Journal of the Royal Statistical Society Series C: Applied Statistics, 2026, **00**, 1–24
https://doi.org/10.1093/jrsssc/qlag015

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series C Applied Statistics

Original Article

Analysis of higher education dropouts dynamics through multilevel functional decomposition of recurrent events in counting processes

Alessandra Ragni[1], Chiara Masci[2] and Anna Maria Paganoni[1]

[1]MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan 20133, Italy
[2]Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Via Conservatorio 7, Milan 20122, Italy

Address for correspondence: Alessandra Ragni, MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan 20133, Italy. Email: alessandra.ragni@polimi.it

Abstract

This paper analyses higher education dropouts dynamics through an innovative approach that integrates recurrent events modelling and point process theory with functional data analysis. We propose a novel methodology that extends existing frameworks to accommodate hierarchical data structures, demonstrating its potential within a simulated setting. By analysing administrative data from student careers at Politecnico di Milano, we explore freshmen dropout patterns across different bachelor's degree programmes and schools. Specifically, we model dropouts as recurrent events occurring across both programmes and schools using a Cox-based recurrent events model. Additionally, we leverage functional data analysis and multilevel principal component analysis to unravel the latent effects of degree programmes and schools on dropout trends, offering valuable insights for institutions seeking to implement strategies aimed at reducing dropout rates. The proposed methodology offers a groundbreaking approach to dropout analysis, opening a new perspective and avenues for modelling its dynamics.

Keywords: functional data analysis, multilevel principal component analysis, recurrent events, students dropout

1 Introduction

The higher education system is worldwide affected by high dropout rates. In this context, 'dropout' refers to students leaving the university world without completing their degree. From the perspective of a single university, dropout occurs when a student exits their academic programme before earning the final qualification (Tinto, 1982). Despite efforts by European governments to expand access to higher education, ensuring successful degree completion remains a challenge, and dropout rates persist at around 30% across OECD member countries (OECD, 2019). In Italy, this issue is particularly pronounced, with a significant proportion of students discontinuing their studies, often within the first two years of enrolment. More than half of those who begin higher education fail to complete their degrees (Aina et al., 2018). Indeed, the percentage of adults with a higher education degree in Italy is below the OECD average (Cannistrà, 2024; OECD, 2019). These high dropout rates not only lower the average skill levels of the workforce (Atzeni et al., 2022), but they are also linked to a growing wage-skill gap (Katz & Murphy, 1992).

From an institutional perspective, high dropout rates represent a waste of resources. In fact, the long-term returns—both in terms of human capital development and credentials awarded—are

Received: February 24, 2025. **Revised:** February 19, 2026. **Accepted:** February 21, 2026
© The Royal Statistical Society 2026.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

lost when students exit without completing their degree, despite the considerable investments made by universities in teaching, recruitment, and student support. As a result, analyzing and reducing university dropout rates has become a critical challenge for higher education institutions.

What makes managing this issue even more complex is the significant variation in dropout behaviour across degree programmes and schools. Even within the same university, dropout patterns differ widely across academic disciplines. For instance, some programmes may experience higher dropout rates during early semesters due to difficult coursework, while others might see students leave later in their studies, near graduation. Additionally, dropout rates can vary between schools within the same university, influenced by factors such as faculty engagement, availability of student support services, and workload.

In the statistical literature, dropout has been mainly studied from a student perspective, by modelling the student dropout event through classification methods (Hegde & Prageeth, 2018; Kehm et al., 2019) or, more recently, by modelling the time to dropout through survival analysis tools (Barragán et al., 2022; Gury, 2011; Min et al., 2011; Patacsil, 2020; Singer & Willett, 1993). The latent effects of degree courses or different educational providers have also been investigated through the use of multilevel and frailty models (Cannistrà, 2024; Cannistrà et al., 2022). In this paper, we introduce a novel approach that conceptualizes dropout as a time-dependent, degree course-specific phenomenon, shifting the focus from the individual to the degree course. By exploiting and combining students data, we aim to reconstruct the dropout curve by modelling dropout intensity as a counting process, capturing the temporal dynamics of dropout risk in terms of cumulative hazard. We then propose a procedure to decompose this dropout curve into contributions from different academic units, highlighting how the various data levels influence the evolution of dropout.

To serve this purpose, we analyse administrative data from Politecnico di Milano (PoliMi) to examine dropout patterns across its bachelor's degree programmes. PoliMi comprises four distinct *schools*: Architecture, Design and Engineering, further divided into the School of Civil, Environmental, and Land Management Engineering and the School of Industrial and Information Engineering, offering 23 different undergraduate programmes, referred to as degree courses or simply *courses*. Our focus is on understanding dropout rates across these programmes, exploring how they vary by degree programme and school over the first three semesters span, as it is highly predictive of dropout risk, as seen in previous works (Cannistrà, 2024; Cannistrà et al., 2022; Masci et al., 2024).[1]

Our approach builds on methodologies that integrate recurrent events modelling, point process theory, and functional data analysis, extending the techniques proposed by Baraldo et al. (2013) and Spreafico and Ieva (2021). In these studies, hospital readmissions and drug consumption over time are analysed to predict outcomes related to heart failure telemonitoring in the former and time-to-death in the latter. We extend and generalize this framework to account for the *hierarchical structure* of the data (Pinheiro & Bates, 2000) given by units (students) organized within nested groups or levels (degree programmes and schools), enabling a more detailed exploration of dropout dynamics across different academic units. Specifically, the analysis comprises two phases. In the first phase, we utilize historical dropout data to fit a counting process model (Daley & Vere-Jones, 2002), enabling us to compute the realized trajectories of the cumulative hazard process (compensators) underlying the dropout counting process. While many alternatives are available for the modelization of recurrent events (Amorim & Cai, 2015), such as extensions of Cox models (see, for instance, the Prentice, Williams and Peterson model Prentice et al., 1981, Wei, Lin and Weissfeld model Wei et al., 1989, frailty models Therneau et al., 2000), models for the mean number of events or their occurrence rate (Diao et al., 2014; Lin et al., 2000), multistate models (Andersen & Keiding, 2002), and virtual (effective) age models (Beutner, 2023; Kijima et al., 1988; Peña & Hollander, 2004), we employ the Andersen-Gill (AG) model (Andersen & Gill, 1982), following the choice made in Spreafico and Ieva (2021). The AG model extends the Cox proportional hazards model by incorporating the increments in event counts over time, assuming that correlations between event times can be explained by prior occurrences, as

[1] We focus on a single university as it provides a naturally heterogeneous setting, encompassing multiple schools and degree programmes with known various dropout patterns (Masci et al., 2024). This allows us to analyse diverse academic environments while ensuring data consistency and comparability, which would be challenging in a scenario encompassing multi-university data due to differences in data availability, institutional policies, and programme structures.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

well as through the specification of appropriate time-varying covariates, such as the count of previous occurrences (Amorim & Cai, 2015). In contrast to the alternative approaches, the AG model represents recurrent events as a single counting process for each group, with a common baseline hazard across events and without explicit modelling of event order. These features align with our objective of estimating cumulative hazard trajectories, whereas the other approaches differ in their handling of event dependence, ordering, or post-intervention dynamics, making them less suitable for our study. This framework allows us to represent these events as non-stationary stochastic counting processes that may depend on specific characteristics or labels, referred to as *marks* (Daley & Vere-Jones, 2002; Spreafico & Ieva, 2021). At this stage, the longitudinal trajectory of instantaneous dropout risk over time within a degree programme is treated as a function, and functional data analysis techniques (Ramsay & Silverman, 2005) are employed to extract insights from repeated dropout events as two-level functional covariates. These covariates are derived through dimensionality reduction using Multilevel Functional Principal Component Analysis (MFPCA) (Cui et al., 2023; Di et al., 2009), preserving most of the historical information while effectively managing variability across two levels. In this first phase, our aims are twofold: (i) to reconstruct the dropout curve in terms of cumulative hazard for each degree programme, and (ii) to decompose this dropout curve into contributions from both the degree programme and school levels using MFPCA, highlighting how different levels are associated to the evolution of dropout risk. In the second phase, we adopt a predictive framework to investigate how this information influences the subsequent risk of dropout among students, incorporating information specific to the dropout risk associated with each course or school. The aim of this second phase is to assess the predictive value of the extracted functional covariates in forecasting future dropout events at the student level. To model the association between functional and scalar predictors and time-to-event outcomes, we adopt the functional linear Cox regression framework proposed by Kong et al. (2018), which incorporates functional predictors within a high-dimensional Cox model.

Previous studies focused on PoliMi data have addressed various aspects of student dropout prediction and quantification (Cannistrà et al., 2022; Diaz Lema et al., 2024; Masci et al., 2024; Ragni et al., 2025), with some of them specifically examining the impact of hierarchical structures (i.e. grouping factors, such as degree programmes) on the time to dropout within the first few semesters of enrolment up to the full three years of bachelor degree. The tools commonly employed in this context include shared frailty Cox proportional hazard models (Cook et al., 2007; Kleinbaum & Klein, 1996), where the frailty term represents a constant factor shared among clusters (e.g. degree programmes), which affects the baseline hazard multiplicatively, accounting for unobserved heterogeneity within clusters and allowing for a more detailed understanding of the dropout risk across different academic programmes. Our study extends previous research by offering a more refined analysis of dropout behaviour over time, specifically examining how dropout dynamics evolve both within and between degree programmes and schools. A key advancement is the incorporation of the dropout history into the predictive framework differently from the shared frailty model, which simplifies this information into a single measure. By modelling dropout behaviour over time—identifying peaks at different stages of the academic journey—and analyzing how these patterns vary across degree programmes and schools while considering student characteristics, we can generate insights that can help institutions develop targeted strategies to reduce dropout rates.

In this perspective, two key elements of novelty need to be highlighted. First, we introduce the use of multilevel FPCA to decompose the dropout curve constructed from the compensator function. This novel application allows us to separate the contributions of different academic units (e.g. programmes and schools) to the overall dropout dynamics, providing a richer understanding of how these factors influence dropout risk over time. Second, and more importantly, our approach represents a completely new perspective in the dropout literature, where most studies focus on classification-based predictive models or, in more sophisticated cases, time-to-event models, almost exclusively at the student level (see, for instance, Arulampalam et al., 2004; Gury, 2011; Min et al., 2011; Patacsil, 2020; Plank et al., 2008; Vallejos & Steel, 2017, and Masci et al., 2024 for a discussion). In contrast, our approach models dropout dynamics at the degree course level, capturing historical temporal patterns that can later be included for predictions at student level.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

The paper is structured as follows. Section 2 outlines the employed methodology, providing a recap of the framework and extending it within a multilevel context. Section 3 reports on a simulation study that illustrates the application of the proposed methodology within the multilevel framework. In Section 4, we introduce the PoliMi dataset, detailing the cohort selection and study design. Section 5 presents the results obtained from applying the proposed methodology to the PoliMi case study. This study is complemented by a nonparametric bootstrap analysis designed to quantify uncertainty and assess the influence of the first estimation step on the second, as well as by a benchmark analysis comparing the results with natural alternatives. Discussion and concluding remarks are provided in Section 6. The code[2] is openly accessible at https://github.com/alessandragni/DropoutMultilevelFunct.

2 Methodology

In this section, we outline the methodology following three consecutive steps: recap on model formulation for recurrent events and compensator reconstruction (Subsection 2.1), compensator decomposition through multilevel principal component analysis (Section 2.2) and the development of a predictive model (Section 2.3). The core methodological contribution of this work regards the extension to the multilevel setting of the decomposition of recurrent events in counting processes.

2.1 Recap on the model formulation for recurrent events and compensator reconstruction

Let $N_{ij}(t)$, with $t \in [0, T]$, denote the stochastic process counting the dropout events observed up to time t, where $j = 1, \ldots, J_i$ indexes the lower-level units and $i = 1, \ldots, I$ indexes the higher-level units, or clusters, with the total number of lower-level units given by $\sum_{i=1}^{I} J_i = n$ (Cook et al., 2007). The process $N_{ij}(t)$ is adapted to the filtration $\{\mathcal{F}_{t,ij}\}_{t\in[0,T]}$, that is the history of realizations of the process itself. Assuming $N_{ij}(t)$ is a class D submartingale, the Doob-Meyer (D-M) decomposition theorem (Meyer, 1962) states that $M_{ij}(t) = N_{ij}(t) - \Lambda_{ij}(t)$ is a zero-mean, uniformly integrable martingale. Here, $\Lambda_{ij}(t) := \int_0^t Y_{ij}(s)\lambda_{ij}(s)\,ds$ is the unique predictable non-decreasing cadlag[3] and integrable *compensator* (or cumulative hazard process), with $\lambda_{ij}(t)$ being the hazard function and $Y_{ij}(t)$ the at-risk process (i.e. equals value 1 when unit j in cluster i is at risk at time t and 0 otherwise, under independent censoring); their product defines the intensity process $\lambda^*_{ij}(t) = Y_{ij}(t)\lambda_{ij}(t)$.

Building on the formulation for marked point processes described in Spreafico and Ieva (2021), we consider events whose cumulative number up to time t are recorded by the counting process $N_{ij}(t)$, where events can be further characterized by *marks*, represented by random variables $\boldsymbol{\omega}_{ij}$. These marks provide additional information about each event, such as the size or magnitude (Daley et al., 2003; Last & Brandt, 1995). In this framework, the intensity process depends not only on time t but also on the mark $\boldsymbol{\omega}_{ij}$. Assuming conditional independence of jump times and marks, the intensity function factorizes as follows $\lambda^*_{ij}(t, \boldsymbol{\omega}_{ij}) = \lambda^*_{ij,g}(t) f_{ij}(\boldsymbol{\omega}_{ij})$, where $\lambda^*_{ij,g}(t)$ is the ground intensity process of the counting process and f_{ij} is the multivariate density function of the marks $\boldsymbol{\omega}_{ij}$. Proper modelling of compensators and particularly of $\lambda^*_{ij}(t, \boldsymbol{\omega}_{ij})$ allows for an accurate reconstruction of $N_{ij}(t)$, as $M_{ij}(t)$ represents the residual of the process in the D-M decomposition.

Several models for $\lambda^*_{ij}(t)$ are available in the literature on counting processes (Aalen et al., 2008; Andersen et al., 2012; Peña & Hollander, 2004). Employing the model introduced by (Andersen & Gill, 1982), and assuming that $f_{ij}(\boldsymbol{\omega}_{ij})$ depends on $\mathbf{z}_{ij}(t)$ (which represents time-dependent features related to the marks $\boldsymbol{\omega}_{ij}$), we obtain:

$$\begin{aligned}\lambda^*_{ij}(t, \boldsymbol{\omega}_{ij}) &= Y_{ij}(t)\ \lambda_0(t)\ \exp\{\boldsymbol{\beta}^T \mathbf{x}_{ij}\}\ \exp\{\boldsymbol{\theta}^T \mathbf{z}_{ij}(t)\} \\ &= Y_{ij}(t)\ \lambda_0(t)\ \exp\{\boldsymbol{\beta}^T \mathbf{x}_{ij} + \boldsymbol{\theta}^T \mathbf{z}_{ij}(t)\}.\end{aligned} \tag{1}$$

Here, $\mathbf{x}_{ij}$ is the column vector of covariates of the j^{th} unit in i^{th} cluster, $\lambda_0(t)$ is the baseline hazard function, and $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are Q- and P-dimensional column vectors of coefficients, respectively, and ^{T}

[2] Analysis are performed through the statistical software R (R Core Team, 2022).

[3] i.e. right-continuous with left limits.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

stands for the transpose. Following Spreafico and Ieva (2021), the mark density $f_{ij}(\boldsymbol{\omega}_{ij})$ is implicitly incorporated through the exponential term involving $\mathbf{z}_{ij}(t)$, which captures the influence of the marks on the process. Thus, both $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}(t)$ serve as covariates. The former contains baseline covariates that can be observed throughout the process, while the latter is composed by the time-dependent covariates associated with the marks and is non-measurable with respect to the filtration at time zero. Then, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated in the model fitting by partial likelihood maximization (Andersen & Gill, 1982), while the baseline cumulative hazard can be estimated through Breslow estimator (Breslow, 1975) as a step-function $\hat{\Lambda}_0(t)$ and then smoothed into $\tilde{\Lambda}_0(t)$ as described in Baraldo et al. (2013).

Let now $[t_k^{(ij)}, t_{k+1}^{(ij)}]$ for $k = 0, \ldots, N_{ij}(T)$ be the intervals whose extremes are the jump times for each unit j in cluster i, being $t_0^{(ij)} = 0$ and $t_{N_{ij}(T)+1}^{(ij)} = T$. Then $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}^*(s, \boldsymbol{\omega}_{ij})\, ds$ can be estimated by approximation as follows (see computation in the online supplementary material, Section S1):

$$\hat{\Lambda}_{ij}(t) = \sum_{k=0}^{N_{ij}(t^-)} \exp\left(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij} + \hat{\boldsymbol{\theta}}^T \mathbf{z}_{ij}(t_k^{(ij)})\right)\left[\tilde{\Lambda}_0(t_{k+1}^{(ij)} \wedge t) - \tilde{\Lambda}_0(t_k^{(ij)})\right]. \tag{2}$$

where $a \wedge b = \min\{a, b\}$, $N_{ij}(t^-)$ represents the number of occurrences that have happened strictly before time t, and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are the estimated vectors of coefficients.

2.2 Multilevel functional principal component analysis for compensator decomposition

After reconstructing compensators through a marked point process formulation for recurrent events, $\hat{\Lambda}_{ij}(t)$ can be regarded as functional data objects, allowing the application of functional data analysis techniques (Ramsay & Silverman, 2005).

Given the high-dimensional nature of these data and the hierarchical setting given by the fact that each unit j is nested within a cluster i, we aim to decompose functional variability and reduce dimensionality, while getting insights. To achieve this, we apply MFPCA (Cui et al., 2023; Di et al., 2009). MFPCA integrates classical FPCA (Ramsay & Silverman, 2005), which selects only the relevant components of an appropriate orthonormal basis expansion, with standard multilevel mixed models. This approach effectively decomposes the observed data according to two levels of functional variation. Specifically, from the one-way functional ANOVA (Di et al., 2009) follows that

$$\hat{\Lambda}_{ij}(t) = \mu(t) + V_i(t) + W_{ij}(t) + \epsilon_{ij}(t) \tag{3}$$

$$= \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k^{(1)}(t) + \sum_{l=1}^{\infty} \zeta_{ijl}\phi_l^{(2)}(t) + \epsilon_{ij}(t) \tag{4}$$

where, in Eq. (3), $\mu(t)$ is a fixed functional effect, $V_i(t)$ and $W_{ij}(t)$ are mean 0 stochastic processes (uncorrelated between each other) and ϵ_{ij} is a white noise process with variance assumed to be σ^2, that appears only when functional data are observed with errors. Eq. (4) follows from Karhunen-Loève (KL) expansion (Karhunen, 1947; Loeve, 1948), where $\phi_k^{(1)}(t)$ and $\phi_l^{(2)}(t)$ are respectively level 1 (i.e. cluster level) and level 2 (i.e. unit level) eigenfunctions (fixed functional effects), and ξ_{ik} and ζ_{ijl} are respectively level 1 and 2 principal component scores (zero mean random variables, uncorrelated between each other).

Moreover, one may truncate the decomposition by pre-specifying at both levels the Percentage of Variance Explained (PVE) as explained in Di et al. (2009), resulting into

$$\hat{\Lambda}_{ij}(t) \simeq \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k^{(1)}(t) + \sum_{l=1}^{L} \zeta_{ijl}\phi_l^{(2)}(t) + \epsilon_{ij}(t) \tag{5}$$

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

being K and L the number of principal components finally identified respectively at level 1 (cluster) and level 2 (unit). Other interesting indicators defined in Di et al. (2009) are the total explained variance between-clusters and within-clusters, and the proportion of variability explained by level 1.

2.3 Cox regression model with functional compensators

The compensator decomposition in previous sections allows to extract dropout information focussing on the *observation period* $S = [0, T]$, where the units are at level j, clustered within level i. As a last step, we include this functional information into a functional Cox regression model Kong et al. (2018) which considers another cohort, where the units are now at the level h, for $h = 1, \ldots, H_{ij}$, nested within level j again nested within level i, so that $n_{\text{tot}} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} H_{ij}$ is the total number of units.

Let (T_{ijh}, δ_{ijh}) denote the observed pair of variables representing the time-to-event data, where $T_{ijh} = min\{C_{ijh}, \tilde{T}_{ijh}\}$, with $\tilde{T}_{ijh}$ being the true event time and C_{ijh} the censoring time. The inicator variable is defined as $\delta_{ijh} = 1$ if $\tilde{T}_{ijh} \leq C_{ijh}$, and 0 otherwise. Then, the functional compensators are included in the Cox model as follows:

$$\begin{aligned}
\eta_{ijh}\left(t \mid \mathbf{w}_{ijh}, \hat{\Lambda}_{ij}\right) &= \eta_0(t) \exp\left\{\boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \textstyle\int_S \hat{\Lambda}_{ij}(s)\alpha(s)\, \mathrm{d}s\right\} \\
&= \eta_0^*(t) \exp\left\{\boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \textstyle\int_S \left[\sum_{k=1}^{K} \xi_{ik}\phi_k^{(1)}(s) + \sum_{l=1}^{L} \zeta_{ijl}\phi_l^{(2)}(s)\right]\alpha(s)\, \mathrm{d}s\right\} \\
&= \eta_0^*(t) \exp\left\{\boldsymbol{\gamma}^T \mathbf{w}_{ijh} + \sum_{k=1}^{K} \xi_{ik}\alpha_k^{(1)} + \sum_{l=1}^{L} \zeta_{ijl}\alpha_l^{(2)}\right\}
\end{aligned} \tag{6}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J_i$ and $h = 1, \ldots H_{ij}$ where $\eta_{ijh}(t \mid \mathbf{w}_{ijh}, \hat{\Lambda}_{ij})$ denotes the hazard function for unit h at time t, η_0 is the baseline hazard function, $\boldsymbol{\gamma}$ is a q-dimensional vector of parameters to be estimated, $\mathbf{w}_{ijh}$ is a vector of covariates available at unit level h, $\alpha : S \to \mathbb{R}$ is a functional parameter. The second line follows from Eq. (5), so that $\eta_0^*(t) = \eta_0(t) \exp\{\int_S \mu(t)\alpha(s)\, \mathrm{d}s\}$ and last equality is given by rewriting $\alpha(s)$ according to different representations into the two different orthonormal bases $\phi_k^{(1)}$ and $\phi_l^{(2)}$, thanks to the orthonormality property; the subscripts are added in order to distinguish the two projections.

3 Simulation framework

In this section, we aim to test the methodology described above, in Sections 2.1 and 2.2. Specifically, after simulating unit-level intensities with shapes based on specific similarities within clusters and generating the Non-Homogeneous Poisson Processes (NHPPs) from these intensities, we show that compensator reconstruction using AG models recovers the within-cluster similarities.

In Subsection 3.1, we begin by simulating intensities $\lambda_{ij}^*(t)$ following a procedure similar to Cui et al. (2023) and Di et al. (2009). We employ a one-way functional ANOVA model to capture similarities within clusters and integrate $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}^*(u)\, \mathrm{d}u$ to obtain compensator-like shapes, translating the intensity functions into cumulative hazard functions, and we simulate event times from the $\lambda_{ij}^*(t)$. In Subsection 3.2.1, we fit AG models to the simulated event data, and reconstruct the compensator $\hat{\Lambda}_{ij}(t)$. Finally, in Subsection 3.2.2 we evaluate the consistency of the information captured by MFPCA before and after NHPPs extraction. We aim to show that cumulative hazard reconstruction using AG models preserves the essential information captured by the MFPCA.

The Cox regression model with functional compensators is not included in the simulation study, as it closely follows the framework developed in Kong et al. (2018) for the functional linear Cox regression model.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

3.1 Data-generating process

Let $\lambda^*_{ij}(t)$ be an intensity function measured over a continuous variable $t \in [0, 1]$ for observation j within cluster i, for $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$, generated by a modified one-way functional ANOVA model (Morris et al., 2003) as follows:

$$\lambda^*_{ij}(t) := \mu(t) + 2 \cdot i \cdot \left(\sum_{k=1}^{K} \xi_{ik} \phi^{(1)}_k(t) + \sum_{l=1}^{L} \zeta_{ijl} \phi^{(2)}_l(t) + \epsilon_{ij}(t) \right) \tag{7}$$

where $\mu(t) = 200$, $\xi_{ik} \sim \mathcal{N}(0, \lambda^{(1)}_k)$, $\zeta_{ijl} \sim \mathcal{N}(0, \lambda^{(2)}_l)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

For our data generation, we draw inspiration from the simulation study described in Section 4 of Di et al. (2009), introducing a few modifications to make the generation of the intensities more suitable for our specific context. Firstly, we include a constant $\mu(t)$ to increase the frequency of events in the NHPP generated by $\lambda^*_{ij}(t)$. Additionally, we scale level 1 and 2 components by a cluster-dependent constant, enhancing the differentiation between-groups and, consequently, the cumulative intensities. Lastly, we assign a higher standard deviation to the true eigenvalues at level 1 compared to level 2, further distinguishing clusters and reducing within-cluster variability.

The decision to simulate the intensities rather than directly simulating the cumulative hazard function stems from the specific characteristics required for the cumulative hazard (increasing monotonicity and ensuring that $\Lambda_{ij}(0) = 0$). Simulating scores from a normal distribution while maintaining these properties is not possible. Therefore, we opt to simulate the intensities to ensure these essential characteristics are preserved.

As a baseline scenario, we assume $J_i = 4$ units for each cluster i, with $i = 1, \ldots, I$, with $I = 20$, leading to $n = \sum_{i=1}^{I} J_i = 80$ units, and $K = L = 4$. The chosen value of the eigenvalues are $\lambda^{(1)}_k = 0.9^{k-1}$ for $k = 1, \ldots, K$ and $\lambda^{(2)}_l = 0.2^{l-1}$ for $l = 1, \ldots, L$, while the chosen value of the eigenfunctions, chosen following (Cui et al., 2023; Di et al., 2009), are

$$\{\phi^{(1)}_1(t), \phi^{(1)}_2(t), \phi^{(1)}_3(t), \phi^{(1)}_4(t)\} = \{\sqrt{2}\sin(2\pi t), \sqrt{2}\cos(2\pi t), \sqrt{2}\sin(4\pi t), \sqrt{2}\cos(4\pi t)\} \tag{8}$$

$$\{\phi^{(2)}_1(t), \phi^{(2)}_2(t), \phi^{(2)}_3(t), \phi^{(2)}_4(t)\} = \{1, \sqrt{3}(2t-1), \sqrt{5}(6t^2-6t+1), \sqrt{7}(20t^3-30t^2+12t-1)\} \tag{9}$$

at levels 1 and 2, respectively, observed on an equally spaced grid $\{1/T, 2/T, \ldots, (T-1)/T, 1\}$, with $T = 10^3$. Moreover, we assume $\mu(t) = 200$ and $\sigma = 0$ (absence of noise or measurement error). Afterwards, we compute the cumulative hazard function as $\Lambda_{ij}(t) = \int_0^t \lambda^*_{ij}(u)\, du$. In Figure 1, we display the simulated intensities $\lambda^*_{ij}(t)$ (panel (a)) and the cumulative hazards $\Lambda_{ij}(t)$ (panel (b)) on the y-axis, evaluated over the time grid $\{1/T, 2/T, \ldots, (T-1)/T, 1\}$ shown on the x-axis, representing time t.

Following Pasupathy (2010), through the thinning method we simulate event times for a NHPP over [0, 1], where the process is characterized by the time-varying intensity function $\lambda^*_{ij}(t)$, replicating the procedure 100 times.

Furthermore, we consider alternative settings of J_i, I and T to the baseline scenario, in order to quantify the impact of sample size and the clustering scheme on the proposed simulation. Specifically, we select in turn $I = \{10, 30, 50\}$, $J_i = \{4, 10, 50\}$ and $T = \{200, 1000, 5000\}$. Furthermore, following Cui et al. (2023), we also consider unbalanced design, where the number of units per clusters is drawn from Poisson(J_i) with a minimum of 1 unit per cluster.

The estimation accuracy is quantified for each of the 100 replications using the mean integrated squared error (MISE), defined in Cui et al. (2023) as $\text{MISE}(\hat{\boldsymbol{\phi}}^{(1)}) = (2T)^{-1} \|\hat{\boldsymbol{\phi}}^{(1)}_{\text{true}} - \hat{\boldsymbol{\phi}}^{(1)}_{\text{fitted}}\|^2_F$ and $\text{MISE}(\hat{\boldsymbol{\phi}}^{(2)}) = (2T)^{-1} \|\hat{\boldsymbol{\phi}}^{(2)}_{\text{true}} - \hat{\boldsymbol{\phi}}^{(2)}_{\text{fitted}}\|^2_F$ for the two eigenfunctions at levels 1 and 2, respectively, where $\|\cdot\|_F$ stands for the Frobenius norm.

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026

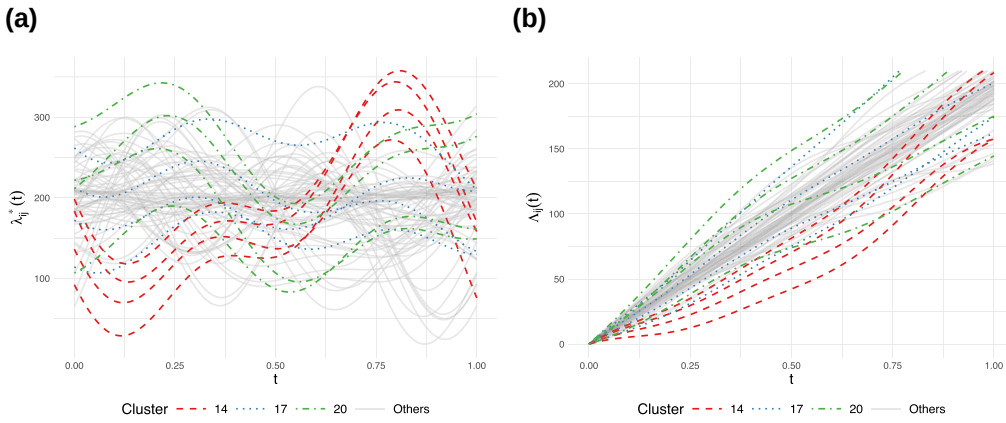

Figure 1. Simulated intensities $\lambda^*_{ij}(t)$ (a) and cumulative hazards $\Lambda_{ij}(t)$ (b), with clusters 14, 17, and 18 highlighted using different colours and line types due to their outlying shapes, which stand out from the general patterns observed in other clusters. (a) $\lambda^*_{ij}(t)$. (b) $\Lambda_{ij}(t) = \int_0^t \lambda^*_{ij}(u)\,\mathrm{d}u$.

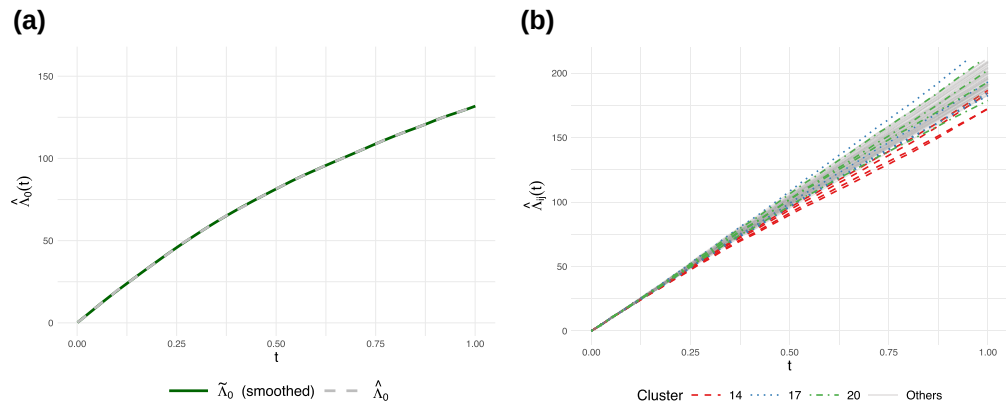

Figure 2. Estimated $\hat{\Lambda}_0(t)$ (a) and $\hat{\Lambda}_{ij}(t)$ (b). In (b), clusters 14, 17, and 18 are highlighted using different colours and line types due to their outlying shapes, which stand out from the general patterns observed in other clusters. (a) $\hat{\Lambda}_0(t)$. (b) $\hat{\Lambda}_{ij}(t)$.

3.2 Results

3.2.1 Compensator estimation

We apply the pipeline described in Section 2.1, where in the AG model in Eq. (1) we employ the number of events recorded up to that time interval for unit j in cluster i as a time-dependent covariate.

The baseline cumulative hazard, $\hat{\Lambda}_0(t)$, along with $\tilde{\Lambda}_0(t)$ is estimated. Additionally, $\hat{\Lambda}_{ij}(t)$ is derived according to Eq. (2). These functions are depicted in Figure 2 for a single replicate of the event time generation, in the baseline scenario, where the x-axis represents time t and the y-axis reports $\hat{\Lambda}_0(t)$ (panel (a)) and $\hat{\Lambda}_{ij}(t)$ (panel (b)). By visual inspection, we observe some information loss due to the stochastic nature of NHPPs in the simulation of event times. Nonetheless, cluster behaviours remain distinguishable and can still be effectively recognised and characterized.

3.2.2 Multilevel functional principal component analysis

As final step of our simulation study, we implement the decomposition described in Section 2.2. We recall that in Eq. (7) we simulate intensities $\lambda^*_{ij}(t)$ employing the eigenfunctions in Eqs.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

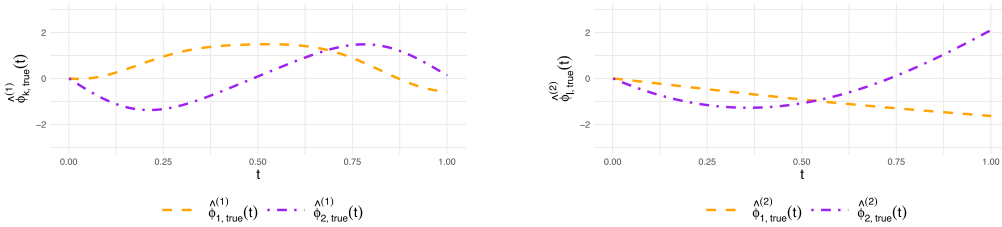

Figure 3. First two eigenfunctions at levels 1 and 2 (left and right panels, respectively), computed from simulated $\Lambda_{ij}(t)$ represented in Figure 1 (b).

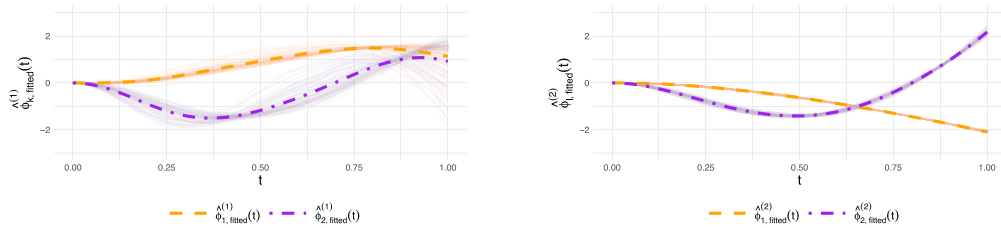

Figure 4. Mean of the two estimated eigenfunctions at levels 1 and 2 (left and right panels, respectively), computed from the reconstructed $\hat{\Lambda}_{ij}(t)$. The solid curves represent the mean eigenfunctions averaged across 100 simulation replicates, while the thin transparent lines correspond to the eigenfunctions estimated in each individual replicate.

(8)–(9). However, our primary interest lies in $\Lambda_{ij}(t)$. Therefore, we first decompose $\Lambda_{ij}(t)$ using Eq. (5), to get $\hat{\phi}^{(1)}_{k,\text{true}}$ and $\hat{\phi}^{(2)}_{l,\text{true}}$. At this stage, we obtain 4 functional principal components at level 1 and 2 at level 2. To determine the number of principal components at both levels, we set the PVE to 0.99, following the default setting in the `mfpca.face` function from Cui et al. (2023). In Figure 3, we show results of MFPCA on functional compensators related to the first and second principal components in the baseline scenario, respectively, for levels 1 and 2. Specifically, the x-axis represents time t, while the y-axis shows the estimated eigenfunctions $\hat{\phi}^{(1)}_{k,\text{true}}(t)$ (left panel) and $\hat{\phi}^{(2)}_{l,\text{true}}(t)$ (right panel).

On the other hand, after the simulation of the NHPP as described in previous section and having computed $\hat{\Lambda}_{ij}(t)$ according to Eq. (2), we apply the same multilevel functional decomposition to $\hat{\Lambda}_{ij}(t)$, to get $\hat{\phi}^{(1)}_{k,\text{fitted}}$ and $\hat{\phi}^{(2)}_{l,\text{fitted}}$. This step is replicated 100 times, to better measure the variability due to the event time simulations. Here, across the 100 replicates of the event time generation in the baseline scenario, we consistently obtain 3 functional principal components at level 1 and 2 at level 2, and the magnitude of the eigenvalues reduce. If we analyse the eigenfunctions $\hat{\phi}^{(1)}_{k,\text{fitted}}(t)$ and $\hat{\phi}^{(2)}_{l,\text{fitted}}(t)$ shown on the y-axis of Figure 4, respectively in the left and right panels, with the x-axis representing time t, similar pattern can be observed, both for levels 1 and 2. Moreover, the estimated eigenfunctions in Figure 4 seem to successfully recover the main temporal features of the true eigenfunctions shown in Figure 3. At level 1, the first eigenfunction exhibits an inverted U-shaped pattern with a mid-interval peak, while the second captures a negative sinusoidal behaviour. At level 2, the first eigenfunction reproduces a linear downward trend, and the second shows a pronounced increase toward the end of the interval. Overall, the reconstructed compensators closely match the original simulated shapes up to sign alignment, despite minor discrepancies arising from event times sampling and model fitting. In particular, the principal trends, as well as the locations of maxima and minima, are well preserved after reconstruction, although small phase shifts are observed, slightly displacing peaks to the right in the reconstructed eigenfunctions of Figure 4. Furthermore, we notice that $\hat{\phi}^{(1)}_{2,\text{fitted}}$ is the eigenfunction that accumulates the most variability over the 100 replications, where the noise is to be imputed to the event times generation.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

Table 1 presents the simulation results for various combinations of J_i, I and T, under both balanced and unbalanced designs. For each configuration, the mean, standard deviation (SD), and median of the MISE($\hat{\phi}^{(1)}$) and MISE($\hat{\phi}^{(2)}$) are reported across 100 replications. The results show that the MISE decreases noticeably as I increases, indicating that having more groups leads to more accurate estimation due to the larger amount of independent information. In contrast, the effect of increasing J_i is less systematic, suggesting that additional within-group observations contribute less to improving estimation accuracy, with the MISE in fact tending to increase slightly as J_i increases. The impact of T (number of time points) is generally smaller, suggesting that the estimators stabilize at moderate T; this result is in line with those in Table 3 of fast-MFPCA paper (Cui et al., 2023). Across all configurations, $\hat{\phi}^{(2)}$ consistently achieves lower MISE values and smaller variability than $\hat{\phi}^{(1)}$, confirming that most of the randomness induced by the thinning process is mainly captured by the first level. As expected, the balanced design generally yields slightly better performance than the unbalanced one, although the differences diminish for large I and J_i.

Overall, the results presented in this section indicate that cumulative hazard reconstruction via AG models preserves the essential information captured by MFPCA, demonstrating that for a NHPP with cluster-similar intensities, the cumulative hazard reconstruction using AG models effectively retrieves the simulated shapes, retaining the crucial information captured by MFPCA prior to process extraction.

4 Dataset

The data employed in this study were obtained from the administrative records of PoliMi, which collect the academic progress of students enrolled in bachelor's degree programmes (Mussida & Luca Lanzi, 2022). These records encompass various aspects of students' academic *careers*, including enrolment and end-of-study dates, any changes in their enrolled degree programmes, and eventually incidents of dropout.[4] Additionally, information regarding the student's history of passed and attempted *exams* at different time points (semesters) is contained, including credits earned within the European Credit Transfer and Accumulation System (ECTS) and weighted Grade Point Average (GPA). While we are aware that psychological and social factors such as self-confidence, motivation, parental education and family support are known to have impact on dropout behaviour, as highlighted in Rumberger (2011) and Tinto (1975), our analysis is constrained to the available administrative records, which do not include such information. As a result, our study focuses on academic and institutional factors that can be directly observed and measured within the university's data infrastructure. In this section, we delineate the cohort selection criteria for our study (Subsections 4.1) alongside the study design (Subsection 4.2).

4.1 Cohort selection

For our analysis, we focus on bachelor's students enrolled in academic years 2016/2017 and 2017/2018, who maintained a consistent degree programme throughout their academic paths. We assume that students enrolled in the 2017/2018 academic year were only marginally impacted by Covid-19, which occurred during the last semester of their final year. Moreover, we omit fully remote or single-cycle degree programmes, as the analysis focuses solely on traditional bachelor's degrees, which typically last between three and three and a half years.

In the first phase of the analysis, we utilize data from students with `career_start_ay` = '2016' to construct the compensators. For these students, we track the dropout events occurring within each `course` and `school` during a period $S = [T_0, T_1]$ that approximately corresponds to the first three semesters since enrolment. In the second phase, we shift to the '2017' cohort, using data from the end of the first semester and historical information to enhance time-to-dropout predictions. The selection of the main variables to be included is informed by prior research, such as Masci et al. (2024). Table 2 provides an overview of the key variables used in the second phase, organized into four categories:

[4] At PoliMi, students can withdraw at any point during their academic journey, rather than being restricted to doing so only at the time of enrolment renewal, as is common in some other universities.

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

Table 1. Simulation results for different I, J_i and T in terms of Mean, Standard Deviation (SD) and Median across 100 replications, for balanced and unbalanced designs

I	J_i	T	Balanced Design						Unbalanced Design					
			$\text{MISE}(\hat{\phi}^{(1)})$			$\text{MISE}(\hat{\phi}^{(2)})$			$\text{MISE}(\hat{\phi}^{(1)})$			$\text{MISE}(\hat{\phi}^{(2)})$		
			Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
10	4	200	1.497	0.250	1.462	0.072	0.022	0.069	1.496	0.174	1.517	0.049	0.011	0.048
		1,000	1.458	0.280	1.443	0.073	0.022	0.070	1.495	0.176	1.524	0.050	0.011	0.047
		5,000	1.445	0.313	1.435	0.073	0.022	0.070	1.492	0.170	1.520	0.050	0.011	0.048
	10	200	1.377	0.307	1.359	0.071	0.011	0.070	1.402	0.368	1.497	0.079	0.011	0.077
		1,000	1.409	0.358	1.476	0.072	0.011	0.071	1.346	0.403	1.449	0.080	0.011	0.080
		5,000	1.408	0.370	1.472	0.072	0.011	0.071	1.343	0.396	1.444	0.081	0.011	0.080
	50	200	1.556	0.329	1.596	0.086	0.005	0.086	0.852	0.134	0.821	0.081	0.004	0.081
		1,000	1.412	0.362	1.453	0.088	0.005	0.088	0.818	0.082	0.813	0.082	0.004	0.082
		5,000	1.405	0.363	1.422	0.088	0.005	0.088	0.818	0.082	0.812	0.082	0.004	0.083
30	4	200	0.900	0.099	0.916	0.111	0.031	0.125	0.453	0.086	0.464	0.092	0.017	0.095
		1,000	0.896	0.096	0.910	0.113	0.030	0.126	0.448	0.085	0.457	0.093	0.017	0.096
		5,000	0.895	0.095	0.911	0.112	0.030	0.124	0.448	0.085	0.457	0.093	0.018	0.097
	10	200	1.290	0.139	1.304	0.114	0.012	0.116	0.703	0.049	0.707	0.085	0.010	0.086
		1,000	1.295	0.138	1.311	0.114	0.012	0.115	0.704	0.049	0.708	0.086	0.012	0.088
		5,000	1.296	0.137	1.313	0.114	0.012	0.115	0.704	0.049	0.708	0.086	0.012	0.088
	50	200	1.189	0.024	1.190	0.112	0.003	0.112	0.552	0.012	0.552	0.131	0.003	0.131
		1,000	1.182	0.023	1.183	0.112	0.003	0.112	0.550	0.012	0.550	0.132	0.003	0.132
		5,000	1.181	0.023	1.181	0.112	0.003	0.112	0.549	0.012	0.550	0.133	0.003	0.133

(*continued*)

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

Table 1. Continued

I	J_i	T	Balanced Design						Unbalanced Design					
			MISE($\hat{\phi}^{(1)}$)			MISE($\hat{\phi}^{(2)}$)			MISE($\hat{\phi}^{(1)}$)			MISE($\hat{\phi}^{(2)}$)		
			Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
50	4	200	0.464	0.048	0.467	0.139	0.027	0.147	0.634	0.030	0.631	0.078	0.020	0.072
		1,000	0.452	0.046	0.453	0.138	0.027	0.146	0.635	0.028	0.632	0.083	0.026	0.073
		5,000	0.449	0.046	0.450	0.138	0.027	0.146	0.635	0.028	0.633	0.086	0.028	0.073
	10	200	0.741	0.037	0.746	0.072	0.020	0.067	0.838	0.029	0.839	0.112	0.026	0.123
		1,000	0.730	0.037	0.734	0.075	0.026	0.067	0.832	0.029	0.834	0.118	0.022	0.125
		5,000	0.727	0.036	0.731	0.074	0.026	0.066	0.830	0.029	0.831	0.118	0.022	0.125
	50	200	0.849	0.011	0.850	0.067	0.007	0.067	1.099	0.008	1.099	0.058	0.001	0.058
		1,000	0.838	0.011	0.839	0.068	0.012	0.066	1.103	0.008	1.103	0.058	0.001	0.058
		5,000	0.835	0.011	0.836	0.067	0.012	0.065	1.103	0.008	1.103	0.058	0.001	0.058

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

Table 2. Overview of the variables considered in the analysis.

Variable	Description	Type
Measured at enrolment		
- studentID	Student's unique identifier (anonymized)	Categorical {1,2,...}
- origins	Student's geographic origins	Categorical {Onsite, Commuter, Offsite}
- gender	Student's gender	Categorical {Male, Female}
- highschool_type	Type of attended high school	Categorical {Scientific, Classical, Others, Technical}
- highschool_grade	Student's normalized high school grade	Real number [60, 100]
- income	University fee bracket	Categorical {Medium, Grant, High, Low}
- age19	Equals 1 if student's age at enrolment > 19	Categorical {0,1}
- admission_score	PoliMi entrance test's admission score	Real number [60, 100]
- career_start_ay	Student's enrolment year	Categorical {2016, 2017}
Measured at end of 1st semester		
- ECTS1sem	ECTS gained by end of 1st semester	Natural number >0
Grouping factors		
- course	Undergraduate programme	Categorical {P01, P02,..., P23}
- school	A larger organizational unit grouping courses	Categorical {sA, sB, sC, sD}
Outcome		
- time	Time at which a student drops out after 1st sem	Real number [1,6]
- dropout	Equals 1 if after 1st semester a student drops within 6 sem, 0 otherwise	Categorical {0, 1}

Note. in categorical variables, the first reported class represents the reference level.

- Variables measured at enrolment capture essential demographic and background characteristics. These include geographic origin (origins, i.e. whether a student lives onsite, offsite, or commutes to Milan), gender (gender), and age at enrolment (age19, a binary variable which equals 1 for students older than 19, and 0 otherwise). Socio-economic status is approximated by the university fee bracket (income), which classifies students based on their family's financial situation into categories such as low, medium, high, or those receiving grants. Educational background is represented by the type of high school attended (highschool_type) and the obtained highschool grade (normalized) (highschool_grade). Lastly, the PoliMi admission test score (admission_score) reflects academic readiness at the time of university entry, although students may take this test up to a year prior to their actual enrolment.
- Variables measured at the end of the first semester focus on academic progress, particularly the number of credits earned (ECTS1sem), a key predictor of dropout risk.
- Grouping factors include the undergraduate programme (course) and broader organizational structure (school).
- The time-to-event outcome variables pair (time, dropout) indicate whether a student dropped out and, if so, the time at which the dropout occurred within three years after the first

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026

Table 3. Descriptive statistics for considered covariates after data pre-processing for `career_start_ay` = '2017', according to the dropout status.

Variable			`dropout` = 0	`dropout` = 1
Type	**Name**		**Mean (sd)**	**Mean (sd)**
Numerical	`admission_score`		66.99 (11.49)	63.52 (11.65)
	`highschool_grade`		83.90 (11.61)	76.62 (11.15)
	`ECTS1sem`		19.49 (9.87)	4.67 (7.55)
		Category	**N (Frequency)**	**N (Frequency)**
Categorical	`origins`	`Onsite*`	1031 (21.54%)	234 (26.53%)
		`Commuter`	3422 (71.49%)	590 (66.89%)
		`Offsite`	334 (6.97%)	58 (6.58%)
	`gender`	`Male*`	3197 (66.79%)	678 (76.87%)
		`Female`	1590 (33.21%)	204 (23.13%)
	`highschool_type`	`Scientific*`	3175 (66.33%)	546 (61.91%)
		`Classical`	326 (6.81%)	54 (6.12%)
		`Others`	607 (12.68%)	93 (10.54%)
		`Technical`	679 (14.18%)	189 (21.43%)
	`income`	`Medium*`	986 (20.60%)	125 (14.17%)
		`Grant`	1404 (29.33%)	287 (32.54%)
		`High`	1766 (36.89%)	388 (43.99%)
		`Low`	631 (13.18%)	82 (9.30%)
	`age19`	`0*`	4267 (89.14%)	681 (77.21%)
		`1`	520 (10.86%)	201 (22.79%)

*Reference category.

semester. The variable `time` is measured in semesters, starting from the end of the first semester, and ranges from 1 to 6, with three decimal places indicating the fraction of the semester at which the event occurred. If a student does not drop out or graduate within the six-semester window (three years following the end of the first semester), we assign `time` a value of six and `dropout` a value of zero, treating the observation as right-censored at six semesters (measured from the conclusion of the first semester).

4.2 Study design

With focus on students enrolled in `career_start_ay` = '2016', we include a three-semesters *observation period*, denoted by $S = [T_0, T_1]$, with T_0 = '1 October of`career_start_ay`' and T_1 = '1 March of (`career_start_ay` + 2)'. The date of October 1 is chosen to exclude students who dropped out within the first two weeks, potentially due to waiting for other university entrance test results,[5] to ensure they do not affect the analysis. Instead, March 1 is selected as the closing date, since at PoliMi the first-semester exams are generally completed by this time and the new lectures have just begun. Moreover, we focus exclusively on the first three semesters (516 days, approximately 17 months) because this period is highly predictive of dropout risk, as seen in previous works (Cannistrà, 2024; Cannistrà et al., 2022; Masci et al., 2024). Additionally, analyzing this time frame allows us to move the cohort into the subsequent academic year and perform predictions starting from the end of the first semester, avoiding overlap between cohorts. This ensures that future information is not used, ensuring practical implementation of the proposed methodology. During this period, dropouts of students enrolled in the chosen `career_start_ay`, `course` and `school` are daily monitored. We thus create a longitudinal dataset, in a counting process fashion, tracking student dropout events across different `courses` and `schools`. Each row represents a time interval between consecutive dropout events, where the start and stop columns indicating the time (in days) between these events.

[5] At PoliMi, lectures typically begin in mid-September.

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026

Following this, the focus is thus moved to the cohort of students with `career_start_ay` = '`2017`', particularly at the end of the first semester, and the primary outcome of interest is the couple of variables (`time`, `dropout`). To predict this outcome, we employ data collected at enrolment, at the end of the first semester, and at the level of grouping factors, as shown in Table 2. The choice is based on previous studies' results (Masci et al., 2024), which indicate that the optimal prediction window occurs within the first few semesters, as the inclusion of data from later semesters provides minimal improvement in accuracy. Following preprocessing as described in previous Section 4.1, our dataset consists of 5669 students, of which 882 dropped out. Descriptive statistics for the over-described covariates after data pre-processing, are reported in Table 3, according to the dropout status. Moreover, the model will incorporate information derived from the analysis of the previous academic year's data, adding valuable historical context to enhance prediction.

5 Case study

We apply the proposed methodology to the case study involving the administrative dataset of PoliMi. First, we present the compensator reconstruction and decomposition at degree `course` and `school` levels (Subsection 5.1), then effectively implement a predictive model at `studentID`-level (Subsection 5.2).

5.1 Compensator reconstruction and decomposition

For the analysis of dropouts as a marked point process, after cohort selection described in Subsection 4.1 and having filtered the data to focus on a specified academic year (in our case, `career_start_ay` = '2016'), we establish start and stop dates for each dropout event that happened on distinct days and enumerate the cumulative occurred dropout distinct days (`enum`), as well as the number of events (`dropout_count_std`) standardized by the number of students remained enrolled in that course, that will perform as the mark of the counting process. Afterwards, these covariates are employed for the AG model for recurrent events describing the dropouts, with $\hat{\theta}_{\texttt{enum}} = 0.018$ (p-value = 0, HR = 1.021 [1.014; 1.029]) and $\hat{\theta}_{\texttt{dropout\_count\_std}} = 0.094$ (p-value = 0, HR = 1.099 [1.071; 1.128]), where HR stands for the hazard ratio and the values in square brackets represent the associated 95% confidence interval.

The compensators are then reconstructed as described in Eq. (2). In Figure 5, we display the baseline cumulative hazard $\hat{\Lambda}_0(t)$ (panel (a)) and the reconstructed compensators $\hat{\Lambda}_{ij}(t)$ (panel (b)), on two different scales, with the x-axis representing time t measured in months over $S = [T_0, T_1]$. The behaviour of the curves is notable: there is a steep increase in dropout counts at the beginning and end of the first year, particularly pronounced for specific degree courses. This pattern can be explained by several factors. Early in the first year, high dropout rates are often observed as students realize that the degree course they have chosen does not meet their expectations, leading them to switch programmes or drop out. Additionally, many dropouts may occur by the end of the first year because students find the coursework too challenging or the degree programme not aligned with their career aspirations. This combination of early and end-of-year dropouts contributes to the distinct peaks observed in the cumulative hazard curves. Notable is the case of a degree course in school `sC`. Although the pointwise distributions of $\hat{\Lambda}_{ij}(t)$ exhibit slight right-skewness, the AG model accounts for the number of events standardized by the number of students remaining enrolled in each course, already mitigating much of this heterogeneity. A log-transformation could, in principle, further reduce this residual skewness before applying MFPCA; however, we retained the original (untransformed) scale to preserve interpretability. This choice does not adversely affect the stability of the decomposition of Eq. (4) and ensures optimal predictive performance of the model fitted in Section 5.2.

Afterwards, the compensators are decomposed as in Eq. (4). The number of principal components at both levels is chosen by setting a proportion of variance explained equal to 0.99 (default). As a result, two principal components are retained at both levels. At the higher hierarchical level (denoted as level 1, corresponding to the `school`), the eigenvalues obtained are $\hat{\lambda}_1^{(1)} = 89.007$ and $\hat{\lambda}_2^{(1)} = 0.900$, explaining, respectively, the 98.99% and 1.00% of variability, while at the lower hierarchical level (level 2, corresponding to the `course`), the eigenvalues are $\hat{\lambda}_1^{(2)} = 61.467$ and

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

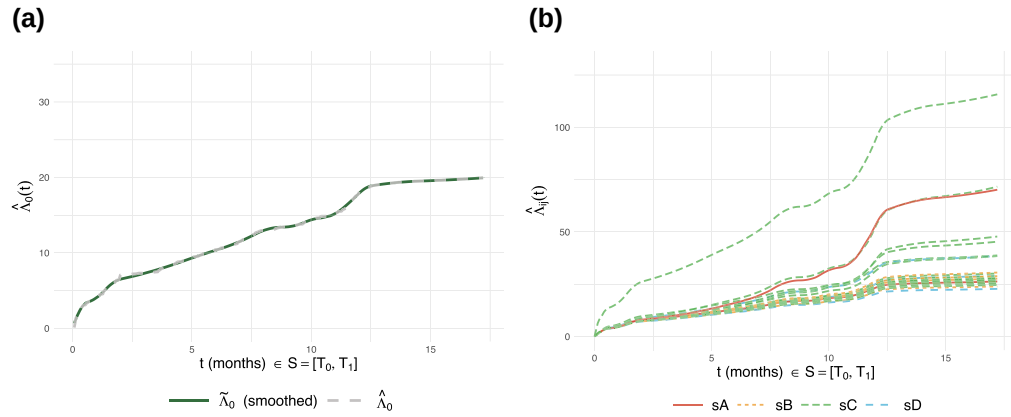

Figure 5. In panel (a) we show the baseline cumulative hazard of the AG model for recurrent events for the marked stochastic processes describing the dropouts. In panel (b), we represent the reconstructed compensators as in (2) of the latter processes, each line representing a different degree `course` and each colour and line type representing a different `school`, as indicated in legend. (a) Baseline cumulative hazard $\hat{\Lambda}_0(t)$. (b) Reconstructed compensators $\hat{\Lambda}_{ij}(t)$.

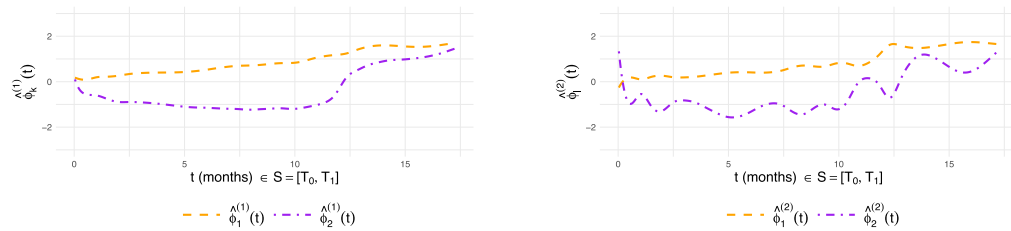

Figure 6. First two eigenfunctions at levels `school` and `course` levels (left and right panels, respectively), computed from the obtained $\hat{\Lambda}_{ij}(t)$.

$\hat{\lambda}_2^{(2)} = 1.901$, explaining, respectively, 95.93% and 3.00% of variability. Furthermore, the proportion of variability explained by the higher hierarchical level, defined in Di et al. (2009) as $\sum_{k=1}^{\infty} \hat{\lambda}_k^{(1)} / (\sum_{k=1}^{\infty} \hat{\lambda}_k^{(1)} + \sum_{l=1}^{\infty} \hat{\lambda}_l^{(1)})$, equals 58.39%.

Figure 6 illustrates the first and second eigenfunctions for each of the two levels, with the x-axis representing time t over S and the y-axis showing the corresponding estimated eigenfunctions, $\hat{\phi}_k^{(1)}(t)$ and $\hat{\phi}_l^{(2)}(t)$, displayed in the left and right panels for the first and second components, respectively. To improve interpretability, as suggested in Ramsay and Silverman (2005), all panels in Figure 7 show the mean compensator functions $\hat{\mu}(t)$ (solid line) on the y-axis along with perturbation curves (dashed lines for positive perturbations and dot-dashed lines for negative) based on the MFPCA performed on $\hat{\Lambda}_{ij}(t)$, representing the eigenfunctions within one or three standard deviation (i.e. the square roots of the eigenvalues) from the mean, respectively for first and second level. Indeed, $\hat{\lambda}_1^{(i)}$ is much higher than $\hat{\lambda}_2^{(i)}$, for both $i = 1, 2$. This difference in magnitude is reflected in Figure 7 with respect to the mean compensator function $\hat{\mu}(t)$. Moreover, the second principal component for level 2 exhibits significant variability, which may be attributed to the estimation process involving B-spline basis functions in the MFPCA framework. These basis functions can potentially introduce artificial fluctuations, especially in higher-order components. On the other hand, when the proportion of explained variance is reduced to 0.9, only a single principal component is retained at each level, suggesting that most of the relevant dropout dynamics are captured by the first component alone.

The distribution of dropouts over time reveals distinct patterns across `schools` and degree `courses`, each associated with varying dropout risks. As a contribution to the existing literature,

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

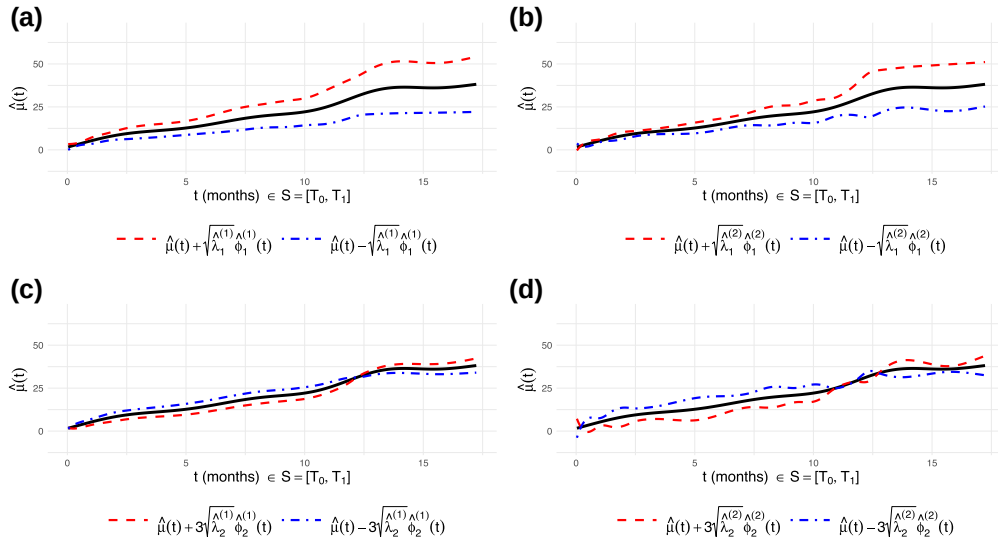

Figure 7. Average compensators curves $\hat{\mu}(t) = \frac{1}{n}\sum_{i,j}\hat{\Lambda}_{ij}(t)$ and their perturbations as indicated in legends, for 1^{st} principal component for `school` level in left panel (a) and `course` level in right panel (b), and 2^{nd} principal component for `school` level in left panel (c) and `course` level in right panel (d). (a) 1^{st} PC level 1 perturbation of $\hat{\mu}(t)$. (b) 1^{st} PC level 2 perturbation of $\hat{\mu}(t)$. (c) 2^{nd} PC level 1 perturbation of $\hat{\mu}(t)$. (d) 2^{nd} PC level 2 perturbation of $\hat{\mu}(t)$.

our analysis aims to disentangle the effects of schools from those of degree programmes. As result, the first eigenfunctions (represented by the dashed lines in Figure 6) at both hierarchical levels are positive and increasing, implying a deviation from the average dropout intensity (represented by the solid line in Figure 7) along that eigenfunction's shape, according to the sign of the computed scores. Specifically, this means that schools and degree courses with a positive score on the first principal component (dashed lines in Figure 7) are likely to experience a higher-than-average dropout rate, while those with a negative score (dot-dashed lines in Figure 7) are likely to see fewer dropouts than average. Interestingly, the eigenfunction shape capturing dropout patterns differ between the `school` and `course` levels: at the `school` level, there is a smoother increase toward $t \approx 13$ months, while at the `course` level, small earlier peaks emerge around $t \approx 8$, 10 months and another sharper one around $t \approx 12$ months. A possible explanation for these patterns lies in the exam schedule at PoliMi. For the first-semester classes, students have two exam sessions in January–February, two more in June–July, and a final opportunity in September. Similarly, for the second-semester classes, there are two exam sessions in June–July, one in September, and two more in January–February of the following year. As a result, students typically have several chances to pass exams over the year. However, if students fail to pass the majority of their exams by September, they may decide to drop out without renewing their enrolment and paying the new instalment of the tuition fees. The second eigenfunctions at both hierarchical levels (represented by the dot-dashed lines in Figure 6), though associated with less explained variance and thus potentially more susceptible to estimation uncertainty, highlight additional temporal contrasts. At the `school` level, those students enrolled in schools with a positive score (dashed curve in Figure 7) tend to experience fewer dropouts during the first $t \approx 12$ months but more dropouts in the third semester. A similar pattern emerges at the `course` level, though with greater oscillations, suggesting that some noise may also be captured. These patterns may reflect periods when students enrolled in certain schools or courses encounter particularly challenging exams or key academic milestones, which can lead to concentrated dropout events at specific stages of their studies.

5.2 Prediction: Cox regression model with functional compensators

Given results in previous section, we now include information derived from compensators into a predictive model, having first filtered the data to focus on `career_start_ay` =

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

'2017', and considered variables shown in Table 2 and preprocessed in Table 3, as described in Section 4.

We aim to model the time-to-dropout up to 3 years after the first semester, using covariates at the `studentID`-level and functional principal component scores derived from the cumulative hazard of historical dropouts over time. We consider two principal components ($K = 2$) at the `school` level, and two principal components ($L = 2$) at the `course` level, even though the second principal components carry low explained variability. The time-to-event data (T_{ijh}, δ_{ijh}) correspond to the couple (`time`, `dropout`) described earlier; the functional Cox model is thus the one in Eq. (6). The vector of covariates $\mathbf{w}_{ijh}$ at the `studentID`-level includes demographic and academic information that could influence dropout risk, i.e. `origins`, `gender`, `highschool_type`, `highschool_grade`, `income`, `age19`, `admission_score`, `ECTS1sem`.

We fit the Cox regression model using a stepwise variable selection procedure based on the Akaike Information Criterion (AIC) Bozdogan (1987), employing bidirectional selection. Indeed, AIC-based selection is used in the functional regression literature and serves as an effective alternative to thresholding the percentage of variance explained when determining the number of retained FPC scores in estimation. Existing simulation studies and theoretical results provide a general justification for its effectiveness (Kong et al., 2018; Yao et al., 2005).

The reduced model retains the variables `origins`, `highschool_type`, `income`, `highschool_grade`, `ECTS1sem` and the principal components except for the second at the `course` level. The results are further validated by fitting a Lasso-penalized (Tibshirani, 1996) Cox regression model. The exclusion of the second principal component at the `course` level is consistently supported by this alternative approach. Following notation in Eq. (6) explicit formulation of fitted model is

$$
\begin{aligned}
&\eta_{ijh}\big(t \mid [\texttt{origins}, \texttt{highschool\_type}, \texttt{income}, \texttt{highschool\_grade}, \texttt{ECTS1sem}], \hat{\Lambda}_{ij}\big) \\
&\quad = \eta_0^*(t) \exp\Big\{\gamma_1 \mathbb{I}(\texttt{origins} = \texttt{Commuter}) \\
&\qquad + \gamma_2 \mathbb{I}(\texttt{origins} = \texttt{Offsite}) + \gamma_3 \mathbb{I}(\texttt{highschool\_type} = \texttt{Classical}) \\
&\qquad + \gamma_4 \mathbb{I}(\texttt{highschool\_type} = \texttt{Others}) + \\
&\qquad + \gamma_5 \mathbb{I}(\texttt{highschool\_type} = \texttt{Technical}) + \gamma_6 \mathbb{I}(\texttt{income} = \texttt{Grant}) + \gamma_7 \mathbb{I}(\texttt{income} = \texttt{High}) + \\
&\qquad + \gamma_8 \mathbb{I}(\texttt{income} = \texttt{Low}) + \gamma_9 \texttt{highschool\_grade} + \gamma_{10} \texttt{ESCS1sem} + \sum_{k=1}^{2} \xi_{ik} \alpha_k^{(1)} + \sum_{l=1}^{1} \zeta_{ijl} \alpha_l^{(2)} \Big\},
\end{aligned}
\tag{10}
$$

being $\mathbb{I}(\cdot)$ the indicator function. The estimated coefficients obtained from the analysis are presented in Table 4.

The obtained results align closely with expectations. Specifically, earning a higher number of ECTS credits in the first semester and achieving a higher high school grade significantly reduce the risk of dropout within the first three years of a bachelor's programme (first semester excluded). This finding is consistent with previous research, such as Masci et al. (2024), which underscores the strong predictive power of first-semester credit accumulation in determining dropout risk. Conversely, students from 'Others' and 'Technical' high schools, as well as high-income students, exhibit a higher likelihood of dropping out. The model selection led to the exclusion of the `admission_score` variable, as its effect on dropout risk is likely mediated by the number of ECTS credits earned in the first semester (`ECTS1sem`).

Notably, all coefficients associated with the retained score variables (two at the `school` level and one at the `course` level) are statistically significant. This confirms that the information captured by compensators and multilevel principal component analysis effectively enhances time to dropout prediction. In particular, the positive coefficient $\hat{\alpha}_1^{(2)}$ indicates that students enrolled in courses with high scores on the first principal component at the `course`-level face an increased risk of dropping out within three years. This result aligns with the trend observed in Figure 7b, suggesting that courses characterized by above-average dropout rates are associated with a higher probability of students leaving their programmes. On the other hand, both principal component

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

Table 4. Estimates, standard errors, and *P*-values for the Cox regression model with functional compensators defined in Eq. (10)

Parameter	Estimate	exp(Estimate)	Std. Error	*p*-value
$\hat{\gamma}_1$ (origins = Commuter)	0.099	1.104	0.078	0.206
$\hat{\gamma}_2$ (origins = Offsite)	−0.215	0.807	0.151	0.155
$\hat{\gamma}_3$ (highschool_type = Classical)	−0.030	0.970	0.146	0.836
$\hat{\gamma}_4$ (highschool_type = Others)	0.280	1.324	0.128	0.028
$\hat{\gamma}_5$ (highschool_type = Technical)	0.176	1.192	0.088	0.045
$\hat{\gamma}_6$ (income = Grant)	0.014	1.014	0.108	0.896
$\hat{\gamma}_7$ (income = High)	0.256	1.292	0.103	0.013
$\hat{\gamma}_8$ (income = Low)	−0.123	0.884	0.143	0.391
$\hat{\gamma}_9$ (highschool_grade)	−0.007	0.993	0.003	0.040
$\hat{\gamma}_{10}$ (ECTS1sem)	−0.145	0.865	0.005	0.000
$\hat{\alpha}_1^{(1)}$ (ξ_{i1} - school-level score 1)	−0.020	0.980	0.005	0.000
$\hat{\alpha}_2^{(1)}$ (ξ_{i2} - school-level score 2)	−0.885	0.413	0.172	0.000
$\hat{\alpha}_1^{(2)}$ (ζ_{ij1} - course-level score 1)	0.006	1.006	0.003	0.027

coefficients at the school level are negative and statistically significant. In particular, the coefficient $\hat{\alpha}_2^{(1)}$ suggests that students from schools with relatively high dropout rates in the first year but lower-than-average dropout rates by the third semester are less likely to drop out, as shown in Figure 7c. This protective effect is consistent with the construction of compensators, which are based on the first three semesters, while dropout predictions are made from the end of the first semester for the subsequent cohort. The sign of the coefficient $\hat{\alpha}_1^{(1)}$ may appear counterintuitive; however, this could be due to its potential overlap with the information captured by $\hat{\alpha}_1^{(2)}$. As a result, its weight in time-to-dropout prediction may reflect underlying similarities between these components.

5.2.1 Model evaluation and comparison against natural alternatives

The selected model, whose results are reported in Table 4, achieves a loglikelihood of −6717.687, a BIC (Schwarz, 1978) of 13523.54, an AIC of 13461.37 and a concordance index[6] of 0.847.

With the aim of quantifying uncertainty and examining the effect of the first estimation step on the second, we apply a nonparametric bootstrap. This approach aims at better characterizing the main distributions and drivers of the second-phase outcomes by resampling in the first phase. Specifically, in the first phase of the analysis, which focuses on career_start_ay = '2016', we sample with replacement within clusters from the original administrative dataset, which contains information on whether students dropped out, before aggregating to the course level. We repeat the estimation procedure 1000 times to assess the stability and variability of our estimates. The results are summarized through boxplots in Figure 8. Specifically, the *y*-axis shows the bootstrap values, with separate panels corresponding to different components and levels. These plots indicate that, even though the actual dropout counting process is not considered as bootstrap is applied, the proportion of variability across components and levels (Figure 8a) remains coherent with the non-bootstrapped case study results (red-dashed lines), even though the second component at level 2 receives slightly less weight than the first. The same holds when we move forward fitting the Cox regression model in the final predictive step, and we analyse the C-index and the coefficients for school and course-level scores in Figure 8b. However, we observe some noise

[6] The concordance index, or C-index (Harrell et al., 1982), is a commonly used method adapted for time-to-event models and also accounting for censored data (Harrell's C-index). It measures a model's ability to correctly provide a reliable ranking of the survival times based on the individual risk scores. It ranges from 0.5 (indicating random prediction) to 1 (perfect prediction) and quantifies the probability that, for a randomly selected pair of comparable units, the one with the higher predicted risk experiences the event earlier.

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026

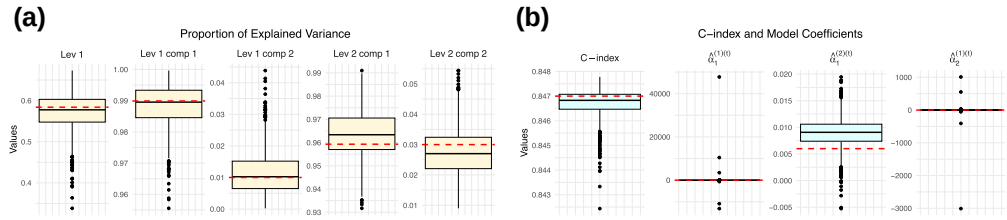

Figure 8. Results of the nonparametric bootstrap procedure across 1000 resamples. Boxplots represent the distribution of (a) level- and component-specific proportion of explained variance and (b) the Cox model outputs, including the C-index and selected coefficients (`school` and `course`-level scores). Red-dashed lines indicate the estimates from the original, non-bootstrapped case study.

Table 5. Model comparison metrics for the proposed model of Table 4 and simpler alternatives.

Model	log-Likelihood	*p*-value	BIC	AIC	C-index
Benchmark (Table 4)	**−6717.687**		**13523.54**	**13461.37**	**0.847**
a - Without both $\sum_{k=1}^{K} \xi_{ik}\alpha_k^{(1)}$ and $\sum_{l=1}^{L} \zeta_{ijl}\alpha_l^{(2)}$	−6742.974	0	13553.77	13505.95	0.842
b - Only `school`-level $\sum_{k=1}^{K} \xi_{ik}\alpha_k^{(1)}$	−6723.302	0	13514.43	13466.60	0.846
c - Only `course`-level $\sum_{l=1}^{L} \zeta_{ijl}\alpha_l^{(2)}$	−6731.864	0	13545.11	13487.73	0.845
d - Only `school`-level random-effects	−6739.671	—	13565.71	13504.81	0.845
e - Only `course`-level random-effects	−6718.186	—	13605.80	13486.33	0.846

Note. Metrics reported include log-likelihood, AIC, concordance index, and *p*-values of likelihood ratio tests (relative to simpler models).

in the model coefficients compared to Table 4 and lower median C-index value, reflecting the expected information loss introduced by the bootstrap procedure.

To contextualize the performance of the model proposed in Table 4, we benchmark it against a series of conceptually natural and simpler alternatives. We start with model *a*, that excludes both `school`-level and `course`-level principal components. To disentangle the individual contributions of each hierarchical component to the reported model, we then fit: model *b* that includes only `school`-level principal components and not `course`-level ones, and model *c* including only `course`-level principal components and not `school`-level ones. Table 5 summarizes the performance of these models in terms of log-likelihood, *p*-value of the likelihood ratio test comparing *a*, *b*, and *c* against the baseline model, in which they are nested, along with BIC, AIC and C-index. Removing both `school`- and `course`-level effects (model *a*) significantly worsens model fit (higher BIC/AIC, lower C-index). Including only `school`-level effects (model *b*) performs better than only `course`-level effects (model *c*). Overall, the baseline model (with both effects) fits best, with `school`-level effects contributing slightly more to model performance. Furthermore, we compare our approach with a parametric frailty model, which is conceptually more appropriate than treating the grouping structure merely as a categorical variable, and is also more computationally feasible in scenarios involving a very large number of groups. Models *d* and *e* include, respectively, a `school`- and a `course`-level random effect, together with the other covariates based on student information available at the end of the first semester, as adopted in previous studies (Masci et al., 2024). A gamma frailty model was employed for this comparison. Our proposed model demonstrates superior predictive performance in terms of BIC, AIC, and C-index. Although the C-index values are comparable, this is consistent with prior findings indicating that `ECTS1sem` is the primary predictive covariate. The improvement observed in our benchmark model stems from incorporating dropout information across courses and schools over the first three semesters of the previous academic-year cohort, rather than relying solely on static first-semester data.

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026

A further comparison aimed at checking whether the functional covariates effectively capture more aspects of timing than a simple dropout rate at the course level is addressed in the online supplementary material, Section S2.

6 Discussion

Addressing student dropouts is a critical concern for universities, both academically and financially. Each dropout represents an inefficient use of institutional resources allocated to recruitment, teaching, and student support. Reducing dropout rates directly impacts both financial stability and the overall effectiveness of educational systems.

One of the complexities in tackling this issue lies in the heterogeneous nature of dropout behaviour across degree programmes and schools. Different academic disciplines present unique challenges - some programmes may experience high dropout rates early on due to demanding foundational courses, while others see increased dropouts as students' careers progress. Similarly, the dropout patterns can vary considerably across schools even within the same university, influenced by factors such as faculty engagement and available student support.

In this paper, we present a novel approach to modelling dropout behaviour by examining occurrences over time within both degree programmes and schools. Our work has two main goals: (i) to estimate the dropout trends over time and examine its variability across different degree programmes and schools, and (ii) to leverage this information at the student level to enrich dropout prediction. To achieve these objectives, we utilize Cox-based regression for recurrent events to model the counting process of the student dropouts, seen from the point of view of individual faculties. In this initial phase, we employ an AG model, as supported by existing literature (Spreafico & Ieva, 2021). However, it is important to note that other modelling choices, such as those proposed by Baraldo et al. (2013), which build on Peña et al. (2007), are also possible. Selecting the appropriate model can be challenging; consequently, this first step of the analysis could be replicated using alternative modelling approaches, allowing for further exploration of our findings. On the other hand, by decomposing dropout patterns within programmes and schools through multilevel functional principal component analysis, we provide a detailed view of time periods in which dropout rates tend to spike. This approach offers a new perspective by providing both visual and quantitative insights into when students are at the highest risk of leaving their studies, allowing institutions to identify periods of high dropout rates throughout the first three semesters and pinpoint the schools and faculties most affected by these trends.

Our predictive model incorporates historical dropout data on current time-to-dropout. By integrating information from previous cohorts, our approach provides an innovative tool to forecast future dropout risks, which can support the implementation of targeted interventions. Indeed, this procedure enables educational institutions to identify at-risk students earlier in their academic journeys, based on a combination of baseline characteristics such as academic performance in first semester, socioeconomic status, and prior high schools attended, alongside historical dropout trends specific to each course and school. With these insights, degree programmes and schools can implement targeted interventions aimed at reducing dropout rates and improving overall student retention and success. For example, by implementing an early warning system, as suggested in Cannistrà et al. (2022), students with low credit accumulation in the first semester or those at high risk of dropping out, as highlighted by our model, can be identified and proactively contacted via targeted emails promoting support services, such as tutoring programmes. At PoliMi, tutoring sessions focused on specific exams, particularly in peer to peer format, have already demonstrated positive effects on student performance and retention. In addition, other strategies are being explored, including enhanced orientation initiatives designed to help students better understand the professional opportunities associated with their chosen degree programmes. Such efforts can facilitate more informed enrolment decisions, increase student motivation, and reduce dropout.

While our model presents promising results, there is space for future refinement. First, the analysis could be validated across multiple academic years. Indeed, the previous academic year's cohort is used to reconstruct compensators for the current cohort, reflecting the practical need to inform predictions early in the academic trajectory. This approach implicitly assumes that dropout patterns remain relatively stable from year to year, particularly during the first three semesters. This assumption is supported by a quantitative stability check we performed, which showed

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

that the estimated compensator patterns remain consistent across past cohorts. Nevertheless, we acknowledge that external shocks, such as major curriculum changes or events like the COVID-19 pandemic, could invalidate this assumption and would require recalibrating the model. Additionally, since the multilevel functional principal component analysis may find issue in disentangle effects across schools and degree programmes, particularly given the limited number of schools, different hierarchies could be evaluated. Finally, focussing on the first three semesters —highly predictive of dropout risk—enables real-time forecasting of incoming cohorts, as the compensator must be integrated over historical data. Extending the observation period beyond this point did not significantly enhance predictions. However, while this approach effectively captures early dropouts at both the course and school levels, future research could examine the full three-year undergraduate cycle for a more comprehensive understanding of dropout patterns.

Acknowledgments

Alessandra Ragni and Anna Maria Paganoni acknowledge the support by the Italian Ministry of University and Research (MUR), Italy, grant 'Dipartimento di Eccellenza 2023–2027'. Chiara Masci acknowledges the support from the MUR under the Department of Excellence 2023–2027 grant agreement 'Centre of Excellence in Economics and Data Science' (CEEDS). The authors thank the anonymous reviewers and the associate editor for helping improving the paper.

Conflicts of interest: None declared.

Funding

The authors received no funding for this research.

Data availability

Due to privacy and confidentiality issues we cannot release the data used in this article. We refer to Section 5 for details on the data used in this article. Supporting code is available at https://github.com/alessandragni/DropoutMultilevelFunct.

Author contributions

A.R. conceived the pipeline, set up the simulation and case studies, implemented and analysed the results, prepared the figures, and wrote the manuscript. C.M. conceived the pipeline, set up the simulation and case studies, and wrote the manuscript. A.M.P. conceived the pipeline, supervised the analyses and the manuscript drafting. All authors reviewed the manuscript.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Aalen O., Borgan O., & Gjessing H. (2008). *Survival and event history analysis: A process point of view*. Springer Science & Business Media.

Aina C., Baici E., Casalone G., & Pastore F. (2018). The economics of university dropouts and delayed graduation: A survey.

Amorim L. D. A. F., & Cai J. (2015). Modelling recurrent events: A tutorial for analysis in epidemiology. *International Journal of Epidemiology*, *44*(1), 324–333. https://doi.org/10.1093/ije/dyu222

Andersen P. K., Borgan O., Gill R. D., & Keiding N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen P. K., & Gill R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, *10*(4), 1100–1120. https://doi.org/10.1214/aos/1176345976

Andersen P. K., & Keiding N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*(2), 91–115. https://doi.org/10.1191/0962280202SM276ra

Arulampalam W., Naylor R. A., & Smith J. P. (2004). A hazard model of the probability of medical school dropout in the uk. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *167*(1), 157–178. https://doi.org/10.1046/j.0964-1998.2003.00717.x

Downloaded from https://academic.oup.com/jrsssc/advance-article/doi/10.1093/jrsssc/qlag015/8529790 by guest on 20 March 2026

Atzeni G., Deidda L. G., Delogu M., & Paolini D. (2022). *Drop-out decisions in a cohort of italian universities*. In *Teaching, research and academic careers: An analysis of the interrelations and impacts* (pp. 71–103). Springer International Publishing Cham.

Baraldo S., Ieva F., Paganoni A. M., & Vitelli V. (2013). Outcome prediction for heart failure telemonitoring via generalized linear models with functional covariates. *Scandinavian Journal of Statistics*, *40*(3), 403–416. https://doi.org/10.1111/sjos.2013.40.issue-3

Barragán S., González L., & Calderón G. (2022). Modelling student dropout risk using survival analysis and analytic hierarchy process for an undergraduate accounting program. *Interchange*, *53*(3–4), 407–427. https://doi.org/10.1007/s10780-022-09463-7

Beutner E. (2023). A review of effective age models and associated non-and semiparametric methods. *Econometrics and Statistics*, *28*(1), 105–119. https://doi.org/10.1016/j.ecosta.2021.12.005

Bozdogan H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. https://doi.org/10.1007/BF02294361

Breslow N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review = Revue Internationale De Statistique*, *43*(1), 45–57. https://doi.org/10.2307/1402659

Cannistrà M. (2024). *Reducing dropout rates: the challenge of learning analytics in higher education institutions* [PhD thesis] Politecnico di Milano.

Cannistrà M., Masci C., Ieva F., Agasisti T., & Paganoni A. M. (2022). Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, *47*(9), 1935–1956. https://doi.org/10.1080/03075079.2021.2018415

Cook R. J., & Lawless J. F. (2007). *The statistical analysis of recurrent events*. Springer.

Cui E., Li R. L., Crainiceanu C. M., & Xiao L. (2023). Fast multilevel functional principal component analysis. *Journal of Computational and Graphical Statistics*, *32*(2), 366–377. https://doi.org/10.1080/10618600.2022.2115500

Daley D. J., & Vere-Jones D. (2002). *An introduction to the theory of point processes*. Springer.

Daley D. J., & Vere-Jones D. (2003). *An introduction to the theory of point processes: Volume I: Elementary theory and methods*. Springer.

Di C.-Z., Crainiceanu C. M., Caffo B. S., & Punjabi N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics*, *3*(1), 458. https://doi.org/10.1214/08-AOAS206SUPP

Diao L., Cook R. J., & Lee K.-A. (2014). Statistical analysis of recurrent adverse events. In *Statistical methods for evaluating safety in medical product development* (pp. 180–192). Wiley Online Library.

Diaz Lema M., Vooren M., Cannistrà M., van Klaveren C., Agasisti T., & Cornelisz I. (2024). Predicting dropout in higher education across borders. *Studies in Higher Education*, *49*(1), 141–156. https://doi.org/10.1080/03075079.2023.2224818

Gury N. (2011). Dropping out of higher education in France: A micro-economic approach using survival analysis. *Education Economics*, *19*(1), 51–64. https://doi.org/10.1080/09645290902796357

Harrell F. E., Califf R. M., Pryor D. B., Lee K. L., & Rosati R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*(18), 2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

Hegde V., & Prageeth P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (pp. 694–699). IEEE.

Karhunen K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarun Fennicae Series A1: Mathematia Physica*, *37*, 1–79.

Katz L. F., & Murphy K. M. (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *The Quarterly Journal of Economics*, *107*(1), 35–78. https://doi.org/10.2307/2118323

Kehm B. M., Larsen M. R., & Sommersel H. B. (2019). Student dropout from universities in Europe: A review of empirical literature. *Hungarian Educational Research Journal*, *9*(2), 147–164. https://doi.org/10.1556/063.9.2019.1.18

Kijima M., Morimura H., & Suzuki Y. (1988). Periodical replacement problem without assuming minimal repair. *European Journal of Operational Research*, *37*(2), 194–203. https://doi.org/10.1016/0377-2217(88)90329-3

Kleinbaum D. G., & Klein M. (1996). *Survival analysis a self-learning text*. Springer.

Kong D., Ibrahim J. G., Lee E., & Zhu H. (2018). Flcrm: Functional linear Cox regression model. *Biometrics*, *74*(1), 109–117. https://doi.org/10.1111/biom.12748

Last G., & Brandt A. (1995). *Marked point processes on the real line: The dynamical approach*. Springer Science & Business Media.

Lin D. Y., Wei L.-J., Yang I, & Ying Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *62*(4), 711–730. https://doi.org/10.1111/1467-9868.00259

Loeve M. (1948). Functions aleatoires du second ordre. In *Processus stochastique et mouvement Brownien* (pp. 366–420). Gauthier-Villars.

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026

Masci C., Cannistrà M., & Mussida P. (2024). Modelling time-to-dropout via shared frailty cox models. A trade-off between accurate and early predictions. *Studies in Higher Education*, *49*(4), 763–781. https://doi.org/10.1080/03075079.2023.2252833
Meyer P.-A. (1962). A decomposition theorem for supermartingales. *Illinois Journal of Mathematics*, *6*(2), 193–205. https://doi.org/10.1215/ijm/1255632318
Min Y., Zhang G., Long R. A., Anderson T. J., & Ohland M. W. (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of Engineering Education*, *100*(2), 349–373. https://doi.org/10.1002/jee.2011.100.issue-2
Morris J. S., Vannucci M., Brown P. J., & Carroll R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, *98*(463), 573–583. https://doi.org/10.1198/016214503000000422
Mussida P., & Lanzi P. L. (2022). A computational tool for engineer dropout prediction. In *2022 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1571–1576). IEEE.
OECD (2019). Education at a glance 2019.
Pasupathy R. (2010). Generating homogeneous poisson processes. In *Wiley encyclopedia of operations research and management science*. Wiley Online Library.
Patacsil F. F. (2020). Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. *Universal Journal of Educational Research*, *8*(9), 4036–4047. https://doi.org/10.13189/ujer.2020.080929
Peña E. A., & Hollander M. (2004). Models for recurrent events in reliability and survival analysis. In R. Soyer, T.A. Mazzuchi, N.D. Singpurwalla (Eds.), *Mathematical reliability: An expository perspective. International series in operations research & management science* (Vol. 67). Springer.
Peña E. A., Slate E. H., & González J. R. (2007). Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference*, *137*(6), 1727–1747. https://doi.org/10.1016/j.jspi.2006.05.004
Pinheiro J. C., & Bates D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.
Plank S. B., DeLuca S., & Estacion A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education*, *81*(4), 345–370. https://doi.org/10.1177/003804070808100402
Prentice R. L., Williams B. J., & Peterson A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, *68*(2), 373–379. https://doi.org/10.1093/biomet/68.2.373
Ragni A., Romani G., & Masci C. (2025). 'TimeDepFrail: Time-dependent shared frailty Cox models in R', arXiv, arXiv:2501.12718, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2501.12718
Ramsay J. O., & Silverman B. W. (2005). Principal components analysis for functional data. In *Functional data analysis*. Springer.
R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rumberger R. W. (2011). *Dropping out: Why students drop out of high school and what can be done about it*. Harvard University Press.
Schwarz G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136
Singer J. D., & Willett J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, *18*(2), 155–195. https://doi.org/10.3102/10769986018002155
Spreafico M., & Ieva F. (2021). Functional modeling of recurrent events on time-to-event processes. *Biometrical Journal*, *63*(5), 948–967. https://doi.org/10.1002/bimj.202000374
Therneau T. M., Grambsch P. M., Therneau T. M., & Grambsch P. M. (2000). *The Cox model*. Springer.
Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Tinto V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089
Tinto V. (1982). Defining dropout: A matter of perspective. *New Directions for Institutional Research*, *1982*(36), 3–15. https://doi.org/10.1002/ir.v1982:36
Vallejos C. A., & Steel M. F. J. (2017). Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(2), 613–631. https://doi.org/10.1111/rssa.12211
Wei L.-J., Lin D. Y., & Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, *84*(408), 1065–1073. https://doi.org/10.1080/01621459.1989.10478873
Yao F., Müller H.-G., & Wang J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*(470), 577–590. https://doi.org/10.1198/016214504000001745

Downloaded from https://academic.oup.com/jrssc/advance-article/doi/10.1093/jrssc/qlag015/8529790 by guest on 20 March 2026