Latent Dirichlet Allocation to Study Territorial Identity via in-depth Interviews

Roberto Ascari[1], Alice Giampino[1], and Consuelo Rubina Nava[3]

[1] Univerisità degli Studi Milano Bicocca, Milano, Italy,
roberto.ascari@unimib.it and alice.giampino@unimib.it
[2] Univerisità della Valle d'Aosta - Université de la Vallée d'Aoste, Aosta, Italy,
c.nava@univda.it

Abstract. Territories shifting from industrial to tourism-based development face complex processes of identity redefinition. Although in-depth interviews are commonly used to investigate these dynamics, their analysis often remains largely interpretative, limiting reproducibility and the systematic detection of latent themes. To fill this gap, we combine qualitative interviewing with the latent Dirichlet allocation (LDA) probabilistic topic model. By modeling each interview as a mixture of latent topics, the approach enables a data-driven extraction of recurrent themes while preserving the richness of narrative material. The results identify three main thematic dimensions: (i) cultural identity and everyday territorial belonging, (ii) tourism promotion and strategic repositioning, and (iii) development challenges and governance concerns. These topics reveal both continuity with the industrial past and tensions associated with emerging tourism-oriented futures.

Keywords: Latent variables, Textual data, Topic model, Tourism

1 Introduction

Fragile mountain territories undergoing structural economic transition, particularly from an industrial to a tourism-oriented development model, face complex challenges in redefining and consolidating their territorial identity [10]. In such contexts, identity cannot be reduced to institutional branding strategies or policy-driven narratives; rather, it emerges as a stratified and negotiated construct shaped by productive history, collective memory, landscape representations, and everyday practices [11]. When long-standing industrial vocations decline, and tourism is promoted as an alternative development pathway, communities frequently experience tensions between inherited meanings and projected futures [10].

In fragile mountain areas, in many cases characterized by demographic decline, limited accessibility, and fragmented settlements, these transformations can generate identity dissonance, social ambivalence, and internal differentiation. Official narratives often emphasize regeneration and opportunity, yet they may overlook the emotional and cultural dimensions of transition [5, 1].

In destination marketing and participatory place branding, in-depth qualitative interviews represent a crucial methodological tool for identifying both the core elements of territorial identity and the latent criticalities that remain invisible in quantitative data or institutional discourse [9]. Through semi-structured or narrative interviews, researchers can access the lived dimension of transition. Unlike standardized surveys, in-depth interviews allow for the exploration of ambiguity, contradiction, and emotional nuance-dimensions that are central to understanding identity formation in contexts of structural change [4].

Qualitative inquiry also supports an inductive reconstruction of territorial identity, allowing themes to emerge from actors' narratives rather than from predefined categories. Recurring motifs, such as attachment to landscape, work culture, or perceptions of fragility, can be identified as elements of collective identity, while silences or contested memories may reveal underlying tensions. In this sense, in-depth interviews serve not only descriptive purposes but also help detect structural vulnerabilities and fractures in territorial cohesion during periods of economic and symbolic transition [2].

Although in-depth interviews are central to the study of territorial identity in contexts of structural transition, their analysis often remains largely interpretative and manually coded, limiting comparability and the identification of latent thematic patterns. This creates a gap between rich qualitative insight and systematic analytical rigor. Our approach addresses this limitation by combining qualitative interviewing with probabilistic topic modeling. Through the application of latent Dirichlet allocation (LDA) [3], which represents each interview as a mixture of latent topics, we enable a data-driven extraction of recurring themes, enhancing transparency and reproducibility while preserving the depth of narrative material. In this way, we bridge interpretive territorial research and computational text analysis. Moreover, to the best of our knowledge, this study represents the first application of LDA to support participatory place branding strategies and the identification of place identity.

The remainder of the paper is structured as follows. Section 2 presents the methodological framework introducing the LDA model. Section 3 applies the proposed approach to a case study related to a portion of the Aosta Valley territory. Section 4 concludes by summarizing the main contributions of this study.

2 Latent Dirichlet Allocation model

The LDA [3] is the most widely used latent topic model. Although several alternatives and extensions have been proposed in the statistical literature, LDA remains a strong benchmark and a competitive baseline against which new latent topic models are typically evaluated.

The LDA is a hierarchical Bayesian topic model that represents each document (i.e., a single interview) as a mixture of latent topics, while each topic is characterized by a probability distribution over the vocabulary (i.e., the set of unique words observed in the interviews).

Formally, considering a collection of documents $\mathcal{C}$, let K and V denote the number of topics and the size of the vocabulary, respectively. Each document d is represented by a K-dimensional vector $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,K})^\intercal$, where the generic element $\theta_{d,k}$ denotes the proportion of words in document d generated from topic k. Conversely, each topic k is characterized by a V-dimensional vector $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,V})^\intercal$, where $\phi_{k,v}$ represents the probability assigned by topic k to the v-th word in the vocabulary. By construction, $\boldsymbol{\theta}_d$ and ϕ_k lie in the K- and V-part simplex, respectively. The LDA model assumes that

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha}), \qquad \boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta}), \quad k = 1, \ldots, K, \tag{1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ and $\boldsymbol{\beta} \in \mathbb{R}_+^V$ are two sets of hyperparameters controlling both the expected vector and the covariance matrix of the proper Dirichlet distribution.

Thus, the generative process of LDA can be described as follows:

(i) For each topic k, draw the word distribution $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$;
(ii) For each document d, draw the topic proportions $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$;
(iii) For each word $n = 1, \ldots, N_d$ in document d:
   1. draw a latent topic assignment $Z_{d,n} \sim \mathrm{Categorical}(\boldsymbol{\theta}_d)$;
   2. draw the observed word $w_{d,n} \mid Z_{d,n} = k \sim \mathrm{Categorical}(\boldsymbol{\phi}_k)$.

The Dirichlet distributions assumed for $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ imply several theoretical properties, such as the conjugacy between the Dirichlet prior and the multinomial likelihood, which one can use to define an efficient estimation procedure based on the collapsed Gibbs sampling (CGS) [7] algorithm. The latter differs from standard Gibbs sampling approaches since it integrates the parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ out from the posterior distribution, sampling directly from the full conditional distribution of the latent topic assignments, which can be expressed as

$$\Pr(Z_{d,n} = k \mid \mathbf{z}_{-(d,n)}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \left(n_{d,k}^{-(d,n)} + \alpha_k\right) \frac{n_{k,v}^{-(d,n)} + \beta_v}{n_k^{-(d,n)} + \sum_{v=1}^V \beta_v}, \tag{2}$$

where v denotes the index of the observed word $w_{d,n}$. Here, $n_{d,k}^{-(d,n)}$ denotes the number of words in document d currently assigned to topic k, excluding the n-th word, $n_{k,v}^{-(d,n)}$ is the number of times word v is assigned to topic k in the corpus, again excluding the current word, and $n_k^{-(d,n)} = \sum_{v=1}^V n_{k,v}^{-(d,n)}$ represents the total number of words assigned to topic k. Consequently, estimates of $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ can be obtained as the posterior expected vector of $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$, obtained by relying on the above mentioned conjugacy property. These estimates are given by

$$\hat{\theta}_{d,k} = \frac{n_{d,k} + \alpha_k}{N_d + \sum_{k=1}^K \alpha_k} \qquad \text{and} \qquad \hat{\phi}_{k,v} = \frac{n_{k,v} + \beta_v}{n_k + \sum_{v=1}^V \beta_v}, \tag{3}$$

where N_d denotes the total number of words in document d.

3 A case study

The Lower Aosta Valley (LAV), comprising 17 small and medium-sized municipalities in the eastern Aosta Valley (northwestern Italy), is characterized by a longitudinal alpine morphology with relatively accessible valley floors that historically supported industrial development and demographic concentration [10]. It represents a paradigmatic mountain area undergoing a gradual shift from an industrial vocation, rooted in manufacturing and hydroelectric production, toward tourism and service-oriented activities. While material and symbolic traces of its industrial past remain embedded in infrastructure, settlement patterns, and collective memory, current development strategies increasingly emphasize landscape valorization, cultural heritage, and outdoor tourism. To investigate how this transition is locally perceived and narrated, we conducted in-depth semi-structured interviews with diverse territorial actors. This approach captures the experiential dimensions of identity reconstruction, shedding light on how economic reconversion is socially interpreted and collectively managed in a fragile mountain context.

After an initial preprocessing phase to clean and standardize the textual data and a division of the corpus into a training corpus and a test corpus (by randomly selecting 80% and 20% of the words in each document, respectively), the LDA model was applied to the training corpus of 15 interviews, comprising a vocabulary of 497 unique terms. Estimation was performed in `R` using the `EFLDA` package (available at the GitHub repository AliceGiampino/EFLDA).

We specified a symmetric Dirichlet prior for the document–topic proportions, setting $\alpha_k = 50/K$ for $k = 1, \ldots, K$, and a symmetric Dirichlet prior for the topic–word distributions with $\beta_v = 1$ for $v = 1, \ldots, V$. Posterior inference was carried out via Markov Chain Monte Carlo with 10,000 iterations, discarding the first half of the chain as burn-in to mitigate the effect of initialization.

To assess model fit and to guide the selection of the number of topics K, we rely on two complementary criteria: perplexity and topic coherence. Perplexity is a standard predictive performance measure in probabilistic topic models, which evaluates how well a trained model predicts unseen data and is defined as the exponential of the negative average log-likelihood of a held-out corpus. Lower perplexity values indicate better generalization performance. Topic coherence, instead, measures the degree of semantic relatedness among the most probable words within each topic, typically based on their co-occurrence patterns in the corpus; higher values indicate more interpretable and internally consistent topics [8].

Table 1 reports both perplexity and coherence values across different choices of K. Perplexity is minimized at $K = 6$ in the training set and at $K = 3$ in the test set, while coherence is maximized at $K = 3$ for both training and test data. Together, these results suggest that the three-topic specification achieves the best balance between predictive performance and semantic interpretability.

The analysis of the interviews reveals several recurrent thematic dimensions that structure the discourse in the LAV. In Figure 1, the wordclouds associated with the estimated topic–word distributions highlight several recurrent thematic

		Number of topics (K)												
		3	4	5	6	7	8	9	10	11	12	13	14	15
Perplexity	Training	220	215	212.39	**211.00**	211	212	215	216	217	217	218	218	219
	Test	**245**	251	253	256	258	255	257	258	259	260	260	261	261
Coherence	Training	**-0.04**	-0.09	-0.08	-0.12	-0.18	-0.29	-0.32	-0.30	-0.40	-0.45	-0.38	-0.56	-0.42
	Test	**-0.27**	-0.33	-0.32	-0.40	-0.50	-1.40	-3.32	-2.93	-4.56	-3.65	-4.54	-4.79	-4.75

Table 1: Perplexity and coherence for LDA model across different numbers of topics.

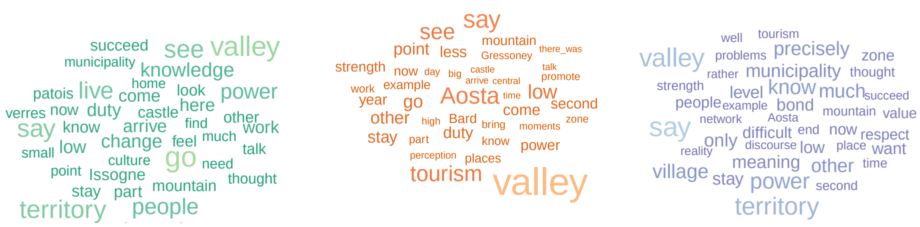

Fig. 1: Word clouds of the estimated proportions of words for each latent topic.

dimensions emerging from the interviews. The first topic is predominantly centred on cultural identity and everyday territorial experience. Highly salient terms such as *culture*, *patois*, *municipality*, *valley*, *people*, *home*, *live*, *territory*, and references to specific localities (e.g., *Issogne*, *Verrès*) underscore the strong embeddedness of narratives in local traditions, community life, and place-based belonging. The prominence of words such as *feel*, *understand*, and *knowledge* further suggests an experiential and relational dimension of identity construction. The second topic appears more closely related to tourism and territorial promotion. Keywords such as *tourism*, *promote*, *mountain*, *Aosta*, *Bard*, *Gressoney*, *zone*, and *central* indicate a focus on attractiveness, visibility, and the strategic positioning of the region. The recurrence of terms like *bring*, *strength*, and *perception* suggests reflections on development opportunities and the role of tourism in shaping the external image of the territory. The third topic emphasizes development challenges and governance-related concerns. Words such as *problems*, *difficult*, *network*, *power*, *value*, and *territory* point to structural constraints, institutional dynamics, and debates around local capacity building. The presence of terms like *meaning*, and *respect* also indicates a normative and forward-looking dimension, reflecting aspirations and tensions connected to future development trajectories.

4 Conclusions

From a methodological perspective, a known limitation of LDA stems from the Dirichlet prior imposed on the document–topic proportions $\boldsymbol{\theta}_d$. Conditional on the total concentration parameter, the Dirichlet distribution induces necessarily negative covariances among components, proportional to the product of their means. This structure implies limited dependence among topic proportions and may restrict modeling flexibility in contexts where themes strongly co-occur within the same narrative [6]. In applications such as the present one, where territorial identity, development challenges, and cultural dimensions are intrinsically intertwined, more flexible prior specifications could potentially capture richer dependence patterns among topics.

Acknowledgments Partially funded by the European Union – NextGenerationEU, Mission 4 Component 1.5 – ECS00000036 – CUP B63B22000010001.

References

[1] Alaoui, Y., Abba, R.: The r (evolution) of territorial marketing: Towards an identity marketing. Journal of Marketing Research and Case Studies 944163 (2019)
[2] Bartůněk, M., Marek, P.: Territorial identity of a region: A review of data collection techniques. Hungarian Geographical Bulletin 74(1), 37–55 (2025)
[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
[4] Della Porta, D.: 10 in-depth interviews. Methodological practices in social movement research p. 228 (2014)
[5] Dredge, D., Jenkins, J.: Destination place identity and regional tourism policy. Tourism Geographies 5(4), 383–407 (2003)
[6] Giampino, A., Ascari, R., Migliorati, S.: A flexible mixed-membership model for community and enterotype detection for microbiome data. Computational Statistics & Data Analysis 210, 108181 (2025)
[7] Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(suppl. 1), 5228–5235 (2004)
[8] Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 conference on empirical methods in natural language processing. pp. 262–272 (2011)
[9] Naipaul, S., Wang, Y., Okumus, F.: Regional destination marketing: A collaborative approach. Journal of Travel & Tourism Marketing 26(5-6), 462–481 (2009)
[10] Pioletti, A.M., Nava, C.R.: Territori montani fragili tra sfide e opportunità: Il caso della Bassa Valle d'Aosta. FrancoAngeli (2025)
[11] Storey, D.: Territories: The claiming of space. Routledge (2024)