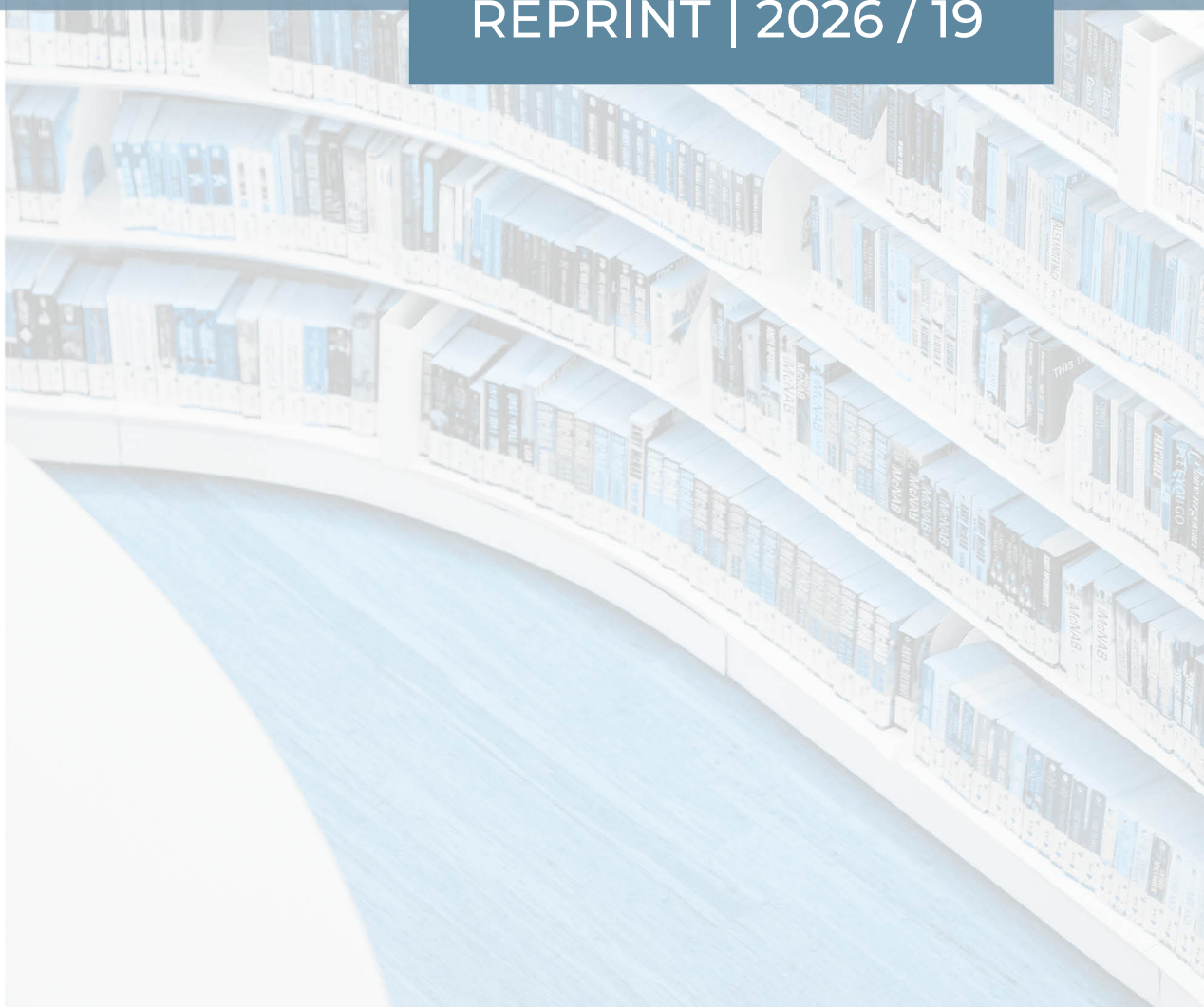
LATENT DIRICHLET ALLOCATION FOR STRUCTURED INSURANCE DATA

Charlotte Jamotton, Donatien Hainaut

REPRINT | 2026 / 19



UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

Latent Dirichlet Allocation for structured insurance data

Charlotte Jamotton[1]*, and Donatien Hainaut[1†]

[1]Université catholique de Louvain (UCLouvain)
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA)
Louvain Institute of Data Analysis and Modeling (LIDAM)
Louvain-la-Neuve, Belgium

Abstract

This article explores the application of Latent Dirichlet Allocation (LDA) to structured tabular insurance data. LDA is a probabilistic approach initially developed in Natural Language Processing (NLP) to uncover the underlying structure of (unstructured) textual data. It was designed to represent textual documents as a mixture of latent topics, and topics as mixtures of words. This study introduces LDA's document-topic distribution as a soft clustering tool for unsupervised learning tasks in actuarial science. By defining each topic as a risk profile, each insurance policy as a document, and each modality of categorical covariates as a word, we extend LDA's application beyond textual data. Our experimental results and analysis highlight how the modelling of policies based on topic cluster membership, and the identification of dominant modalities within each risk profile, can give insights into the prominent risk factors contributing to higher or lower claim frequencies.

Keywords: latent dirichlet allocation, topic modelling, soft clustering, insurance data, risk profile, natural language processing

*charlotte.jamotton@uclouvain.be
†donatien.hainaut@uclouvain.be

1. Introduction

Insurance companies manage large portfolios of policies, each described by numerous policyholder features and policy information. These datasets are typically structured in tabular form, where rows represent individual policies and columns correspond to covariates. Tabular insurance data often consist of a heterogeneous mix of numerical variables (e.g., insured age, policy duration) and categorical variables (e.g., geographic location, gender, vehicle type). This mixture presents several challenges. Numerical variables may require scaling, discretization, or standardization, depending on the model's distributional assumptions. Categorical variables also pose unique challenges due to their qualitative, often high-cardinality nature [1]. Missing values further complicate preprocessing. Accurately representing these variables in machine-processable form is critical, as they directly influence risk assessment and pricing decisions.

Current practice in actuarial machine learning often relies on one-hot encoding to preprocess categorical variables. While simple and interpretable, this approach can lead to high-dimensional sparse representations, making models harder to interpret and potentially limiting predictive performance. Embeddings [2] and other dimensionality reduction methods have been proposed in actuarial applications as an alternative to encode categorical variables directly into numeric, machine-processable representations [3]. These methods also address challenges of high cardinality by capturing similarities between categories in a lower-dimensional space. For example, [4] first addressed the challenges of hierarchically structured nominal risk factors by proposing a top-down partitioning algorithm that groups similar categories and generates features for each hierarchical level. Their approach reduces the effective number of categories, and improves differentiation between high and low risk groups. This method has been extended by [5] using multi-dimensional, multi-level random-effects embeddings, regularized according to the hierarchical structure of the risk factor. Similarly, [6] apply entity embeddings in a hierarchical setting combined with a top-down clustering algorithm to reduce both within-level dimensionality and overall granularity.

In parallel, clustering and segmentation methods have been applied to identify groups of similar policyholders and define homogeneous risk profiles. For instance, [7] applied clustering to telematics-based motor insurance data to uncover latent driving behavior segments, [8] proposed an integrated clustering approach combining policy attributes and claims experience for risk differentiation, and [9] developed a clustering framework to identify natural groupings of policyholders for improved predictive modelling. More recent studies have explored clustering for mixed-type insurance data, including [10], who applied k-prototypes to life insurance portfolios with numeric, categorical, and spatial variables, and [11], who compared multiple clustering algorithms for auto insurance customer segmentation. However, traditional approaches such as K-means or hierarchical clustering assign each policy to a single cluster, failing to model uncertainty in cluster membership and ignoring the overlapping nature of policyholders' risk profiles. Probabilistic and soft clustering approaches have been applied to actuarial problems to address this. For example, nonparametric approaches for actuarial data compression [12], topic-based finite-mixture models that jointly cluster textual and loss information [13], and fuzzy C-means variants adapted for fraud detection or semi-supervised segmentation [14, 15, 16].

The Latent Dirichlet Allocation (LDA) introduced by [17] as an extension to [18] proba-

bilistic Latent Semantic Analysis (pLSA), offers a promising framework that accommodates heterogeneous portfolios, variable-length policies, missing values, and soft cluster membership to generate interpretable latent risk profiles. Originally developed for topic modelling, LDA represents each document as a mixture of topics, where each topic is a distribution over words. Under the Bag-of-Words (BoW) assumption, the sequential arrangement of words within a document, as well as the order of documents within the corpus can be disregarded [19]. According to de Finetti's theorem, the distribution of any set of exchangeable random variables can be represented as a mixture of distributions of independent and identically distributed random variables. Therefore, assuming exchangeability for documents and words within documents amounts to considering a mixture model for both. The three-level Bayesian LDA generative model [17] is built around this exchangeability assumption. By modelling each document as a mixture of latent (hidden) topics and each topic as a mixture of words, LDA attempts to capture the underlying distributional structure of textual documents. This probabilistic, soft clustering framework allows documents to simultaneously belong to multiple topics in varying proportions.

While [20] demonstrated that applying LDA to unstructured textual claim descriptions improves fraud detection in automobile insurance, its potential for structured categorical insurance data remains largely unexplored. In their survey on categorical data processing, [21] recognised the research potential of applying [20] technique to other types of insurance data. The ongoing research to develop more efficient and effective encoding methods for handling categorical data in machine learning lead us to consider the application of the LDA model on categorical and discretized numeric rating factors.

Using the terminology specific to the LDA model, each policy $\mathbf{x}$ characterised by N_d categorical (and discretized numeric) variables can be viewed as a set of words or modalities $\{x_1, x_2, \ldots, x_{N_d}\}$. An insurance portfolio (corpus) X could thus be interpreted as a collection of D policies (documents) $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_D}\}$ depicting K different risk profiles (topics) denoted $\{\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_K}\}$. Instead of modelling topics over words as in text data, LDA would model risk profiles over modalities. A risk profile (topic) would then be defined as a probability distribution over modalities (words). The so-called topic-word $\boldsymbol{\beta}_k$ and document-topic $\boldsymbol{\theta}_d$ distributions could be respectively renamed risk profile-modality and policy-risk profile distributions. The order of policies within a portfolio is inconsequential, and the modalities characterising each policy are also exchangeable (it does not matter if the gender of the insured appears before or after its geographic location). Therefore, the BoW's exchangeability still holds for both policies (documents) and modalities (words within a document).

The policy-topic $\boldsymbol{\theta}_d$ distribution can be used to summarise and organise insurance policies into a smaller set of prominent risk profiles (topics), defined by a list of most-likely (or dominant) modalities, e.g., young drivers of high-power vehicles in the riskiest profile. This corresponds to the definition of cluster analysis [22]. LDA is to textual documents what soft clustering is to numeric observations. In actuarial literature, clustering has been applied to identify groups of policyholders with similar technical attributes (e.g., their car model) and build predictive models (see e.g., [7, 8, 9]). However, when dealing with insurance policies, the crisp cluster assignments of standard clustering algorithms amounts to classifying each policy into a single group. This approach fails to reflect the policy attributes that may overlap with different groups. Whereas LDA's probability distribution $\boldsymbol{\theta}_d$ over K risk profiles, given a policy d, assigns cluster membership weights to each policy topic. This is the definition of

fuzzy (or soft) clustering [23].

Compared to deep learning embeddings, LDA offers several practical advantages for actuarial applications. While neural embeddings [2] can capture complex interactions between categories, they are typically optimized for predictive performance in downstream tasks. LDA, by contrast, provides interpretable topic distributions, requires fewer hyperparameters, can be applied directly to categorical and discretized numeric variables without requiring additional encoding or neural network architectures, and naturally supports soft clustering, which aligns with the overlapping risk characteristics often observed in insurance portfolios. LDA is therefore directly suited for policy segmentation.

Moreover, in LDA the number of words (modalities) in each document (policy) can differ. This property allows LDA to be flexible enough to model documents (policies) of varying lengths, while still effectively discovering topics (risk profiles) that are common across the entire corpus (insurance portfolio). This characteristic of LDA's hierarchical model is of interest to model policies containing missing values. While incomplete data analysis has already been addressed in the literature by [24], who introduced grouped Dirichlet distributions to analyse incomplete categorical data, or [25], who proposed a clustering algorithm to group mixed numerical and categorical data with missing values, to name a few, the management of missing values in actuarial applications is not strongly developed.

In this article, we contribute to the actuarial literature by (i) extending LDA to structured categorical (instead of unstructured textual) insurance data, (ii) introducing the LDA document-topic distribution as an unsupervised soft clustering tool for policy segmentation, and (iii) demonstrating its empirical performance on realistic insurance portfolio scenarios. To the best of our knowledge, it is the first application of LDA to structured, non-textual, insurance data.

The article is organised as follows. Section 2 introduces the LDA model as a soft clustering tool for structured insurance data. Section 2.1 sets the notations. Section 2.2 and Section 2.3 detail the generative process and the variational inference method [26] used to derive the model's approximate posterior. Section 3 presents our experiment settings and results and compares our approach with the Burt-adapted fuzzy clustering [27]. Section 4 concludes with insights and potential extensions.

2. Methods

Figure 1 illustrates the workflow of the LDA modelling framework. First, the insurance portfolio X is represented as a collection of policies, each expressed as a sequence of modalities. Second, LDA is applied to uncover latent risk profiles (topics) explaining the distribution of observed modalities. Finally, the posterior distributions of the latent variables are approximated using variational Bayes (VB) optimisation, yielding interpretable per-policy risk profile (topic) mixtures. The framework produces a soft clustering of policies in a low-dimensional latent risk space, which can be used for predictive premium modelling.

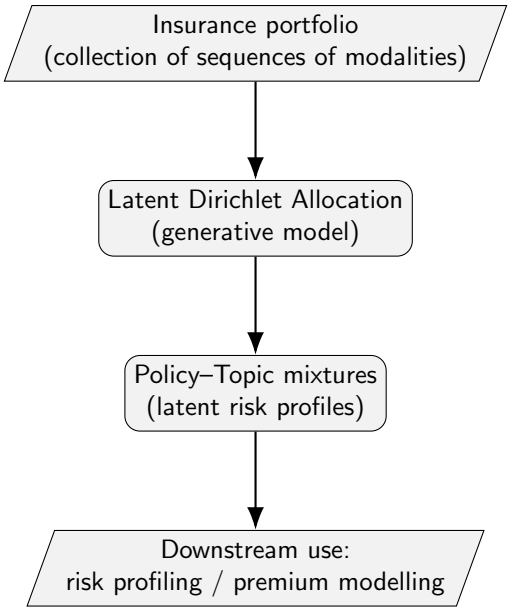


Figure 1. Overview of the proposed modelling framework. The insurance portfolio is represented as categorical policy data, encoded into modalities, processed through the LDA generative model, and inferred via VB. The resulting topic mixtures define latent risk profiles for downstream actuarial use.

2.1 Notations

Before introducing the Latent Dirichlet Allocation (LDA) framework, we define the main notations and concepts used throughout this article. For reference, Table 11 summarises the key mathematical symbols used throughout the paper, together with their actuarial interpretation.

Let X denote a tabular insurance portfolio (corpus) of dimension $D \times C$ consisting of a collection of D insurance policies (documents) and C categorical variables (e.g., policyholder's gender, geographic location, vehicle colour, etc.):

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_D\} \ . \tag{1}$$

Each policy (document) $\mathbf{x}$ in the insurance portfolio can be described as a finite sequence of N_d words, representing the categorical or discretized numeric rating factors of the policy, often referred to as modalities (e.g., female, west, grey, etc.):

$$\mathbf{x} = \left\{x_1, \ldots, x_{N_d}\right\} \ , \tag{2}$$

The insurance portfolio (corpus) can further be characterised by its vocabulary size, $V = \sum_{c=1}^{C} m_c$, which is equal to the number of unique modalities across all categorical variables.

Each policy $\mathbf{x}$ is modelled as a mixture $\boldsymbol{\theta}^{(1\times K)}$ over the K latent risk profiles (topics), each capturing a co-occurrence pattern of modalities across policies.

We denote $x_{d,n}$ the nth modality of the dth policy. $\boldsymbol{\theta}_d$ denotes the topic distribution (or mixture of risk profiles) for policy d. Similarly, $z_{d,n}$ is a latent variable indicating the risk profile (topic) assigned to the nth modality of the dth policy. Each risk profile $k = 1, \ldots, K$ will be described by a mixture $\boldsymbol{\beta}_k^{(1\times V)}$ over all V modalities (words). In other

words, $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{k|\mathcal{V}|})$ denotes the multinomial distribution of modalities under risk profile k, where $\sum_v \boldsymbol{\beta}_{kv} = 1$. By doing so, LDA estimates which latent topics (risk profiles) are important for a particular policy, and which modalities (words) are important for a particular risk profile (topic).

Under these notations, the goal of the LDA model is to estimate the posterior distributions of the latent variables $\mathbf{z}$ and $\boldsymbol{\theta}$ given the observed modalities X and hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\eta})$. $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are the Dirichlet hyperparameters controlling, respectively, the concentration of $\boldsymbol{\theta}_d$ (policy–topic mixtures) and $\boldsymbol{\beta}_k$ (topic–modality distributions). The next section formalizes the corresponding generative process.

2.2 Generative process

The LDA generative process, illustrated in Figure 2, works as follows. For K pre-specified risk profiles (topics), LDA assumes the dth policy $\mathbf{x}_d \in X$ with length N_d is generated in the following steps [28, 29, 30]:

1. For each risk profile $k = 1, \ldots, K$, $\boldsymbol{\beta}_k \sim \mathcal{D}_V(\boldsymbol{\eta})$
2. For each policy $d = 1, \ldots, D$
   (a) $\boldsymbol{\theta}_d \sim \mathcal{D}_K(\boldsymbol{\alpha})$
   (b) for each modality $x_{d,n}$
      i. $z_{d,n} \sim \mathcal{M}_K(1, \boldsymbol{\theta}_d)$
      ii. $x_{d,n}|z_{d,n} \sim \mathcal{M}_V\left(1, \boldsymbol{\beta}_{z_{d,n}}\right)$

In the first step, a risk profile-modality (topic-word) distribution $\boldsymbol{\beta}_k$ is drawn for each risk profile k over the vocabulary V from a V-dimensional Dirichlet distribution $\mathcal{D}_V(\boldsymbol{\eta})$ with hyperparameter $\boldsymbol{\eta}$. Each element $\boldsymbol{\beta}_{k,n}$ in $\boldsymbol{\beta}_k$ indicates a probability of a modality n for this topic k. The higher the $\boldsymbol{\eta}$ value, the more likely each risk profile (topic) is to contain a mixture of most of the modalities and not a single modality specifically. Note that a slight drawback of LDA is that we have to specify upfront the number of topics to group the set of documents into. Setting a small number of topics K typically leads to fairly general risk profiles that are mixtures of more specific sub-risk profiles [31]. [32] addressed this by introducing a topic alignment method to find an optimal number of topics.

In the second step, the risk profiles proportions $\boldsymbol{\theta}_d$ are sampled for each policy d over the available K risk profiles from a K-dimensional Dirichlet distribution $\mathcal{D}_K(\boldsymbol{\alpha})$ with corpus-level hyperparameter $\boldsymbol{\alpha}$. Each element $\boldsymbol{\theta}_{d,k}$ in $\boldsymbol{\theta}_d$ indicates a probability of a risk profile k for this policy d. The higher the $\boldsymbol{\alpha}$ value, the more likely each policy is to contain a mixture of most of the risk profiles and not a single risk profile specifically. This policy-specific risk profile matrix $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D) \in \mathbb{R}^{D \times K}$ can be further interpreted as a low-dimensional representation of each policy in the risk profiles space, and ultimately as a soft clustering tool.

A latent topic (risk profile) assignment $z_{d,n} \in \{1, \ldots, K\}$ is then sampled for each modality $x_{d,n}$, $n = 1, \ldots, N_d$ in the dth policy from a K-dimensional Multinomial distribution $\mathcal{M}_K(1, \boldsymbol{\theta}_d)$ over the K topics parameterised by the topics weights $\boldsymbol{\theta}_d$.

Finally, the observed modality $x_{d,n}$ is sampled given the sampled risk profile $z_{d,n}$ from a V-dimensional Multinomial distribution $\mathcal{M}_V\left(1, \boldsymbol{\beta}_{z_{d,n}}\right)$ parameterised by $\boldsymbol{\beta}_{z_{d,n}}$, following the

standard LDA generative process. In insurance data, modalities within a categorical variable are mutually exclusive. Each modality $x_{d,n}$ should then be sampled from a subset $V^{(n)} \subset V$ containing only the unique modalities of the corresponding categorical variable. Formally, this can be written as:

$$x_{d,n}|z_{d,n} \sim \mathcal{M}_{V^{(n)}}\left(1, \beta_{z_{d,n}}^{V^{(n)}}\right) . \tag{3}$$

For example, if the first categorical variable, $n = 1$, is the geographic location (West, East, or Central) of the policyholder, $V^{(1)}$ should only contain the identifiers of the modalities West, East, and Central. While Eq. (3) reflects the practical restriction of sampling for insurance data, LDA is not used in this work as a generative model for data synthesis but rather as a latent representation framework to encode categorical insurance variables and derive risk profiles. Consequently, we never actually sample $x_{d,n}$ from the model. Therefore, all subsequent joint and marginal distribution formulas are defined over the full vocabulary V, ensuring formal consistency with the standard LDA framework.

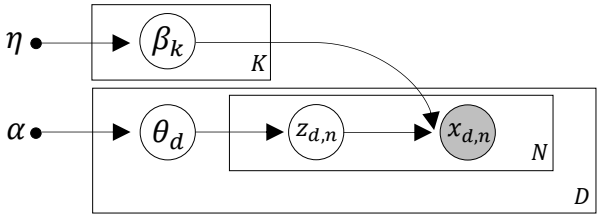


Figure 2. Representation of the smoothed LDA generative process. The top rectangle defines the topic-level parameters, the inner rectangle defines the modality-level parameters, and the outer rectangle defines the policy-level parameters. We can only observe x (filled with grey).

The LDA generative process described and illustrated above results in the following joint distribution [28, 29, 30]:

$$p\left(x, z, \theta, \beta|\alpha, \eta\right) = \prod_{k=1}^{K} p\left(\beta_k|\eta\right) \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(x_{d,n}|\theta_d, \beta\right) , \tag{4}$$

where the modalities x are the only observable data and z, θ, β are the latent (hidden) variables. When marginalising over topic assignments z, the modalities (words) distribution is given by a mixture of weights $p\left(z_{d,n}|\theta_d\right)$ and components $\mathrm{p}\big(x_{d,n}|z_{d,n}, \beta\big)$:

$$p\left(x_{d,n}|\theta_d, \beta\right) = \sum_{z} p\left(x_{d,n}|z_{d,n}, \beta\right) p\left(z_{d,n}|\theta_d\right) , \tag{5}$$

such that the joint distribution in Eq. (4) becomes:

$$p\left(x, z, \theta, \beta|\alpha, \eta\right) = \prod_{k=1}^{K} p\left(\beta_k|\eta\right) \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(x_{d,n}|z_{d,n}, \beta\right) p\left(z_{d,n}|\theta_d\right) , \tag{6}$$

where β_k and θ_d are sampled using Dirichlet probability distributions (see Eq. (23) and (24) in Appendix 2 for detailed formulas) and $z_{d,n}$ and $x_{d,n}|z_{d,n}$ are sampled using Multinomial probability distributions (see Eq. (25) and (26) in Appendix 2).

While LDA's goal is to uncover the latent topics z, we can only observe x. Therefore, we need to reverse the generative process described above to learn the posterior distribution $p\left(z, \theta, \beta | x, \alpha, \eta\right)$ of the latent variables z_d, θ_d, and β_k given the evidence (i.e., the observed data x) and the Dirichlet hyperparameters α, η. This amounts to solving the following equation [30]:

$$p\left(z, \theta, \beta | x, \alpha, \eta\right) = \frac{p\left(x, z, \theta, \beta | \alpha, \eta\right)}{\int_{\theta,\beta} \sum_z p\left(x, z, \theta, \beta | \alpha, \eta\right)} . \tag{7}$$

The numerator of this conditional Eq. (7) can be expanded given Eq. (23), (24), (25), and (26).

$$\begin{aligned} p\left(x, z, \theta, \beta | \alpha, \eta\right) &= \prod_{k=1}^{K} p\left(\beta_k | \eta\right) \prod_{d=1}^{D} p\left(\theta_d | \alpha\right) \prod_{n=1}^{N_d} p\left(x_{d,n} | z_{d,n}, \beta_{1:K}\right) p\left(z_{d,n} | \theta_d\right) \\ &= c \prod_{k=1}^{K} \left(\prod_{v=1}^{V} \beta_{k,v}^{\eta_v - 1} \right) \prod_{d=1}^{D} \left(\prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1} \right) \prod_{n=1}^{N_d} \prod_{k=1}^{K} \prod_{v=1}^{V} \left(\beta_{k,v} \theta_{d,k}\right)^a , \end{aligned} \tag{8}$$

where $c = \frac{\Gamma\left(\sum_{v=1}^{V} \eta_v\right)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}$ is a constant and $a = \mathbb{I}\left\{x_{d,n} = v, z_{d,n} = k\right\}$ is an indicator function.

The denominator of the conditional Eq. (7) is intractable due to the coupling of the latent variables θ and β in the summation over topics in the marginal likelihood:

$$\int_\beta \int_\theta \sum_z p\left(x, z, \theta, \beta | \alpha, \eta\right) = c \int_\beta \prod_{v=1}^{V} \beta_{k,v}^{\eta_v - 1} \int_\theta \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \prod_{v=1}^{V} \left(\beta_{k,v} \theta_{d,k}\right)^a . \tag{9}$$

There are two families of techniques to derive a (tractable) approximate posterior distribution: the Markov Chain Monte Carlo (MCMC) sampling approaches and the VB optimisation approaches [28, 29, 30]. Although the choice of a simplified parametric distribution introduces bias, VB is empirically shown to be faster than and as accurate as MCMC whose stochastic nature and iterative sampling process often require longer convergence time [26]. To efficiently fit models to larger collections of documents, [26] developed an online VB algorithm based on online stochastic optimisation. It is implemented in the Scikit-learn library in Python [33] and detailed hereafter.

2.3 Variational posterior and ELBO

A variational posterior distribution $q\left(z, \theta, \beta | \phi, \gamma, \lambda\right)$, illustrated in Figure 3, is introduced to approximate the intractable posterior $p\left(z, \theta, \beta | x, \alpha, \eta\right)$. Under the mean-field assumption, this joint distribution of latent variables $q\left(z, \theta, \beta | \phi, \gamma, \lambda\right)$ can be factorised into three mutually independent marginal distributions:

$$q\left(z, \theta, \beta | \gamma, \lambda\right) = \prod_{k=1}^{K} q\left(\beta_k | \lambda_k\right) \prod_{d=1}^{D} q\left(\theta_d | \gamma_d\right) \prod_{n=1}^{N_d} q\left(z_{d,n} | \phi_{d,n}\right) , \tag{10}$$

parameterised by three variational parameters:

- $\overrightarrow{\lambda_k}$ of length V for topic $k = 1, \ldots, K$;
- non-negative (and does not sum to 1) $\overrightarrow{\gamma_d}$ of length K for policy $d = 1, \ldots, D$;
- non-negative ϕ of size $D \times V \times K$ with $\sum_{k=1}^{K} \phi_{dvk} = 1, \forall d, v$.

where β_k and θ_d are sampled using Dirichlet probability distributions (see Eq. (27) and (28) in Appendix 2 for detailed formulas) and $z_{d,n}$ is sampled using a Multinomial probability distribution (see Eq. (29) in Appendix 2).

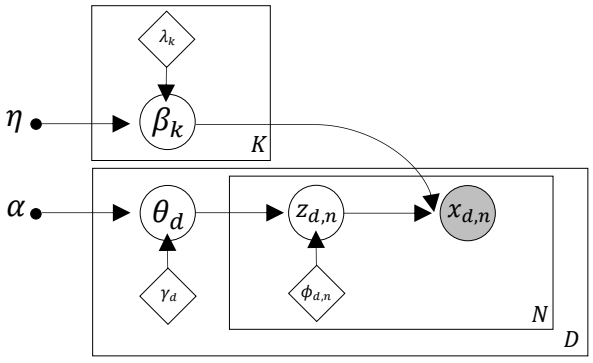


Figure 3. Representation of the variational distribution to approximate the posterior in the smoothed LDA model. We can only observe x (filled with grey). The variational parameters are framed with lozenge-shapes.

The variational parameters are estimated by minimising the Kullback-Leibler (KL) divergence between the variational approximation q and the true posterior distribution p:

$$\min_{q} \left\{ KL\left(q\left(z, \theta, \beta | \phi, \gamma, \lambda\right) \,||\, p\left(z, \theta, \beta | x, \alpha, \eta\right)\right) = \mathbb{E}_q \left[\log \frac{q\left(z, \theta, \beta | \phi, \gamma, \lambda\right)}{p\left(z, \theta, \beta | x, \alpha, \eta\right)} \right] \right\} . \tag{11}$$

Minimising Eq. (11) is equivalent to maximising the evidence lower bound (ELBO):

$$\mathcal{L}(\phi, \gamma, \lambda; \alpha, \eta) = \mathbb{E}_q[\ln p(x, z, \theta, \beta \mid \alpha, \eta)] - \mathbb{E}_q[\ln q(z, \theta, \beta \mid \phi, \gamma, \lambda)], \tag{12}$$

derived from the marginal log-likelihood $\ln p\left(x | \alpha, \eta\right)$ using Jensen's inequality (see Appendix 3 for detailed derivations).

To maximise the ELBO with respect to the variational parameters ϕ, γ and λ, we first maximise $\mathcal{L}$ with respect to ϕ_{dvk}. This is a constraint maximisation since $\sum_k \phi_{dvk} = 1 \, \forall d, v$. We form the Lagrangian by isolating the terms which contain ϕ_{dvk} and by adding the appropriate Lagrange multipliers, detailed in

$$\begin{aligned} \mathcal{L}_{[\phi_{dvk}]} &= \phi_{dvk} \left(\mathbb{E}_q \left[\ln \beta_{k,v} \right] + \mathbb{E}_q \left[\ln \theta_{d,k} \right] - \ln \phi_{dvk} \right) + \lambda_k^{Lagrange} \left(\sum_{k=1}^{K} \phi_{dvk} - 1 \right) \\ \frac{\partial \mathcal{L}}{\partial \phi_{dvk}} &= \mathbb{E}_q \left[\ln \beta_{k,v} \right] + \mathbb{E}_q \left[\ln \theta_{d,k} \right] - \left(\ln \phi_{d,v,k} + 1 \right) + \lambda_k^{Lagrange} . \end{aligned} \tag{13}$$

Setting the gradient to zero and solving for ϕ_{dvk}, we get the update rule for ϕ_{dvk}:

$$\phi_{dvk} \propto \exp\left(\mathbb{E}_q \left[\ln \beta_{k,v} \right] + \mathbb{E}_q \left[\ln \theta_{d,k} \right] \right) . \tag{14}$$

We then maximise the ELBO with respect to γ_{dk}:

$$\begin{aligned}
\mathcal{L}_{[\gamma_{dk}]} &= \left(\sum_{v=1}^{V} n_{d,v}\phi_{d,v,k} + \left(\alpha - \gamma_{d,k}\right)\right)\left(\Psi\left(\gamma_{dk}\right) - \Psi\left(\sum_{k=1}^{K}\gamma_{dk}\right)\right) + a\left(\gamma_{dk}\right) + \text{cst} \\
\frac{\partial\mathcal{L}}{\partial\gamma_{dk}} &= \left(\Psi'\left(\gamma_{dk}\right) - \Psi'\left(\sum_{k=1}^{K}\gamma_{dk}\right)\right)\left(\sum_{v=1}^{V} n_{d,v}\phi_{d,v,k} + \left(\alpha - \gamma_{d,k}\right)\right) .
\end{aligned} \tag{15}$$

Setting the gradient to zero and solving for γ_{dk}, we get the update rule for γ_{dk}:

$$\gamma_{d,k} = \alpha + \sum_{v=1}^{V} n_{d,v}\phi_{d,v,k} \,. \tag{16}$$

The number of times policy d is assigned to topic k is thus equal to the Dirichlet prior α plus the weighted sum of the number of times each modality in d is assigned to topic k, where the weight ϕ_{dvk} is the probability that modality v in policy d belongs to topic k.

We finally maximise the ELBO with respect to λ_{kv}:

$$\begin{aligned}
\mathcal{L}_{[\lambda_{kv}]} &= \left(\sum_{d=1}^{D} n_{d,v}\phi_{d,v,k} + \left(\eta - \lambda_{kv}\right)\right)\left(\Psi\left(\lambda_{kv}\right) - \Psi\left(\sum_{v=1}^{V}\lambda_{kv}\right)\right) + b\left(\lambda_{k,v}\right) + \text{cst} \\
\frac{\partial\mathcal{L}}{\partial\lambda_{kv}} &= \left(\sum_{d=1}^{D} n_{d,v}\phi_{d,v,k} + \left(\eta - \lambda_{kv}\right)\right)\left(\Psi'\left(\lambda_{kv}\right) - \Psi'\left(\sum_{v=1}^{V}\lambda_{kv}\right)\right) .
\end{aligned} \tag{17}$$

Setting the gradient to zero and solving for λ_{kv}, we get the update rule for λ_{kv}:

$$\lambda_{kv} = \eta + \sum_{d=1}^{D} n_{d,v}\phi_{d,v,k} \,. \tag{18}$$

Similarly to the interpretation of γ_{dk}, the count of modality v in topic k is equal to the Dirichlet prior η plus the weighted sum of modality count v in each policy d, where the weight $\phi_{d,v,k}$ is the probability that modality v in document d belongs to topic k.

Given the updates rules for the variational parameters ϕ, γ and λ (see Eq. (18), (16), and (14)), the online VB EM steps for solving LDA are detailed in Algorithm 1 [26].

3. Results and discussion

3.1 Experimental design

We aim to assess the validity, consistency, and interpretability of the proposed LDA-based soft clustering approach for structured insurance data, and to benchmark its performance against both a Generalized Linear Model (GLM) and a Burt adapted fuzzy K-means. The evaluation focuses on two distinct insurance portfolios differing in size, target distribution, and data structure.

We consider a loan default portfolio containing 25,917 policies, and a motor insurance portfolio containing 62,289 policies. Each portfolio includes a combination of categorical, binary, and continuous covariates, as detailed in Tables 1 and 2. Continuous variables are discretized into four groups based on their empirical quartiles (25th, 50th, and 75th percentiles), ensuring comparability with categorical variables within the LDA framework. The text descriptions in the loan dataset are not used for LDA training. No resampling or data augmentation is applied to preserve the original portfolio structure.

For both datasets, the observations are randomly partitioned into an 80% training subset and a 20% test subset (Table 3). All model fitting is performed exclusively on the training subset, and the test subset is used to evaluate the out-of-sample goodness of fit.

For each portfolio, we train multiple LDA models with the number of topics K varying from 2 to 40. Each model is initialized using symmetric Dirichlet priors on the topic–word and document–topic distributions. The hyperparameters are held constant across runs to ensure comparability. For reference, a GLM is fitted using the same training data, where categorical (and discretized) variables are one-hot encoded. The Burt adapted fuzzy K-means is trained under the same conditions, with K set to match the LDA topic count.

Model performance is evaluated on the test subsets using the deviance of the assumed target distribution, that is, Binomial deviance for the loan portfolio and Poisson deviance for the motor portfolio. Each observation is assigned to the topic (or cluster) of maximum posterior probability $\max(\theta_d)$, from which the predicted frequency $\hat{y}_d$ is derived as the empirical mean of targets within the same cluster. The resulting cluster-level predictions are compared across models to assess fit quality.

To assess interpretability, the top modalities and average numeric values per topic are analyzed and visualized. All results (deviance values, boxplots, and cluster compositions) are reported on the test sets, unless otherwise stated.

3.2 Datasets

We use two distinct insurance portfolios to evaluate the validity and consistency of the LDA model across various types of covariates with a variety of marginal distributions and different assumptions regarding the target variable's distribution. The first dataset is a loan default portfolio with 25 917 policies described by six categorical variables (discretized into four distinct groups based on its 25th, 50th, and 75th percentile), two binary variables, four continuous variables (i.e., structured data) and a loan description provided by the borrower (i.e., unstructured data) detailed in Table 1. The target variable takes two possible values (Charged Off or Fully Paid) and is assumed to follow a Binomial distribution.

Table 1. Descriptions of the loan portfolio attributes.

Categorical variables	**Description**	**Modalities**	**Proportions**		
Inquiries	The number of inquiries in the past 6 months.	0	0.48%		
		1	0.27%		
		2	0.15%		
		≥ 3	0.10%		
Region	The region of the state provided by the borrower in the loan application.	East	0.51%		
		West	0.29%		
		Central	0.20%		
Purpose	A category provided by the borrower for the loan request.	Debt consolidation	0.47%		
		Other	0.19%		
		Credit card	0.13%		
		Home improvement	0.07%		
		Major purchase	0.05%		
		Small business	0.05%		
Grade	LC assigned loan subgrade.	1–10	0.56%		
		11–20	0.34%		
		21–35	0.10%		
Home ownership	The home ownership status provided by the borrower during registration.	Rent	0.48%		
		Mortgage	0.44%		
		Own	0.08%		
Verification status	Indicates if the income (source) was verified by LC.	Not verified	0.45%		
		Verified	0.32%		
		Source verified	0.23%		
Term	The number of payments on the loan.	36 months	0.76%		
		60 months	0.24%		
Public records	Number of derogatory public records.	0	0.95%		
		1, 2 or 3	0.05%		
Target variable					
Loan status	Current status of the loan.	Fully Paid	0.85%		
		Charged Off	0.15%		
Continuous variables		**min**	**max**	**mean**	**std**
Loan amount	The listed amount of the loan applied for by the borrower.	500	35 000	11 369.9	7 274.08
Interest rate	Interest rate on the loan.	5.42	24.4	11.95	3.6
Annual income	The self-reported annual income provided by the borrower during registration.	400	3 900 000	68 854	59 173
Revolving line utilization rate	The amount of credit the borrower is using relative to all available revolving credit.	0	99.9	48.49	28.22
Claim description	Loan description provided by the borrower.				

To validate the results obtained with this first database and check their reproducibility, we use a second insurance portfolio that contains information about motorcycle insurance policies over the period 1994–1998 [34]. Table 2 gives the description and basic summary statistics of the two quantitative variables and the three categorical variables describing each policy. The target variable (number of claims) is assumed to follow a Poisson distribution. Given the exposure of each policy, we will predict its claim frequency (i.e., the ratio between the number of claims and the contract duration). In the following analyses, the policyholder's age and its vehicle's age are categorised into respectively nine (≤ 19, $20-29$, $30-39$, $\ldots$, $80-89$, ≥ 90) and four ($0-4$, $5-10$, $11-20$, ≥ 21) distinct groups. This motorcycle portfolio contains 62 289 insurance policies, which is double the number of observations in the loan dataset.

Table 2. Descriptions of the motorcycle portfolio attributes [34].

Categorical variables	**Description**	**Modalities**		**Proportions**	
Gender	Male (ma)	M		0.85%	
	Female (kvinnor)	K		0.15%	
Geographic area	Central and semi-central parts of Sweden's three largest cities.	1		0.13%	
	Suburbs plus middle-sized cities.	2		0.18%	
	Lesser towns, except those in 5 or 7.	3		0.20%	
	Small towns and countryside.	4		0.39%	
	Northern towns.	5		0.04%	
	Northern countryside.	6		0.06%	
	Gotland (Sweden's largest island).	7		0.01%	
Vehicle power	EV ratio -5	1		0.11%	
	EV ratio 6–8	2		0.08%	
	EV ratio 9–12	3		0.30%	
	EV ratio 13–15	4		0.19%	
	EV ratio 16–19	5		0.18%	
	EV ratio 20–24	6		0.13%	
	EV ratio 25-	7		0.01%	
Continuous variables		**min**	**max**	**mean**	**std**
Owner's age	Age of the policyholder.	16	92	42.54	12.95
Vehicle age	Age of the policyholder's vehicle.	0	99	12.56	9.68
Target variable					
Claim frequency	The ratio between the number of claims and the contract duration.	0	182.51	0.028	0.947
Exposure	The contract duration.	0.002	11.99	1.008	1.06
Number of claims	The number of motor claims.	0		0.9898%	
		1		0.0101%	
		2		0.0004%	

3.3 LDA soft clustering for structured data

As explained in Section 2, by considering each distinct value of the categorical (and discretized numeric) rating factors as a word, we can extend the scope of application of the LDA model to structured data and use the document-topic matrix $\boldsymbol{\theta}^{(D\times K)}$ as a soft clustering tool where $k = 1, \ldots, K$ are the topic clusters and $d = 1, \ldots, D$ are the observations we want to cluster.

For each insurance portfolio introduced above, a LDA model is trained using 80% of the dataset. The model's goodness of fit is then evaluated on the remaining 20% of the dataset. This train-test split, displayed in Table 3, is essential for assessing how well the model generalises to unseen data. Given the learned parameters from the trained model (i.e., the topic-word and policy-topic distributions), we can infer the risk profiles distributions $\boldsymbol{\theta}_{d^{new}}$ of these unseen (new) insurance policies (in the test set) using the optimised posterior distribution $q(\boldsymbol{\theta})$.

Table 3. Number of observations (insurance policies) in the training-testing subsets.

	Train set (80%)	Test set (20%)
Loan	20 733	5 184
Motor	49 831	12 458

To evaluate the goodness of fit of the models, each policy is assigned to the topic for which it has the highest relative probability $\max(\boldsymbol{\theta}_d)$. The quality of the topic clustering can then be evaluated with the Binomial deviance for the loan dataset and with the Poisson deviance for the motor dataset. The Binomial deviance is defined as:

$$D_{Bin}(y, \hat{y}) = 2\sum_{d=1}^{D}\left(y_d \ln\left(\frac{y_d}{\hat{y_d}}\right) + (1-y_d)\ln\left(\frac{1-y_d}{1-\hat{y_d}}\right)\right), \tag{19}$$

where y_d is the observed loan default probability of the dth policy, and $\hat{y_d}$ is the predicted loan default probability of the dth policy (which is equal to the average number of defaults of all policies assigned to the same (topic) cluster as the dth policy). The Poisson deviance is defined as:

$$D_{Poi}(y, \hat{y}) = 2\sum_{d=1}^{D} N_d\left(\frac{\boldsymbol{\nu}_d}{N_d}\hat{\boldsymbol{\lambda}_d} - \left(\ln\left(\frac{\boldsymbol{\nu}_d}{N_d}\hat{\boldsymbol{\lambda}_d}\right) + 1\right)\mathbb{I}_{\{N_d \geq 1\}}\right), \tag{20}$$

where N_d and $\boldsymbol{\nu}_d$ are respectively the number of claims and the exposure of the dth policy, and $\hat{\boldsymbol{\lambda}_d}$ is the predicted claim frequency (which is equal to the average number of claim frequencies of all policies assigned to the same (topic) cluster as the dth policy).

For each portfolio, we trained around 40 different LDA models (by increasing the number of topics K) on the structured covariates in the respective training subset (80%) of each portfolio. Figure 4 shows the evolution of the Binomial (left plot) and Poisson (right plot) deviances with respect to the number of topics $K \in (2, 40)$ used to fit these LDA models. The orange lines represent the evolution of the deviances on the train sets, whereas the green lines (of interest) represent the evolution of the deviances on the test sets. We observe a general (but not smooth) decrease in the deviances when the number of topics K increases, especially

for the motor portfolio (see right plot of Figure 4) which contains more observations (see Table 3). In the following analyses, all results (including goodness of fit or deviance values) are evaluated on the test sets, unless otherwise specified.

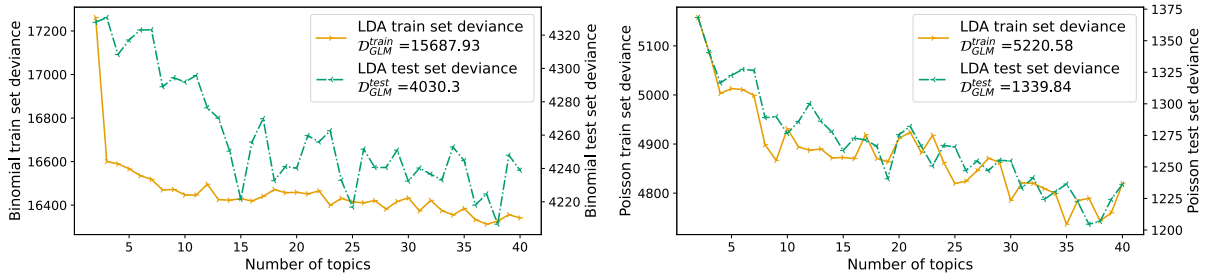


Figure 4. Evolution of the LDA Binomial (left plot) and Poisson (right plot) deviances with respect to the number of LDA topics $K \in (2, 40)$.

If we chose to train our LDA model with $K = 15$ topics on a subset (80%) of the loan portfolio (in Table 1), we get a Binomial deviance equal to 4 221.52 (with 15 distinct $\hat{y_d}$). In comparison, a GLM trained on the same subset (80%) of the loan portfolio (with the difference that the discretized and categorical covariates are one-hot encoded) gives a lower (and better) deviance equal to 4 030.3. The granularity of the GLM (4 809 distinct $\hat{y_d}$) is however much higher than LDA clustering (15 distinct $\hat{y_d}$). The choice of setting the number of topics K to 15 was determined through visual analysis of the evolution of the deviance (green line) in the left plot of Figure 4. Moreover, this number is sufficiently large to capture the data underlying structure, while still being small enough to allow for a concise analysis of the composition of the different clusters.

While the use of LDA's document-topic distribution as an unsupervised soft clustering tool is not as competitive in terms of deviance as a standard GLM on our loan portfolio, we still observe in Table 4 that the LDA approach gives some insights into the prominent risk factors contributing to higher or lower claim frequencies. This table 4 gives a summary of some of the most common modalities (in columns titled `term` and `sub_grade`) and average numeric values (in columns `int_rate` and `revol_util`) in each topic cluster. For ease of analysis, the topic clusters (as well as the corresponding figures) are ordered by increasing (average) number of loan defaults (see `Frequency` column). The categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec` are omitted from Table 4 as none of their respective categories (modalities) has a consistent association with higher or lower loan default rates (see Figure 14 in Appendix). The continuous variables `loan_amnt` and `annual_inc` are also omitted as no increasing (or decreasing) trend with the default rate is observed (see Figure 15 in Appendix).

The column titled `% of policies` displays the size of each cluster (i.e., number of policies relative to the portfolio size). We observe that nearly 50% of the total observations are in the first three clusters with the lowest number of defaults. Insurance claims are often considered as rare events due to their low probability of occurrence. In the context of clustering, the data may naturally exhibit an imbalanced distribution, where the majority of policies are assigned to lower-risk clusters with a lower likelihood of defaults.

Table 4. Some of the most common modalities and average numeric values in each topic cluster obtained with the LDA soft clustering model on the test set (20%) of the loan portfolio. The table is ordered by increasing (average) number of loan defaults.

Topic	% of policies	term	sub_grade	int_rate	revol_util	Frequency (%)
1	0.77	36 months	1–10	6.86	37.0	0.05
2	37.21	36 months	1–10	8.55	36.61	0.1032
3	11.94	36 months	1–10	9.53	45.77	0.1066
4	2.37	36 months	1–10	9.64	30.24	0.1138
5	4.53	36 months	1–10	9.61	46.26	0.1404
6	2.22	60 months	1–10	10.53	43.29	0.1565
7	3.49	36 months	11–20	11.68	55.09	0.1602
8	10.07	36 months	['11–20', '1–10']	12.78	67.04	0.1686
9	1.58	60 months	11–20	11.86	51.17	0.1829
10	2.72	36 months	21–35	15.04	62.26	0.1844
11	7.99	36 months	['11–20', '1–10']	12.34	57.26	0.1932
12	9.43	36 months	11–20	12.54	51.81	0.2331
13	0.33	36 months	11–20	15.17	63.44	0.2353
14	1.6	60 months	['21–35', '11–20']	15.22	61.08	0.253
15	3.74	60 months	21–35	15.33	72.8	0.2835

We observe in Table 4 and in the left plot of Figure 5 that the two topic clusters with the highest number of defaults have the highest number of long-term (60 months) loans. Moreover, the topic cluster characterised by the lowest number of loan defaults exclusively consists of short-term loans. We also observe in the right plot of Figure 5 a decline in the number of policies characterised by lower grades (1–10) and an increase in the number of policies characterised by higher grades (11–20, 21–35), when the default rate increases.

Figure 5. Topic cluster distribution in the categorical variables `term` and `sub_grade`. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

With regard to numeric covariates, we observe in the left plot of Figure 6 an upward trend where higher interest rates are associated with topic clusters characterised by higher default rates. This increase with the probability of loan default is also observable, but less pronounced, in the right plot of Figure 6 for the revolving line utilisation rate.

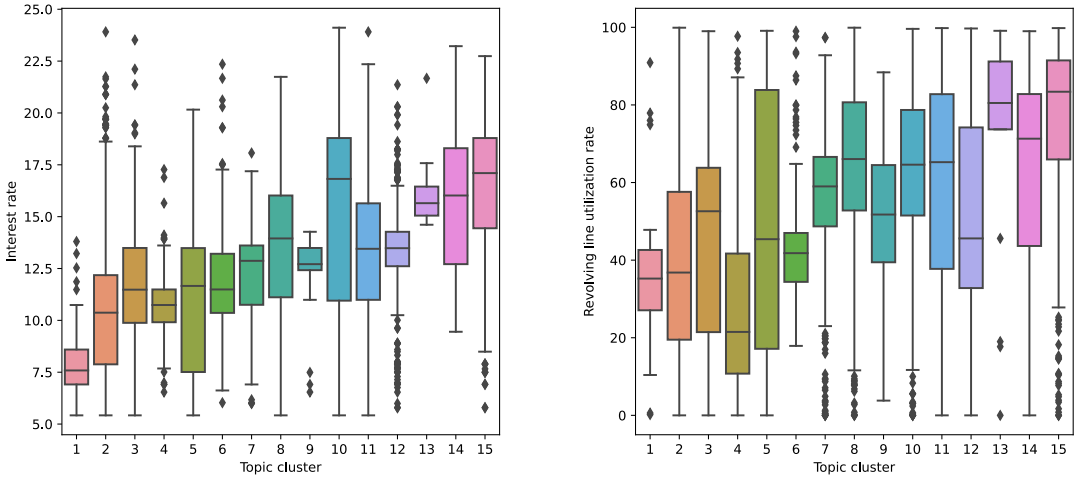


Figure 6. Boxplot of the numeric variables `int_rate` and `revol_util` by topic cluster. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

To confirm LDA's ability to model the underlying structure of structured insurance data, we train our LDA model on a subset (80%) of the motor portfolio introduced above (in Table 2). To strike a balance between capturing the underlying data structure and maintaining a manageable number of clusters for analysis, and after visually analysing the evolution of the deviance on the test set (green line) in the right plot of Figure 4, we chose to set the number of topic clusters K to 10. With $K = 10$, the Poisson deviance is equal to 1 276.51 (with 10 distinct $\hat{y_d}$). In comparison, a GLM trained on the same subset (80%) of the motor portfolio (with the difference that the discretized and categorical covariates are one-hot encoded) gives a deviance equal to 1 339.84 (with more than 1 269 distinct $\hat{y_d}$). With a quite lower level of granularity (10 versus 1 269 distinct $\hat{y_d}$), the LDA clustering model results in a significantly lower deviance than the one obtained with a GLM, which was not the case in our loan portfolio analysis. This suggests that LDA is better at accurately capturing the underlying structures of larger insurance portfolios. The following analysis of the topic clusters composition, summarised in Table 5, further highlights LDA's ability to model the data underlying structure and give insights into the prominent risk factors contributing to higher or lower claim frequencies. This Table 5 gives a summary of some of the most common modalities (in columns titled `Zone`, `Class` and `Gender`) and average numeric values (in columns `OwnersAge` and `VehicleAge`) in each topic cluster. For ease of analysis, the topic clusters (as well as the corresponding figures) are ordered by increasing (average) motor claim frequencies (see `Frequency` column).

Table 5. Most common modalities and average numeric values with LDA soft clustering (motor portfolio). The table is ordered by increasing (average) motor claim frequencies.

Topic	% of policies	Zone	Class	Gender	OwnersAge	VehicleAge	Frequency (%)
1	7.53	4	1	M	48.17	32.19	0.13
2	14.44	4	['3', '4']	M	46.2	14.5	0.29
3	6.01	['4', '1']	3	K	45.0	6.62	0.52
4	15.61	4	['6', '3']	M	56.46	14.0	0.61
5	8.35	['4', '2']	4	M	46.5	13.0	0.79
6	8.01	['4', '1']	['3', '4']	M	32.8	15.0	1.2
7	13.42	['2', '3']	3	M	50.79	2.0	1.38
8	7.71	['3', '6']	5	M	47.18	15.0	1.54
9	10.05	['4', '3']	['6', '3']	M	35.69	6.0	2.03
10	8.87	['4', '2']	['5', '4']	M	26.75	8.0	4.89

Cluster 3, with the third lowest number of claims, has an interesting gender make-up. While all other clusters have a male dominating gender, cluster 3 has the largest proportions of females (see right plot of Figure 7). It is also the cluster with the largest number of insureds located in Sweden's three largest cities (see the proportion of `Zone=1` in the left plot of Figure 7).

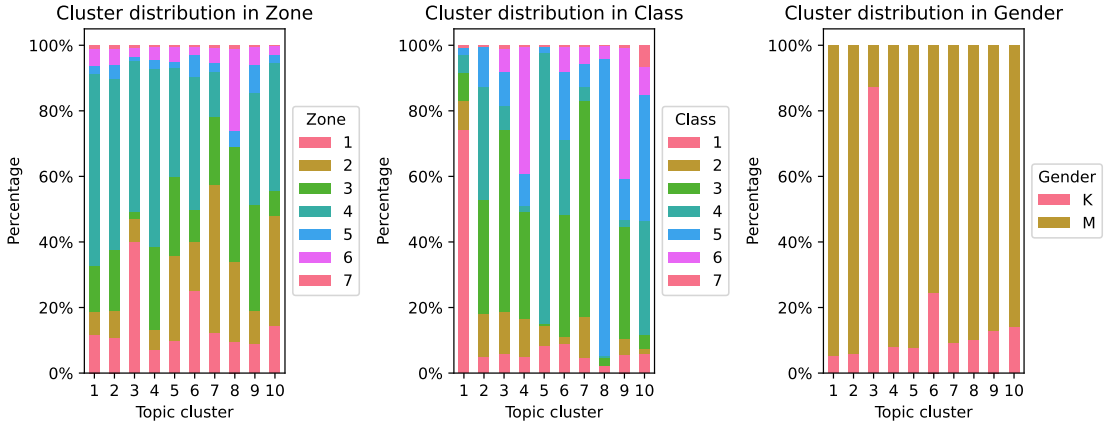


Figure 7. Topic cluster distribution in the categorical variables `Zone`, `Class` and `Gender`. The topic clusters are ordered by increasing (average) motor claim frequencies, as in Table 5.

The left plot of Figure 8 also highlights the differences in age distribution across topic clusters. It is interesting to note that the topic cluster associated with the highest claim frequency is exclusively characterised by younger (less experienced) policyholders (26–27 years old), and is amongst the most concentrated age distribution. Insurance premiums for less experienced drivers are typically higher due to their perceived higher risk profile. We also observe in the right plot of Figure 8 that the topic cluster associated with the lowest claim frequency consists of policies with the oldest vehicle's ages (can refer to collection vehicle) with a very large majority of low EV ratio ≤ 5 (see middle plot of Figure 7). Collection

vehicles might be driven less frequently, reducing the likelihood of accidents.

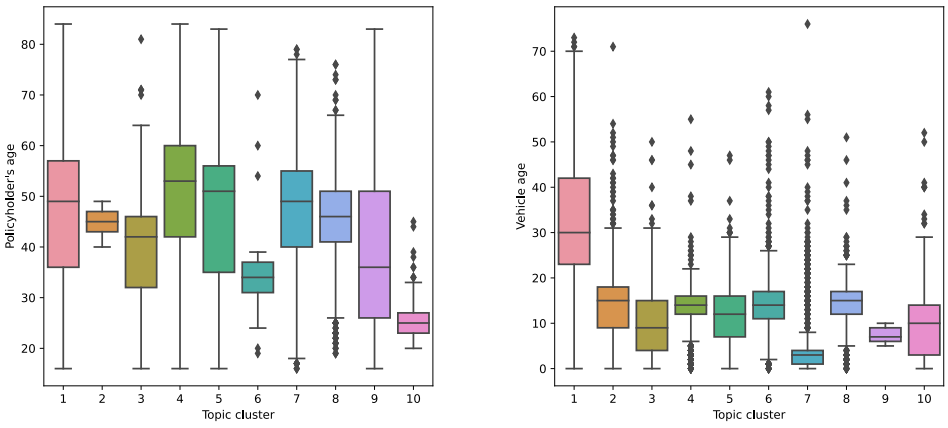


Figure 8. Boxplot of the numeric variables `OwnersAge` and `VehicleAge` by topic cluster. The topic clusters are ordered by increasing (average) motor claim frequencies, as in Table 5.

The following Table 6 provides a summary comparison of the goodness of fit obtained with our LDA soft clustering approach on both (test sets of) the insurance portfolios in Table 1 (loan portfolio) and Table 2 (motor portfolio).

Table 6. Summary of the LDA goodness of fit results on the test sets.

	loan	**motor**
# of (topic) clusters K	15	10
# of policies in the test set	5 184	12 458
Deviance	Binomial	Poisson
LDA deviance	4 221.52	1 276.51
GLM deviance	4 030.3	1 339.84

3.4 Burt adapted fuzzy (or soft) K-means

In this section, we compare our LDA soft clustering approach to the Burt adapted fuzzy (or soft) K-means [27]. In this Burt framework, categorical data are mapped to numeric vectors as follow:

$$\mathbf{B}^{\mathbf{W}} = \frac{1}{q}\mathbf{CBC}\,, \tag{21}$$

where q is equal to the number of categorical (and discretized numeric) variables, $\mathbf{B}$ is equal to the product of the disjunctive matrix $\mathbf{D}$ and its transpose, and the diagonal weight matrix $\mathbf{C} = diag\left(n_{11}^{-1/2} \dots n_{mm}^{-1/2}\right)$ contains on its diagonal the number of policies characterised by the ith modality, denoted by n_{ii}. For more details, we refer to [27]. For a survey of other state-of-the-art mixed data clustering algorithms, see [35]. The following Figure 9 shows the evolution of the Binomial and Poisson deviances with respect to the number of fuzzy K-means

clusters. The decrease in deviance appears smoother with the Burt approach than with LDA for the loan portfolio (see left plot of Figure 4).

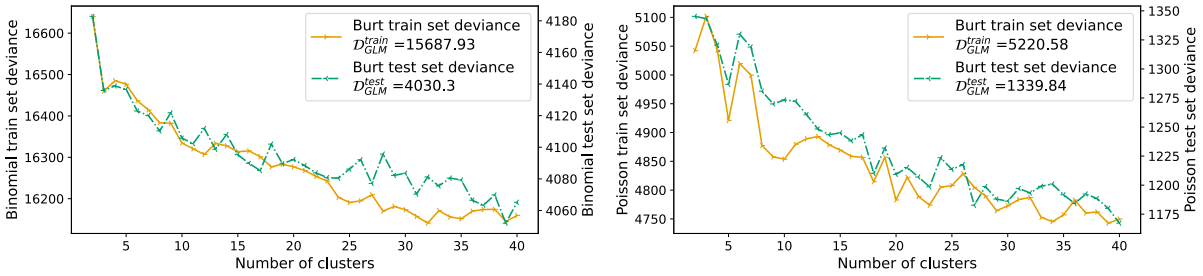


Figure 9. Evolution of the Binomial (left plot) and Poisson (right plot) deviances with respect to the number of K-means clusters $K \in (2, 40)$ in our Burt adapted fuzzy K-means.

If we train our Burt adapted fuzzy K-means on the same subset (80%) of the loan insurance portfolio (in Table 1) as in the LDA approach, with a number of fuzzy cluster $K = 15$ (equals to the number of topic clusters in the LDA approach), we get a Binomial deviance (on the test set) equal to 4 095.31 (with 15 distinct $\hat{y_d}$). In comparison, the GLM deviance was equal to 4 030.3 (with 4 809 distinct $\hat{y_d}$) and the LDA deviance was equal to 4 221.52 (with 15 distinct $\hat{y_d}$). The following Table 7 reports some of the most common modalities and average numeric values in each topic cluster.

Table 7. Dominant modalities and average numeric values (from the loan database) of the continuous variables with K-means Burt adapted soft clustering. The table is ordered by increasing (average) number of loan defaults.

Topic	% of policies	term	sub_grade	int_rate	revol_util	Frequency (%)
1	7.14	36 months	1–10	6.42	17.85	0.0459
2	6.17	36 months	1–10	6.64	33.28	0.0531
3	8.2	36 months	1–10	7.53	19.3	0.0847
4	4.98	36 months	1–10	9.16	36.56	0.093
5	5.59	60 months	1–10	9.91	33.14	0.1034
6	8.16	36 months	1–10	10.31	51.08	0.1064
7	6.67	36 months	1–10	9.86	53.32	0.1098
8	7.64	36 months	1–10	9.88	49.55	0.1111
9	8.55	36 months	11–20	13.09	71.45	0.158
10	7.79	36 months	11–20	12.79	55.82	0.1683
11	6.96	36 months	11–20	13.64	61.82	0.1717
12	6.23	36 months	11–20	12.82	53.21	0.2415
13	5.63	60 months	11–20	15.62	70.47	0.2432
14	6.15	60 months	21–35	17.71	71.01	0.3009
15	4.13	60 months	21–35	16.52	53.79	0.3178

While the Burt and GLM approaches outperform LDA in terms of deviance in this loan portfolio analysis, we observe that the Burt and LDA models give similar insights into the

prominent risk factors contributing to higher or lower claim frequencies. Similarly to LDA, the left plot of Figure 10 highlights the majority (minority) of long-term loans in the clusters associated with higher (lower) default rates. Unlike what we observed in the LDA approach, the first cluster, associated with the lowest default rate, is not exclusively characterised by long-term loans (but still consists of a large majority of long-term loans). Conversely, Burt appears to discriminate better than LDA between different grades. The column `sub_grade` in Table 7 displays a consistent evolution of the grade sub-classes with the increase of default rate. This majority of lower (higher) grades in clusters associated with lower (high) default rate is visually confirmed by the right plot of Figure 10.

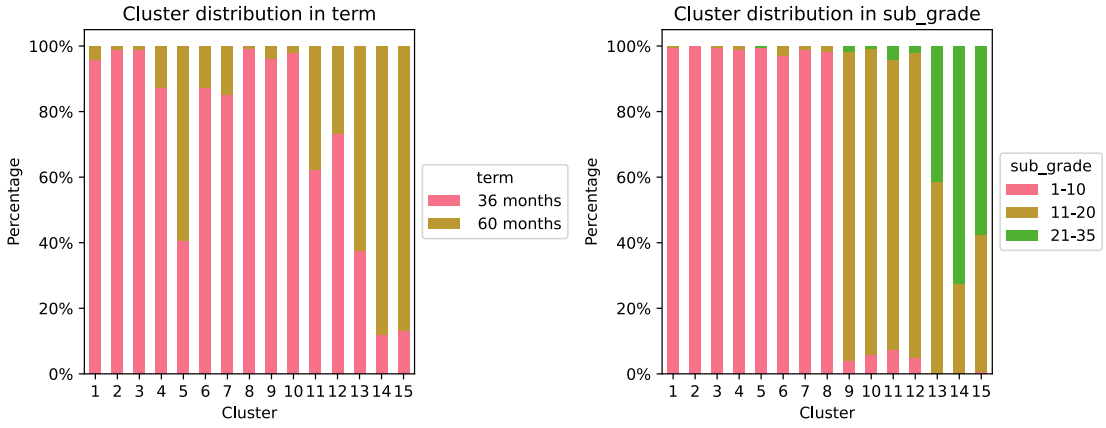


Figure 10. Cluster distribution in the categorical variables `term` and `sub_grade`. The clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

In Figure 11, the upward trend in both the interest rate and revolving line utilisation rate appears to be more consistent than the one observed in the LDA approach in Figure 6.

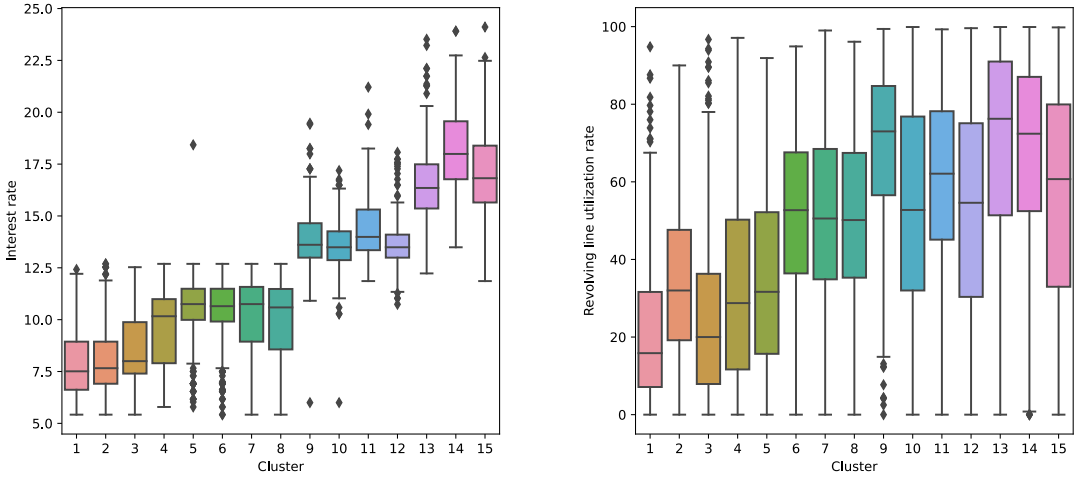


Figure 11. Boxplot of the numeric variables `int_rate` and `revol_util` by cluster. The clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

However, note that the size of the different clusters is more or less the same (see column titled `% of policies` in Table 1). Whereas in the LDA approach, the first three topic clusters already accounted for 50% of the entire portfolio (and reflected the rarity of insurance claims).

To be consistent with the LDA analysis in Section 3.3, we also train our Burt adapted fuzzy K-means model on a subset (80%) of the motor portfolio (in Table 2), with the same number of fuzzy clusters $K = 10$ as the number of LDA topic clusters. It results in a Poisson deviance equal to 1 273.37 (with 10 distinct $\hat{y_d}$), which is quite close to the one obtained with the LDA approach (1 276.51 with 10 distinct $\hat{y_d}$) and still better than the one obtained with the GLM (1 339.84 with more than 1 269 distinct $\hat{y_d}$). While both approaches seem to be able to distinguish high-risk groups from the low-risk groups, the LDA approach produces a higher degree of heterogeneity between the clusters with significantly different claim frequencies. While the average claim frequency of the first four clusters are quite similar in the Burt approach, ranging from 0.42% up to 0.49% (see Table 8), we observe in the LDA approach in Table 5 an average claim frequency of 0.13%, 0.29%, 0.52%, and 0.61% in the first four clusters.

Table 8. Dominant modalities and average numeric values of the continuous variables (from the motor database) with K-means Burt adapted soft clustering. The table is ordered by increasing (average) motor claim frequencies.

Cluster	% of policies	Zone	Class	Gender	OwnersAge	VehicleAge	Frequency
1	9.35	4	1	M	42.94	29.46	0.42
2	14.54	['4', '3']	['3', '4']	K	41.53	11.25	0.45
3	7.59	4	['3', '4']	M	46.2	7.5	0.46
4	13.78	4	['5', '3', '4']	M	47.0	14.0	0.49
5	7.87	3	['5', '3']	M	44.94	15.0	0.57
6	6.62	['4', '2']	['3', '4']	M	53.5	6.0	1.01
7	15.77	['2', '1']	['3', '5']	M	47.4	15.5	1.16
8	7.12	['4', '3', '2']	3	M	46.24	2.67	1.2
9	9.91	['4', '3', '2']	['4', '5']	M	36.5	2.0	3.35
10	7.45	['4', '3']	['6', '5']	M	25.5	7.0	4.8

Nevertheless, the dominant categories and average numeric values obtained across the hardened fuzzy clusters in Table 8 are highly similar to our LDA topic clusters analysis in Table 5. The evolution of the numeric variables with respect to the clusters average claim frequency in Figure 12, as well as the cluster distribution in the categorical variables in Figure 13 are also quite similar to that of the LDA approach. For example, the first cluster (associated with the smallest claim frequency) is once again characterised by older vehicles (see right plot of Figure 13) with the lowest power class (see middle plot of Figure 12). Whereas, the cluster with the highest claim frequency exclusively consists of young (less experienced) drivers (around 25), as illustrated in the left plot of Figure 13 and has the largest proportions of higher power vehicles (see middle plot of Figure 12). The cluster distribution in the `Gender` variable (see right plot of the following Figure 12) is particularly interesting as one cluster (with the second lowest claim frequency on average) contains the majority of the female policyholders in the portfolio, while all other clusters predominantly (if not exclusively) consist of male policyholders.

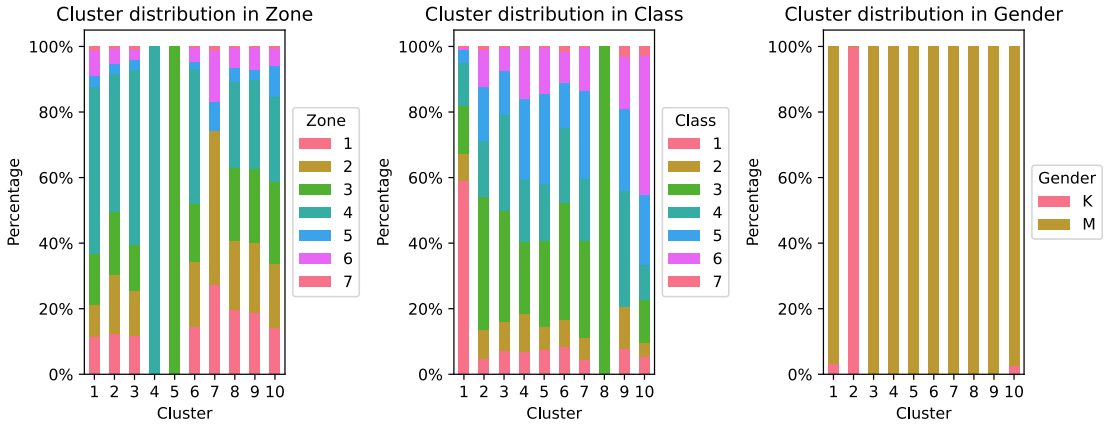


Figure 12. Cluster distribution in the categorical variables `Zone`, `Class` and `Gender`. The clusters are ordered by increasing (average) motor claim frequencies, as in Table 8.

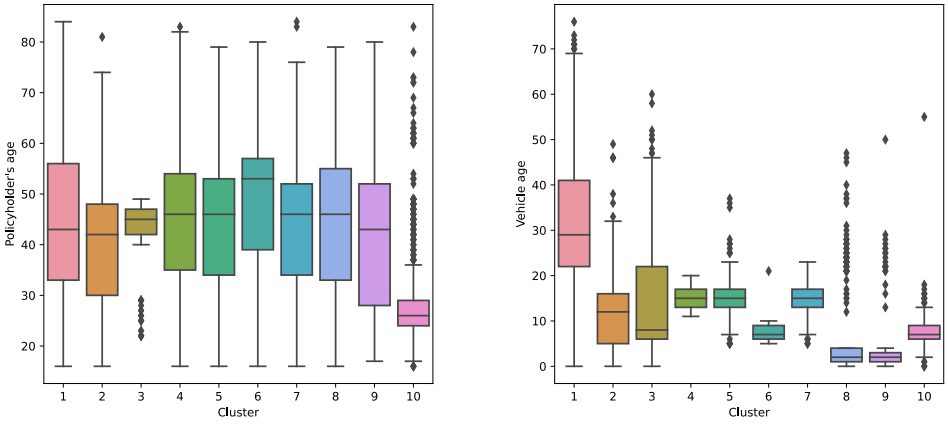


Figure 13. Boxplot of the numeric variables `OwnersAge` and `VehicleAge` by cluster. The clusters are ordered by increasing (average) motor claim frequencies, as in Table 8.

The following Table 9 provides a summary of the goodness of fit obtained in our LDA and Burt soft clustering approaches on both (test sets of) our insurance portfolios.

Table 9. Summary of the goodness of fit results on the test sets.

	loan	**motor**
# of (topic) clusters K	15	10
# of policies in the test set	5 184	12 458
Deviance	Binomial	Poisson
LDA deviance	4 221.52	1 276.51
Burt deviance	4 095.31	1273.37
GLM deviance	4 030.3	1 339.84

While the LDA-based soft clustering approach does not systematically outperform classical models such as the GLM or the Burt adapted fuzzy K-means in terms of deviance, its practical relevance resides in the interpretability and dimensional reduction it provides. For instance, in the loan portfolio, the LDA model exhibits a higher Binomial deviance than the GLM. However, it reduces the prediction space to only 15 interpretable topic clusters. This compression facilitates the identification of high and low-risk groups and the visualisation of prominent risk factors, providing insights into the latent structure of insurance portfolios.

To formally assess whether the increase in the LDA deviance for the loan portfolio is statistically significant, we compute the scaled deviance, adjusting for empirical dispersion $\hat{\phi}$ (theoretically equal to 1) estimated on the training set:

$$\hat{\phi} = \frac{D}{n - p_{\text{model}}}, \tag{22}$$

where n is the number of observations in the training set, p_{model} is the number of regression coefficients for each model. p_{GLM} includes the intercept and all one-hot encoded features, p_{LDA} is approximated as the number of topic clusters.

The difference in scaled deviance is then tested using an approximate chi-square statistic with the null hypothesis H_0 that both models (LDA and GLM) have equivalent predictive performance.

Table 10. Summary of empirical dispersions, train and test deviance, scaled test deviance, and chi-square comparison between GLM and LDA models.

Metric	**GLM**	**LDA**
Empirical dispersion $\hat{\phi}_{\text{train}}$	0.7627	0.8039
Train deviance	15 788.63	16 654.41
Test deviance	4 030.3	4 221.52
Scaled test deviance	5 283.61	5 250.65
Chi-square test on Δ deviance		
Δ deviance	32.96	
Degrees of freedom (approx.)	17	
p-value	0.0114	

For the loan portfolio, the LDA model (with $K = 15$ topic clusters) exhibits a scaled deviance of 5 250.65, compared to the GLM's scaled deviance of 5 283.61, giving $\Delta\text{Dev} = 32.96$ with 17 degrees of freedom and a p-value of 0.0114. Using a significance level $\alpha = 0.01$, we do not reject the null hypothesis, indicating no statistically significant difference in predictive performance between the models, although LDA shows a slightly lower scaled deviance.

The residual heatmaps Figures 16 and 19 in Appendix further illustrate model performance beyond overall deviance. Figure 16 corresponds to LDA residuals, and Figure 19 to Burt residuals. The marginal histograms (top and right) show similar marginal distributions (e.g., policyholder's age, and vehicle age), meaning both models are evaluated on the same range of input features. Differences in the residual pattern stem from the model behavior, not from data imbalance. In the visualizations, the left plot corresponds to the loan portfolio and the right

plot to the motor portfolio, with the continuous predictors densities highlighting regions with more concentrated data where residuals are smaller and more stable. No systematic patterns in residual $y - \hat{y}$ magnitude are apparent in either portfolio across covariate combinations, indicating that predictions remain well-calibrated across the feature space. Larger residuals are observed mainly in low-density regions, reflecting the reduced number of observations rather than model misspecification. These visual diagnostics support the quantitative deviance comparisons reported in Table 9.

4. Conclusions

In this article, we introduced LDA's document-topic distribution as an unsupervised soft clustering tool and extended LDA's probabilistic framework, initially developed for unstructured text data, to tabular insurance data. To model (structured) discrete covariates in tabular insurance portfolios, we defined each topic as a risk profile, each insurance policy as a document and each modality of the categorical and discretized numeric rating factors as a word. Much like the original fuzzy K-means, the soft clustering ability of LDA allows to model each policy by a topic cluster membership. Through the identification of dominant modalities within each topic cluster, insurers can gain insights into risk factors contributing to higher or lower claim frequencies (e.g., default rates).

Our experimental results and analysis showed that the use of LDA's document-topic distribution as an unsupervised soft clustering tool is effective at organising insurance policies characterised by categorical (and discretized numeric) variables into a smaller set of prominent risk profiles (topics). While it did not show to always be as competitive in terms of deviance, comparative analyses with the Burt adapted soft clustering approach still supports the effectiveness of LDA in capturing the underlying structure within structured insurance data.

Building on the findings of this study, several prospects for future research are apparent. First, hybrid approaches that integrate LDA with embedding-based methods (e.g., word or entity embeddings) could enhance the representation of complex interactions among risk factors, capturing subtle dependencies across modalities that pure LDA may overlook. Such approaches may allow insurers to leverage both probabilistic topic structures and continuous latent representations for more accurate risk profiling. Finally, the inherent flexibility of LDA in handling varying document lengths suggests potential for addressing heterogeneous datasets with missing values, or combining discrete and continuous variables via hybrid discretization or embedding techniques. Moreover, dynamic or sequential LDA variants could be applied to better understand how risk profiles develop over time, capturing evolving risk patterns. Collectively, these extensions could strengthen LDA's utility for predictive modelling and actuarial insights.

Funding Statement This research was supported by the project ASTeRISK under research grant ID 40007517, from the Excellence of Science (EoS) program call by FWO and FNRS.

Competing Interests No potential conflict of interest was reported by the authors.

Ethical Approval All authors approved this article.

Data availability statement The loan data set used to support the findings of this study is available at: https://www.kaggle.com/datasets/imsparsh/lending-club-loan-dataset-2007-2011/data. The motor data set used to support the findings of this study is available on the companion website of the book "Non-Life Insurance Pricing with Generalized Linear Models" by Ohlsson and Johansson (2010). Link: https://staff.math.su.se/esbj/GLMbook/case.html

References

[1] Benjamin Avanzi et al. "Machine Learning with High-Cardinality Categorical Features in Actuarial Applications". In: *ASTIN Bulletin* 54.2 (Apr. 2024), pp. 213–238. ISSN: 1783-1350. DOI: 10.1017/asb.2024.7. URL: http://dx.doi.org/10.1017/asb.2024.7.

[2] Ronald Richman. "AI in actuarial science – a review of recent advances – part 1". In: *Annals of Actuarial Science* 15.2 (Aug. 2020), pp. 207–229. ISSN: 1748-5002. DOI: 10.1017/s1748499520000238. URL: http://dx.doi.org/10.1017/S1748499520000238.

[3] Peng Shi and Kun Shi. "Non-Life Insurance Risk Classification Using Categorical Embedding". In: *North American Actuarial Journal* 27.3 (Nov. 2022), pp. 579–601. ISSN: 2325-0453. DOI: 10.1080/10920277.2022.2123361. URL: http://dx.doi.org/10.1080/10920277.2022.2123361.

[4] Bavo D.C. Campo and Katrien Antonio. "On clustering levels of a hierarchical categorical risk factor". In: *Annals of Actuarial Science* 18.3 (Feb. 2024), pp. 540–578. ISSN: 1748-5002. DOI: 10.1017/s1748499523000283. URL: http://dx.doi.org/10.1017/S1748499523000283.

[5] Ronald Richman and Mario V. Wüthrich. "High-cardinality categorical covariates in network regressions". In: *Japanese Journal of Statistics and Data Science* 7.2 (Feb. 2024), pp. 921–965. ISSN: 2520-8764. DOI: 10.1007/s42081-024-00243-4. URL: http://dx.doi.org/10.1007/s42081-024-00243-4.

[6] Paul Wilsens, Katrien Antonio, and Gerda Claeskens. "Reducing the dimensionality and granularity in hierarchical categorical variables". In: *Advances in Data Analysis and Classification* 19.4 (Nov. 2024), pp. 1087–1118. ISSN: 1862-5355. DOI: 10.1007/s11634-024-00614-5. URL: http://dx.doi.org/10.1007/s11634-024-00614-5.

[7] Guojun Gan and Emiliano A. Valdez. "Data Clustering with Actuarial Applications". In: *North American Actuarial Journal* 24.2 (June 2019), pp. 168–186. ISSN: 2325-0453. DOI: 10.1080/10920277.2019.1575242. URL: http://dx.doi.org/10.1080/10920277.2019.1575242.

[8] Malcolm Campbell. "An Integrated System for Estimating the Risk Premium of Individual Car Models in Motor Insurance". In: *ASTIN Bulletin* 16.2 (1986), pp. 165–183. DOI: 10.2143/AST.16.2.2015006.

[9] Ji Yao. "Clustering in General Insurance Pricing". In: *Predictive Modeling Applications in Actuarial Science*. Cambridge University Press, July 2016, pp. 159–179. DOI: 10.1017/cbo9781139342681.007. URL: http://dx.doi.org/10.1017/CBO9781139342681.007.

[10] Shuang Yin et al. "Applications of Clustering with Mixed Type Data in Life Insurance". In: *Risks* 9.3 (Mar. 2021), p. 47. ISSN: 2227-9091. DOI: 10.3390/risks9030047. URL: http://dx.doi.org/10.3390/risks9030047.

[11] Kai Zhuang, Sen Wu, and Xiaonan Gao. "Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms". In: *Tehnički vjesnik* 25.6 (2018), pp. 1783–1791. ISSN: 1848-6339. DOI: 10.17559/tv-20180720122815. URL: http://dx.doi.org/10.17559/TV-20180720122815.

[12] Adrian O'Hagan and Colm Ferrari. "Model-Based and Nonparametric Approaches to Clustering for Data Compression in Actuarial Applications". In: *North American Actuarial Journal* 21.1 (Nov. 2016), pp. 107–146. ISSN: 2325-0453. DOI: 10.1080/10920277.2016.1234398. URL: http://dx.doi.org/10.1080/10920277.2016.1234398.

[13] Yanxi Hou, Xiaolan Xia, and Guangyuan Gao. *Combining Structural and Unstructured Data: A Topic-based Finite Mixture Model for Insurance Claim Prediction*. 2024. DOI: 10.48550/ARXIV.2410.04684. URL: https://arxiv.org/abs/2410.04684.

[14] Oguz Koc, Furkan Baser, and A. Sevtap Selcuk-Kestel. "Clustering based fuzzy classification with a noise cluster in detecting fraud in insurance". In: *Applied Soft Computing* 167 (Dec. 2024), p. 112430. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2024.112430. URL: http://dx.doi.org/10.1016/j.asoc.2024.112430.

[15] Seyed Emadedin Hashemi, Fatemeh Gholian-Jouybari, and Mostafa Hajiaghaei-Keshteli. "A fuzzy C-means algorithm for optimizing data clustering". In: *Expert Systems with Applications* 227 (Oct. 2023), p. 120377. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120377. URL: http://dx.doi.org/10.1016/j.eswa.2023.120377.

[16] Amin Golzari Oskouei et al. "SSFCM-FWCW: Semi-Supervised Fuzzy C-Means method based on Feature-Weight and Cluster-Weight learning". In: *Software Impacts* 21 (Sept. 2024), p. 100678. ISSN: 2665-9638. DOI: 10.1016/j.simpa.2024.100678. URL: http://dx.doi.org/10.1016/j.simpa.2024.100678.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435. DOI: 10.5555/944919.944937.

[18] Thomas Hofmann. "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR99. ACM, Aug. 1999, pp. 50–57. DOI: 10.1145/312624.312649. URL: http://dx.doi.org/10.1145/312624.312649.

[19] Junyu Xuan et al. "Release 'Bag-of-Words' Assumption of Latent Dirichlet Allocation". In: *Foundations of Intelligent Systems*. Springer Berlin Heidelberg, 2014, pp. 83–92. ISBN: 9783642549243. DOI: 10.1007/978-3-642-54924-3_8. URL: http://dx.doi.org/10.1007/978-3-642-54924-3_8.

[20] Yibo Wang and Wei Xu. "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud". In: *Decision Support Systems* 105 (Jan. 2018), pp. 87–95. ISSN: 0167-9236. DOI: 10.1016/j.dss.2017.11.001. URL: http://dx.doi.org/10.1016/j.dss.2017.11.001.

[21] John T. Hancock and Taghi M. Khoshgoftaar. "Survey on categorical data for neural networks". In: *Journal of Big Data* 7.1 (2020). ISSN: 2196-1115. DOI: 10.1186/s40537-020-00305-w. URL: http://dx.doi.org/10.1186/s40537-020-00305-w.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review". In: *ACM Computing Surveys* 31.3 (Sept. 1999), pp. 264–323. ISSN: 1557-7341. DOI: 10.1145/331499.331504. URL: http://dx.doi.org/10.1145/331499.331504.

[23] M.-S. Yang. "A survey of fuzzy clustering". In: *Mathematical and Computer Modelling* 18.11 (Dec. 1993), pp. 1–16. ISSN: 0895-7177. DOI: 10.1016/0895-7177(93)90202-a. URL: http://dx.doi.org/10.1016/0895-7177(93)90202-A.

[24] Kai Wang Ng et al. "Grouped Dirichlet distribution: A new tool for incomplete categorical data analysis". In: *Journal of Multivariate Analysis* 99.3 (Mar. 2008), pp. 490–509. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2007.01.010. URL: http://dx.doi.org/10.1016/j.jmva.2007.01.010.

[25] Duy-Tai Dinh, Van-Nam Huynh, and Songsak Sriboonchitta. "Clustering mixed numerical and categorical data with missing values". In: *Information Sciences* 571 (Sept. 2021), pp. 418–442. ISSN: 0020-0255. DOI: 10.1016/j.ins.2021.04.076. URL: http://dx.doi.org/10.1016/j.ins.2021.04.076.

[26] Matthew D. Hoffman, David M. Blei, and Francis Bach. "Online learning for Latent Dirichlet Allocation". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'10. Vancouver, British Columbia, Canada: Curran Associates Inc., 2010, pp. 856–864. DOI: 10.5555/2997189.2997285.

[27] Charlotte Jamotton, Donatien Hainaut, and Thomas Hames. "Insurance Analytics with Clustering Techniques". In: *Risks* 12.9 (Sept. 2024), p. 141. ISSN: 2227-9091. DOI: 10.3390/risks12090141. URL: http://dx.doi.org/10.3390/risks12090141.

[28] Yi Wang. "Distributed gibbs sampling of latent dirichlet allocation: The gritty details". In: (2011). URL: https://api.semanticscholar.org/CorpusID:6015396.

[29] Bob Carpenter. "Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling". In: *Rapport Technique* 4 (2010), p. 464. URL: https://api.semanticscholar.org/CorpusID:2104516.

[30] William M Darling. "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011, pp. 642–647. URL: https://api.semanticscholar.org/CorpusID:5790714.

[31] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. "Latent dirichlet allocation for tag recommendation". In: *Proceedings of the third ACM conference on Recommender systems*. RecSys '09. ACM, Oct. 2009, pp. 61–68. DOI: 10.1145/1639714.1639726. URL: http://dx.doi.org/10.1145/1639714.1639726.

[32] Julia Fukuyama, Kris Sankaran, and Laura Symul. "Multiscale analysis of count data through topic alignment". In: *Biostatistics* 24.4 (June 2022), pp. 1045–1065. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxac018. URL: http://dx.doi.org/10.1093/biostatistics/kxac018.

[33] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[34] Esbjörn Ohlsson and Björn Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer Berlin Heidelberg, 2010. ISBN: 9783642107917. DOI: 10.1007/978-3-642-10791-7. URL: http://dx.doi.org/10.1007/978-3-642-10791-7.

[35] Amir Ahmad and Shehroz S. Khan. "Survey of State-of-the-Art Mixed Data Clustering Algorithms". In: *IEEE Access* 7 (2019), pp. 31883–31902. ISSN: 2169-3536. DOI: 10.1109/access.2019.2903568. URL: http://dx.doi.org/10.1109/ACCESS.2019.2903568.

Appendix 1. Notations

Table 11. Summary of notation and actuarial interpretation.

Symbol	Mathematical definition	Actuarial interpretation
$\mathbf{X}$	Set of all policies $\{\mathbf{x}_1, \ldots, \mathbf{x}_D\}$	Insurance portfolio
$\mathbf{x}_d$	dth policy with N_d modalities	Individual policy characteristics
x_{dn}	nth modality of policy d	Value of a risk factor (e.g., driver age group)
$\mathcal{V}$	Set of all unique modalities	Collection of all covariate levels
K	Number of latent topics	Number of latent risk profiles
β_k	Topic–modality distribution	Distribution of risk factor combinations in risk profile k
θ_d	Policy–topic mixture	Policy's mixture of latent risk profiles
z_{dn}	Latent topic assignment	Risk profile linked to modality x_{dn}
α	Dirichlet prior for θ_d	Regularity of portfolio-level risk heterogeneity
η	Dirichlet prior for β_k	Smoothness of risk factor representation within a profile

Appendix 2. Sampled distributions

$$p\left(\theta_d|\alpha\right) = \mathcal{D}\left(\theta_d|\alpha\right) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k-1}, \tag{23}$$

$$p\left(\beta_k|\eta\right) = \mathcal{D}\left(\beta_k|\eta\right) = \frac{\Gamma\left(\sum_{v=1}^{V}\eta_v\right)}{\prod_{v=1}^{V}\Gamma\left(\eta_v\right)}\prod_{v=1}^{V}\beta_{k,v}^{\eta_v-1}, \tag{24}$$

$$p\left(x_{d,n}|z_{d,n},\beta_{1:K}\right) = \mathcal{M}\left(x_{d,n}|z_{d,n},\beta_{1:K}\right) = \beta_{z_{d,n},x_{d,n}}, \tag{25}$$

$$p\left(z_{d,n}|\theta_d\right) = \mathcal{M}\left(z_{d,n}|\theta_d\right) = \theta_{d,z_{d,n}}. \tag{26}$$

$$p\left(\theta_d|\gamma_d\right) = \mathcal{D}\left(\theta_d|\gamma_d\right) = \frac{\Gamma\left(\sum_{k=1}^{K}\gamma_{d,k}\right)}{\prod_{k=1}^{K}\Gamma\left(\gamma_{d,k}\right)}\prod_{k=1}^{K}\theta_{d,k}^{\gamma_{d,k}-1}, \tag{27}$$

$$q\left(\beta_k|\lambda_k\right) = \mathcal{D}\left(\beta_k|\lambda_k\right) = \frac{\Gamma\left(\sum_{v=1}^{V}\lambda_{k,v}\right)}{\prod_{v=1}^{V}\Gamma\left(\lambda_{k,v}\right)}\prod_{v=1}^{V}\beta_{k,v}^{\lambda_{k,v}-1}, \tag{28}$$

$$q\left(z_{d,n}=k\right) = \mathcal{M}_K\left(\overrightarrow{\phi}_{d,n}\right) = \phi_{dx_{d,n}k}. \tag{29}$$

Appendix 3. ELBO derivation

The marginal log-likelihood is defined as:

$$\ln p\left(x|\alpha,\eta\right) = \ln\int\sum_{z}p\left(x,z,\theta,\beta|\alpha,\eta\right)d\theta, \tag{30}$$

where $p\left(x, z, \theta, \beta | \alpha, \eta\right)$ is the joint distribution of observed variables x and latent variables z, θ, β. We introduce the variational distribution $q\left(z, \theta, \beta | \phi, \gamma, \lambda\right)$ defined in Eq. (10) and obtain the lower bound from Jensen's inequality:

$$\begin{aligned}\ln p\left(x|\alpha,\eta\right) &= \ln \int \sum_{z} p\left(x,z,\theta,\beta|\alpha,\eta\right) \frac{q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)}{q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)} d\theta = \ln \mathbb{E}_q\left[\frac{p\left(x,z,\theta,\beta|\alpha,\eta\right)}{q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)}\right] \\ &\geq \mathbb{E}_q\left[\ln \frac{p\left(x,z,\theta,\beta|\alpha,\eta\right)}{q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)}\right] = \mathbb{E}_q\left[\ln p\left(x,z,\theta,\beta|\alpha,\eta\right)\right] - \mathbb{E}_q\left[\ln q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)\right] \\ &= \mathcal{L}\left(\phi,\gamma,\lambda;\alpha,\eta\right) ,\end{aligned} \tag{31}$$

where the under-script q is a short form for $q\left(z, \theta, \beta | \phi, \gamma, \lambda\right)$. From Bayesian rules, we have:

$$\mathcal{L}\left(\phi,\gamma,\lambda;\alpha,\eta\right) = -KL\left(q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)||p\left(z,\theta,\beta|x,\alpha,\eta\right)\right) + \ln p\left(x \mid \alpha,\eta\right) \tag{32}$$

The variational form of the marginal likelihood in Eq. (30) is then equal to:

$$\ln p\left(x \mid \alpha,\eta\right) = \mathcal{L}\left(\phi,\gamma,\lambda;\alpha,\eta\right) + KL\left(q\left(z,\theta,\beta|\phi,\gamma,\lambda\right)||p\left(z,\theta,\beta|x,\alpha,\eta\right)\right) . \tag{33}$$

To facilitate calculations (and coding), the tractable lower bound $\mathcal{L}\left(\phi,\gamma,\lambda;\alpha,\eta\right)$ in Eq. (31) is reorganised using the complete model in Eq. (6) and the variational distribution in Eq. (10):

$$\begin{aligned}\mathcal{L}\left(x;\phi,\gamma,\lambda;\alpha,\eta\right) = &\mathbb{E}_q\left[\ln p\left(\beta|\eta\right)\right] + \mathbb{E}_q\left[\ln p\left(\theta|\alpha\right)\right] + \mathbb{E}_q\left[\ln p\left(x|z,\beta\right)\right] + \mathbb{E}_q\left[\ln p\left(z|\theta\right)\right] \\ &- \mathbb{E}_q\left[\ln q\left(\beta|\lambda\right)\right] - \mathbb{E}_q\left[\ln q\left(\theta|\gamma\right)\right] - \mathbb{E}_q\left[\ln q\left(z|\phi\right)\right] .\end{aligned} \tag{34}$$

We now expand the expectations above to get the complete objective function (ELBO) as functions of the variational parameters. In [26], the per-corpus terms are brought into the summation over policies, and are thus divided by the number of policies D, such that the ELBO is equal to:

$$\begin{aligned}\mathcal{L} = &\sum_{d=1}^{D}\left\{\sum_{v=1}^{V} n_{d,v} \sum_{k=1}^{K} \phi_{d,v,k}\left(\mathbb{E}_q\left[\ln \theta_{d,k}\right] + \mathbb{E}_q\left[\ln \beta_{k,v}\right] - \ln \phi_{d,v,k}\right)\right\} \\ &+ \sum_{d=1}^{D}\left\{\sum_{k=1}^{K} a\left(\gamma_{d,k}\right) + \left(\alpha - \gamma_{d,k}\right)\mathbb{E}_q\left[\ln \theta_{d,k}\right]\right\} \\ &+ \sum_{d=1}^{D}\left\{\sum_{k=1}^{K}\left(\sum_{v=1}^{V} b\left(\lambda_{k,v}\right) + \left(\eta - \lambda_v\right)\mathbb{E}_q\left[\ln \beta_{k,v}\right]\right)/D\right\} \\ &+ \sum_{d=1}^{D}\left\{\left(\ln \Gamma\left(K\alpha\right) - K\ln\Gamma\left(\alpha\right)\right) + K\left(\ln\Gamma\left(V\eta\right) - V\ln\Gamma\left(\eta\right)\right)/D\right\} ,\end{aligned} \tag{35}$$

where $a\left(\gamma_{d,k}\right) = \ln\Gamma\left(\gamma_{d,k}\right) - \ln\Gamma\left(\sum_{k=1}^{K}\gamma_{d,k}\right)$, $b\left(\lambda_{k,v}\right) = \ln\Gamma\left(\lambda_{k,v}\right) - \ln\Gamma\left(\sum_{v=1}^{V}\lambda_{k,v}\right)$, $\mathbb{E}_q\left[\ln\theta_{d,k}\right] = \Psi\left(\gamma_{d,k}\right) - \Psi\left(\sum_{k=1}^{K}\gamma_{d,k}\right)$, and $\mathbb{E}_q\left[\ln\beta_{k,v}\right] = \Psi\left(\lambda_{k,v}\right) - \Psi\left(\sum_{v=1}^{V}\lambda_{k,v}\right)$.

Appendix 4. LDA soft clustering

Algorithm 1: Online variational Bayes for LDA [26].

Define iteration weight $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$ where $\tau_0 > 1$ is the learning offset, t is the iteration number of the EM step, and $\kappa \in (0.5, 1.0)$ is the learning decay.
Define convergence condition $K^{-1} \sum_{k=1}^{K} |\text{change in } \gamma_{tk}| < 1e-5$.
Randomly initialise variational parameter $\boldsymbol{\lambda}$.
while *not converged* **do**
  `// E-step`
  Initialize γ_{tk} to an arbitrary constant (e.g., $\gamma_{tk} = 1$);
  repeat
    Compute $\phi_{tvk} \propto \exp\left(\mathbb{E}_q[\ln \beta_{kv}] + \mathbb{E}_q[\ln \theta_{tk}]\right)$;
    Set $\gamma_{tk} = \alpha + \sum_{v=1}^{V} n_{tv} \phi_{tvk}$;
  until *convergence*;
  `// M-step`
  Compute $\tilde{\lambda}_{kv} = \eta + D n_{tv} \phi_{tvk}$;
  Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}}$;
end
Output : β, θ
Compute $\mathcal{L}$ for diagnostics.

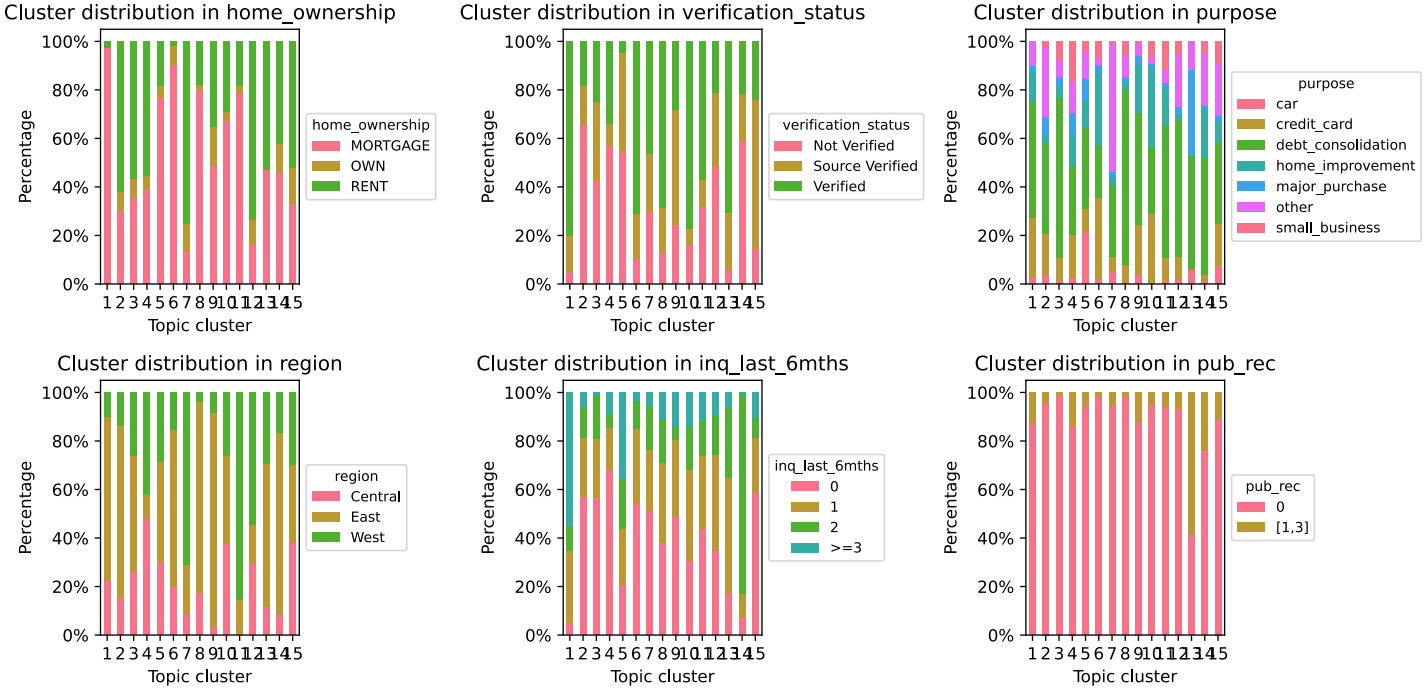


Figure 14. Topic cluster distribution (from the LDA approach) in the categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec`. The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

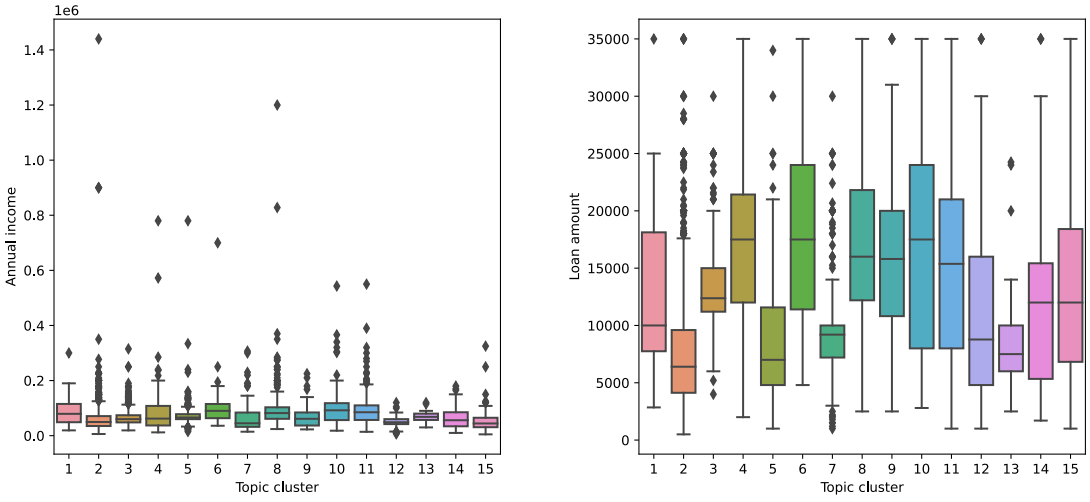


Figure 15. Boxplot of the numeric variables `annual_inc` and `loan_amnt` by topic cluster (LDA clustering approach). The topic clusters are ordered by increasing (average) number of loan defaults, as in Table 4.

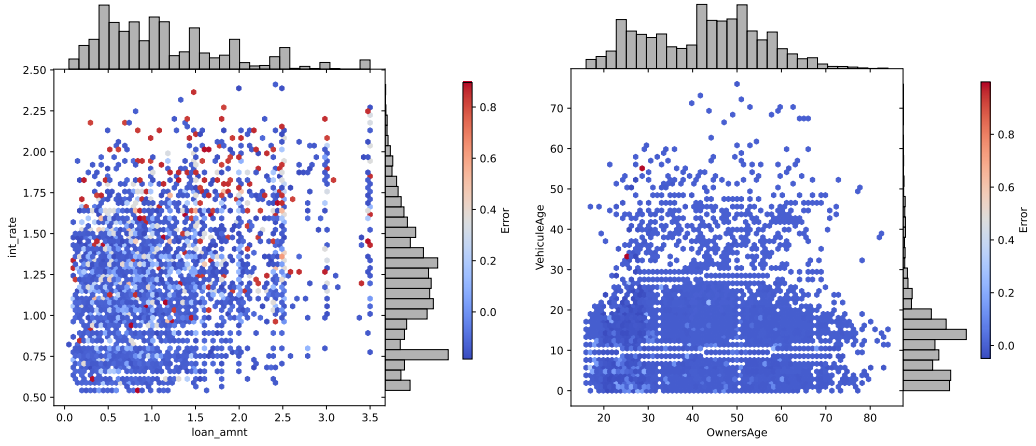


Figure 16. Heatmap of residuals $(y - \hat{y})$ across two continuous predictors, overlaid with the density of the predictors. No systematic pattern in residual magnitude is apparent. Larger residuals are more frequent in low-density regions due to fewer observations. Left plot: loan portfolio. Right plot: motor portfolio.

Appendix 5. Burt-adapted soft clustering

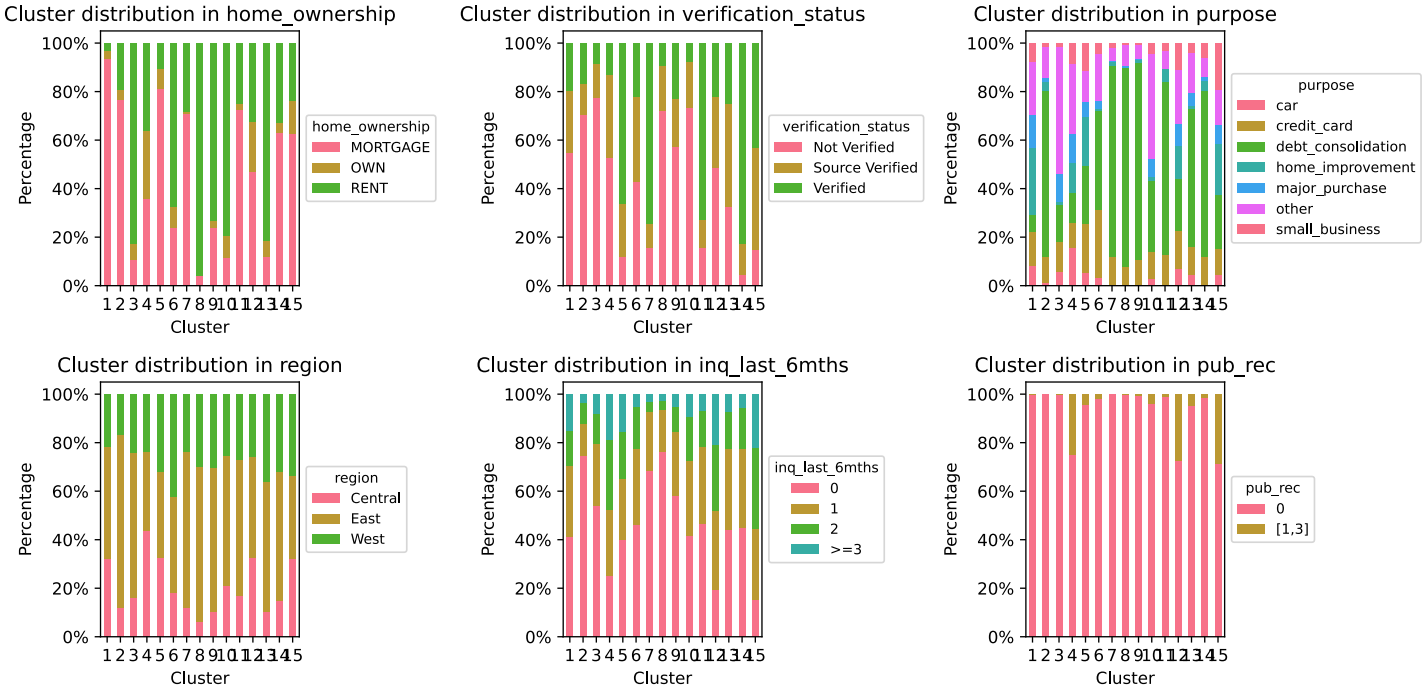


Figure 17. Cluster distribution (from the Burt approach) in the categorical variables `home_ownership`, `verification_status`, `purpose`, `region`, `inq_last_6mths`, and `pub_rec`. The clusters are ordered by increasing (average) number of loan defaults, as in Table 7.

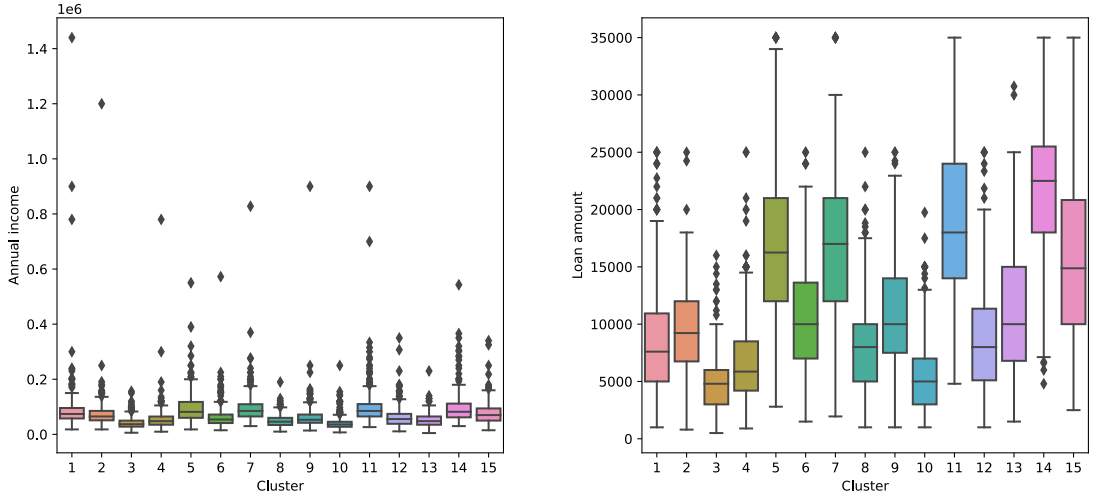


Figure 18. Boxplot of the numeric variables `annual_inc` and `loan_amnt` by cluster (Burt adapted clustering approach). The clusters are ordered by increasing (average) number of loan defaults, as in Table 7.

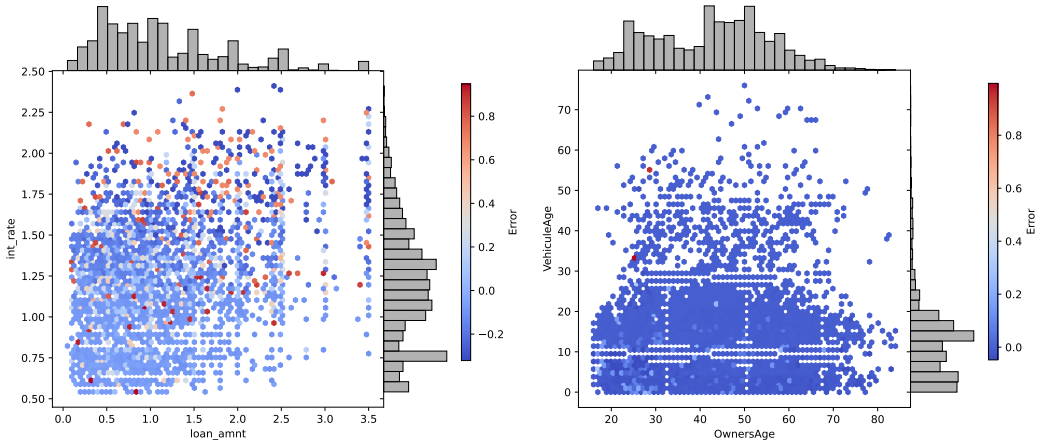


Figure 19. Heatmap of residuals $(y - \hat{y})$ across two continuous predictors, overlaid with the density of the predictors. No systematic pattern in residual magnitude is apparent. Larger residuals are more frequent in low-density regions due to fewer observations. Left plot: loan portfolio. Right plot: motor portfolio.