The core of the clustering process, as we described, involves updating γ_{it} and ρ_t. Their update step is inherently complex, as it necessitates checking for compatibility issues. The approach entails simulating the assignment of each unit i, currently belonging to cluster j, to either one of the existing clusters or to a new singleton cluster. For each scenario, we compute the probability of this assignment to occur, from which we derive weights to inform the sampling decision for the next iteration. Key elements influencing the definition of these weights include spatial cohesions and, in our JDRPM updated formulation, covariate similarities. We will now explore both these functions.

1.2 Spatial cohesions

rimosso "analysis", come indicato nelle sue annotazioni

Based on our generic joint probability model of (1.7), it is straightforward to incorporate additional information into the partition model such as space or covariates. The incorporation of spatial information can be effectively accommodated through the EPPF in our framework, resulting in spatially informed clusters that evolve over time (Garrit L. Page et al., 2022).

To introduce this extension, let $\boldsymbol{s}_i$ denote the spatial coordinates of the i-th (noting that these coordinates do not change over time), and let $\boldsymbol{s}^\star_{jt}$ denote the subset of spatial coordinates of the units belonging to cluster S_{jt}. Then, we can express the EPPF for the t-th partition in the following product form

$$P(\rho_t|M,\mathcal{S}) \propto \prod_{j=1}^{k_t} C(S_{jt}, \boldsymbol{s}^\star_{jt}|M,\mathcal{S}) \tag{1.19}$$

Compared to the original formulation of (1.5), where $P(\rho_t|M) \propto \prod_{j=1}^{k_t} c(S_{jt}|M)$, (1.19) incorporates a spatial component into the partition weights through the spatial cohesion function $C(S_{jt}, \boldsymbol{s}^\star_{jt}|M,\mathcal{S})$. The original term $c(S_{jt}|M)$ describes how units inside cluster S_{jt} are likely to be clustered together a priori, while the cohesion function $C(S_{jt}, \boldsymbol{s}^\star_{jt}|M,\mathcal{S})$, parametrized by a set of parameters $\mathcal{S}$, measures the compactness of the spatial coordinates $\boldsymbol{s}^\star_{jt}$.

With these two functions established, we transition to a spatially informed dependent random partition model, so that in models (1.8) and (1.9) we can replace tRPM($\boldsymbol{\alpha}, M$) with stRPM($\boldsymbol{\alpha}, M, \mathcal{S}$) to denote our spatio-temporal random partition model (1.7) parametrized by $\alpha_1, \ldots, \alpha_T$ and the EPPF in (1.19).

In this section, we will present the different cohesions functions available in both CDRPM and JDRPM implementations. We will discuss their definition and conduct experiments on each cohesion function, to observe how their tuning parameters influence the computed values. For these experiments, we will consider the clusters configuration illustrated in Figure 1.2, which represents the spatial coordinates of the units of the spatio-temporal dataset that will be used in Chapter 3. For visualization purposes, these values will be presented in a log-transformed form to better highlight differences among the weights assigned to each cluster. Moreover, for clarity, we will use the notation S_h to refer to a generic h-th cluster, rather

All these cohesion functions appeared to agree on the ranking of the clusters shown in Figure 1.2. Cohesion 3 and 4, corresponding to Figures 1.5 and 1.6, clearly indicate the order of clusters as orange, green, purple, and blue, sorted from highest to lowest cohesion weight. This ranking is also reflected in the results of the other cohesions; however, different evaluations can emerge by adjusting their associated tuning parameters. For instance, Figures 1.3, 1.7, and 1.8, corresponding to C_1, C_5 and C_6, show how the singleton (purple) cluster tends to receive the highest weight as the penalization parameters increase. In contrast, cohesion 2, illustrated in Figure 1.4, ranks the singleton cluster at the top, followed by the green cluster, being the first among the non-singletons that activates C_2 when increasing its threshold parameter a, and lastly by clusters orange and blue.

1.3 Covariates similarities idem

The incorporation of covariates information, a characteristic feature of our generalized model, can be integrated into the EPPF (1.19) in a way similar to that used for the spatial information. To introduce this extension, let $X^\star_{jt}$ denote the $p \times |S_{jt}|$ matrix that contains the covariates of the units belonging to cluster S_{jt}, i.e. $X^\star_{jt} = \{\boldsymbol{x}^\star_{it} = (x_{it1}, \dots, x_{itp})^T : i \in S_{jt}\}$. In the current implementation of JDRPM we chose to treat each covariate individually. Therefore, the new term in the definition of the EPPF for $P(\rho_t)$ will be a function of the vector $\boldsymbol{x}^\star_{jtr}$ that collects the values of the r-th covariate for the units inside cluster S_{jt}, i.e. row r of matrix $X^\star_{jt}$. Then, each contribution of the covariates will be considered independently, leading to an EPPF in the form

$$P(\rho_t|M,\mathcal{S},\mathcal{C}) \propto \prod_{j=1}^{k_t} C(S_{jt}, \boldsymbol{s}^\star_{jt}|M,\mathcal{S}) \left(\prod_{r=1}^{p} g(S_{jt}, \boldsymbol{x}^\star_{jtr}|\mathcal{C}) \right) \tag{1.26}$$

This approach is convenient as it can seamlessly accommodate numerical and categorical covariates. Nonetheless, a unified and multidimensional treatment of the covariates would be possible, with appropriate adjustments to the similarity functions, and would yield an EPPF in the form

$$P(\rho_t|M,\mathcal{S},\mathcal{C}) \propto \prod_{h=1}^{k_t} C(S_{jt}, \boldsymbol{s}^\star_{jt}|M,\mathcal{S}) g(S_{jt}, X^\star_{jt}|\mathcal{C}) \tag{1.27}$$

We therefore transition to a spatially and covariates-informed dependent random partition model, so that in model (1.9) we can replace tRPM$(\boldsymbol{\alpha}, M)$ with stRPMx$(\boldsymbol{\alpha}, M, \mathcal{S}, \mathcal{C})$ to denote our spatio-temporal covariates-informed random partition model (1.7) parametrized by $\alpha_1, \dots, \alpha_T$ and the EPPF defined in (1.26).

As in the previous section, we will now present the covariates similarity functions implemented in the JDRPM model, discussing their definition and conducting experiments on each function. These experiments refer to the test case partition illustrated in Figure 1.9, which considers the `Altitude` covariate from the spatio-temporal dataset that will be used in Chapter 3. For consistency, we will again employ a simplified notation by omitting spatio-temporal indicators.

Chapter 3

Analysis of the models

In the following analyses, we will make use of the Adjusted Rand Index (ARI) (Hubert et al., 1985) to compare the partitions generated by the models. The ARI index serves as a correlation metric that quantifies the similarity between two clusterings. Specifically, for two partitions ρ_1 and ρ_2, the function $\text{ARI}(\rho_1, \rho_2)$ produces a value within the range $[-1, 1]$ where higher values indicating greater agreement between the partitions. A perfect match $\rho_1 = \rho_2$ is represented with the limit case $\text{ARI}(\rho_1, \rho_2) = 1$.

We will employ this index to analyse the temporal evolution of the partitions, examining whether ρ_{t+k} correlates with ρ_t, and to evaluate the level of agreement between the two models, by comparing clusters estimates generated by CDRPM and JDRPM. These cluster estimates will be computed using the `salso` function, with the binder loss, using the associated `salso` library (David B. Dahl et al., 2022) on R.

All analyses of this Chapter were conducted on a laptop equipped with 8 GB of RAM and a 1.80 GHz CPU base clock speed. The software used was R (R Core Team, 2024), interfaced with Julia through the `JuliaConnectoR` library (Lenz et al., 2022). This library handled all the communication between R and Julia, where the JDRPM's MCMC algorithm is implemented. The CDRPM implementation is also accessible from R via a dedicated package, `drpm`, developed by (Garrit L. Page et al., 2022), which similarly employs a wrapper to invoke the C code where the MCMC algorithm is implemented.

3.1 Comparing the two algorithms

Our model, along with its corresponding Julia implementation, represents an enhancement over the original DRPM and its associated C implementation. The improvements, as outlined in previous chapters, include the ability to incorporate covariates into both the prior and likelihood levels of the Bayesian model, the possibility of allowing for missing data in the response variable, and the guarantee of greater computational efficiency. In this regard, our updates serve as extensions to the original model. Therefore, when tested under equivalent hyperparameters

Acknowledgements

"This is to be my haven for many long years, my niche which I enter with such a mistrustful, such a painful sensation... And who knows? Maybe when I come to leave it many years hence I may regret it!"

— Fëdor Dostoevskij, *The House of the Dead*

I would like to thank my advisors Alessandra Guglielmi and Alessandro Carminati, primarily for having made me feel, during the entire course of the thesis, in a very calm, lovely atmosphere, in which I have been able to feel perfectly at ease (which is notoriously a NP-hard problem). The epigraphs of my beloved author Dostoevskij and the title inspired by the Star Wars films, as well as some poetic licenses and easter eggs scattered across the chapters (which they have kindly allowed me to keep) I think prove this feeling of sweet but respectful confidence.

qui riscritto in un inglese più corretto

I thank professor Alessandra Guglielmi for having guided me through the fog of uncertainty and puzzlement often associated to the thesis or in general to the final period of university. From the beginning she demonstrated a genuine interest in the topic of the thesis and a confidence in my chances of carrying it out.

I thank Alessandro Carminati for having assisted me in many of the most critical phases of the thesis. His action always precise and effective has been necessary to solve the various theoretical puddles in which I got stuck. Moreover, we share the same passion for Julia and, I believe, the delight for having overthrown the not-so-missed C of the old model implementation.

I thank my family for having helped and supported me during all these years of university, and my friends and fellows who have always and wisely brought to light my highest potential and resources. The passion for solving problems, inventing creative solutions, getting lost in calculations, wandering through the magical world of the most advanced mathematics, sharing doubts and reasonings: all this has always found in you a wonderful and unforeseen companion.

I also thank all the boys and girls I had the pleasure of working and playing with in the various summer camps, for consistently reminding me of the cheerful and mild spirit with which all the various challenges can be faced.

Ringraziamenti

> *"Ecco il mio ponte d'approdo per molti lunghi anni, il mio angoletto, nel quale faccio il mio ingresso con una sensazione così diffidente, così morbosa... Ma chi lo sa? Forse, quando tra molti anni mi toccherà abbandonarlo, magari potrei anche rimpiangerlo!"*
>
> — Fëdor Dostoevskij, *Memorie da una casa di morti*

Vorrei ringraziare innanzitutto i miei relatori Alessandra Guglielmi e Alessandro Carminati, principalmente per avermi fatto sentire, durante l'intero svolgimento della tesi, in un clima molto tranquillo, sereno, in cui ho avuto modo di trovarmi pienamente a mio agio. Le epigrafi del mio caro autore Dostoevskij e il titolo ispirato alla saga di Star Wars, nonché alcune licenze poetiche sparse nei vari capitoli (che loro mi hanno gentilmente concesso di tenere) credo dimostrino questa atmosfera di piacevole ma rispettosa confidenza.

anche qui un attimo rifinita la sintassi italiana

Ringrazio la professoressa Alessandra Guglielmi per avermi guidato attraverso la nebbia di incertezza e confusione che spesso accompagna la tesi o in senso lato gli ultimi mesi di università. Fin dall'inizio ha infatti condiviso un sincero interesse per il lavoro che avremmo dovuto svolgere e una fiducia nelle mie possibilità di condurlo a termine.

Ringrazio Alessandro Carminati per avermi aiutato in molte delle fasi più critiche della tesi. Il suo intervento sempre preciso e puntuale è stato necessario per risolvere le varie pozzanghere teoriche in cui mi ero impantanato. Inoltre, condividiamo la stessa passione per Julia e, credo, la soddisfazione per aver battuto il non-così-compianto C della vecchia implementazione del modello.

Ringrazio la mia famiglia per avermi aiutato e supportato durante tutti questi anni di università, e i miei amici e compagni di studio, i quali hanno sempre sapientemente fatto emergere le mie migliori potenzialità e risorse. La passione per risolvere problemi, ideare soluzioni creative, perdersi nei calcoli, addentrarsi nel magico mondo della matematica più avanzata, condividere dubbi e ragionamenti: tutto questo ha sempre trovato in voi un'ottima e inaspettata compagnia.

Ringrazio anche tutti i bambini e ragazzi con cui ho avuto modo di lavorare e giocare nei vari campi estivi, per farmi sempre ricordare dello spirito gioioso e leggero con cui si possono approcciare tutte le varie sfide.