(j)} = \exp\Big(\mu_c + \sigma_c\, \varepsilon_{i,t}^{(j)}\Big), \qquad \varepsilon_{i,t}^{(j)} \sim \mathcal{N}(0,1), \tag{10}$$

where (μ_c, σ_c) are the mean and standard deviation of $\log(Y_{i,t}^{(j)})$ computed within stratum c. For variables such as $\Delta P_{i,t}$ and $\Delta P_{i,t+1}^{e}$, which may take positive or negative values, we use a normal specification,

$$\tilde{Y}_{i,t}^{(j)} = \mu_c + \sigma_c\, \varepsilon_{i,t}^{(j)}, \qquad \varepsilon_{i,t}^{(j)} \sim \mathcal{N}(0,1), \tag{11}$$

with (μ_c, σ_c) defined as the empirical mean and standard deviation of $Y_{i,t}^{(j)}$ in stratum c. This method is applied separately to the cross-sectional and panel components, as follows.

Cross–section. Cross–sectional strata are defined as $c(A_i, S_i, t)$, so that the location and scale parameters reflect area–sector-year heterogeneity.

Panel. Panel strata are defined by the firm identifier $c(\iota_i)$, and all variables in $Y_{i,t}$ are synthesised independently over $t \in T_i$ using firm–specific empirical moments.

3.5 Measurement errors

To mitigate disclosure risk, we add controlled measurement error after the generation of the synthetic variables (Brand, 2002; Kim and Winkler, 2003; Domingo-Ferrer et al., 2004; Shlomo, 2010). Let $\tilde{\mathbf{Y}}_{i,t}$ denote the value produced by the synthesis model (sequential or copula). The noisy version is denoted by $\tilde{\mathbf{Y}}_{i,t}^{*}$.

The noise is applied independently across (i,t) and is based on the empirical variability of the synthetic values themselves. Specifically, for strictly positive variables we use multiplicative lognormal noise:

$$\tilde{Y}_{i,t}^{*(j)} = \tilde{Y}_{i,t}^{(j)}\, \exp\big(\psi\, \sigma_{\log(\tilde{Y}^{(j)})}\, \eta_{i,t}^{(j)}\big), \qquad \eta_{i,t}^{(j)} \sim \mathcal{N}(0,1), \tag{12}$$

where $\sigma_{\log(\tilde{Y}^{(j)})}$ is the empirical standard deviation of $\log(\tilde{Y}^{(j)})$. The scalar $\psi > 0$ is a global tuning parameter. In our implementation, ψ is calibrated to 0.20 so that the standard deviation of $\log(\tilde{Y}_{i,t}^{*(j)})$ is approximately 2% larger than the standard deviation of $\log(\tilde{Y}_{i,t}^{(j)})$. As discussed in Subsection 4.3, introducing a measurement error of this magnitude allows the sequential tree and Gaussian copula methods to achieve a level of protection against re-identification comparable to that provided by the naïve approach.

For variables which may take positive or negative values (such as $\Delta P_{i,t}$ and $\Delta P_{i,t+1}^{e}$), we add Gaussian noise:

$$\tilde{Y}_{i,t}^{*(j)} = \tilde{Y}_{i,t}^{(j)} + \psi\, \sigma_{\tilde{Y}^{(j)}}\, \eta_{i,t}^{(j)}, \qquad \eta_{i,t}^{(j)} \sim \mathcal{N}(0,1), \tag{13}$$

where $\sigma_{\tilde{Y}}$ is the empirical standard deviation of $\tilde{Y}^{(j)}$. The same calibration principle applies: $\psi = 0.20$ is chosen such that the noisy series has a standard deviation roughly 2% larger than the unperturbed synthetic series.

3.6 Post–processing

A post–processing step restores the structural coherence of the synthetic panel. In the copula–based and naïve methods, the founding year F_i is reconciled to a firm–specific constant, using the median of its synthetic draws, in order to ensure time–invariance and plausibility.[6] In all methods, lagged variables are reconstructed to ensure temporal consistency within the panel structure. For each panel unit, at time t, we generate both the contemporaneous value $Y_{i,t}^{(j)}$ and its lagged counterpart $Y_{i,t-1}^{(j)}$. Since records for the same unit are generated separately, this latter value will not match with the corresponding contemporaneous value $Y_{i,t}^{(j)}$ in the previous time period. In order to restore temporal coherence, all lagged values are rebuilt ex-post setting them equal to the synthetic contemporaneous value in the previous year.[7]

Integer-valued variables, such as employment $L_{i,t}$ and hours worked $H_{i,t}$ (and their lagged or forward-looking counterparts), are rounded to the nearest integer to preserve their natural support.

As a final remark, it is important to stress that both synthetic-data generation approaches naturally preserve the missing-data structure observed in the original dataset. In the sequential tree-based approach, missing values are treated as part of the conditional distribution and are reproduced during synthesis. In the copula-based approach, dependence is estimated on complete cases, while the quantile-inversion step propagates missing values without enforcing imputation. In both cases, the resulting synthetic data replicate the extent and structure of missingness present in the original data.

4 Analysis

There is no single measure of synthetic data quality, as different applications may require different statistical properties to be preserved. Since the intended use of the data by researchers is not known ex ante, the objective is to provide a comparative assessment of alternative methods across multiple relevant dimensions. To this end, we assess the synthetic datasets produced by the three methods—sequential trees, Gaussian copulas, and the naïve generator—along three dimensions: (i) the preservation of marginal and conditional distributions, (ii) the stability of multivariate relationships under randomized regression specifications, and (iii) disclosure risk measured using nearest–neighbour metrics.

[6] This also ensures that the synthetic founding year precedes the years of observation.

[7] It is worth stressing that this reconstruction procedure applies only to panel units. For cross-sectional units we keep the synthetic version of lag values.

4.1 Comparison of distributions

Comparing synthetic and original distributions serves a dual purpose: it provides an absolute assessment of data quality while also enabling a relative evaluation of alternative methods. Indeed, the original distribution provides a benchmark against which the synthetic counterparts can be directly evaluated, making it possible to assess not only how methods perform relative to each other, but also how closely they reproduce the true underlying distributions.

Univariate densities for log–turnover, log–employment, log–exports, log–hours worked, log–investment, and price variation in Fig. 2 show that the sequential tree and copula-based generators perform very well in absolute terms. Across all variables, their synthetic densities closely track—and in most cases almost entirely overlap with—the original distributions, indicating a high degree of fidelity to the empirical data. Adding noise mechanically increases dispersion for both sequential trees and copulas, widening the distributions and producing smoother, flatter tails without altering their fundamental shape.

In marked contrast, the naïve generator fails to reproduce the asymmetry and the dispersion of the true empirical distributions. Because it relies on stratum-specific means and variances, the naïve model generates approximately lognormal shapes that are too regular and too symmetric to overlap perfectly with the original data. Moreover, since the sample variance is particularly sensitive to outliers, the naïve method may also overestimate dispersion, leading to synthetic distributions that are spuriously more spread out than their original counterparts. These distortions are especially evident in variables such as employment, where the synthetic distribution based on the naïve method fails to capture the asymmetry of the original one, and in variables such as exports, investment, and price variation, where the naïve generator produces distributions that are more dispersed than the original data.

The sequential tree and copula-based generators are also able to track conditional distributions reasonably well. As an example, Fig. 3 shows the distribution of the log-employment across geographical areas (North–West, North–East, Centre, South). Sector-level comparisons across the eleven NACE classes reveal the same pattern, as depicted in Fig. 4 for the log-employment.

Finally, boxplots constructed over adjacent six-year periods confirm that trees and copulas also preserve the temporal evolution of dispersion in employment, turnover, exports, hours and investment and they always outperform the naïve method (Fig. 5). Adding measurement error increases variability uniformly but does not distort temporal patterns nor the relative ranking of the three generative models[8].

[8]The downward shift observed in the boxplots during the second six-year period reflects changes in sample composition driven by the expansion of survey coverage. In particular, the inclusion of smaller firms (with 10 to 49 employees in 2001) brought in units characterized by systematically lower values of key variables.

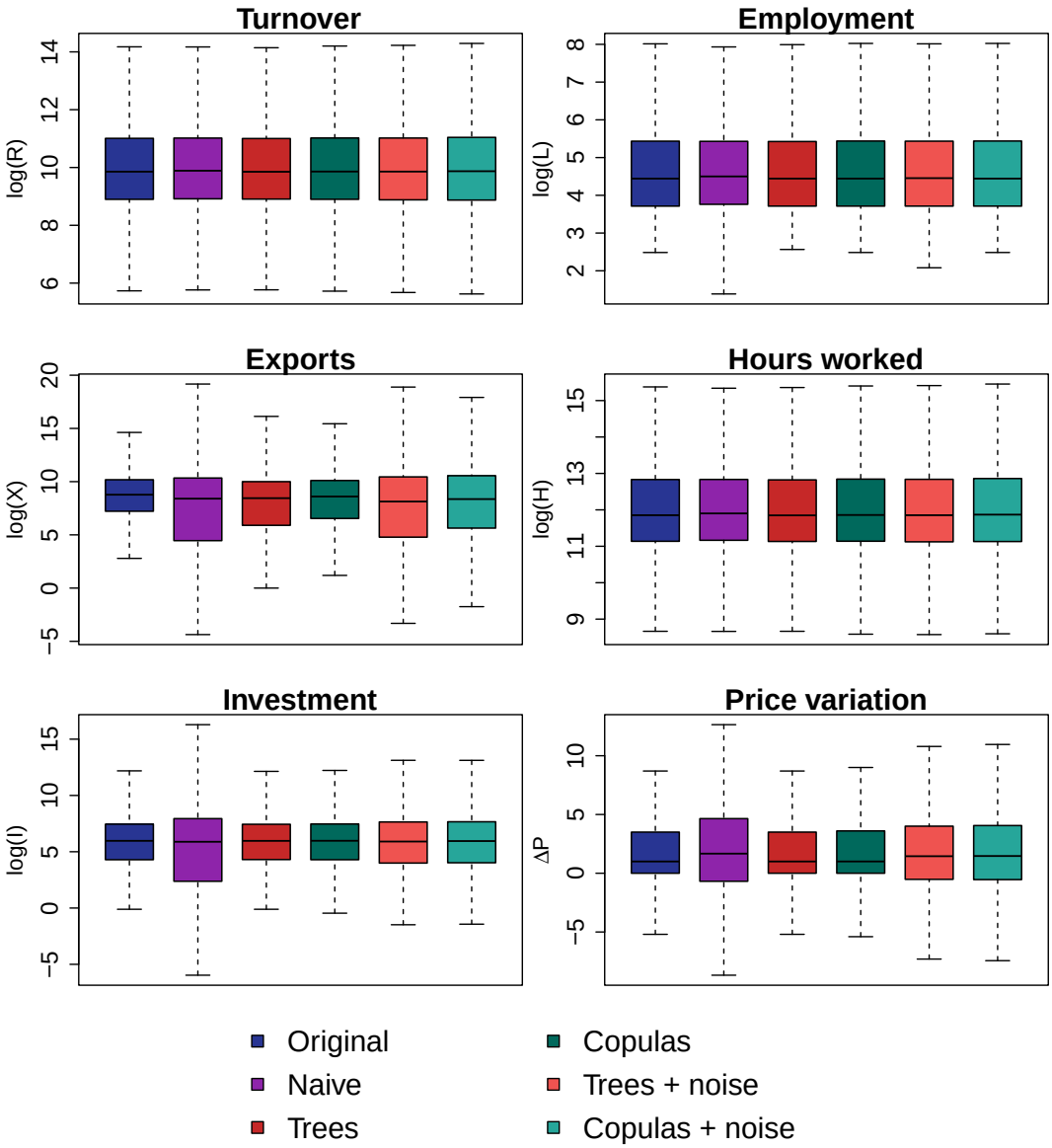

Figure 2: Comparison between original and synthetic univariate distributions.

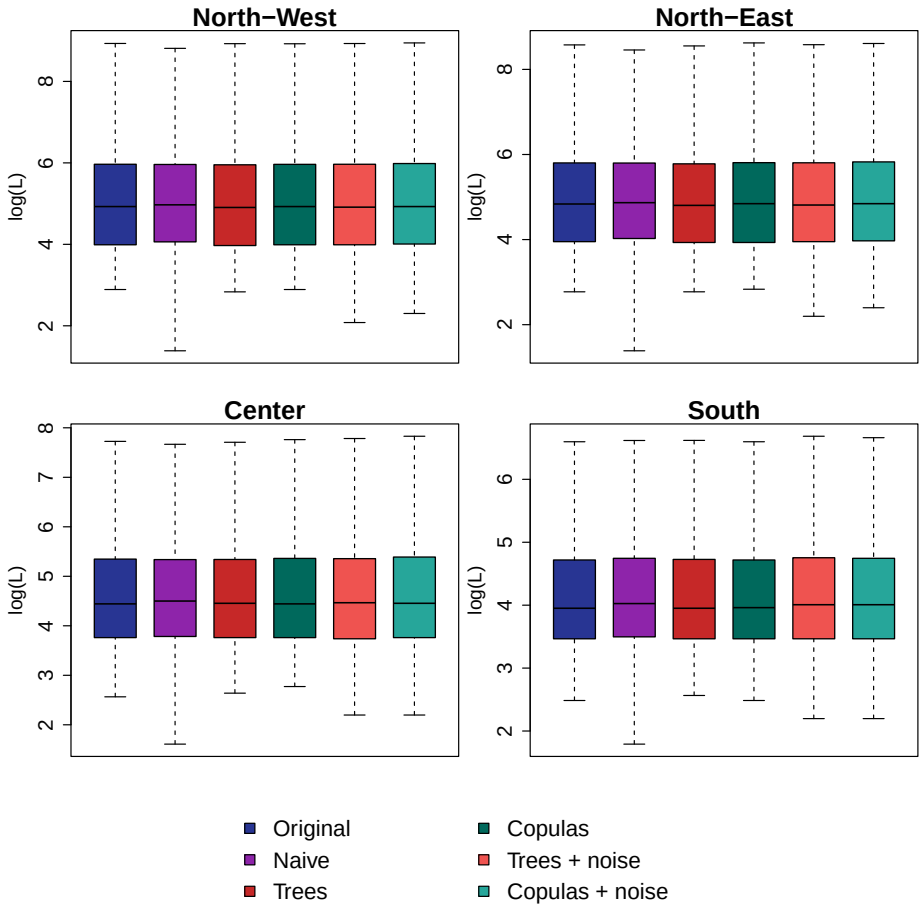

Figure 3: Comparison between original and synthetic conditional distributions of employment by geographic area.

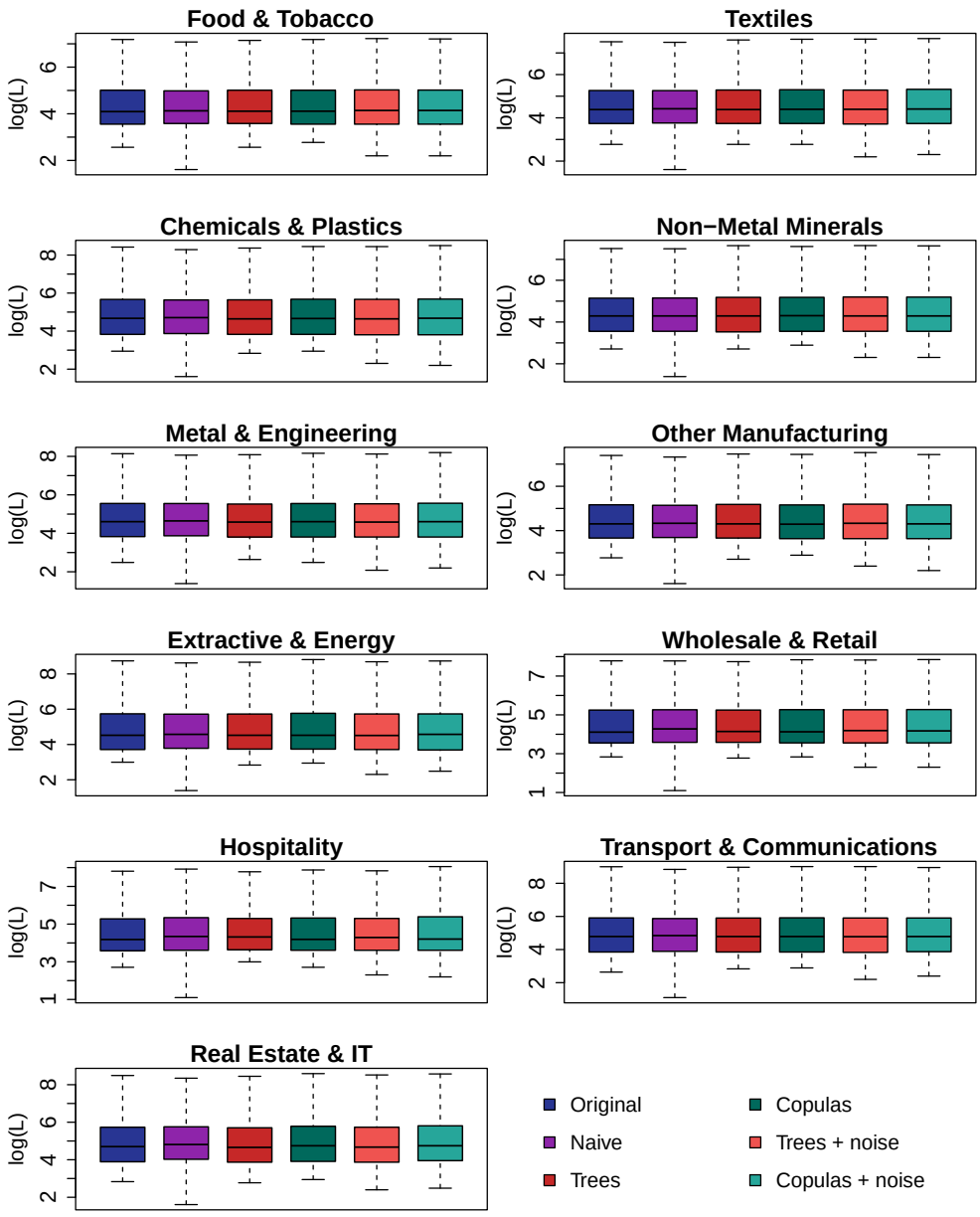

Figure 4: Comparison between original and synthetic conditional distributions of employment by economic sector.

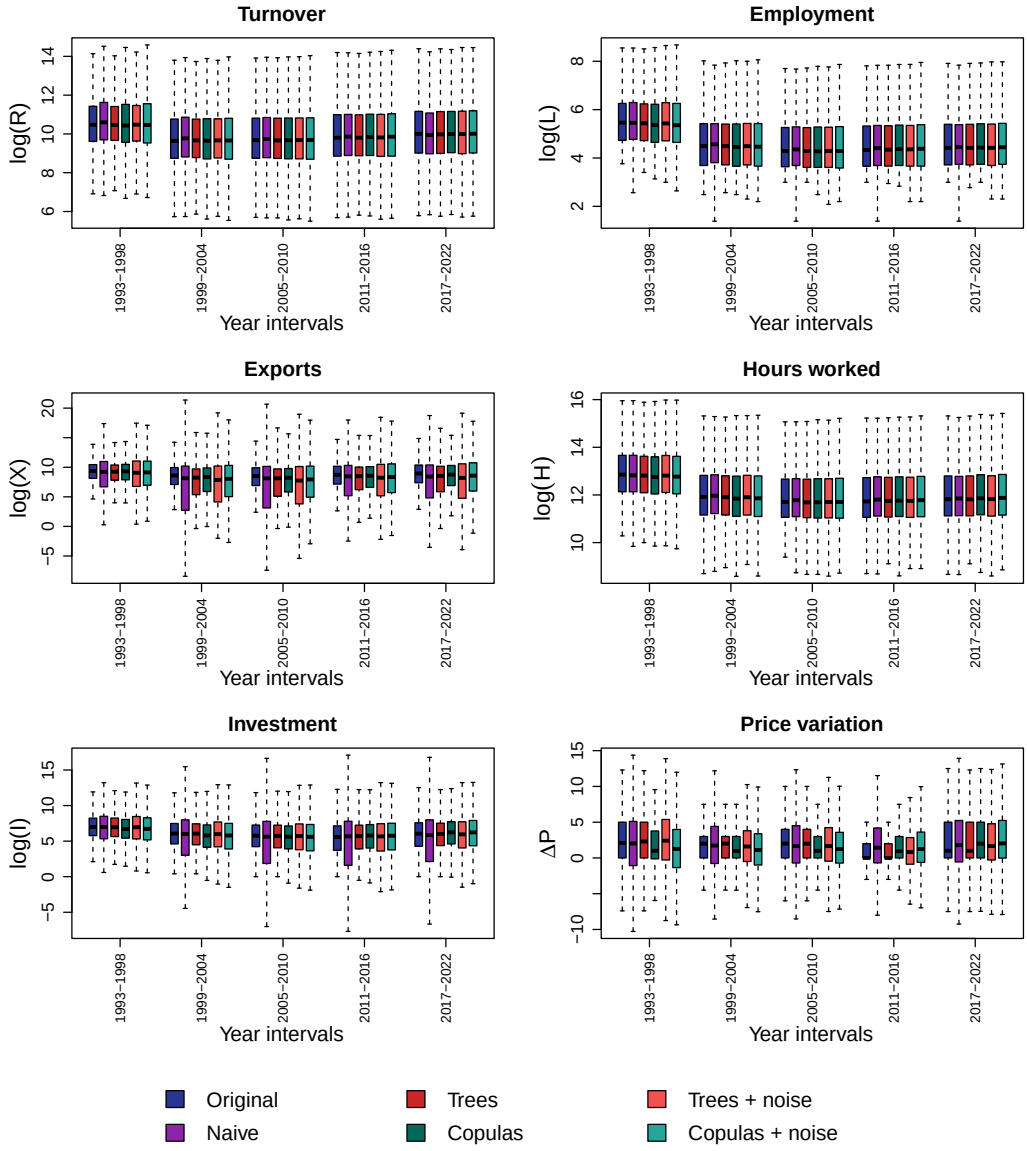

Figure 5: Comparison between original and synthetic conditional distributions by time period (six-year intervals).

4.2 Random regression experiments

To evaluate how well the synthetic datasets preserve multivariate relationships, we run 100 random regressions as follows. In each iteration, we draw one variable $Y^{(j)}$ at random from the set of all quantitative variables $\mathcal{Y}$ (which may include contemporaneous, lagged, and expectation elements) and use it as the dependent variable. We then select a random subset of regressors from the remaining variables $\mathcal{Y} \setminus \{Y^{(j)}\}$. A random subset of fixed effects is drawn from $\mathcal{F} = \{\text{area}, \text{sector}, \text{year}, \text{firm}\}$; when firm fixed effects are selected, the regression is estimated in deviations from firm means using only panel observations.

Each randomly generated specification s therefore takes the form:

$$\log\big(Y_{i,t}^{(j)}\big) = \alpha_s + \sum_{k \neq j} \beta_{k,s} \, \log\big(Y_{i,t}^{(k)}\big) + \sum_{f} \gamma_{f,s} \, \mathrm{FE}_f + \varepsilon_{i,t}. \tag{14}$$

Variables that may take zero or negative values (e.g. price variation) enter in levels rather than logs. Standard errors are clustered at firm level.

The purpose of this exercise is not to focus on economically meaningful models; some randomly generated specifications may lack a clear theoretical interpretation. This is intentional, as the objective is to assess whether a wide range of random specifications yields synthetic coefficients that are broadly consistent with those obtained from the original data.

For each coefficient k that appears in a given specification s, we compute the standardised error (Woo and Slavkovic, 2015; Snoke et al., 2018):

$$E_{k,s} = \frac{\widehat{\beta}_{k,s}^{\text{syn}} - \widehat{\beta}_{k,s}^{\text{orig}}}{\mathrm{sd}\big(\widehat{\beta}_{k,s}^{\text{orig}}\big)}, \tag{15}$$

where the numerator captures the difference between the estimates on the synthetic and original data, while the denominator scales it by the clustered-robust standard error from the original data. The resulting distribution of $E_{k,s}$ is examined across 100 randomly generated specifications, comparing around 3,500 coefficients. This distribution provides a direct assessment of the performance of the different methods. In particular, goodness can be judged by how tightly the distribution is concentrated around zero and by the absence of systematic distortions: a distribution that is both narrowly dispersed and centred at zero indicates high fidelity in reproducing the original multivariate relationships, while wider dispersion and shifts away from zero signal poorer performance.

A clear ranking emerges from Fig. 6. The tree-based generator performs best overall: its distribution is the most tightly concentrated around zero, with limited dispersion and no evidence of systematic bias. In addition, more than 50% of the mass lies within the interval $[-2, 2]$, i.e. within two standard errors of the original estimates, further confirming the strong absolute performance of the tree-based method. The copula-based method ranks second, showing a somewhat wider distribution but still reasonably concentrated around zero. By contrast, the naïve

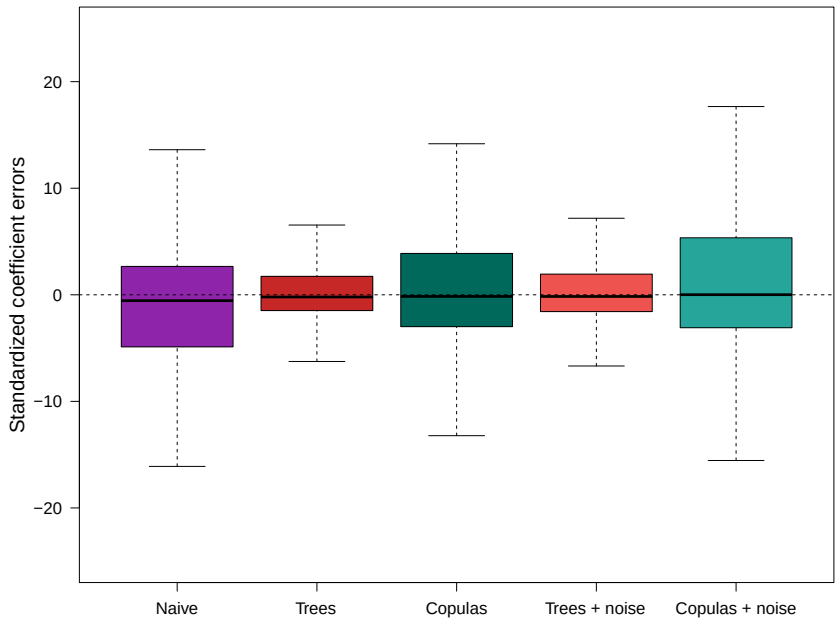

Figure 6: Standardized coefficients errors of the random regressions.

approach performs worst: its box is substantially wider and the tails of the distribution extend much further, consistent with the absence of multivariate structure in that method. Introducing measurement error in the tree-based and copula-based methods increases dispersion in a mechanically predictable way. However, this affects the two approaches differently. The tree-based method remains very robust, with only a moderate increase in dispersion and no meaningful emergence of bias. In contrast, the copula-based method deteriorates substantially: once measurement error is added, its dispersion becomes comparable to that of the naïve approach. As a result, the performance gap between the copula-based and naïve methods largely disappears, while the tree-based generator continues to clearly dominate.

4.3 Re-identification risk

As already mentioned, synthetic microdata must balance two objectives: achieving a high degree of statistical similarity with the original dataset, while keeping the risk of re-identification sufficiently low. Methods that reproduce the joint distribution more faithfully tend to generate synthetic records that lie closer to real observations, thereby increasing disclosure risk; conversely, generators that introduce greater variability tend to provide safer outputs at the cost of reduced analytical validity. Understanding this trade-off is crucial for evaluating the over-

all quality of the synthetic datasets.

Disclosure risk is assessed by measuring how close each synthetic record $\mathbf{Y}_{i,t}^S$ is to its nearest neighbor in the original data. We compute distances over the following set of quasi-identifiers: year, geographic area, economic sector (NACE), employment, turnover, exports, and investment. For each synthetic record we compute the distance to its nearest ($\mathbf{Y}_{j,s}^{R1}$) and second-nearest ($\mathbf{Y}_{l,m}^{R2}$) neighbors in the original data and summarise the local disclosure risk using the nearest neighbor distance ratio (NNDR, see Lowe, 2004):

$$\text{NNDR}(\mathbf{Y}_{i,t}^S) = \frac{d\left(\mathbf{Y}_{i,t}^S, \mathbf{Y}_{j,s}^{R1}\right)}{d\left(\mathbf{Y}_{i,t}^S, \mathbf{Y}_{l,m}^{R2}\right)} \in (0,1). \tag{16}$$

Values of NNDR close to one indicate that the synthetic record lies in a region of the data space where at least two original observations are similarly close, implying a low risk of unique re-identification. Conversely, values close to zero indicate that the synthetic observation is unusually close to a single original record relative to nearby alternatives, suggesting a higher disclosure risk. Distances are computed using both Euclidean and Gower's metrics.

4.3.1 Euclidean distance

The Euclidean distance $d_E(\cdot,\cdot)$ requires all variables involved in its computation to be represented in numerical form.[9] We also standardise the quasi-identifiers using the mean and standard deviation computed from the original dataset:

$$d_E\Big(\widetilde{\mathbf{Y}}_{i,t}, \widetilde{\mathbf{Y}}_{j,s}\Big) = \sqrt{\sum_{k=1}^{K}\left(\widetilde{Y}_{i,t}^{(k)} - \widetilde{Y}_{j,s}^{(k)}\right)^2}, \tag{17}$$

where $\widetilde{\mathbf{Y}}_{i,t} = (\widetilde{Y}_{i,t}^{(1)}, \ldots, \widetilde{Y}_{i,t}^{(K)})$ denotes the vector of standardised quasi-identifiers. Each component is given by $\widetilde{Y}_{i,t}^{(k)} = (Y_{i,t}^{(k)} - \bar{Y}^{(k)})/\sigma^{(k)}$, with $\bar{Y}^{(k)}$ and $\sigma^{(k)}$ denoting the mean and standard deviation computed from the original dataset, respectively.

The empirical NNDR distributions reveal clear differences across synthesis methods (Fig. 7). The naïve generator attains the highest NNDR values because it does not reproduce the joint structure of the original data: its synthetic records are more dispersed and therefore remain farther from any specific real observation. This reduced fidelity comes with the advantage of naturally providing stronger protection against re-identification. The sequential tree method exhibits the lowest NNDR values, consistent with its ability to reproduce multivariate structure more faithfully and thus place synthetic records closer to real ones. The copula-based generator lies in between: it preserves dependence more effectively than the naïve method, but the resulting synthetic points remain further from the originals than those produced by the sequential approach.

[9] Hence, we assign numerical codes to the geographic area and the economic sector.

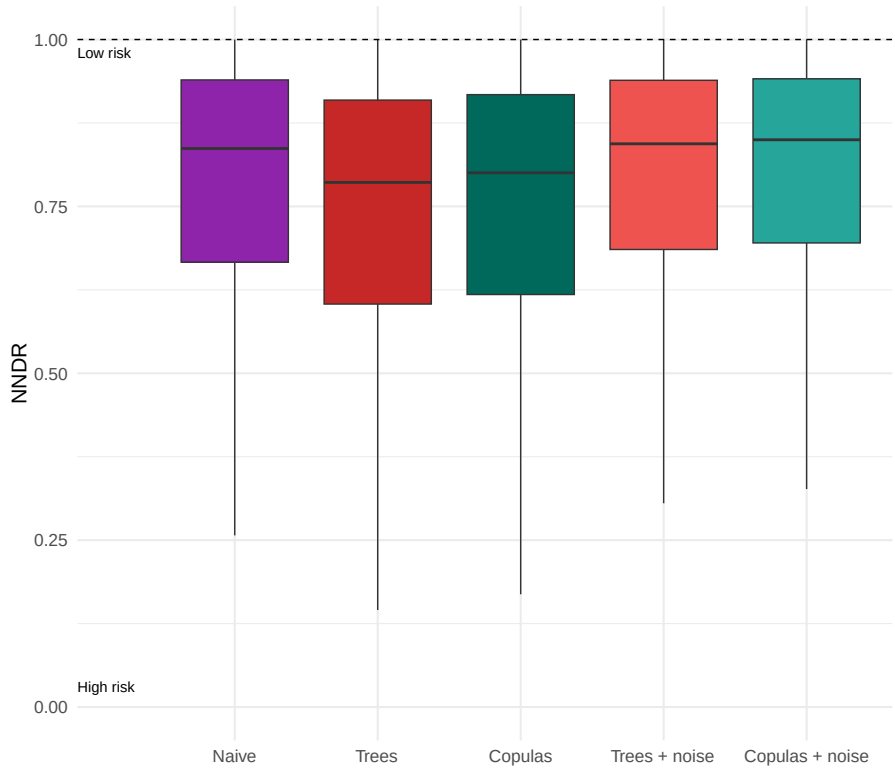

Figure 7: NNDR computed using the Euclidean distance, by data synthesis method.

Introducing measurement error shifts the NNDR distributions of both trees and copulas upward, restoring levels of disclosure protection comparable to those of the naïve generator. The added noise increases the local dispersion of synthetic records without altering the qualitative ordering of the methods in terms of statistical fidelity, thereby improving confidentiality while preserving analytical usefulness. Importantly, the NNDR distributions of the noise-augmented tree and copula generators remain concentrated around relatively high values, indicating a satisfactory level of disclosure protection also in absolute terms.

4.3.2 Gower's distance

Since quasi-identifiers comprise both continuous and categorical variables, we also use the weighted Gower's distance $d_G(\cdot,\cdot)$, which can handle mixed quantitative and qualitative data:

$$d_G\left(\widehat{\mathbf{Y}}_{i,t}, \widehat{\mathbf{Y}}_{j,s}\right) = \sum_{k=1}^{K} \omega^{(k)} \delta\left(\widehat{Y}_{i,t}^{(k)}, \widehat{Y}_{j,s}^{(k)}\right), \tag{18}$$

where $\widehat{\mathbf{Y}}_{\mathbf{i,t}} = (\widehat{Y}_{i,t}^{(1)}, ..., \widehat{Y}_{i,t}^{(K)})$ denotes the vector of normalized quasi-identifiers, $\omega^{(k)}$ is a variable-specific weight, and $\delta(\widehat{Y}_{i,t}^{(k)}, \widehat{Y}_{j,s}^{(k)})$ is the component-wise distance

associated with the $k-$th variable. Quasi-identifiers are normalized as

$$\widehat{Y}_{i,t}^{(k)} = \frac{Y_{i,t}^{(k)} - \min\left(Y^{(k)}\right)}{\max\left(Y^{(k)}\right) - \min\left(Y^{(k)}\right)}$$

Following the standard specification the component-wise distance is defined as

$$\delta\left(\widehat{Y}_{i,t}^{(k)}, \widehat{Y}_{j,s}^{(k)}\right) = \begin{cases} |\widehat{Y}_{i,t}^{(k)} - \widehat{Y}_{j,s}^{(k)}| & \text{for continuous variables} \\ \\ \mathbb{1}(\widehat{Y}_{i,t}^{(k)} \neq \widehat{Y}_{j,s}^{(k)}) & \text{for categorical variables} \end{cases}$$

The weight vector $\omega = (\omega^{(1)}, ..., \omega^{(K)})$ is constructed to reflect the empirical discriminative power of each quasi-identifier in the original dataset, following an approach already adopted in re-identification risk literature (see, e.g. Dankar et al., 2012). Specifically, weights are derived through a variable-type-specific measure of attribute distinctiveness, following a procedure that treats continuous and categorical variables differently. For continuous variables, the weight is defined as the coefficient of variation:

$$\omega^{(k)} = \frac{\sigma^{(k)}}{\bar{Y}^{(k)}},$$

where $\bar{Y}^{(k)}$ and $\sigma^{(k)}$ denote the mean and standard deviation computed from the original dataset, respectively. Variables exhibiting greater relative dispersion are thus assigned larger weights, reflecting their stronger capacity to distinguish between records. For categorical variables, the weight is defined as the normalized Shannon entropy:

$$\omega^{(k)} = \frac{-\sum_{n=1}^{J^{(k)}} p_n^{(k)} \log\left(p_n^{(k)}\right)}{\log\left(J^{(k)}\right)},$$

where $p_n^{(k)}$ denotes the empirical relative frequency of the $n-$th category of variable $Y_{i,t}^{(k)}$ and $J^{(k)}$ is the total number of distinct categories. Normalized entropy attains its maximum value when all categories are equally represented, indicating the highest discriminative potential. Conversely, a variable dominated by a single category receives a lower weight. All weights are rescaled to the unit interval $[0, 1]$ and then used to compute the weighted Gower's distance, ensuring that variables with higher empirical discriminative power contribute proportionally more to the overall inter-record dissimilarity measure.

When inter-record dissimilarities are computed using the weighted Gower's distance, NNDR values are uniformly higher across all synthesis methods (Fig. 8). This systematic increase likely reflects the inclusion of categorical variables, as well as the use of variable-specific weights. Despite this overall shift, the relative ranking of the methods remains largely consistent: the sequential tree generator continues to yield lower NNDR values than the copula-based approach, while the naïve method still provides stronger baseline disclosure protection than both

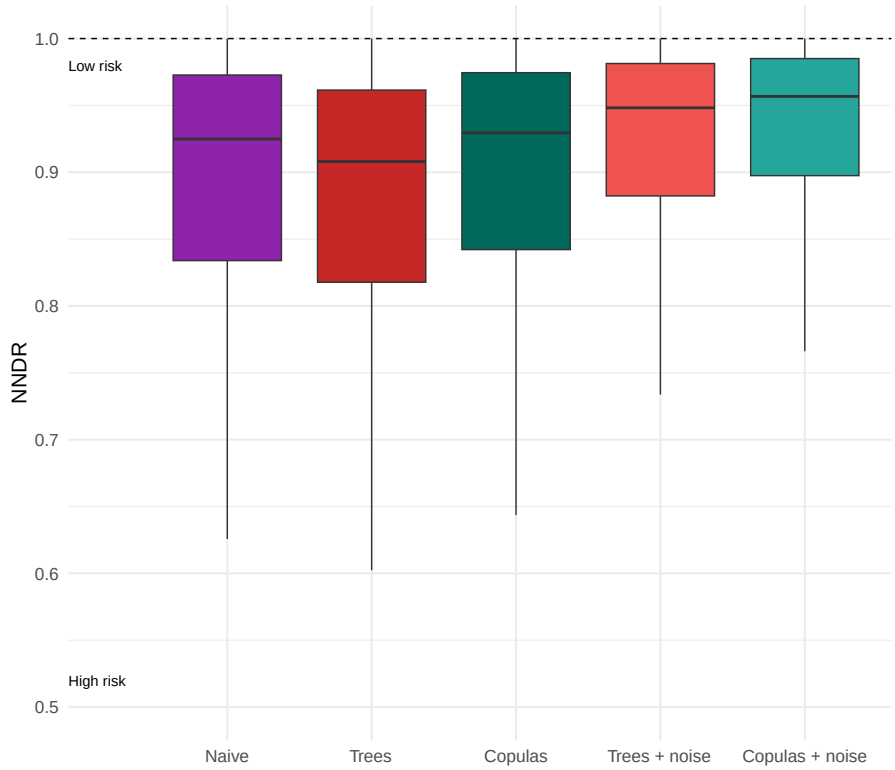

Figure 8: NNDR computed using the weighted Gower's distance, by data synthesis method.

structure-preserving techniques. As in the Euclidean case, introducing measurement error shifts the NNDR distributions of both tree- and copula-based methods upward, indicating enhanced protection against re-identification. Notably, however, using the weighted Gower's metric, noise-based methods exhibit higher NNDR values than the naïve generator.

This result suggests that, once the discriminative power of the variables is explicitly taken into account, incorporating controlled measurement errors can outperform random dispersion.

4.3.3 Further discussion

The NNDR metric can be used not only as an aggregate indicator of disclosure risk, but also as a diagnostic tool to identify potentially high-risk observations in the extreme lower tail of the NNDR distribution.

A specific disclosure challenge arises for large panel firms, which are economically important units and may be inherently recognizable due to their size and distinct combinations of characteristics. For reasons of realism and analytical credibility, we deliberately avoid synthesising some key stratification variables for these firms—namely geographic area and economic sector. Fully randomising

these dimensions could generate implausible firm trajectories, such as large manufacturing firms originally located in the North being reassigned to the South or to service industries in the synthetic data. Preserving these attributes ensures that the synthetic panel retains a close resemblance to the real economic structure and remains interpretable for empirical analysis.

For these units, alternative ad-hoc disclosure-control actions can be considered. These include excluding selected records from dissemination or further perturbing specific attributes through targeted noise injection or additional synthesis. Such interventions should be evaluated on a case-by-case basis, balancing analytical relevance against disclosure risk within the dissemination process.

5 Conclusions

This paper develops and evaluates a framework for generating and validating synthetic microdata for large, unbalanced firm-level panels, using the INVIND Survey as a case study.

By extending synthetic-data methodologies originally designed for cross-sectional settings, taking into account the time dimension in the data-generation mechanism, our approach addresses key challenges posed by heterogeneous firm histories, rich multivariate structures, and temporal dependence.

Our empirical results point to three central findings. First, both the sequential tree-based generator and the Gaussian copula approach are able to reproduce remarkably well the marginal, conditional, and temporal distributions of the original data, substantially outperforming a naïve generative benchmark.

Second, the two multivariate approaches preserve complex multivariate relationships: random regression experiments show that synthetic coefficients closely track their original counterparts across a wide range of specifications and fixed-effects structures. Among the methods considered, the sequential fully conditional specification based on regression and classification trees delivers the highest analytical fidelity, while the copula-based approach provides a robust and interpretable alternative with slightly lower but still strong performance.

Third, our analysis highlights the intrinsic trade-off between statistical accuracy and disclosure risk. Methods that better approximate the joint distribution of the original data tend to place synthetic records closer to real observations, thereby increasing re-identification risk under nearest-neighbor metrics. We show that this tension can be effectively mitigated through the introduction of controlled measurement error. Adding a carefully calibrated noise substantially increases local dispersion in the synthetic data, raising disclosure-protection levels to those achieved by much less informative generators, while preserving the key statistical properties required for exploratory analysis and model development. This result supports the view that noise injection should be understood not as a crude last-resort anonymization device, but as a flexible and transparent component of a modern synthetic-data pipeline.

Our findings have direct implications for confidential data disclosure. High - quality synthetic panels can serve as a powerful intermediate access layer, enabling researchers to explore data structures, design empirical strategies, and debug code in an interactive environment before submitting final analyses for execution on confidential microdata.

Overall, this paper shows that synthetic data for large, unbalanced firm-level panels are both feasible and analytically valuable when generated with appropriate methods and accompanied by systematic validation and risk assessment. We view the framework proposed here as a practical step toward the broader adoption of synthetic microdata by confidential data providers.

References

Brand, Ruth, "Microdata protection through noise addition." in "Inference control in statistical databases: From theory to practice," Springer, 2002, pp. 97–116.

Breiman, Leo, Jerome Friedman, R.A. Olshen, and Charles J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

Dankar, Fida Kamal, Khaled Emam, Angelica Neisa, and Tyson Roffey, 2012, "Estimating the re-identification risk of clinical data sets." *BMC Medical Informatics and Decision Making*, *12* (1).

Domingo-Ferrer, Josep, Francesc Sebé, and Jordi Castella-Roca, "On the security of noise addition for privacy in statistical databases." in "International Workshop on Privacy in Statistical Databases" Springer 2004, pp. 149–161.

Drechsler, Jörg, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, New York: Springer New York, 2011.

Drechsler, Jörg, 2012, "New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey." *Journal of Applied Statistics*, *39* (2), 243–265.

Gambs, Sébastien, Frédéric Ladouceur, Antoine Laurent, and Alexandre Roy-Gaumond, 2021, "Growing synthetic data through differentially-private vine copulas." *Proceedings on Privacy Enhancing Technologies.*

Hofert, Marius and Martin Mächler, 2011, "Nested Archimedean Copulas Meet R: The nacopula Package." *Journal of Statistical Software*, *39* (9), 1–20.

Hofert, Marius, Ivan Kojadinovic, Martin Maechler, and Jun Yan, *copula: Multivariate Dependence with Copulas* 2026. R package version 1.1-7.

Jahangir Alam, M., Benoit Dostie, Jörg Drechsler, and Lars Vilhuber, 2020, "Applying data synthesis for longitudinal business data across three countries." *Statistics in Transition New Series*, *21* (4), 212–236.

Kim, Jay and William Winkler, 2003, "Multiplicative noise for masking continuous data." *Statistics*, *1* (9), 1–18.

Kinney, Satkartar K., Jerome P. Reiter, and Javier Miranda, 2014, "SynLBD 2.0: Improving the synthetic Longitudinal Business Database." *Statistical Journal of the IAOS*, *30* (2), 129–135.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd, 2011, "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review*, *79* (3), 362–384.

Kojadinovic, Ivan and Jun Yan, 2010, "Modeling Multivariate Distributions with Continuous Margins Using the copula R Package." *Journal of Statistical Software*, *34* (9), 1–20.

Lowe, David G, 2004, "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, *60* (2), 91–110.

Meyer, David, Thomas Nagler, and Robin J. Hogan, 2021, "Copula-based synthetic data augmentation for machine-learning emulators." *Geoscientific Model Development*, *14* (8), 5205–5215.

Nowok, Beata, Gillian M. Raab, and Chris Dibben, 2016, "synthpop: Bespoke Creation of Synthetic Data in R." *Journal of Statistical Software*, *74* (11), 1–26.

Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni, "The synthetic data vault." in "2016 IEEE international conference on data science and advanced analytics (DSAA)" IEEE 2016, pp. 399–410.

Reiter, Jerome P., 2005, "Using CART to Generate Partially Synthetic, Public Use Microdata." *Journal of Official Statistics*, *21* (3), 441–462.

Rubin, Donald B., *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons, 1987.

Rubin, Donald B., 1993, "Discussion Statistical Disclosure limitation." *Journal of Official Statistics*, pp. 462–468.

Shlomo, Natalie, "Measurement error and statistical disclosure control." in "International Conference on Privacy in Statistical Databases" Springer 2010, pp. 118–126.

Sklar, Abe, 1959, "Fonctions de répartition à n dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Université de Paris, 8*, pp. 229–231.

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic, 2018, "General and specific utility measures for synthetic data." *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181* (3), 663–688.

Woo, Yong Ming Jeffrey and Aleksandra Slavkovic, 2015, "Generalised linear models with variables subject to post randomization method." *Statistica Applicata-Italian Journal of Applied Statistics*, *24* (1), 29–56.

Yan, Jun, 2007, "Enjoy the Joy of Copulas: With a Package copula." *Journal of Statistical Software*, *21* (4), 1–21.