Spatial Autoregressive Model on a Dirichlet distribution

Teo Nguyen[1,2] & Sarat Moka[3] & Kerrie Mengersen[1,4] & Benoit Liquet[1,2]

[1] *Université de Pau et des Pays de l'Adour, Anglet, France*
[2] *Macquarie University, Sydney, Australia*
[3] *University of New South Wales, Sydney, Australia*
[4] *Queensland University of Technology, Brisbane, Australia*

teo.nguyen@univ-pau.fr
s.moka@unsw.edu.au
k.mengersen@qut.edu.au
benoit.liquet@univ-pau.fr

Résumé. Les données de composition sont largement utilisées dans divers domaines tels que l'écologie, la géologie, l'économie et la santé publique, car elles représentent les proportions ou les pourcentages des différents éléments composant un ensemble. Toutefois, en raison de leur nature relative et de leur contrainte de vivre dans un simplexe, les méthodes statistiques traditionnelles ne sont pas directement applicables aux données de composition (Aitchison 1982). Des dépendances spatiales existent souvent dans les données de composition, en particulier lorsque les composantes représentent des différentes répartitions des terres ou bien des variables écologiques. L'auto-corrélation spatiale peut résulter de conditions environnementales communes ou de la proximité géographique. Il est donc essentiel d'incorporer des informations spatiales dans l'analyse statistique des données de composition afin d'obtenir des résultats précis et fiables. Pour traiter les données compositionnelles, la distribution de Dirichlet est couramment utilisée car son support est un vecteur compositionnel. Maier (2014) a proposé un modèle de régression pour les données à distribution de Dirichlet, mais ce modèle ne prend pas en compte les dépendances spatiales, ce qui limite son applicabilité aux problèmes spatiaux. Dans cette étude, nous présentons un modèle autorégressif spatial pour des données suivant une distribution de Dirichlet qui incorpore des dépendances spatiales entre les observations. Nous développons un estimateur du maximum de vraisemblance sur une fonction de densité de Dirichlet qui inclut un terme dit de "spatial lag". Nous comparons ce modèle autorégressif spatial au même modèle sans "spatial lag" et testons les deux modèles sur des ensembles de données synthétiques et réelles. Différentes matrices de poids spatiales sont utilisées pour tenir compte de leur effet sur l'ensemble des données synthétiques. Les résultats démontrent que l'incorporation des dépendances spatiales peut améliorer la performance du modèle et confirment que l'efficacité dépend de la définition de la matrice des poids (Anselin 1988). En tenant compte des relations spatiales entre les observations, notre modèle fournit des résultats plus précis et plus fiables pour l'analyse des données de composition. Les recherches futures pourraient explorer plus en détail l'application du modèle proposé dans différents domaines et étudier d'autres matrices de poids pour l'analyse des données de composition dans divers contextes spatiaux.

Mots-clés. données de composition, modèle autoregressif spatial, régression de dirichlet.

Abstract. Compositional data are widely utilized in various fields, such as ecology, geology, economics, and public health, as they effectively represent proportions or percentages of different components in a whole. However, due to their relative nature and the constraint of lying on a simplex, traditional statistical methods are not directly applicable to compositional data (Aitchison 1982). Spatial dependencies often exist in compositional data, particularly when the components represent different land uses or ecological variables. Spatial autocorrelation can arise from shared environmental conditions or geographical proximity. Therefore, it is essential to incorporate spatial information into the statistical analysis of compositional data to obtain accurate and reliable results. To handle compositional data, the Dirichlet distribution is commonly used as its support is a compositional vector. Maier (2014) proposed a regression model for Dirichlet-distributed data, but this model does not consider spatial dependencies, which limits its applicability in spatial problems. In this study, we introduce a spatial autoregressive model for Dirichlet-distributed data that incorporates spatial dependencies between observations. We develop a maximum likelihood estimator on a Dirichlet density function that includes a spatial lag term. We compare this spatial autoregressive model with the same model without spatial lag and test both models on synthetic and real datasets. Different spatial weights matrices are employed to account for their effect on the synthetic dataset. The results demonstrate that incorporating spatial dependencies can improve the performance of the model and confirm that the efficiency depends on the definition of the spatial weights matrix (Anselin 1988). By considering the spatial relationships among observations, our model provides more accurate and reliable results for the analysis of compositional data. Future research could further explore the application of the proposed model in different fields and investigate alternative spatial weights matrices for compositional data analysis in diverse spatial contexts.

Keywords. compositional data, dirichlet regression, spatial autoregressive model.

1 Introduction

Compositional data, widely used in different fields such as ecology, geology or economics, are data able to represent proportions or percentages of different components in a whole. We define a D-part compositional vector as a vector $y = (y_1, y_2, \ldots, y_D) \in \mathbb{R}^D$ such that,

$$\begin{cases} y_i \geq 0, \quad \forall\, i \in \{1, 2, \ldots, D\}, \\ \sum_{i=1}^{D} y_i = 1. \end{cases}$$

Compositional vectors lie on a simplex S^D, where traditional statistical methods cannot be applied directly (Aitchison 1982).

One of the most commonly used probability distributions for compositional data is the Dirichlet distribution, as a Dirichlet distribution of parameter $\alpha \in \mathbb{R}^D$ will generate a D-part

compositional vector. Maier (2014) proposed a regression model for Dirichlet-distributed data, but this model does not take spatial dependencies into account.

Over the past decades, spatial autoregressive (SAR) models have emerged as powerful tools for analyzing spatially correlated data in various fields, including economics, ecology, and epidemiology. The fundamental idea behind SAR models is that the value of a variable at a particular location is influenced not only by its own characteristics but also by the characteristics of neighboring locations. These models explicitly account for the spatial interdependencies among the observed variables, allowing for a more comprehensive understanding of the underlying spatial processes. While spatial dependencies are often present in compositional data, particularly when the components represent different land uses or ecological variables, only a few studies have developed a SAR model for such data. In these studies, the authors employed either a Bayesian estimation approach to estimate the parameters of a spatial multinomial logit model (Krisztin et al. 2022) or transform the data into the Euclidian space before applying a multivariate regression model (Nguyen et al. 2021), a Gaussian Markov random field (Pirzamanbein et al. 2018) or a multivariate conditionally autoregressive model (Leininger et al. 2013).

Here, we present a spatial autoregressive model for Dirichlet-distributed data. We develop a maximum likelihood estimator that handles the spatial interdependency and demonstrate the effectiveness of our model on one synthetic dataset and two real-world datasets.

2 Materials and Methods

We consider the case where the labels of the dataset are compositional. Let K be the number of features, J the number of classes, n the sample size of the dataset. We denote the features of a sample i as $x_i \in \mathbb{R}^K$ and its label $y_i \in S^J$ (i.e., y_i is a compositional vector of dimension J). The features (resp., labels) of the whole dataset are then denoted by $X \in \mathbb{R}^{n\times K}$ (resp., $Y \in \mathbb{R}^{n\times J}$). If for a given data row i, the label y_i follows a Dirichlet of parameter $\alpha_i \in \mathbb{R}^J$, then the probability density function is

$$f(y_i|\alpha_i) = \frac{\Gamma(\sum_{j=1}^J \alpha_{ij})}{\prod_{j=1}^J \Gamma(\alpha_{ij})} \prod_{j=1}^J y_{ij}^{\alpha_{ij}-1},$$

where Γ is the gamma function, α_i is such that $\alpha_{ij} > 0$ for every class j. The parameters α_i can be parametrized by $\alpha_i = \phi_i \mu_i$ where $\phi_i \in \mathbb{R}$ is called the precision parameter (or dispersion parameter) and the compositional vector $\mu_i \in S^J$ represents the individual expected values. Hence, the model's predictions $\hat{y}_i$ is given by the estimated values $\hat{\mu}_i$. The parameter ϕ_i has an effect on the distinction of the classes. For a fixed μ_i, the smaller ϕ_i is, the more likely the point will be distributed around extreme values (the edges of the simplex), while with a high ϕ_i, the point is more likely to be close to the value of μ_i (see Figure 1).

All the parameters α_i can be stacked into a matrix $\alpha \in \mathbb{R}^{n\times J}$. Similarly, we can stack all the μ_i (resp. ϕ_i) in a matrix $\mu \in \mathbb{R}^{n\times J}$ (resp. a vector $\phi \in \mathbb{R}^n$).

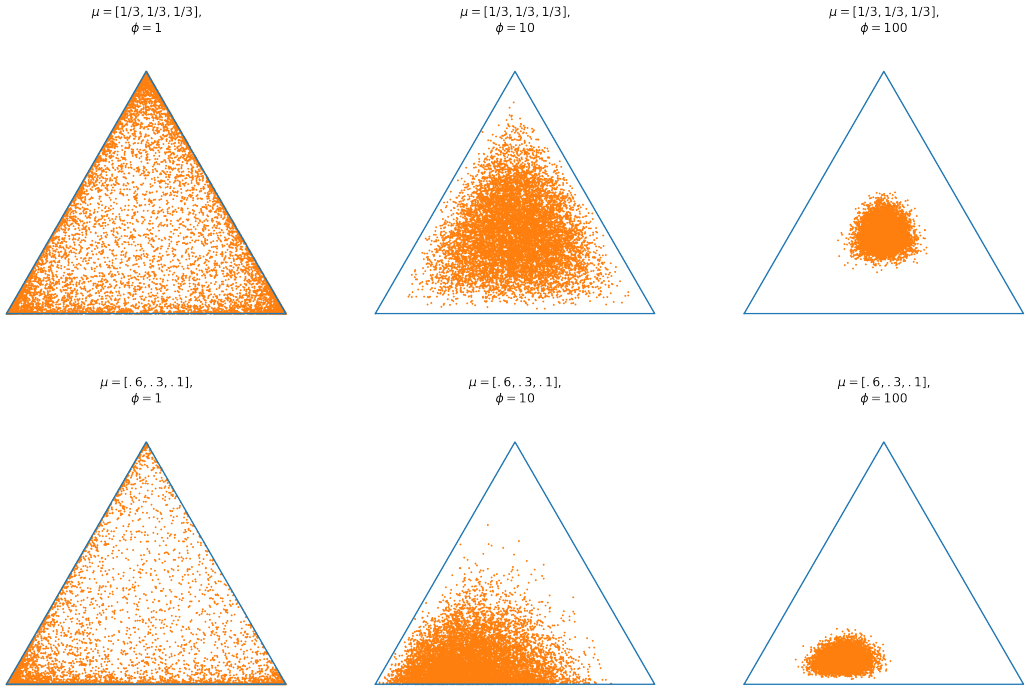

Figure 1: Distribution of 10000 points drawn from a Dirichlet distribution, for different values of $\mu \in S^3$ and $\phi \in \mathbb{R}$.

2.1 Maximum likelihood regression without spatial lag

Let $\beta \in \mathbb{R}^{K \times J}$ be a matrix of coefficients. The parameter $\mu \in \mathbb{R}^{n \times J}$ can be defined as depending of β and the features X such as

$$\forall i \in [1, \dots, n],\ \forall j \in [1, \dots, J], \quad \mu_{ij} = \frac{\exp(\sum_{k=1}^{K} X_{ik}\beta_{kj})}{\sum_{j'=1}^{J} \exp(\sum_{k=1}^{K} X_{ik}\beta_{kj'})}. \tag{1}$$

Then, let $K_Z \in \mathbb{N}$ where $\mathbb{N}$ denotes the set of nonnegative integers. We introduce a matrix $Z \in \mathbb{R}^{n \times K_Z}$ and a vector $\gamma \in \mathbb{R}^{K_Z}$, that allow to define $\phi \in \mathbb{R}^n$ as

$$\forall i \in [1, \dots, n], \quad \phi_i = \exp([Z\gamma]_i).$$

For any row i and class j, we set $\alpha_{ij=}\phi_i\mu_{ij}$, which implies that $\phi_i = \sum_j \alpha_{ij}$. This parametrization is referred to as the *alternative* parametrization in Maier (2014), as opposed to the *common* parametrization where each ϕ_i is set to 1.

To ensure the unicity of the solution when maximizing the likelihood, the mapping $\beta \mapsto \mu$ has to be injective, which is ensured by setting a column of β as 0, for instance the first column, as done in Maier (2014).

The density function can be rewritten depending on μ and ϕ,

$$f(y_i|\mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\prod_{j=1}^{J} \Gamma(\phi_i\mu_{ij})} \prod_{j=1}^{J} y_{ij}^{\phi_i\mu_{ij}-1}. \tag{2}$$

And thus, the log-likelihood of the Dirichlet distribution is,

$$\ell(y|\mu,\phi) = \sum_{i=1}^{n}\left(\ln\Gamma(\phi_i) - \sum_{j=1}^{J}\ln(\Gamma(\phi_i\mu_{ij})) + \sum_{j=1}^{J}((\phi_i\mu_{ij}-1)\ln(y_{ij}))\right) \tag{3}$$

$$= \sum_{i=1}^{n}\left(\ln\Gamma(\phi_i) - \sum_{j=1}^{J}\ln\left(\Gamma\big(\phi_i\frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([X\beta]_{ij'})}\big)\right)\right.$$
$$\left. + \sum_{j=1}^{J}\left((\phi_i\frac{\exp([X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([X\beta]_{ij'})} - 1\right)\ln(y_{ij}))\right). \tag{4}$$

Because of the $\ln(y_{ij})$ term, y_{ij} needs to be strictly positive. This issue is addressed by using the transformation $y^* = \frac{y(n-1)+1/J}{n}$ (Maier 2014), which ensures that the transformed values are positive and has the property that $\lim_{n\to+\infty} y^* = y$. In the following, we will still denote the data as y and assume it does not contain any zero values, but note that the transformation can be applied if necessary.

Because μ and ϕ are parameterized by β and γ, maximum likelihood estimators $\hat{\beta}$ and $\hat{\gamma}$ are used to estimate these parameters, which then allow us to predict the label of an unseen data point $\tilde{x} \in \mathbb{R}^K$. This prediction is the compositional vector $\tilde{\mu} \in S^J$, computed from (1). The probability vector $\tilde{\mu}$ is considered as being the estimated value of the label.

2.2 Maximum likelihood regression with spatial lag

We now introduce a *spatial lag* term through the matrix $M = I_n - \rho W$, where I_n is the identity matrix of size n, $\rho \in \mathbb{R}$ is the strength of spatial correlation and $W \in \mathbb{R}^{n\times n}$ is the spatial weights matrix (Anselin 1988). This spatial lag term will allow us to introduce spatial effect in the model. It is common to apply row-normalization on W in order to make its rows sum to 1. Because of this, this matrix is often asymmetric, even though the original non-normalized weights matrix was symmetric.

For a given matrix β, we redefine μ as:

$$\forall(i,j), \quad \mu_{ij} = \frac{\exp([M^{-1}X\beta]_{ij})}{\sum_{j'=1}^{J}\exp([M^{-1}X\beta]_{ij'})} = \frac{\exp(\sum_{i'=1}^{n}\sum_{k=1}^{K}M^{-1}_{ii'}X_{i'k}\beta_{kj})}{\sum_{j'=1}^{J}\exp(\sum_{i'=1}^{n}\sum_{k=K}^{n}M^{-1}_{ii'}X_{i'k}\beta_{kj'})}. \tag{5}$$

The introduction of M modifies the computation of the vector μ, and by multiplying $X\beta$ with the inverse of M, we introduce the explanatory variables of the neighboring observations.

The value of the spatial correlation parameter ρ needs to be estimated from the data, while W is fixed and has to be defined beforehand. Common choices include distance-based weights or contiguity-based weights (Cliff et Ord 1970). In distance-based weights, the weight between each pair of points is determined by the inverse of the distance between them, while in contiguity-based weights, for each points, the same weight is given to each of its nearest neighbours.

Finally, if we set a matrix $\tilde{X} \in \mathbb{R}^{n\times K}$ such that $\tilde{X} = M^{-1}X$, the loglikelihood remains the same as in (4) provided that we replace the term X with $\tilde{X}$ in this expression.

3 Results

We present here the results obtained from applying both the spatial lag model and the non-spatial model to the different synthetic and real datasets. Each dataset is described and some results are presented for each of them.

3.1 Synthetic dataset

The synthetic spatially-correlated dataset is generated with 2 features and 3 classes, by varying the number of samples n (50, 200, or 1000) and the values of ρ (0.1, 0.5, or 0.9). The values of β and γ are predetermined at the beginning of the simulation. In the presented results, we used

$$\beta = \begin{matrix} \\ \text{features} \end{matrix} \overset{\text{classes}}{\begin{bmatrix} 0 & 0 & 0.1 \\ 0 & 1 & -2 \\ 0 & -1 & -2 \end{bmatrix}}, \qquad \gamma = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

These specific values were selected to ensure certain properties in the generated data. The value of β has been chosen to ensure that the classes are balanced in μ, i.e., that no class is significantly more frequent or rarer than others. The values of γ are chosen in a way that the precision parameter ϕ is sufficiently high, so that the distribution of the points is relatively concentrated around their class probabilities, which ensures well-defined class patterns that are more distinguishable.

First, the features matrix $X \in \mathbb{R}^{n \times 2}$ is created by drawing n samples from a multivariate normal distribution with two covariates. We build the matrix Z with one covariate drawn from a uniform distribution. The matrix W is created by assigning each sample to its row index and identifying its nearest neighbors based on these indices. Here, we considered 5 neighbors, but also tried with more (10 or 20) and the results were suggesting similar performances.

Then, the parameters μ and ϕ are computed from the matrices X and Z and the parameters β and γ. The response matrix Y is finally generated by drawing in the Dirichlet distribution of parameter α_i, defined through μ and ϕ, for each row i.

We perform 100 repetitions of the following: we create the data, and compute the bias of the estimated parameters. In regard to the non-spatial parameters, both models demonstrate similar behavior, with their bias, variance, and mean squared error being asymptotically unbiased. In the spatial models, we also observe the expected behavior, where the bias and mean squared error of the estimated $\hat{\rho}$ decrease as the number of samples increases. Interestingly, when $\hat{\rho}$ is biased (which occurs when the sample size n is small), the bias is negative, suggesting that the model tends to underestimate the spatial correlation strength.

Then, the prediction accuracy of the models is assessed as follow. For each value of ρ, we create the test set by generating 1000 new data points using the true parameters β^*. On this test data, the true value μ^* is computed. Then, we compute $\hat{\mu}$ using the parameters $\hat{\beta}$ estimated from the $n = 1000$ simulation. To evaluate the difference between μ^* and each $\hat{\mu}$,

metrics such as R^2, RMSE, cross-entropy and cosine similarity are utilized.

We observe that for a low spatial correlation ($\rho = 0.1$), both models perform equally well. However, as the spatial correlation increases ($\rho = 0.5$ and $\rho = 0.9$), the performances of the non-spatial models decrease, and they are outperformed by the spatial model across all metrics. Actually, the best performances of the spatial model are observed under a moderately high spatial correlation ($\rho = 0.5$). This suggests that the performances of the spatial model are optimal around a moderate level of spatial dependence, but decline at low or extremely high spatial correlation levels.

3.2 Real datasets

3.2.1 Arctic lake

The Arctic Lake dataset (Coakley et Rust 1968) provides data for $n = 39$ sediment samples taken at various water depths in an Arctic lake. The label is compositional because it correspond to the percentage of sand, silt, and clay in the samples. Here, the goal is to analyze the influence of the water depth on the composition of the sediment samples. Here, we propose two regression models: one with a single predictor variable (the depth) with an intercept term, and another model with an intercept term, the depth variable, and its squared value.

The Leave One Out Cross Validation (LOOCV) strategy is particularly adapted to the small size of the dataset. We iteratively exclude one sample (the k-th sample) from the dataset, and use the remaining samples to estimate the parameters of the models. We then compute the predicted odd values $\hat{\mu}_k$ for the excluded sample, and evaluate its proximity to the true compositional label using metrics such as R^2, RMSE, cross-entropy and cosine similarity.

The results, not detailed here in the seek of simplicity, suggest that the utilization of spatial information leads to slight improvements in model performance. However, in terms of variability, the difference may not be statistically significant. This lack of significance could be attributed to the spatial information being derived solely from the depth variable, resulting in the absence of any new information being introduced. Instead, the data is essentially replicated in a different manner. What would be interesting, for instance, would be to have the exact spatial location of the different samples to use them to build the spatial weight matrix W.

3.2.2 Elections

The elections dataset present the votes at the French departmental election of 2015 in the Occitanie region (Goulard et al. 2017), for $n = 207$ cantons. For each canton, the voting distribution (initially between 15 political parties) is categorized into three major political movements: left, right, and extreme right. In our study, we utilized 25 distinct social indicators as features, including age categories, employment fields, and education level, among

others. Initially, the dataset consisted of 283 cantons, but any cantons where one of the classes was not present were removed. This resulted in the exclusion of 76 points, which represents 27% of the data.

The spatial weights matrix W is computed based on the geographic proximity of each canton's center. Two cases are considered: in the first case, the contiguity-based, we consider the 5 nearest neighboring cantons, determined by their center-to-center distances. In the second case, the distance-based, the inverse of the distance between each canton and the others is considered, with a cut-off at a certain value that minimizes the average number of neighbors and to ensure that each canton has at least one neighbor. This cut-off gives 12 neighbors on average.

The matrix Z is defined as a sole intercept, as this choice yielded the best results compared to the case where Z was a copy of the features matrix X.

For the three models (non-spatial and the two spatials), we use the maximum likelihood estimator to retrieve the parameters and compute the performance with our usual metrics. Results are reported in Table 1. The estimated spatial correlation coefficient $\hat{\rho}$ is 0.97 (resp. 0.91) with the distance-based (resp. contiguity-based) matrix.

Table 1: Scores for the Dirichlet models on Elections dataset.

Model	$\boldsymbol{R^2}$	**RMSE**	**Cross-entropy**	**AIC**	**Cos similarity**
No spatial	0.487	0.080	1.048	-862.1	0.975
Spatial (contiguity)	0.582	0.072	1.042	-947.4	0.979
Spatial (distance)	**0.602**	**0.070**	**1.041**	**-965.1**	**0.980**

The spatial models perform better than the non-spatial model across all evaluation metrics, excepted for the AIC which is slightly better for the non-spatial model. Besides, the distance-based spatial model performs slightly better than the contiguity-based. Additionally, we attempted to make predictions using our model through a 10-fold cross-validation technique, where 90% of the data were used to estimate the model parameters, while the remaining 10% (corresponding to 21 values) were reserved for testing the model's performance. However, we observed extremely poor performance on the test set, indicating that the spatial model is highly sensitive to missing values. This could be explained by the fact that 27% of the initial data was already missing, and removing more data might have rendered the spatial information irrelevant.

4 Conclusion

Our study demonstrates that incorporating spatial dependencies in a Dirichlet model increases the performances of the model. Our results from the real-life datasets reveal that a distance-based spatial weight matrix tends to yield better results compared to a contiguity-based matrix. These results underscore the potential advantages of spatial modeling, especially in scenarios where the Dirichlet distribution is well-suited to the data.

The results obtained from the synthetic dataset provide some insights into the behavior of the SAR Dirichlet model. While the spatial model outperforms the non-spatial model in spatially correlated data, the spatial model does not perform optimally under extremely high spatial correlation and provides better results when the spatial correlation is moderate ($\rho = 0.5$).

Overall, our study highlights the importance of considering spatial information when it provides meaningful additional context, as it can significantly enhance the model's effectiveness. It also emphasizes the potential impact of missing data, which should be carefully addressed to avoid adverse effects on model performance.

Bibliographie

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139-160.

Anselin, L. (2013). *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.

Cliff, A. D., et Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46(sup1), 269-292.

Coakley, J. P., et Rust, B. R. (1968). Sedimentation in an Arctic lake. *Journal of Sedimentary Research*, 38(4), 1290-1300.

Goulard, M., Laurent, T., et Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, 12(2-3), 304-325.

Krisztin, T., Piribauer, P., et Wögerer, M. (2022). A spatial multinomial logit model for analysing urban expansion. *Spatial Economic Analysis*, 17(2), 223-244.

Leininger, T. J., Gelfand, A. E., Allen, J. M., et Silander, J. A. (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 314-334.

Maier, M. (2014). DirichletReg: Dirichlet regression for compositional data in R.

Nguyen, T. H. A., Thomas-Agnan, C., Laurent, T., et Ruiz-Gazen, A. (2021). A simultaneous spatial autoregressive model for compositional data.*Spatial Economic Analysis*, 16(2), 161-175.

Pirzamanbein, B., Lindström, J., Poska, A., et Gaillard, M. J. (2018). Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *Spatial statistics*, 24, 14-31.