Lancaster University
Medical School

Bayesian Spatial Modelling of Environmental and Health Data with Applications in Brazilian Amazonia

Linking Environment and Health in Disadvantaged Groups

by

M.Sc. Erick A. Chacón-Montalván

Supervisor(s):

Dr. Benjamin Taylor and Dr. Luke Parry

Thesis submitted for the degree of
Doctor of Philosophy in Statistics and Epidemiology

Lancaster - United Kingdom

August 2019

A los grandes amores de mi vida, Marilú

y Alfredo.

Acknowledgments

My sincere acknowledgment to the Faculty of Health and Medicine of Lancaster University that provided the funding to pursue my Ph.D. studies. I very much appreciate the support of my supervisors, Benjamin Taylor and Luke Parry. Ben helped me in the process of finding a scholarship to pursue a Ph.D. It has been a grateful experience to work with him because he believed in me, listened to the things I had to say, shared his ideas with me and encouraged me to enjoy this long journey. I have enjoyed our, sometimes, complicated discussions about spatial statistics and we have learned that, sometimes, simple things might be what we need. He has been my main supporter even at the moment things were difficult for him. Thanks for all that, Ben. Luke has also been a big support and a friend to me. He helped a lot in the understanding of the Brazilian Amazonia context and to propose research that is relevant for the Brazilian Amazonia. I appreciate that he was always checking my advances, invited me from time to time for a beer and that he was open to using our tools (R, LaTeX, Git, Linux). Luke, thanks for understanding this sometimes crazy statistician.

Thanks to my family and their good wishes for me. It was a hard time to be without you, but even from long distances, I could feel your love. I love you all. A special thanks to my nephews and nieces: David, Adrian, Camila, Rodrigo, Gael, Fernando, and Valeria. You make me want to be a better person every day, I hope each of you finds what makes you happy in this world and pursue your dreams without excitation.

Finally, thanks to my friends here at Lancaster. I have learned from all of you and you made me enjoy my stay in Lancaster. I could say a lot about every one of you, but you already know how much I appreciate you, in random order, Olatunji, Juan, Laura, Claudio, Camila, Fernando, Lisa, Rachel, Irene, Tobias, Jackie, Pierre, and Carlos.

Declaration

No part of this thesis has been submitted in substantially the same form for the award of a higher degree elsewhere. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussions with my supervisors Benjamin Taylor and Luke Parry.

Erick A. Chacón Montalván
Lancaster University, UK

Bayesian Spatial Modelling of Environmental and Health Data with Applications in Brazilian Amazonia - Erick A. Chacón-Montalván - August 2019

Abstract

Impacts of climate change on human health are a major concern for public health. Increase in frequency and intensity of extreme hydro-climatic events (floods and droughts) is one of the main characteristics of climate change. The occurrence of these events can drastically affect the lives of the population through different pathways. For example, by affecting accessibility to sufficient, safe and nutritious food (*food security*), increasing levels of malnutrition or increasing disease incidence. We hypothesize that nutrition might be a relevant pathway through which extreme hydro-climatic events affect human health and that the impacts are worse for vulnerable groups where they exacerbate existing vulnerabilities. Then, to understand and evaluate the effects of extreme hydro-climatic events on human health, we developed three studies. First, we propose a *model-based standardised index* to identify and quantify extreme temporal events and compared it against the classical standardized precipitation index (SPI). We found that our index holds the properties of the SPI, but improves on the methodology by tackling some of its limitations. Second, we used the model-based standardised index to evaluate the effects of exposure to extreme hydro-climatic events during pregnancy on birth weight. We controlled for other social and placed-based factors that could influence birth weight and found out that floods could significantly reduce birth weight. We also detected characteristics of vulnerable groups where birthweight is expected to be lower. Finally, we proposed our denominated *spatial item factor analysis* to model and predict spatially structured latent factors. With our application on predicting food insecurity in a roadless city of the Brazilian Amazonia, we discover that severely food insecure areas were related to flood-prone, poor and marginalized neighbourhoods. In general, our results highlight the importance of policies to reduce the effects of extreme hydro-climatic events on vulnerable populations of the Brazilian Amazonia. Although our methods were motivated by the study of the impacts of extreme hydro-climatic events, they can be applied in more general cases.

Keywords: Brazilian Amazonia, Climate Change, Extreme Hydro-climatic Events, GAMLSS, Item Factor Analysis, Spatial Statistics.

Table of Contents

Chapter 1
Introduction

1.1 Climate Change and Extreme Events

Mitigating the effects of climate change on health and disease is one of the greatest challenges to public health and international development (McMichael, 2013; Watts et al., 2015). One of the main concerns relates to an expected increase in the frequency, intensity and duration of extreme hydro-climatic events such as floods and droughts (Porporato et al., 2006; Lehner et al., 2006). These natural disasters affect human beings across different dimensions by endangering their basic need for food, water, shelter and good health (McGuigan et al., 2002). In particular, human health is affected by the accessibility to sufficient, safe and nutritious food (*food security*) with malnutrition and higher disease incidence among the consequences of this need not being satisfied (Rosenzweig et al., 2001).

Food security, and consequently nutrition, could be affected by extreme events such as floods and droughts because a regular supply of good quality water is arguably the most important factor in food production (McGuigan et al., 2002). Nutrition, therefore, could be one of the health aspects more affected by extreme hydro-climatic events, impeding the normal development of a population. The effects of extreme events are expected to be more pronounced in vulnerable groups like pregnant mothers, with negative consequences to the health of newborns and subsequent generations, affecting longer term outcomes in education, income and morbidity (Makhija et al., 1989; Risnes et al., 2011; Aizer and Currie, 2014).

1.2 Vulnerable Populations

In this thesis, we work with a definition of vulnerability provided by Blaikie et al. (2014): it is a measure of the capability to anticipate, cope, resist or recover from natural hazards. The impact of extreme events as a result of climate change is likely to vary, depending on the underlying vulnerability of the population being affected (McGuigan et al., 2002).

Underdeveloped populations are therefore likely to be the most vulnerable to climate change and, specifically, to extreme climatic events. One of the main reasons for this is that a region's capacity to adapt to extreme climatic events depends on economic resources for adequate infrastructure, technology and social safety nets. Developing populations/countries simply do not have the resources to prevent and cope with these natural disasters, which limits their adaptive capability. The effects of climate change are just likely to act as an additional burden on their available resources, particularly where natural disasters are already a feature of human existence. At the national level, poor countries are the most vulnerable due to their lack of resources, while at the community level, it can depend on socioeconomic class (e.g. education, type of employment), sex, ethnicity, age and access to resources (McGuigan et al., 2002).

1.3 Brazilian Amazonia

Brazilian Amazonia has recently experienced unprecedented level of extreme hydro-climatic events. For example, a rare drought was registered in 2005, a major flood occurred in 2009, in which the main River Solimões-Amazonas channel reached record levels and a large-scale severe drought was observed in 2010 (Zeng et al., 2008; Chen et al., 2010; Filizola et al., 2014; Lewis et al., 2011). These events drastically affect certain populations of the Brazilian Amazonia where around a million citizens live in urban centres that lack access to Brazil's road network (Parry et al., 2017). These urban centers could be more vulnerable because of the difficulty of trading goods under the absence of road networks and inside these urban centers, disadvantaged groups are likely more affected. In this context, it is pertinent investigate how extreme hydro-climatic events such as droughts and floods affect human health in specific populations of Brazilian Amazonia, where the effects could be exacerbated due to major vulnerability. For instance, Smith et al. (2014) found that drought events impacted health in the Amazon, as detected by an increase in hospitalization rates for respiratory infections, linked to forest fires and air pollution.

1.4 Research Questions

We hypothesise that extreme hydro-climatic events can affect population health, in part, by modifying the levels of food insecurity and, consequently, increasing the number of cases of malnutrition. We expect to be able to detect these effects by studying vulnerable urban centers of Brazilian Amazonia, where the effects could be of major consideration in comparison to the more developed cities in Brazil. Hence, in order to better understand the effects of extreme-climatic events on health status on vulnerable populations of Brazilian Amazonia, we aim to answer the following four research questions (R.Q.):

(R.Q. 1) *How should extreme hydro-climatic events be identified and quantified?:* Before trying to study the effects of extreme hydro-climatic events on health related outcomes, it is necessary to have an approach to identify and quantify these extreme events.

(R.Q. 2) *What are the effects of extreme hydro-climatic events on birthweight?:* Once we have an approach to identify and quantify extreme hydro-climatic events, we will use this approach to evaluate the effects of extreme hydro-climatic events on population health. We decided to work with birthweight, because it is a good indicator of newborn health. Our hypothesis is that the effects of extreme hydro-climatic events are likely to have an impact on the nutrition of the population, including pregnant women, and thus on the health of newborns.

(R.Q. 3) *How can we identify areas of high food insecurity, and what characteristics do these areas have?*: A more direct impact of extreme hydro-climatic events can be obtained by evaluating whether or not highly food insecure areas are related to flood or drought prone areas. However, identifying these areas is not simple given that food insecurity is a latent construct, elicited through the use of questionnaires and is thus not directly observable. Since poor neighbourhoods often appear as clusters within urban centres, it is necessary to develop an approach that allows us to model the latent construct while accounting for spatial correlation in the data. This will allow us to map areas with high or low levels of food insecurity, which can be fed back to policy makers to help develop targeted coping strategies.

(R.Q. 4) *How can we predict food insecurity in unobserved urban centers using secondary data?*: Given the relevance of food insecurity to understanding the effects of extreme hydro-climatic events, and the difficulty and expense of obtaining this information in the field, it is desirable to be able to predict the level of food insecurity in urban centers where we were not able to visit due to budget constraints. We seek a way to utilise our primary data along with more readily available secondary data in order to predict food insecurity in a wider number of similarly isolated urban centers.

While providing answers to these four research questions is directly relevant to beginning our scientific understanding of the effects of extreme hydro-climatic events on vulnerable urban centers of Brazilian Amazonia, they can also be used as a basis for developing an early warning system for food insecurity that takes into account environmental and socio-economic effects.

1.5 Flexible Statistical Models

To answer the research questions presented above, we require statistical models that are flexible enough to handle heteroscedasticity, which is common property of precipitation and river level data (Mckee et al., 1993; Erhardt and Czado, 2017). Models with this kind of flexibility are therefore required for R.Q. 1 and R.Q. 2. For R.Q. 3 and R.Q. 4, we require statistical models that can handle the analysis of latent constructs like food insecurity and for these reasons, we explore models from the theory of *distributional regression* and *factor analysis*. More specifically, we will use *generalized additive models for location, scale and shape* (GAMLSS) and *item factor analysis* respectively. We extend these models when required and combine them with models from *spatial statistics*, given that most of the response variables studied have an inherent spatial structure, this is explored in greater detail in subsequent chapters (Chapter 3, 4, and 5).

1.5.1 Generalized Additive Models for Location Scale and Shape

The main characteristic of a generalized additive model for location, scale and shape (GAMLSS) is that, in addition to the location parameter, the scale and shape parameter are also modeled with respect to specific covariates (Rigby and Stasinopoulos, 2005). More generally, if a response variable Y_i (e.g. precipitation) has a probability density function $f(y_i; \theta_{i1}, \dots, \theta_{iK})$, then each parameter θ_{ik} for $k = 1, \dots, K$ is associated with a linear predictor η_{ik} through a monotonic link function $g_k(.)$ such as

$$g_k(\theta_{ik}) = \eta_{ik}. \tag{1.1}$$

The usefulness of these models is their inherent flexibility: they can adequately capture the behaviour of response variables for which the distributional characteristics change with respect to a set of independent variables.

1.5.2 Item Factor Analysis

Item factor analysis is simply an extension to factor analysis, where the response variable Y_{ij} for item $j = 1, 2, \ldots, q$ and subject $i = 1, 2, \ldots, n$ is a binarization around zero of a continuous but unobservable process Z_{ij} that is explained by m latent factors $\theta_{i1}, \ldots, \theta_{im}$ such as

$$Y_{ij} = \begin{cases} 1 & Z_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad Z_{ij} = c_j + \sum_{k=1}^{m} a_{jk}\theta_{ik} + \epsilon_{ij}, \tag{1.2}$$

where $\epsilon_{ij} \sim \mathcal{N}(0, 1)$, $\{c_j\}$ are intercept parameters that take into account the difficulty of items, and the slopes $\{a_{jk}\}$ indicate how well the j-th item can discriminate the k-th ability between the subjects under study (Bock et al., 1988). For instance, the latent factors θ_{ik} could be the dimensions of food insecurity and Y_{ij}, binary responses to the questions from a questionnaire designed to elicit the level of food insecurity in a household. In the context of food insecurity, these models allow the researcher to: (i) identify the level of food insecurity for subject i; (ii) identify which strategies are used to cope with an absence of food (captured through each question's 'difficulty'); and (iii) understand how the responses to the different questions (items) and dimensions of food insecurity are related. More will be said about what we mean by the 'dimensions' of food insecurity.

We extend item factor analysis by incorporating additional structure on the latent factors using a link function $g(.)$, usually an identity function, and a linear predictor η_{ki} as in Equation 1.1.

1.5.3 Types of Effects

Notice that the structure defined for the linear predictor in Equation 1.1 will depend on the characteristics of θ_{ik}. For example, it could have a seasonal and temporal trend as in R.Q. 1, or non-linear effects as in R.Q. 2. It could also have a spatial structure as in R.Q. 3 and R.Q. 4, or even more complex structures.

For this reason we construct our models using a variety of effects including *smooth functions*, *random effects*, *Gaussian processes* and *Gaussian Markov random fields*; when required we build more complex structures on top of these concepts. The reader should refer to Wood (2006) for a comprehensive review on smooth functions and random effects, to Diggle and Ribeiro (2007) for information on Gaussian processes, and to Rue and Held (2005) for information on Gaussian Markov random fields.

1.5.4 Statistical Inference

In flexible models such as GAMLSS and hierarchical models (e.g. model presented in Chapter 4 and Section 5.2.3), there is evidence in the literature to support the idea that the use of asymptotic approximations in quantifying the uncertainty of estimators might not be reliable (Umlauf et al., 2018). For this reason, we prefer to perform Bayesian inference using Markov chain Monte Carlo in our studies, which allows us to easily deal with missing data and deeper model hierarchies. Under Bayesian inference, the way we make predictions for random effects, expected responses or a function of random variables at different levels of a hierarchical model is also arguably more clear and neat given that we simply use probability theory, treating all unknowns in the model as random variables regardless of whether these are parameters, random effects or other quantities (see Skrondal and Rabe-Hesketh, 2009, Section 7).

1.6 Thesis Structure

This thesis is organized as follows. A brief introduction to the context of the studies and the models used throughout this thesis is presented in the present chapter. We propose a model-based approach to identify and quantify extreme hydro-climatic events in Chapter 2 to answer R.Q. 1. Next, we use this approach to evaluate the effects of extreme hydro-climatic events on birthweight in Chapter 3 to answer

R.Q. 2. In Chapter 4, we address R.Q. 3, proposing a novel approach to modelling spatially-structured latent constructs, which we call *spatial item factor analysis*. We use this to study and map food insecurity in a remote city of the Brazilian Amazonia. Finally, we present the general conclusions and contributions of our studies, with respect to the application studies and statistical models, in Chapter 5. In this, we discuss a possible extension to our *spatial item factor analysis* to model food insecurity across different populations and to allow the inclusion of covariates that are available at different spatial scales, which answers R.Q. 4

Bibliography

Aizer, A. and Currie, J. (2014). The intergenerational transmission of inequality: Maternal disadvantage and health at birth. *Science*, 344(6186):856–861.

Blaikie, P., Cannon, T., Davis, I., and Wisner, B. (2014). *At risk: natural hazards, people's vulnerability, and disasters.* Routledge.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement*, 12(3):261–280.

Chen, J. L., Wilson, C. R., and Tapley, B. D. (2010). The 2009 exceptional Amazon flood and interannual terrestrial water storage change observed by GRACE. *Water Resources Research*, 46(12):1–10.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics (Springer Series in Statistics)*, volume 1. Springer.

Erhardt, T. M. and Czado, C. (2017). Standardized drought indices: a novel univariate and multivariate approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics).*

Filizola, N., Latrubesse, E. M., Fraizy, P., Souza, R., Guimarães, V., and Guyot, J. L. (2014). Was the 2009 flood the most hazardous or the largest ever recorded in the Amazon? *Geomorphology*, 215:99–105.

Lehner, B., Döll, P., Alcamo, J., Henrichs, T., and Kaspar, F. (2006). Estimating the impact of global change on flood and drought risks in Europe: A continental, integrated analysis. *Climatic Change*, 75(3):273–299.

Lewis, S. L., Brando, P. M., Phillips, O. L., van der Heijden, G. M. F., and Nepstad, D. (2011). The 2010 Amazon Drought. *Science*, 331(6017):554–554.

Makhija, K., Murthy, G. V., Kapoor, S. K., and Lobo, J. (1989). Socio-biological determinants of birth weight. *Indian Journal of Pediatrics*, 56(5):639–43.

McGuigan, C., Reynolds, R., and Wiedmer, D. (2002). Poverty and climate change: Assessing impacts in developing countries and the initiatives of the international community. *London School of Economics Consultancy Project for the Overseas Development Institute*, pages 1–40.

Mckee, T. B., Doesken, N. J., and Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *AMS 8th Conference on Applied Climatology*, (January):179–184.

McMichael, A. J. (2013). Globalization, Climate Change, and Human Health. *New England Journal of Medicine*, 368(14):1335–1343.

Parry, L., Davies, G., Almeida, O., Frausin, G., de Moraés, A., Rivero, S., Filizola, N., and Torres, P. (2017). Social Vulnerability to Climatic Shocks Is Shaped by Urban Accessibility. *Annals of the American Association of Geographers*, 4452(October):1–19.

Porporato, A., Vico, G., and Fay, P. A. (2006). Superstatistics of hydro-climatic fluctuations and interannual ecosystem productivity. *Geophysical Research Letters*, 33(15):2–5.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Risnes, K. R., Vatten, L. J., Baker, J. L., Jameson, K., Sovio, U., Kajantie, E., Osler, M., Morley, R., Jokela, M., Painter, R. C., Sundh, V., Jacobsen, G. W., Eriksson, J. G., Sørensen, T. I., and Bracken, M. B. (2011). Birthweight and mortality in adulthood: A systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3):647–661.

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., and Chivian, E. (2001). Climate change and extreme weather events. *Global change & human health*, 2(2):90–104.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications.* CRC Press.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 172(3):659–687.

Smith, L. T., Aragão, L. E. O. C., Sabel, C. E., and Nakaya, T. (2014). Drought impacts on children's respiratory health in the Brazilian Amazon. *Scientific reports*, 4:3726.

Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627.

Watts, N., Adger, W. N., Agnolucci, P., Blackstock, J., Byass, P., Cai, W., Chaytor, S., Colbourn, T., Collins, M., Cooper, A., Cox, P. M., Depledge, J., Drummond, P., Ekins, P., Galaz, V., Grace, D., Graham, H., Grubb, M., Haines, A., Hamilton, I., Hunter, A., Jiang, X., Li, M., Kelman, I., Liang, L., Lott, M., Lowe, R., Luo, Y., Mace, G., Maslin, M., Nilsson, M., Oreszczyn, T., Pye, S., Quinn, T., Svensdotter, M., Venevsky, S., Warner, K., Xu, B., Yang, J., Yin, Y., Yu, C., Zhang, Q., Gong, P., Montgomery, H., and Costello, A. (2015). Health and climate change: Policy responses to protect public health. *The Lancet*, 386(10006):1861–1914.

Wood, S. S. (2006). *Generalized Additive Models: An Introduction with R.* CRC Press.

Zeng, N., Yoon, J.-h., Marengo, J. A., Subramaniam, A., Nobre, C. A., Mariotti, A., and Neelin, J. D. (2008). Causes and impacts of the 2005 Amazon drought. *Environmental Research Letters*, 3(1):014002.

Chapter 2

As mentioned in Chapter 1, this thesis focuses on the analysis of the impacts of extreme hydro-climatic events on the Brazilian Amazonia population health. For this reason, it is first necessary to obtain a methodology to identify and quantify the magnitude of extreme hydro-climatic events like the standardized precipitation index (SPI), which is a widely-used and accepted index for these purposes. Unfortunately, this index has certain limitations that we overcome in the present chapter by proposing two model-based alternatives and comparing them against the SPI.

A Model-Based General Alternative to the Standardised Precipitation Index

Erick A. Chacón-Montalván[1], Luke Parry[2,3], Gemma Davies[2], Benjamin M. Taylor[1]

[1]Centre for Health Informatics, Computing, and Statistics (CHICAS), Lancaster Medical School, Lancaster University, United Kingdom.
[2]Lancaster Environment Centre, Lancaster University, United Kingdom.
[3]Núcleo de Altos Estudos Amazônicos, Universidade Federal do Pará, Belém, Brazil

Abstract

In this paper, we introduce two new model-based versions of the widely-used standardized precipitation index (SPI) for detecting and quantifying the magnitude of extreme hydro-climatic events. Our analytical approach is based on generalized additive models for location, scale and shape (GAMLSS), which helps to overcome some limitations of the SPI. We compare our model-based standardised indices (MBSIs) with the SPI using precipitation data collected between January 2004 - December 2013

(522 weeks) in Caapiranga, a road-less municipality of Amazonas State. As a result, it is shown that the MBSI-1 is an index with similar properties to the SPI, but with improved methodology. In comparison to the SPI, our MBSI-1 index allows for the use of different zero-augmented distributions, it works with more flexible time-scales, can be applied to shorter records of data and also takes into account temporal dependencies in known seasonal behaviours. Our approach is implemented in an R package, `mbsi`, available from Github.

Keywords: Droughts, Extreme Events, Flexible Regression Models, Floods, GAMLSS, SPI.

2.1 Introduction

Mitigating the effects of climate change on health and disease is one of the greatest challenges to public health and international development (McMichael, 2013; Watts et al., 2015). One of the main characteristics of the burden of climate change is the expected increase in the frequency, intensity and duration of extreme climate events (Houghton et al., 2001; Rosenzweig et al., 2001). These are events experiencing extreme values of meteorological variables, they often cause damage and are defined as either taking maximum values or exceeding established high thresholds (Stephenson, 2008). In this paper, our focus is on floods and droughts, which are considered extreme hydro-climatic events because they are related to the tails of streamflow distribution (Shelton, 2009).

The impact of extreme hydro-climatic events is not straightforward to understand because they comprise a complex web of direct and indirect impacts on environmental, economic and social areas (Blanka et al., 2017). Floods and droughts, depending on their severity, can produce not only crucial damage to the economy and ecology of a region, but also lives can be endangered (Lehner et al., 2006). Agriculture and associated sectors are highly dependent on surface and ground water; hence, it is common to see major impacts of droughts and floods on these areas (Blanka et al., 2017). The impact of extreme hydro-climatic events will also crucially depend on specific characteristics of the society affected like their vulnerability, adaptive capacity and resilience (Seiler et al., 2002; World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016). Hence, focus should be put on vulnerable societies that are prone to experience extreme hydro-climatic events; for example, on roadless urban centres of the Brazilian Amazonia, where the population is experiencing droughts and floods without precedent (see Zeng et al., 2008; Chen et al., 2010; Filizola et al., 2014; Lewis et al., 2011).

In this context, the importance of being able to identify extreme hydro-climatic events is due to two main reasons. First, it can help to improve the understanding of the effects of floods and droughts by allowing the analysis of extreme hydro-climatic events with respect to different variables or indicators of interest in health, economy or others. For example, Chacón-Montalván et al. (2018) evaluates the effects of these events on newborn health measured through birthweight. Second, the methodology for the identification of extreme events can help to improve monitoring and prediction tools and, potentially, enhancing prevention policies to reduce the impacts of floods and droughts.

There are a large number of indices and indicators for monitoring droughts. World Meteorological Organization (WMO) and Global Water Partnership (GWP) (2016) presented 49 indicators and indices classified among the categories meteorology, soil moisture, hydrology, remote sensing, and composite or modelled. Between these indices, the most common are the standardized precipitation index (SPI), the Palmer drought severity index (PDSI), the crop moisture index, the surface water supply index and the vegetation condition index (Mishra and Singh, 2010). Comparisons between these indices, often, agree that the standardized precipitation index is an appealing index for monitoring droughts because of its simplicity, spatial invariance, probabilistic nature and its flexibility to work with different time-scales (Guttman, 1999; Hayes et al., 1999; Morid et al., 2006; Mishra and Singh, 2010). In addition, the World Meteorological Organization has suggested to use the SPI as a primary meteorological drought index through Hayes et al. (2011) and a user guide for this index has been released in World Metereological Organization (2012).

In the case of flood monitoring, most studies focus on more than one indicator given that flooding is not only related to rainfall, but also to river levels, river discharge and geomorphology. In comparison with the case of droughts, there is not much consensus in which indices or information to use for monitoring floods. Koriche and Rientjes (2016), for example, used rainfall and topography to propose a satellite based index, while Ban et al. (2017) used satellite-based RGB composite imagery. Other approaches applied sensor networks or information from hydrological stations (Keoduangsine and Goodwin, 2012). Despite this variability of methodologies, several studies recognize the potential value of the SPI as a tool for flood monitoring. For instance, Wang et al. (2017) demonstrated that the 2-month SPI is an effective indicator for identifying major floods events in the Minjiang River basin. Similarly, Seiler et al. (2002); Guerreiro et al. (2008); Du et al. (2013); Koriche and Rientjes (2016) have used the SPI for flood predicting systems.

Motivated by the desire to evaluate the impacts of extreme hydro-climatic events on birthweight in the Brazilian Amazon (see Chacón-Montalván et al., 2018), our research initially explored the use of the widely applied standardized precipitation index (SPI). However, although this index has been suggested as the primary meteorological drought index by the World Meteorological Organization and has been shown to be useful for identifying and monitoring droughts and floods, the current methodology for computing it has certain limitations that will be explained in Section 2.2.3. For instance, the SPI can not be computed reliably for series shorter than 30 years. For this reason, we propose two model-based approaches that maintain the desirable characteristics of the SPI but with improved computation and methodology.

Our *model-based standardized indices* (herein, MBSIs) overcomes some of the limitations of the SPI by using generalized additive models for location, scale and shape (GAMLSS). These models are flexible enough to capture the seasonal trend on the parameters of the distribution of rainfall or precipitation data. Our methodology differs from other attempts to improve the SPI by proposing a model-based approach instead of a group of empirical steps to compute the SPI such as presented in Erhardt and Czado (2017). A model-based approach provides a more consistent framework that naturally allows model checking and uncertainty computations. Also, it could allow further extensions; for example, by working on the spatial or spatio-temporal scale, or by taking into account additional structures such us trends and covariates effects. The MBSIs could be applied to other environmental variables of interest, other than precipitation, by choosing an appropriate family of distributions.

This paper is structured as follows. An introduction explaining the motivation for an alternative to the SPI is given in the present section. Then the definition and limitations of the SPI are presented in Section 2.2. In Section 2.3, we provide a short introduction to generalized additive models for location, scale and shape (GAMLSS). In Section 2.4, two model-based approaches to compute the standardized precipitation index are proposed to tackle some of the limitations described in Section 2.2.3 and make possible the use of a theoretically similar index in our study for birthweight (see Chacón-Montalván et al., 2018). After presenting the MBSIs, in Section 2.5, we compare the SPI and MBSIs using precipitation data collected between January 2004 - December 2013 in Caapiranga, a road-less municipality in the Amazonas State. Finally, conclusions and a discussion of the performance of our method is given in Section 2.6.

2.2 Standardised Precipitation Index

The SPI is an index that was proposed by Mckee et al. (1993) to improve drought detection and monitoring capabilities using statistical concepts. This index quantifies how extreme are the observed precipitation values with respect to the mean seasonal behaviour. The main characteristics of this index are simplicity, spatial invariance, probabilistic nature and flexibility to work with different time-scales (Guttman, 1999). This last characteristic allows monitoring of different types of droughts like agricultural (short time-scale) and hydrological (long time-scale) (Mckee et al., 1993).

Therefore, to compute the SPI, it is necessary to choose a time-scale over which to smooth the original precipitation data; this smoothing enables the method to detect extreme events that occur over a period. The computation continues by mapping the empirical cumulative distribution function to a standard normal distribution. The resulting series of values are interpretable as quantiles from a standard normal distribution. For example, an SPI value of 2 indicates that the probability of observing an event at least as extreme as this is 0.0228. In the next sections, we describe the computation of the SPI with further detail (Section 2.2.1), present the approach to monitor floods and droughts using the SPI (Section 2.2.2), and discuss some limitations of the SPI (Section 2.2.3).

2.2.1 Definition of the SPI

In this section, we outline the methodology of Mckee et al. (1993) for computing the SPI for a monthly time series of aggregated precipitation, represented as a discrete-time stochastic process, $\{Z_t : t = 1, \ldots, T\}$. Throughout this section we will refer to $\{Z_t\}$ as the 'monthly precipitation', but the reader should bear in mind that we intend $\{Z_t\}$ to be thought of in more general terms because the methodology can, in theory, be easily applied to other variables such as river levels, river discharge, etc.

We begin by defining $\{X_t^k : t = 1, \ldots, T\}$ as the k-order moving average process of $\{Z_t\}$ such as

$$X_t^k = \frac{1}{k} \sum_{i=0}^{k-1} Z_{t-i}, \quad \text{for } t = 1, \ldots, T, \tag{2.1}$$

i.e. x_t^k is the average of the observed precipitation of the last k months, inclusive of the present month t. In the literature of drought indices, k is referred to as the 'time-scale' under study. The ability to define k prior to analysis is considered one of the appealing characteristics of the SPI (Guttman, 1998).

Rather than employing formal statistical methods for selecting k, the choice of k is determined by the time-scale under consideration by the researcher. For example, if one is interested in detecting droughts that occur over long periods of time (e.g. during a year), then $k = 12$ might be chosen; similarly for analysing quarterly droughts $k = 3$ might be more appropriate. The choice of time-scale can be related to the particular type of drought impact of interest. Different values of k shift the focus of an analysis to different types of extreme events; this is important given that the lack of water in the short, medium or long-term affects different sections of human society and the surrounding ecosystem in different ways (e.g agricultural or hydrological effects) (Mckee et al., 1993). In the interest of disaster prevention, or planning a humanitarian response to a drought, the actions taken will be different for droughts at different time scales. For instance, events occurring on a short time-scale may be important to agricultural decisions whereas events on longer time-scales may be of more relevance for the management of water supplies (Guttman, 1998, 1999).

To continue with the definition of the SPI, it is beneficial to switch notation for the subscript t, replacing Z_t and X_t^k by respectively Z_{ij} and X_{ij}^k, where $i = 1, 2, \ldots, n$ is the year and $j = 1, 2, \ldots, 12$ is the month under study. We next introduce a statistical model for X_{ij}^k, i.e. a parametric density function, $h_j(X_{ij}^k = x; \cdot)$, where x is an arbitrary value on the domain of X_{ij}^k. Notice that the notation

$h_j(\cdot\,;\,\cdot)$ implies that the characteristics of the density function change according to the month of the year, i.e. it has a seasonal behaviour. In the original article, Mckee et al. (1993) suggested a gamma density for $h_j(\cdot\,;\,\cdot)$, but current practice instead makes use of a mixture, a zero-augmented gamma density (ZAGA), which allows X_{ij}^k take zero values (Lloyd-Hughes and Saunders, 2002).

Define $\pi_j = \Pr\left(X_{ij}^k = 0\right)$, the probability that the smoothed precipitation is zero on the month j, and let the density function of X_{ij}^k for $X_{ij}^k > 0$ be $g(X_{ij}^k = x; \boldsymbol{\theta}_j)$, a gamma density with parameters $\boldsymbol{\theta}_j = (\mu_j, \sigma_j)^\intercal$ evaluated at x. Thus the density function of the moving average process X_{ij}^k is a zero-augmented gamma density defined as

$$h_j(X_{ij}^k = x; \pi_j, \boldsymbol{\theta}_j) = \pi_j \mathbb{1}_{(x=0)} + (1-\pi_j) g(X = x; \boldsymbol{\theta}_j) \mathbb{1}_{(x>0)}, \tag{2.2}$$

where $\mathbb{1}_{(.)}$ is an indicator function. Hence, the cumulative distribution function of X_{ij}^k is

$$\Pr\left(X_{ij}^k \leq x\right) = \mathcal{H}_j(x; \pi_j, \boldsymbol{\theta}_j) = \begin{cases} \pi_j & x = 0 \\ \pi_j + (1-\pi_j)\mathcal{G}(x; \boldsymbol{\theta}_j) & x > 0 \end{cases}, \tag{2.3}$$

where $\mathcal{G}(\cdot\,; \boldsymbol{\theta}_j)$ denotes the distribution function for a gamma random variable with parameters $\boldsymbol{\theta}_j$.

A key point we will revisit in the sequel is that the parameters π_j and $\boldsymbol{\theta}_j$ in Equations 2.2 and 2.3 vary from month to month, but not between years, so they are able to capture annual seasonal behaviours. The methodology of Mckee et al. (1993) thus partitions $\left\{X_{ij}^t\right\}$ into twelve independent series of the form $\boldsymbol{X}_{[j]}^k = (X_{1j}^k, X_{2j}^k, \ldots, X_{nj}^k)^\intercal$ for $j = 1, \ldots, 12$. Parameter estimation for each month, $\hat{\pi}_j$ and $\hat{\boldsymbol{\theta}}_j$, is done independently by fitting a realisation of $\boldsymbol{X}_{[j]}^k$, i.e. $\boldsymbol{x}_{[j]}^k = (x_{1j}^k, x_{1j}^k, \ldots, x_{nj}^k)$, to the zero-augmented gamma density $h_j(\cdot\,;\,\cdot)$ in Equation 2.2.

Values of the standardized precipitation index (SPI) are then obtained by computing the quantiles for a standard normal density with probabilities $\mathcal{H}_j(\cdot;\cdot)$. As mentioned before, SPI values are interpreted as quantiles of a standard normal distribution, e.g. values greater than 3 or lower than -3 can be considered extreme values, while values close to zero are likely to happen.

Provided $h_j(\,\cdot\,;\,\cdot\,)$ are independent and fit the data well, the probability integral transform implies we should expect the collection $\Pi = \{\mathcal{H}_j(x_{ij}^k; \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j)\}$ to follow a standard uniform density; the back-transform using the inverse cumulative distribution function of a standard Gaussian is therefore redundant.

Hence, the proposed method of Mckee et al. (1993) to compute the SPI can be summarized as:

1) Define the time-scale k to work with (e.g. 1 month, 3 months, etc).
2) Compute the k-order moving average series $\left\{x_{ij}^k\right\}$ using all the precipitation time series $\left\{z_{ij}^k\right\}$.
3) Split the moving average series $\{x_{ij}^k\}$ into months to obtain $\boldsymbol{x}_{[1]}^k$, $\boldsymbol{x}_{[2]}^k$, $\ldots$, $\boldsymbol{x}_{[12]}^k$.
4) For each month j, obtain the estimates $\hat{\pi}_j$ and $\hat{\boldsymbol{\theta}}_j$ by fitting the realization of $\boldsymbol{X}_{[j]}^k$, i.e. $\boldsymbol{x}_{[j]}^k$, to the density function $h_j(\,\cdot\,;\,\cdot\,)$ on Equation 2.2. Maximum likelihood estimation can be used for this step.
5) Evaluate the cumulative density function $\mathcal{H}(\,\cdot\,;\,\cdot\,)$ to the observed values of the moving average process $\{X_{ij}^k\}$ to obtain the collection $\Pi = \{\mathcal{H}_j(x_{ij}^k; \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j)\}$.
6) Obtain the values for the SPI by computing the quantiles of a standard normal distribution with probabilities $\Pi = \{\mathcal{H}_j(x_{ij}^k; \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j)\}$.

2.2.2 Flood and Drought Monitoring

For drought monitoring, Mckee et al. (1993) defined an episode of *drought* as a period of time in which the SPI is continuously negative reaching at least one value lower than or equal to -1. Then, it is said that the beginning of the drought is the first time that the SPI falls below zero and it finishes when a positive SPI is reached after observing a value lower than or equal to 1 (Mckee et al., 1993).

Similarly, a *flood* can be defined as a period of time where the SPI is continuously positive reaching at least one value greater or equal to 1. Further characteristics of these events, such as *magnitude* and *intensity*, can be computed to improve drought monitoring. For example, the magnitude has been defined as the absolute value of the sum of the SPI during the period of the drought/flood, while the intensity can be classify as shown in table 2.1 (Mckee et al., 1993; Wang et al., 2017).

Table 2.1: Intensity of droughts and floods based on the SPI

Category	**Value**
extreme flood	$\text{SPI} \geq 2$
severe flood	$1.5 \leq \text{SPI} < 2$
moderate flood	$1 \leq \text{SPI} < 1.5$
near normal	$-1 < \text{SPI} < 1$
moderate drought	$-1.5 < \text{SPI} \leq -1$
severe drought	$-2 < \text{SPI} \leq -1.5$
extreme drought	$\text{SPI} \leq 2$

2.2.3 Limitations of the SPI

The standardised precipitation index has the following main limitations (Lim):

(Lim 1) *The zero-augmented gamma distribution might not be a good fit for the precipitation data:* Although in most practical applications the zero-augmented gamma distribution has been observed to be a good choice for precipitation data, there have been cases where it has been found to be inadequate (Guttman, 1999; Mishra and Singh, 2010). While it might be straightforward in theory to extend the standard SPI model to include other distributional choices for $h(\cdot;\cdot)$, it would nevertheless be useful if the methodology itself was more flexible in this regard.

(Lim 2) *The time-scale is based on months:* Theoretically there is no impediment to work with a time-scale other than months, but most published studies do not do this. Additionally, the official SPI user guide recommends working with a time-scale of at least 4 weeks (1 month) stating that lower values will make the SPI behave more erratically (World Metereological Organization, 2012). It would be desirable to develop an index that is flexible enough to allow the use of shorter and more arbitrary time-scales.

(Lim 3) *It requires a long record of precipitation:* In order to compute the SPI, it is recommended that at least 30 years of precipitation records are available, and ideally between 50 and 60 years (Piratheeparajah N and Raveendran S, 2014). The reason for this is the splitting of the complete moving average series into 12 independent subsets corresponding to each month of the year. Each of these twelve subsets has length equal to the number of years n under study, therefore small values of n may not provide reliable estimates of π_j and $\boldsymbol{\theta}_j$. This problem is related to the fact that subsets of data are handled independently.

(Lim 4) *It ignores the temporal correlation and the cyclic nature of Z_t, and hence in X_t^k, (i.e. we would expect $X_{i,12}^k$ to be correlated with $X_{i+1,1}^k$):* It is natural to observe a correlated and cyclic behaviour on precipitation data and the parameters associated with the density function; however, the SPI does not take this into account. This affects parameter estimation for $k = 1$ because an outlier presented in certain month could affect the estimated value of the parameters for that month only; this bias will be reduced for bigger values of k. This way the parameters will not vary smoothly across neighbouring months, which is both an undesirable property, but also affects the reliability of SPI values. When neglecting the temporal correlation inherent in time series such as precipitation, the SPI does not take advantage that time is a continuous variable and that continuous sharing of information across time should improve parameter estimation and allow us to work with shorter time series (which is related to Lim 3).

2.3 Generalized Additive Models for Location, Scale and Shape

In this paper we suggest the use of generalized additive models for location, scale and shape (GAMLSS) to tackle the limitations of the SPI presented in Section 2.2.3. We briefly introduce this type of model in the present section.

A generalized additive model (GAM) is an extension of an generalized linear model (GLM) that allows for the inclusion of smooth functions of covariates in the linear predictor (Hastie and Tibshirani, 1990) and thus they allow complex relationships between predictors and outcomes to be captured. The smooth functions are defined as linear combinations of basis functions, the most common being cubic regression splines, P-splines, thin plate regression splines and tensor product splines (Wood, 2006).

A GAMLSS is an extension of a GAM where, in addition to the location parameter, the scale and shape parameter are also modeled with respect to covariates. More formally, assuming a response variable Y_i with probability density function $f(y_i|\theta_{i1}, \dots, \theta_{iK})$, each parameter θ_{ik} for $k = 1, \dots, K$ is associated with a linear predictor η_{ik} through a monotonic link function g_k such as

$$g_k(\theta_{ik}) = \eta_{ik} = \boldsymbol{x}_{i0k}^{\intercal}\boldsymbol{\beta}_{0k} + f_{1k}(\boldsymbol{x}_{i1k}; \boldsymbol{\beta}_{1k}) + \cdots + f_{J_k k}(\boldsymbol{x}_{iJ_k k}; \boldsymbol{\beta}_{J_k k}), \tag{2.4}$$

where $\boldsymbol{\beta}_{0k}$ represents the fixed effects associated to the covariates $\boldsymbol{x}_{i0k}$ for an individual i, and f_{jk} represent functions able to capture a wide variety of effects with corresponding parameters $\boldsymbol{\beta}_{jk}$ and covariates $\boldsymbol{x}_{ijk}$. Considering $h_{jk}(\cdot)$ a smooth function, $f_{jk}(\cdot)$ can be used to represent: a smooth effect $h_{jk}(x)$, varying coefficient $x_1 \times h_{jk}(x_2)$, a smooth multiple effect $h_{jk}(x_1, \dots, x_L)$, a random intercept b_g, a random slope $x \times b_g$, a spatial effect $h_{jk}(\texttt{lat}, \texttt{lon})$, a temporal effect $h_{jk}(\texttt{time})$, a space-time effect $h_{jk}(\texttt{lat}, \texttt{long}, \texttt{time})$, and others such as seasonal effects (Umlauf et al., 2018). The degree of smoothness of $h_{jk}(\cdot)$ is controlled by additional smoothing parameters $\boldsymbol{\lambda}_{jk}$ (Rigby and Stasinopoulos, 2005).

More generally, for a set of observations $y_1, \ldots, y_n$, parameter vector $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{nk})$ and linear predictor vector $\boldsymbol{\eta}_k = (\eta_{1k}, \ldots, \eta_{nk})$, we can rewrite Equation 2.4 in matrix form as

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_{0k}\boldsymbol{\beta}_{0k} + f_{1k}(\boldsymbol{X}_{1k}; \boldsymbol{\beta}_{1k}) + \cdots + f_{J_k k}(\boldsymbol{X}_{J_k k}; \boldsymbol{\beta}_{J_k k}), \tag{2.5}$$

such as $\boldsymbol{X}_{0k}$ represents the design matrix with fixed effects $\boldsymbol{\beta}_{0k}$ and $\boldsymbol{X}_{jk}$ is the design matrix required to construct the effect f_{jk} with parameters $\boldsymbol{\beta}_{jk}$. The structure of $\boldsymbol{X}_{jk}$ will depend on the type of effects that are desired to be captured by f_{jk} as well as the type of covariates involved. The most common type of effects and the ones included in this study take the form $f_{jk}(\boldsymbol{x}_{ijk}; \boldsymbol{\beta}_{jk}) = \boldsymbol{X}_{jk}\boldsymbol{\beta}_{jk}$. However, Umlauf et al. (2018) allows $f_{jk}(\boldsymbol{x}_{ijk}; \boldsymbol{\beta}_{jk})$ to take more complex structures (e.g. $\beta_1 \exp(-\exp(\beta_2 + \boldsymbol{X}_{jk}\beta_3))$) in the Bayesian approach of GAMLSS, which is called Bayesian additive models for location, scale and shape (BAMLSS).

Estimation usually proceeds using a penalised likelihood approach (Rigby and Stasinopoulos, 2005; Wood, 2006), or a Bayesian approach (Umlauf et al., 2018). Both approaches are similar when obtaining point estimates due to the connection between the posterior mode and the penalised maximum likelihood estimator of $\boldsymbol{\beta}_{jk}$ for fixed values of the smoothing parameters $\boldsymbol{\lambda}_{jk}$ (Rigby and Stasinopoulos, 2005; Umlauf et al., 2018). However, although the Bayesian approach can be more computationally expensive due to the use of Markov chain Monte Carlo sampling, it can provide more reliable uncertainty estimation.

2.4 A Model-Based Method for Evaluating Extreme Hydro-Climatic Events

Having discussed some of the shortcomings of the SPI, in this section we propose two alternatives to the SPI using GAMLSS: these will be model-based approaches which we refer to as *model-based standardised indices* (MBSIs: MBSI-1 in Section 2.4.1 and MBSI-2 in Section 2.4.2); we argue that our indices retain the desirable characteristics of the SPI, but improve the methodology. We discuss some limitations of using GAMLSS in Section 2.4.3. Our model-based standardised indices are more stable, flexible and satisfying (from a modelling perspective) than the SPI as explained in Section 2.6.

Although there have been attempts to improve the methodology of the SPI (e.g. Erhardt and Czado (2017); World Meteorological Organization (WMO) and Global Water Partnership (GWP) (2016)) our method differs because we use a model-based approach, which accounts for the characteristics required to compute a standardized index. In contrast, Erhardt and Czado (2017) proposed a group of steps to compute the SPI including; elimination of seasonality (including variable transformation to reduce skewness, computation of monthly sample and variance mean), elimination of temporal dependence and transformation to the standard normal distribution. The advantage of a model-based approach is that it provides a single framework that naturally allows model checking, model selection, uncertainty computation, and joint incorporation of processes that might influence the index (e.g. seasonality, trends, covariates effects, spatial effects and spatio-temporal effects). Besides, a model-based approach could be used for other interests like interpolation, prediction, or integration with other models when the standardised precipitation values are not the main interest of the study but are required (e.g. to evaluate the effects of extreme hydro-climatic events on newborns health). Specifically, our model-based approach allows us to not only compute the standardized precipitation index appropriately but also enables us to work with short time series, check assumptions, work at different scales (e.g. weeks), work with missing values and obtain further relevant information about the underlying process under study (i.e. precipitation).

2.4.1 Model-based Standardized Index 1 (MBSI-1)

In Section 2.2.1, we saw that the SPI is defined for the moving average process $\left\{X_{ij}^k\right\}$ of a discrete stochastic process $\{Z_{ij}\}$, where i denoted the year and j the month. The MBSI-1 instead uses the initial notation of Equation 2.1, i.e. we work directly with $\{Z_t : t = 1, \ldots, T\}$ and $\left\{X_t^k : t = 1, \ldots, T\right\}$ as the precipitation and moving average process respectively. Note that we are assuming that t and k are on the same scale, which can be an arbitrary one such as daily, weekly, monthly, etc.

For the MBSI-1, we again define the density function of each element of the stochastic process $\left\{X_t^k\right\}$ as a mixture such as

$$h(X_t^k = x; \pi_t, \boldsymbol{\theta}_t) = \pi_t \mathbb{1}_{(x=0)} + (1 - \pi_t) g(X_t^k = x; \boldsymbol{\theta}_t) \mathbb{1}_{(x>0)}, \tag{2.6}$$

where x is an arbitrary value on the domain of X_t^k, while π_t and $\boldsymbol{\theta}_t$ are the parameters associated with the mixture density at time t.

The density function $g(\,\cdot\,;\,\cdot\,)$ can be any distribution defined on the positive real numbers that is adequate for characterizing the moving average precipitation. In this paper, in order to highlight the advantages of our approach with respect to the SPI for the same distribution, we complete the definition of $h(\,\cdot\,;\,\cdot\,)$ by using a gamma density for $g(\,\cdot\,; \boldsymbol{\theta}_t)$ with parameters $\boldsymbol{\theta}_t = (\mu_t, \sigma_t)^\intercal$, defined as follows

$$g(x_t^k; \mu_t, \sigma_t) = \frac{(\sigma_t/\mu_t)^{\sigma_t}}{\Gamma(\sigma_t)} x^{\sigma_t - 1} \exp\left(-\frac{\sigma_t}{\mu_t} x\right). \tag{2.7}$$

However, note that our approach is not limited to this distribution, and a different choice of $g(\cdot; \boldsymbol{\theta}_t)$ may be more suitable in other situations. One consequence of assuming a gamma density is that the mean and variance of $[X_t^k | X_t^k > 0]$ are μ_t and μ_t^2/σ_t respectively.

As mentioned earlier, the SPI tries to quantify the extremity of levels of precipitation by comparing it with the usual seasonal behaviour. For this reason, our approach captures the seasonal behaviour in all the parameters by introducing models for π_t, μ_t and σ_t, as in Equation 2.4, using linear predictors η_{1t}, η_{2t} and η_{3t} such as

$$\begin{aligned} \log\left(\frac{\pi_t}{1-\pi_t}\right) &= \eta_{1t} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + f_1(t;\boldsymbol{\beta}_1), \\ \log(\mu_t) &= \eta_{2t} = \boldsymbol{X}_2\boldsymbol{\alpha}_2 + f_2(t;\boldsymbol{\beta}_2), \\ \log(\sigma_t) &= \eta_{3t} = \boldsymbol{X}_3\boldsymbol{\alpha}_3 + f_3(t;\boldsymbol{\beta}_3), \end{aligned} \tag{2.8}$$

where $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are (optional) design matrices that include information for predicting the process with linear effects $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$; and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are the parameters required to define the flexible non-linear functions $f_1(\cdot\,;\cdot)$, $f_2(\cdot\,;\cdot)$ and $f_3(\cdot\,;\cdot)$ that capture the seasonal effects on π_t, μ_t and σ_t respectively. A common choice for these functions in the generalised additive modelling literature is to represent them using *cyclic cubic splines* because of the nice properties of cubic splines like being the smoothest interpolators (under additional restrictions), able to approximate closely any underlying smooth function, easy to construct and not expensive to compute (Wood, 2006). An alternative to cyclic cubic splines is to use harmonic terms to represent seasonal effects. However, our experience of harmonic models in this context is that they tend to overfit the data because, in part, the fitting method does not include penalties for the harmonic terms, and using stepwise selection to reduce the number of harmonic terms can be a computationally slow process.

Our model, defined with Equations 2.6, 2.7 and 2.8, is a generalized additive model for location, scale and shape (GAMLSS), as explained in Section 2.3, using a zero-augmented gamma likelihood (ZAGA). The computational cost to evaluate the (penalized) likelihood function is $\mathcal{O}(np^2)$, where p is the number of parameters required to define the smooth function $f_k(\cdot\,;\cdot)$; inference can be achieved using standard methods: backfitting or MCMC (Rigby and Stasinopoulos, 2005; Umlauf et al., 2018).

Another option for modelling serial dependence in the parameter vector $\boldsymbol{\theta}_t$

and π_t would be to assume a latent, possibly multivariate, Gaussian process or a moving average process for $f_1(\cdot\,;\,\cdot)$, $f_2(\cdot\,;\,\cdot)$ and $f_3(\cdot\,;\,\cdot)$. We have not explored these options, but they fit into the class of latent Gaussian models, for which there are a range of model fitting options, including INLA, MCMC and particle filtering, if not off-the-shelf software solutions to implement them. If interested in exploring the use of INLA, note that it should be adequately investigated for the particular structure of the model (e.g. see Taylor and Diggle, 2014; Grilli et al., 2015).

Once we have estimated the parameters in our models, we can predict π_t, μ_t and σ_t for any time t and proceed with the computation of the MBSI-1 using steps 5 and 6 of Section 2.2.1. Hence, the computation of standardised precipitation values using MBSI-1 can be summarized with the following steps:

1) Define the time-scale k to work with (e.g. 1 week, 4 weeks, 8 weeks, etc).
2) Compute the k-order moving average series $\{x_t^k\}$ using all the precipitation time series $\{z_t^k\}$.
3) Obtain the parameters estimates $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$, $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$ and $\hat{\boldsymbol{\alpha}}_3$ by fitting the moving average series $\{x_t^k\}$ to the GAMLSS model with zero-augmented gamma distribution (Equations 2.6 and 2.7) and linear predictors defined in Equation 2.8.
4) With the parameters estimated in the previous step ($\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$, $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$ and $\hat{\boldsymbol{\alpha}}_3$), obtain the estimates $\hat{\pi}_t$ and $\hat{\boldsymbol{\theta}}_t$, using Equation 2.8, for $t = 1, \ldots, T$.
5) Evaluate the cumulative density function $\mathcal{H}(\cdot;\cdot)$ of the observed values of the moving average process $\{X_t^k\}$ to obtain the collection $\Pi = \{\mathcal{H}(x_t^k; \hat{\pi}_t, \hat{\boldsymbol{\theta}}_t)\}$.
6) Obtain the values for the SPI by computing the quantiles of a standard normal distribution with probabilities Π.

2.4.2 Model-based Standardized Index 2 (MBSI-2)

One disadvantage of the MBSI-1 is that it requires a separate model for the moving average process $\{X_t^k\}$ for every scale-time of interest k. As an alternative to the MBSI-1, we propose a second approach under which the model fitting is done only once, for $k = 1$. We will refer to this approach as the model-based standardised index 2 (MBSI-2).

For this second approach, instead of imposing a model on the elements of the moving average process $\{X_t^k\}$, we propose a model for the original stochastic process $\{Z_t\}$ that represents the precipitation. Specifically, we assume that Z_t has a zero-augmented gamma distribution, which is defined by Equations 2.6 and 2.7, and the parameters are modelled considering a seasonal behaviour as in Equation 2.8. Unfortunately, the implicit distribution for the moving average variables $X_t^k = \sum_{i=0} Z_{t-i}/n$, for each $t \geq k$, cannot be found analytically, but we can use Monte Carlo methods to obtain the cumulative distribution function $\mathcal{H}(\,\cdot\,;\,\cdot\,)$ evaluated on the observed values of the moving average process $\{X_t^k\}$, obtaining $\Pi = \{\mathcal{H}(x_t^k; \hat{\pi}_t, \hat{\boldsymbol{\theta}}_t)\}$. Finally, we can compute the quantiles of a standard normal distribution associated to these probabilities Π.

Hence, the computation of the MBSI-2 can be summarized as follows:

1) Obtain the parameters estimates $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$, $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$ and $\hat{\boldsymbol{\alpha}}_3$ by fitting the original precipitation series $\{z_t^k\}$ to the GAMLSS model with zero-augmented gamma distribution (Equations 2.6 and 2.7) and linear predictors defined in Equation 2.8.
2) With the parameters estimated in the previous step ($\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$, $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$ and $\hat{\boldsymbol{\alpha}}_3$), obtain the estimates $\hat{\pi}_t$ and $\hat{\boldsymbol{\theta}}_t$, using Equation 2.8, for $t = 1, \ldots, T$.
3) Obtain m realizations $\{z_t^{(l)}\}$, where $l = 1, \ldots, m$, of the precipitation stochastic process $\{Z_t\}$ using $\hat{\pi}_t$ and $\hat{\boldsymbol{\theta}}_t$ for a zero-augmented gamma distribution (Equations 2.6 and 2.7).
4) Define the time-scale k to work with (e.g. 1 week, 4 weeks, 8 weeks, etc).
5) Compute the k-order moving average series $\{x_t^k\}$ of the precipitation time series $\{z_t\}$ and the k-order moving average series $\{x_t^{k^{(l)}}\}$ of the m samples $\{z_t^{(l)}\}$.
6) Evaluate the cumulative density function of the observed values of the moving average process $\{X_t^k\}$ to obtain the collection $\Pi = \{\mathcal{H}(x_t^k; \hat{\pi}_t, \hat{\boldsymbol{\theta}}_t)\}$ considering

that

$$\mathcal{H}(X_t^k = x_t^k; \hat{\pi}_t, \hat{\boldsymbol{\theta}}_t) = \Pr\left(X_t^k \leq x_t^k\right) = \sum_{l=1}^{m} \frac{\mathbb{1}\left\{x_t^{k(l)} < x_t^k\right\}}{m}.$$

7) Obtain the values for the SPI by computing the quantiles of a standard normal distribution with probabilities Π.

Note that this index might be more sensitive to the choice of the density function $g(\,\cdot\,;\,\cdot\,)$. This could happen because the initial error incurred by an inadequate density function for $Z_t = X_t^{\{k=1\}}$ can be compounded when deducing the density function of the moving average X_t^k. However, under an adequate selection of $g(\,\cdot\,;\,\cdot\,)$, it is expected that an appropriate density function will be deduced for X_t^k and therefore adequate standardised precipitation values predicted.

2.4.3 Limitations of GAMLSS

Although generalized additive models are attractive, they have some limitations that are worth exploring. Firstly, there can be a tendency to overfit the data, for example, it is known that the generalized cross-validation criterion used to estimate the smoothing parameters $\boldsymbol{\lambda}_{jk}$ can lead to overfitting; however, this is less likely with a large number of observations and when the values across covariates are very well distributed (Wood, 2006). This problem can worsen when modelling in addition the scale and shape parameters because the model is much more flexible and appropriate precaution should be exercised on small sample sizes.

Another limitation is that prediction outside the range of values observed on the covariates might not be reliable because usually few observations with extreme values in the covariates are observed. Given that the model is very flexible, it will try to adapt to these values. In this way, prediction at the tails of the covariates may vary significantly from one sample to another, indicating that the model has high variance in the tails of covariates. Nevertheless, extrapolation is also problematic in other types of models.

Finally, the interpretability of GAM models is not as easy for GLM models and it is required to visualize the effects in order to understand and interpret them. Despite this, we view the visualization process as actually provide useful information on the effects at different levels. Also, when using credible intervals, insight into the significance of each term is obtained.

In conclusion, GAM and GAMLSS are attractive models, but they should be used with precaution given that their inherent flexibility.

2.5 Comparison Between the SPI and MBSI

In order to illustrate differences between the SPI, MBSI-1 and MBSI-2, we compare parameter estimation, model checking and the resulting standardized precipitation values for different time-scales using data collected between January 2004 - December 2013 (522 weeks) in Caapiranga, a road-less municipality in Amazonas State.

We use our `R` package `mbsi`, created to analyse and visualise extreme events, available from Github, `https://github.com/ErickChacon/mbsi`. It contains the implementation of the SPI, MBSI-1 and MBSI-2 indices used in this section.

2.5.1 Parameter Estimation

In this section we compare the estimated mean and coverage interval of the moving average rainfall X_t^k obtained with the estimated parameters using both the SPI and the MBSI methodologies (Fig. 2.1). Given the density function defined in Equation (2.6) with Gamma density $g(.;\boldsymbol{\theta}_t)$, the 95% coverage interval for a time t is obtained by computing the 0.025 and 0.975 quantiles of the estimated density function $h(x_{ij}^k;\hat{\pi}_j,\hat{\boldsymbol{\theta}}_j)$.

We can see in Figure 2.1 that at a time-scale of 1 week, the mean and coverage interval change quickly for the classical SPI, whereas they change smoothly for the MBSIs. This is an indication that the SPI overfitted the observed precip-

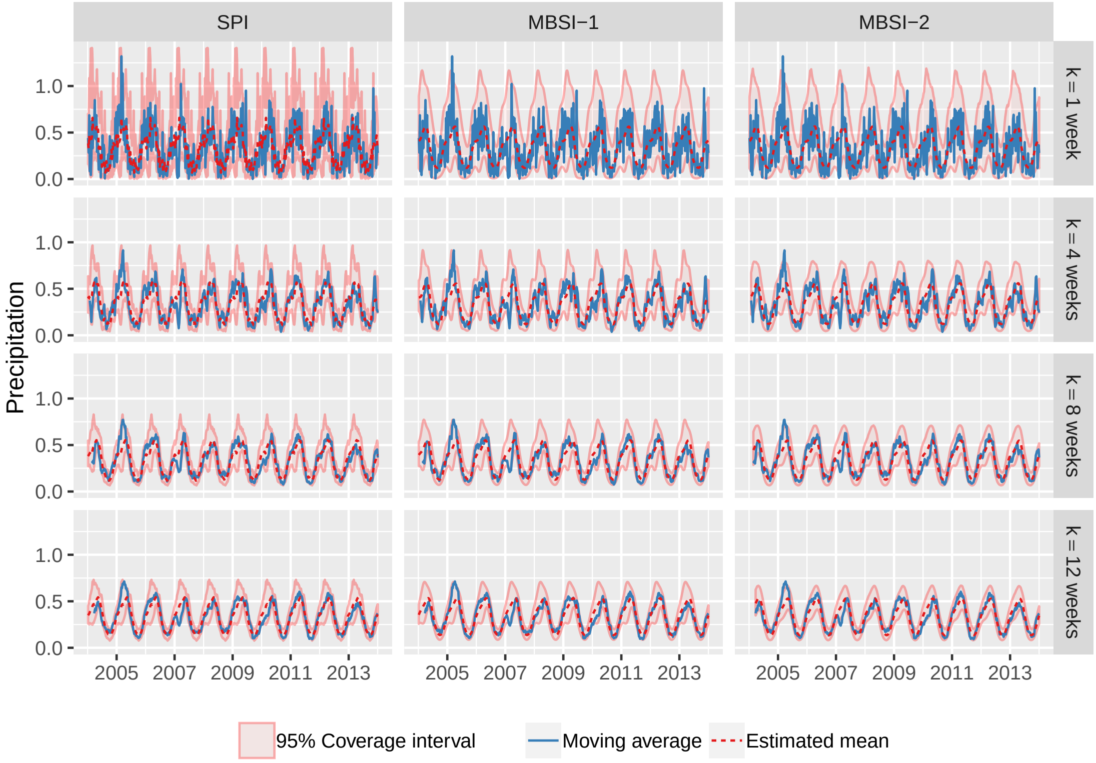

Figure 2.1: Precipitation moving average and 95% coverage interval obtained by the SPI, MBSI-1 and MBSI-2 methodologies for different time-scales (1, 4, 8 and 12 weeks)

itation data. Another characteristic of the SPI at this shorter time-scale is that parameter estimation is strongly affected by extreme short-term values. The coverage interval is highly influenced by these extreme values leading sometimes to much wider coverage intervals (e.g. due to some observations around 2005). This can reduce the ability of the SPI to detect extreme events, e.g. it can be seen in Figure 2.1 that there are more values lying outside the coverage intervals for the MBSIs. Both characteristics happen because parameters in the SPI are independent among months, while the MBSIs explicitly model this dependence using smooth functions.

As the time-scale increases, the difference between the estimated mean and coverage intervals methods decrease, but the coverage intervals are still wider and looser for the SPI.

2.5.2 Model Checking

Provided the assumed density function $h(\,\cdot\,;\,\cdot\,)$ fits the data well and independence, the probability integral transform implies we should expect the collection of the empirical cumulative density values, $\Pi = \{\mathcal{H}(x_{ij}^k; \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j)\}$, to follow a standard uniform density. This should be expected at least for $k = 1$ because independence can not be ensured for $k >> 1$. If this does not hold, then the interpretation of the distribution of the standardized values as a standard normal distribution is misleading since the back-transformed data will not be normally distributed. By inspecting Figure 2.2, we can see that, for $k = 1$ week, the uniformly distributed assumption seems adequate for the three indices, but notice that there are higher deviations for the classical SPI index. It also seems adequate for $k = 4, 8$ weeks, while for $k = 12$ weeks, the probability integral transform is not hold due to the strong correlation on the moving average process. In general, there is no indication of drastic inadequacies for any of the methodologies.

If the uniformity assumption holds, then under the probability integral transform theorem, the obtained standardized precipitation values should follow a standard normal distribution, which can be checked by comparing the empirical quantiles with the theoretical quantiles of a standard normal distribution as shown in Figure 2.3. Although, we can see in Figure 2.3 that there are some small deviations from the identity line for the MBSI-1 and MBSI-2 at small scales, the points lie close to the identity line for the three methodologies and the four time-scales. Something to notice is that the SPI methodology tends to limit the standardized values between 2 and -2 for this data of 522 observations, while we obtain more extreme standardized values with the MBSIs, something highlighted even more for the MBSI-2. This is probably related with the problem of overfitting discussed in Section 2.5.1; given that the SPI tends to overfit the data, it is less likely to obtain extreme values with respect to the estimated parameters. The opposite occurs with the MBSIs indices.

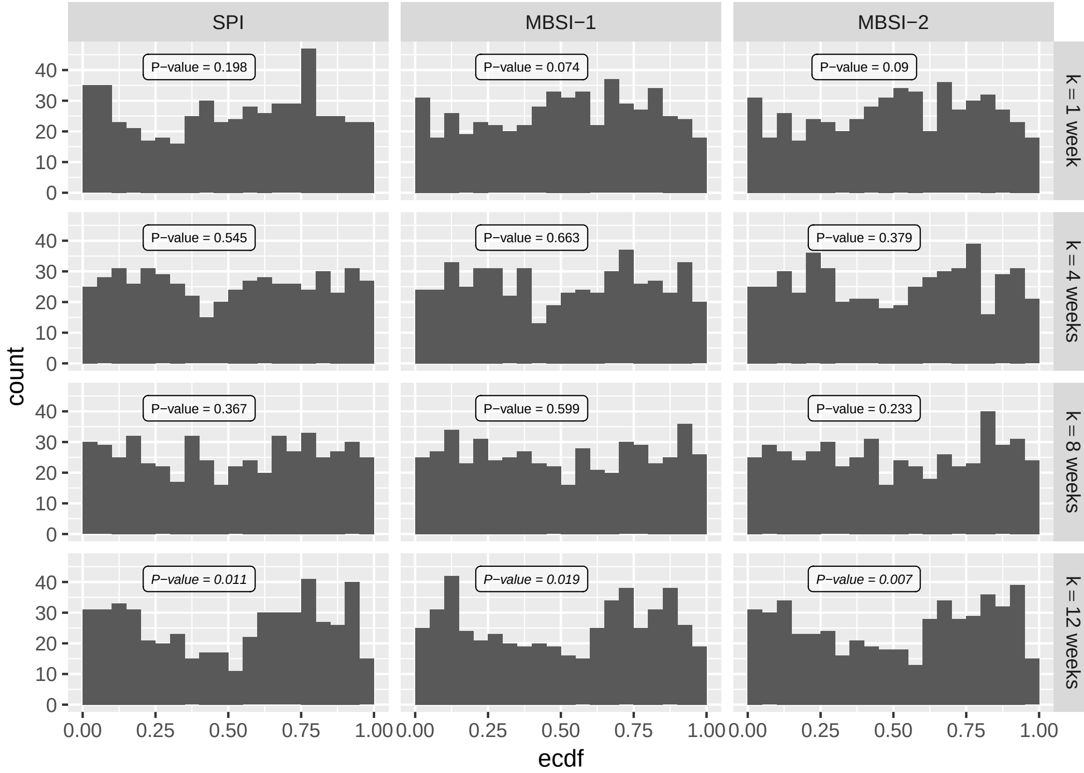

Figure 2.2: Distribution of the empirical cumulative density function $\Pi = \{\mathcal{H}(x_{ij}^{k}; \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j)\}$ for the SPI, MBSI-1 and MBSI-2 methodologies for different time-scales (1, 4, 8 and 12 weeks). P-values are provided to test uniformity using the two-sample Kolmogorov-Smirnov test. P-values with italic fonts correspond to significant tests with 95% confidence.

2.5.3 Standardized Precipitation Values

The general trends of the standardized precipitation values obtained by the three methodologies are similar; however, the actual standardized values corresponding to the identified events differ (Figure 2.4). For example, at the time scale of 1 week, most of the identified droughts have, clearly, greater absolute standardized values when working with the MBSIs. We can also see that the number of identified events varies between the methods. For instance, more droughts are identified with the MBSIs when selecting a threshold of ±1.96 for a time-scale of 8 weeks. Another difference among the methods is that the MBSI-2 tends to intensify more the extreme events. For example, it can be seen that, for time-scales of 8 and 12 weeks, the levels of the standardized precipitation for 2005 and 2007 are more extreme for the MBSI-2 than the SPI and MBSI-1. This could happen because

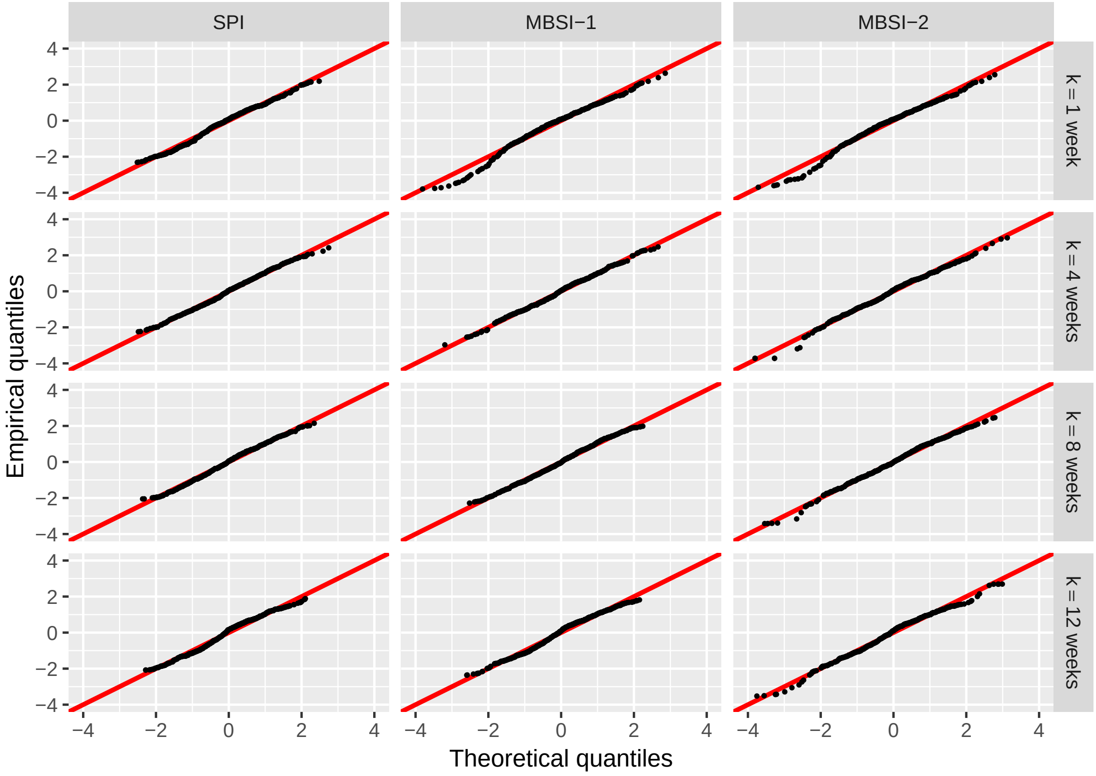

Figure 2.3: Comparison between the empirical quantiles (standardized precipitation values) and theoretical quantiles of a standard normal distribution for the SPI, MBSI-1 and MBSI-2 methodologies for different time-scales (1, 4, 8 and 12 weeks). The points should be close to the identity line (straight line) to hold the assumption of normality.

the MBSI-2 does not overfit the moving average process and is more sensitive to the choice of distribution $g(\,\cdot\,;\,\cdot\,)$; both characteristics can lead to observe more extreme values under this approach.

Amazonas State experienced a well-documented major flood in 2009 and large-scale severe drought in 2010 (Chen et al., 2010; Lewis et al., 2011). The two events are highlighted at 8 and 12 weeks time-scales, but they are more emphasized when using the MBSI-1. For this reason and because it holds properties quite similar to SPI improving the methodology, we preferred to use the MBSI-1 for further studies on cities of the Brazilian Amazonia. However, we encourage the development of indices like the MBSI-2 where the model is imposed on the original process under study and analyse another process of interest (such as the moving average process) that depends on the original one, using theoretical properties derived from the initial model. This avoids the need to re-fit the model at different

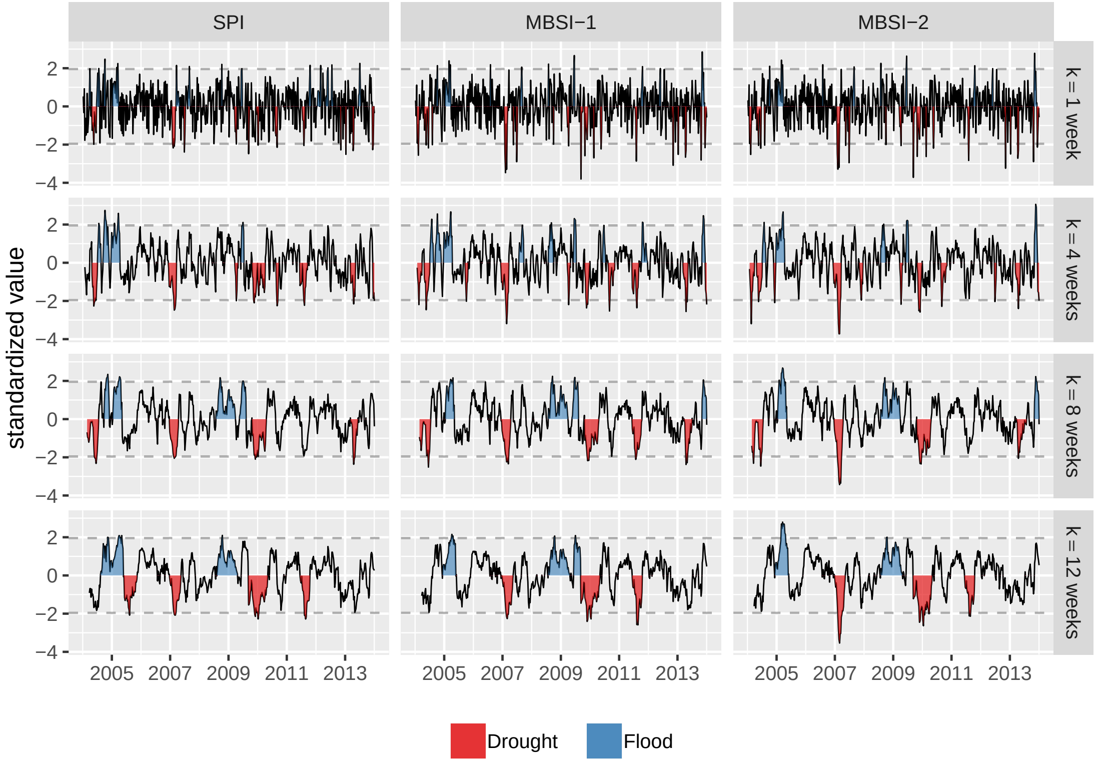

Figure 2.4: Standardized precipitation values and identification of extreme hydroclimatic events at different time-scales using the SPI and MBSI: the threshold to be considered extreme event was ±1.96

time-scales of potential interest.

2.6 Discussion and Conclusions

We compared the SPI with two proposed approaches MBSI-1 and MBSI-2 to obtain standardized precipitation values. It has been seen that the three approaches are adequate in terms of model assumptions; however, we found some differences that leaded as to select the MBSI-1 to be used in our studies conducted in the Brazilian Amazonia. Our results clearly demonstrate that the methodology of the SPI can be adapted and placed in a modelling framework that can resolve some of the disadvantages of this index.

- Because we use the GAMLSS framework, several distributions can be easily applied to compute standardized precipitation values and the diagnostic of the GAMLSS framework can be used to test model adequacy. Alternatively, it is suggested to evaluate the adequacy of the method by checking the property of the probability integral transform.

- The definition of time-scale is generalised in the MBSI-1 and so with this model, it is not necessary to work on the monthly scale. In addition, the observed series of precipitation data (or any other quantity of interest e.g. river levels) could have missing values or it might be observed at irregular intervals. Under the presence of missing values, the MBSIs estimate the parameters with the neighbours of the missing values, while the classical SPI does not take this into account for $k = 1$; for bigger k the problem is reduced. On the other hand, when data is obtained at irregular intervals, the SPI can not be computed given that it requires a collection of observations through the years that correspond to the same seasonal period (e.g. month). This is not a problem for the MBSIs given that they do not require that; the observations could correspond to any time to estimate the parameters and the moving average process could be computed for overlapping intervals of time.

- By borrowing strength from temporal autocorrelation and seasonal patterns, the MBSI-1 can compute standardized precipitation values using a shorter length of records, i.e. less then 30 years, while the SPI usually requires a longer series or a wider time-scale to avoid overfitting.

- The MBSI-1 is a temporally continuous model for precipitation and as such, parameters in the model change more naturally (i.e. smoothly) over time. In addition, the MBSI-1 could be extended to evaluate extreme events, assume trends over the time, or to incorporate spatial effects.

Bibliography

Ban, H. J., Kwon, Y. J., Shin, H., Ryu, H. S., and Hong, S. (2017). Flood monitoring using satellite-based RGB composite imagery and refractive index retrieval in visible and near-infrared bands. *Remote Sensing*, 9(4).

Blanka, V., Ladányi, Z., Szilassi, P., Sipos, G., Rácz, A., and Szatmári, J. (2017). Public Perception on Hydro-Climatic Extremes and Water Management Related to Environmental Exposure, SE Hungary. *Water Resources Management*, 31(5):1619–1634.

Chacón-Montalván, E. A., Parry, L., Torres, P., Orellana, J., Davies, G., and Taylor, B. M. (2018). Evaluating the Effects of Extreme Hydro-climatic Events on Birth-weight in Amazonia.

Chen, J. L., Wilson, C. R., and Tapley, B. D. (2010). The 2009 exceptional Amazon flood and interannual terrestrial water storage change observed by GRACE. *Water Resources Research*, 46(12):1–10.

Du, J., Fang, J., Xu, W., and Shi, P. (2013). Analysis of dry/wet conditions using the standardized precipitation index and its potential usefulness for drought/flood monitoring in Hunan Province, China. *Stochastic Environmental Research and Risk Assessment*, 27(2):377–387.

Erhardt, T. M. and Czado, C. (2017). Standardized drought indices: a novel univariate and multivariate approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics).*

Filizola, N., Latrubesse, E. M., Fraizy, P., Souza, R., Guimarães, V., and Guyot, J. L. (2014). Was the 2009 flood the most hazardous or the largest ever recorded in the Amazon? *Geomorphology*, 215:99–105.

Grilli, L., Metelli, S., and Rampichini, C. (2015). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 85(13):2718–2726.

Guerreiro, M. J., Lajinha, T., and Abreu, I. (2008). Flood Analysis with the Standardized Precipitation Index (SPI). *Revista da Faculdade de Ciênca e Tecnologia. Porto*, 4:8–14.

Guttman, N. B. (1998). Comparing the Palmer drought index and the standardized precipitation index. *Journal Of The American Water Resources Association*, 34(1):113–121.

Guttman, N. B. (1999). Accepting the Standardized Precipitation Index: a Calculation Algorithm1. *JAWRA Journal of the American Water Resources Association*, 35(2):311–322.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, volume 1. CRC Press.

Hayes, M., Svoboda, M., Wall, N., and Widhalm, M. (2011). The Lincoln Declaration on Drought Indices: Universal Meteorological Drought Index Recommended. *Bulletin of the American Meteorological Society*, 92(4):485–488.

Hayes, M. J., Svoboda, M. D., Wilhite, D. A., and Vanyarkho, O. V. (1999). Monitoring the 1996 Drought Using the Standardized Precipitation Index. *Bulletin of the American Meteorological Society*, 80(3):429–438.

Houghton, J. T., Y, D., DJ, G., M, N., PJ, v. d. L., X, D., K, M., and C, J. (2001). Climate Change 2001: The Scientific Basis. *Climate Change 2001: The Scientific Basis*, 57(8):881.

Keoduangsine, S. and Goodwin, R. (2012). An Appropriate Flood Warning System in the Context of Developing Countries. *International Journal of Innovation, Management and Technology*, 3(3):213.

Koriche, S. A. and Rientjes, T. H. M. (2016). Application of satellite products and hydrological modelling for flood early warning. *Physics and Chemistry of the Earth*, 93:12–23.

Lehner, B., Döll, P., Alcamo, J., Henrichs, T., and Kaspar, F. (2006). Estimating the impact of global change on flood and drought risks in Europe: A continental, integrated analysis. *Climatic Change*, 75(3):273–299.

Lewis, S. L., Brando, P. M., Phillips, O. L., van der Heijden, G. M. F., and Nepstad, D. (2011). The 2010 Amazon Drought. *Science*, 331(6017):554–554.

Lloyd-Hughes, B. and Saunders, M. A. (2002). A drought climatology for Europe. *International Journal of Climatology*, 22(13):1571–1592.

Mckee, T. B., Doesken, N. J., and Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *AMS 8th Conference on Applied Climatology*, (January):179–184.

McMichael, A. J. (2013). Globalization, Climate Change, and Human Health. *New England Journal of Medicine*, 368(14):1335–1343.

Mishra, A. K. and Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, 391(1-2):202–216.

Morid, S., Smakhtin, V., and Moghaddasi, M. (2006). Comparison of seven meteorological indices for drought monitoring in Iran. *International Journal of Climatology*, 26(7):971–985.

Piratheeparajah N and Raveendran S (2014). Spatial Variations of the Flood and Drought in the Northern Region of Sri Lanka. *International Research Journal of Earth Sciences*, 2(6):1–10.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., and Chivian, E. (2001). Climate change and extreme weather events. *Global change & human health*, 2(2):90–104.

Seiler, R. A., Hayes, M., and Bressan, L. (2002). Using the standardized precipitation index for flood risk monitoring. *International Journal of Climatology*, 22(11):1365–1376.

Shelton, M. (2009). *Hydroclimatology: perspectives and applications*. Cambridge University Press.

Stephenson, D. B. (2008). Definition, diagnosis and origin of extreme weather and climate events. *Climate Extremes and Society*, page 340.

Taylor, B. M. and Diggle, P. J. (2014). INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes. *Journal of Statistical Computation and Simulation*, 84(10):2266 – 2284.

Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627.

Wang, Y., Chen, X., Chen, Y., Liu, M., and Gao, L. (2017). Flood/drought event identification using an effective indicator based on the correlations between multiple time scales of the Standardized Precipitation Index and river discharge. *Theoretical and Applied Climatology*, 128(1-2):159–168.

Watts, N., Adger, W. N., Agnolucci, P., Blackstock, J., Byass, P., Cai, W., Chaytor, S., Colbourn, T., Collins, M., Cooper, A., Cox, P. M., Depledge, J., Drummond, P., Ekins, P., Galaz, V., Grace, D., Graham, H., Grubb, M., Haines, A., Hamilton, I., Hunter, A., Jiang, X., Li, M., Kelman, I., Liang, L., Lott, M., Lowe, R., Luo, Y., Mace, G., Maslin, M., Nilsson, M., Oreszczyn, T., Pye, S., Quinn, T., Svensdotter, M., Venevsky, S., Warner, K., Xu, B., Yang, J., Yin, Y., Yu, C., Zhang, Q., Gong, P., Montgomery, H., and Costello, A. (2015). Health and climate change: Policy responses to protect public health. *The Lancet*, 386(10006):1861–1914.

Wood, S. S. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.

World Meteorological Organization (WMO) and Global Water Partnership (GWP) (2016). Handbook of drought indicators and indices. *Geneva., IDMP*.

World Metereological Organization (2012). Standardized Precipitation Index User Guide.

Zeng, N., Yoon, J.-h., Marengo, J. A., Subramaniam, A., Nobre, C. A., Mariotti, A., and Neelin, J. D. (2008). Causes and impacts of the 2005 Amazon drought. *Environmental Research Letters*, 3(1):014002.

Chapter 3

In Chapter 2 we proposed a model-based standardised index (MBSI) that overcomes some limitations of the standardised precipitation index (SPI) to identify and quantity extreme hydro-climatic events. In the present chapter, we now use the MBSI to propose three bivariate indices to measure exposure to extreme hydro-climatic events during pregnancy and, consequently, evaluate the effects of floods and droughts on newborn health measured through birthweight.

Evaluating the Effects of Extreme Hydro-climatic Events on Birth-weight in Amazonia

Erick A. Chacón-Montalván[1], Luke Parry[2,5], Marcelo Cunha[3], Jesem Orellana[4], Gemma Davies[2], Benjamin M. Taylor[1]

[1]Centre for Health Informatics, Computing, and Statistics (CHICAS), Lancaster Medical School, Lancaster University, United Kingdom.
[2]Lancaster Environment Centre, Lancaster University, United Kingdom.
[3]Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil
[4]Instituto Leônidas e Maria Deane, Fundação Oswaldo Cruz, Manaus, Brazil
[5]Núcleo de Altos Estudos Amazônicos, Universidade Federal do Pará, Belém, Brazil

Abstract

Climate change poses a major risk to vulnerable populations although major uncertainties remain, such as the health impacts of extreme droughts and floods. Newborn weight is a well-recognized indicator of population health and low birth-weight (< 2500g) has been linked to life-long disadvantage including negative effects on educational attainment, income in adulthood and health. Numerous studies have investigated the social and environmental determinants of low birth-weight, yet the potential impacts of climatic change on birth-weight are poorly understood. In this paper we evaluate the effects of exposure to floods and droughts prior to and during pregnancy on birth-weight in 43 road-less municipalities in Amazonas State, Brazil. The dataset of 191,762 birth registrations from 2006 to 2014, was obtained from the Brazilian Information System of Alive Newborns (Sistema de Informação sobre Nascidos Vivos - SINASC). Our results demonstrate that (i) birth-weight varies according to seasonal changes in river levels; (ii) there was a study-region wide negative effect in 2009, coinciding with a major pan-Amazonian flood event, and a global negative trend on birth-weight; (iii) Birth-weight is lower among vulnerable mothers - those with little or no formal education, limited antenatal care and indigenous Amerindian ethnicity; and (iv) Exposure to extreme hydro-climatic events tends to have a negative impact on birth-weight. We posit that links to birth-weight are mediated by maternal stress, health-care access or inadequate nutritional intake during pregnancy. Overall, this study provides clear evidence that extreme hydro-climatic events - particularly floods - pose a major public health risk by exacerbating existing vulnerabilities in already marginalized areas of Amazonia.

Keywords: Birth-weight, Brazilian Amazonia, Climate Change, Droughts, Floods, GAMLSS, MBSI, Vulnerable Groups, Spatio-Temporal Modelling.

3.1 Introduction

Vulnerability to natural hazards is, to a significant extent, socially-determined and the greatest burden of climate change will be borne by the most disadvantaged (McMichael et al., 2006; Campbell-lendrum et al., 2015). In particular, health risks vary depending on the level of development and pre-existing vulnerabilities. Those most at-risk from extreme events are those with poor health and nutritional status and with low adaptive capacity to prevent or cope with emerging hazards (Rosenzweig et al., 2001). Consequently, climate change acts as a threat multiplier and can exacerbate existing social and spatial inequalities in development (McGuigan et al., 2002; Parry et al., 2017).

Amazonia, for instance, has a population of over 25 million people in the Brazilian part alone yet research on the social and health impacts of climate change in this region is woefully scarce (Brondízio et al., 2016). Investigating how extreme hydro-climatic events such as droughts and floods affect human health is pertinent because under global climatic change these events are predicted to increase in intensity, frequency and duration (Porporato et al., 2006). Smith et al. (2014) found that drought events impacted health in the Amazon, as detected by an increase in hospitalization rates for respiratory infections, linked to forest fires and air pollution. Perhaps the gravest potential health risks of climate change are those that affect newborns and infants and may interact with other social inequities to strongly influence health and well-being throughout an individuals life-course.

Birth-weight is an important predictor of neonatal mortality and post-neonatal mortality and morbidity (McCormick, 1985; McIntire et al., 1999). As well as infant and early childhood outcomes, birth-weight has also been shown to affect longer term outcomes in education, income and morbidity, thus it is an important indicator for population health and development (Makhija et al., 1989; Risnes et al., 2011). Birth-weight is a measure of well-being both for the present generation and also for future generations: if a mother had a low weight at birth then her child is more likely to have a low birth-weight (Currie and Moretti, 2007; Aizer and Currie, 2014). A broad suite of social and environmental factors can affect one or both of the mechanisms by which birth-weight is determined: the *intrauterine growth rate* or *gestational duration*. The most important of these factors include the mother's genetic make-up; her body's ability to sustain a 'normal' pregnancy; her health status (including stress levels); and exposure to toxins.

Genetics can, to some extent, determine the limit of foetal growth: for instance, seven loci have been identified whose combined effect accounts for a proportion of birth-weight variability similar to that accounted for by maternal smoking (Horikoshi et al., 2012). During pregnancy the foetus requires energy and this energy is taken from nutritional stores in the mother (Kramer, 1987). For this reason, the mother's body would ideally be able to store and provide sufficient nutrients to maintain proper foetal development. A body that has not finished growing (i.e. adolescent mothers), in which there are competing demands on nutritional resources between the developing mother and baby, or a body that is under-nourished pre-pregnancy (i.e. an unusually low maternal body mass index) cannot support pregnancy as well as a healthy adult women. Hence, there can be a negative impact on birth-weight among infants born to these mothers (Kramer, 1987).

Maternal morbidity affects foetal development because it can lead to a reduction of energy available and lower uterine blood flow and/or levels of amniotic fluid; some diseases like malaria directly affect the placenta. Maternal stress affects foetal development because it increases levels of Corticotrophin-Releasing Hormone (CRH) in the mother. This hormone regulates the pregnancy duration and foetal maturation (Ludwig and Currie, 2010; Menendez et al., 2000; Kramer, 1987; Camacho, 2008). Finally, toxic exposure due to cigarette smoking, tobacco chewing, alcohol/drug consumption or pollution can also negatively push-down birth-weight (Brooke et al., 1989; Butler et al., 1972; Kramer, 1987; Little, 1977; Dadvand et al., 2013).

There are a range of potential pathways through which extreme climatic events may lead to lower birth-weight and therefore, inter-generational disadvantage. These include barriers of maternal access to food of sufficient quantity, safety or nutritional value (Stephenson, 2002); access to health-care services, especially antenatal care. There are also potential effects of exposure to disease, weather extremes and their consequences (e.g. displacement, famine, disease) through impacts on maternal nutrition, morbidity and stress. As mediators of maternal health, we suggest that, by using an appropriate modelling strategy, insights can be gained into the effect of environmental and socio-economic variables on birth-weight (Aizer and Currie, 2014; Currie and Moretti, 2007; Dadvand et al., 2013). As we have outlined, most existing research on birth-weight determinants has focused on socio-economic disadvantage (Blumenshine et al., 2010; Foster et al., 2000; Danielzik et al., 2004). Research has begun to elucidate how environmental change might affect the odds of low birth-weight by changing levels of pollution, precipitation and temperature (Grace et al., 2015; Stieb et al., 2012; Dadvand et al., 2013).

Nevertheless, the effects of extreme climatic events and other natural hazards on newborn health are not well understood. In particular, floods and droughts, which constitute extreme hydro-climatic events, produce a complex web of direct and indirect environmental, economic, health, and social consequences (Blanka et al., 2017). However, few studies were done to understand their effects on birth-weight in Amazonia.

In this paper we aim to (i) evaluate the impact of floods and droughts on birth-weight in road-less municipalities in Amazonas State, Brazil, and (ii) identify the most vulnerable groups of mothers in these populations. Road-less areas of Amazonia are particularly vulnerable (characterized by high sensitivity and low adaptive capacity) to extreme climatic events, linked to higher food prices, reduced institutional presence and effectiveness, and governance failures (Parry et al., 2017). Moreover, many road-less cities in Amazonas State are also geographically isolated from major urban centres, in some cases by several thousand kilometres of boat travel (Parry et al., 2017). We do not control for gestational age in our analysis because we hypothesized that extreme hydro-climatic events could affect birth-weight through either *intrauterine growth rate* and/or *gestational duration*.

Achieving our study aims was non-trivial, for three main reasons. First, it is surprisingly not straightforward to identify and quantify the size of extreme events and our research in this area has generated a new and generally applicable index for this purpose (see Chacón-Montalván et al., 2018).

Once we have identified the occurrence of extreme events, we must formulate a measure of the exposure to extreme events during pregnancy for both flooding and droughts. This measure must take into consideration the duration of pregnancy so as not to confound the results due to the difference of pregnancy duration, which obviously has a direct effect on birth-weight. We propose bivariate indices to measure the exposure to extreme events during pregnancy and lastly develop Bayesian additive models for location, scale and shape (BAMLSS), which is the Bayesian implementation of GAMLSS, for modelling the impact of floods and droughts. The main reason for using this type of model is because it allows us to include not only non-linear, random, spatial and temporal effect but also non-linear interactions and additionally, the scale and shape parameters can be modelled if required (Rigby and Stasinopoulos, 2005; Umlauf et al., 2018).

This paper is structured as follows. The description of the data being used and the region of analysis are shown in Section 3.2. Our approach to modelling birth-weight and evaluation of the impact of extreme hydro-climatic events is presented in Section 3.3. Finally, the article concludes in Section 3.4 with a discussion on the effect of seasonality on birth-weight, the effects of extreme hydro-climatic events, the characteristics of vulnerable groups and the long trend of birth-weight.

3.2 Data Description

3.2.1 Area and Time Period of Study

Our study area covers 43 road-less municipalities in Amazonas State (Figure 3.1), chosen because extreme hydro-climatic events in these places are more likely to cause harm than compared to road-connected municipalities, where social vulnerability is lower (Parry et al., 2017). These municipalities were classified as road-less based on the connectivity analysis made by Parry et al. (2017). The period under study was January 2006 to December 2013.

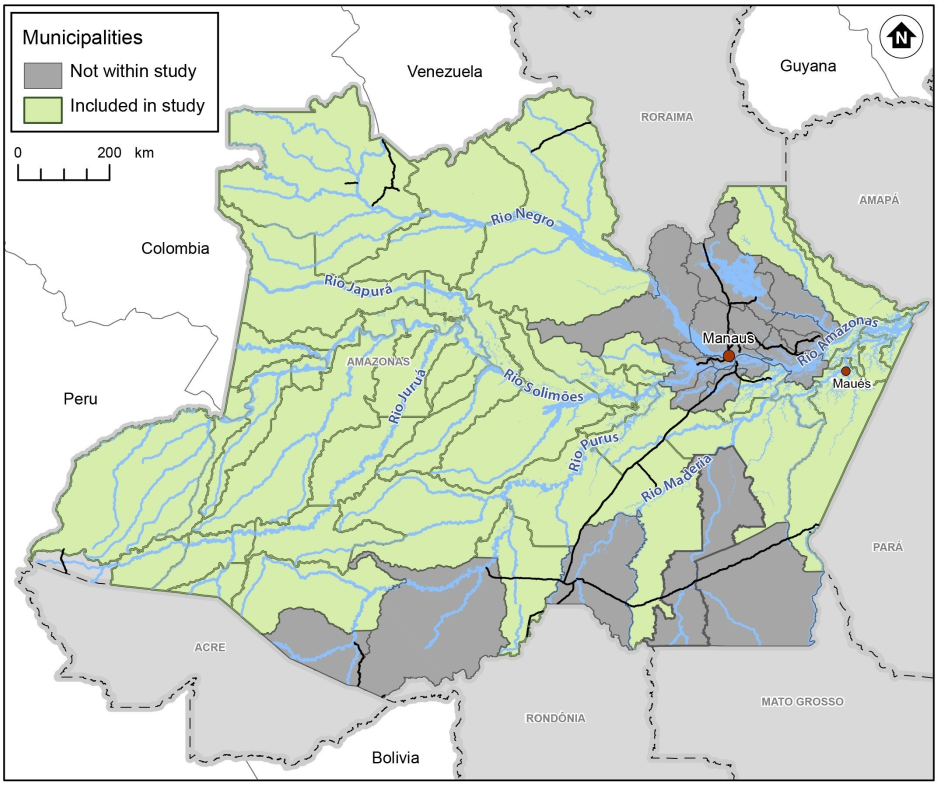

Figure 3.1: Map showing Amazonas State, Brazil, illustrating the municipalities (analagous to US Counties) included in this study. Black lines represent main roads of Amazonas State.

3.2.2 Sources of Information

As mentioned in Section 3.1, myriad factors contribute directly or indirectly to variation in birth-weight. In order to evaluate the impact of extreme hydro-climatic events, we used data related to environmental, social, demographic and genetic covariates. These datasets were obtained from secondary data sources and were measured at differing spatial scales, as explained below.

3.2.2.1 Information System of Alive Newborns (SINASC)

The Sistema de Informação sobre Nascidos Vivos (SINASC) is a Brazilian health information system created in 1994 to register live births. This database contains rich information relating to the alive newborn, the birth circumstances and the mother's social and demographic background. The number of registrations in the municipalities under study during the study period was 191,762. Specifically, we used the following variables: sex of infant, marital status of mother; mother's years of education; ethnicity of new-born; number of antenatal consultations before birth, mother's age and municipality of mother's residence (Table 3.1).

Table 3.1: Meta-data of variables used to predict birth-weight, taken from registrations in Brazil's Information System of Alive newborns (SINAS).

Variable	**Possible Values**
Newborn's sex	Male or female.
Type of birth place	Hospital, home, another health centre and other.
Mother's marital status when she gave birth.	Single, married, widowed and divorced.
Mother's education: number of years the mother spent in formal education.	Categorized: 0, 1-3, 4-7, 8-11 and greater than 12 years.
Newborn's ethnicity.	Non-indigenous and indigenous Amerindian.
Antenatal consultations: Number of antenatal consultations attended during pregnancy.	Categorized: 0, 1-3, 4-6, greater than 7 times and missing values.
Mother's age	Discrete positive value in years.
Mother's residence municipality	Any of the 43 municipalities under study.

The following variables contained missing values; birth-weight (2.07%), mother's marital status (21.79%), type of birth place (0.006%), newborn ethnicity (0.433%), antenatal consultations (1.133%) and mothers'age (0.0005%). Observations where birth-weight had missing values were removed and one observation where mother's age was a missing value was also removed. With respect to the categorical variables, in order to handle the high percentage of missing covariate data, we introduced additional "missing" levels into each of the affected variables.

The Sistema de Informação sobre Nascidos Vivos (SINASC) is likely to present some biases in recording birth-weight. In particular, implausibly large frequencies of specific values, mainly multiples of 500 grams, of birth-weight have been observed in this dataset. This problem is known as heaping and can happen when the mother reports an approximated value because the newborn was not weighed at birth, or because the weight was not recorded with diligence (Blanc and Wardlaw, 2005). Additionally, it is known that the probability of being weighed at birth is higher in urban areas where mothers have, on average, higher education and prenatal care (Blanc and Wardlaw, 2005). Hence, care must be taken in interpreting the results based on the analysis of this data.

3.2.2.2 Municipality-level Data

In order to account for differences in birth-weight between municipalities, we included in our analysis municipality-level predictors, including: (i) socio-economic variables from the 2010 National Brazilian Census, administrated by Instituto Brasileiro de Geografia e Estatística (2010), including proportion of rural people and proportion of population with internal toilet; (ii) an index of malaria exposure (see below for definition) computed from DATASUS (`http://datasus.saude.gov.br/`); and (iii) a weighted index of geographical remoteness, that takes values from 0 (least remote) to 1 (most remote), computed from the shortest travel distances to nearest cities in different levels within a hierarchical urban network (Parry et al., 2017). Regarding the index of malaria exposure, this was computed as the mean weekly rate of hospitalisations per capita due to malaria over the approximate duration of the pregnancy, computed as date of birth minus gestational age (taken as the midpoint of the gestational age factor levels). Malaria hospitalisations from DATASUS appeared under ICD10 codes B50, B500, B508, B509, B51, B510, B518, B519, B52, B520, B528, B529, B53, B531, B538 and B54. This group of municipal-scale variables can be thought of as measuring the urbanity and level of development of the municipalities under study.

3.2.2.3 Rainfall measurement

We obtained a measure of precipitation across our study region from the Tropical Rainfall Measuring Mission (TRMM) project, a joint collaboration between the National Aeronautics and Space Administration (NASA) and the Japan Aerospace

Exploration Agency (JAXA). TRMM produces rainfall products for climate research; these include measures of land surface wetness, derived from satellite images, for an area including Amazonas State. The data from 2004 to 2014 was recorded every 3 hours at a $0.25° \times 0.25°$ spatial resolution; we averaged these measures by week and then for each municipality in order to obtain a measure of weekly-rainfall-per-municipality.

3.2.2.4 River-level measurement

River levels were obtained from the Hidroweb platform from Brazil's National Water Agency (Agência Nacional de Águas [ANA]). Historical river levels were extracted for monitoring stations in and around Amazonas state over the period 2004 to 2014. In order to get a measure of extremeness relative to 'normal' seasonal behaviour, we fitted harmonic regression models to the river levels from each station that had been active for more than 10 years. For each station, we identified the annual period of the year at which rivers reached their highest levels on average. The harmonic terms in our model were computed in relation to this time: weeks numbered 0 and 53 (cyclically) denoting peak wetness and values around 26 are peak dryness. The number of harmonic terms was chosen using forward selection. We then standardised the residuals from these models and interpolated them spatially onto a raster image using ordinary kriging. Lastly, we averaged the resulting pixel-level data in order to obtain an average for each municipality and each week under study. We will refer to this measure as the *seasonal river level index*. Note that because the Amazon is so vast, the calendar week at which peak wetness is attained varies greatly; these differences are particularly pronounced between north and south of our study area.

3.3 Modelling Birth-weight

In this section, we will use a novel model-based standardised index (MBSI), proposed in Chacón-Montalván et al. (2018), along with socio-economic variables to model birth-weight in Amazonia (Figure 3.1). Chacón-Montalván et al. (2018) proposed two approaches to identify extreme hydro-climatic events from which the MBSI-1 was selected. In this paper, we refer to this index simply as the model based standardised index (MBSI).

This section is structured as follows: in Section 3.3.1 we propose bivariate indices for calculating exposure to extreme floods and droughts during pregnancy. In Section 3.3.2 we introduce the proposed model for birth-weight and discuss model selection. Lastly, our results are described in Section 3.3.3.

3.3.1 Quantifying Exposure to Extreme Events During Pregnancy

The MBSI presented in Chacón-Montalván et al. (2018) is an alternative to the classical standardised precipitation index (SPI) to identify and quantify extreme hydro-climatic events. This index uses the k-order moving average process $\left\{X_t^k : t = 1, \ldots, T\right\}$ of the precipitation process $\{Z_t : t = 1, \ldots, T\}$ to quantify how extreme are the values of x_t with respect to the usual seasonal behaviour, which is modelled using generalized additive models for location, scale and shape (GAMLSS). The time-scale k is considered to monitor the type of extreme event; e.g. greater values of k identify longer extreme events. The resulting series of values are interpretable as quantiles from a standard normal distribution. For example, an SPI value of 2 indicates that the probability of observing an event at least as extreme as this is 0.0228.

Here, we use the MBSI to propose three bivariate indices that measure exposure to extreme hydro-climatic events in different ways based on the classification of droughts (and floods) proposed by Mckee et al. (1993), where extreme droughts occur when $MBSI \geq 2$. The first captures exposure to positive and negative deviations from normal seasonality, while the second index captures exposure to floods and droughts. Lastly, the third index captures the exposure to extreme floods and droughts.

3.3.1.1 Exposure to Positive and Negative Deviations in Precipitation

One way of measuring extremity is by evaluating how far the measured precipitation was from usual seasonal behaviour. Based on this definition, we work directly with the weekly precipitation series, which is the same as the moving average precipitation series of order 1. This implies that the time-scale used to compute the MBSI for our first bivariate index is equal to one week.

We will represent the value of the obtained MBSI for mother i at week of pregnancy j as $S_{ij}^{k=1}$, where $i = 1, \ldots, m$ and $j = -12, \ldots, 0, \ldots, d_i$. Note that the possible values of j are from 12 weeks before the mother was pregnant, in order to take into account the pre-pregnancy trimester, until the pregnancy duration d_i. The sum of only positive and only negative deviations during the pre-pregnancy and pregnancy period divided by the number of weeks are the elements our bivariate indicator D_i to measure deviations from the seasonal rainfall such as

$$D_i = \left(\sum_{j=-12}^{d_i} \frac{S_{ij}^{k=1} \mathbb{1}_{\left(S_{ij}^{k=1}<0\right)}}{d_i + 12}, \sum_{j=-12}^{d_i} \frac{S_{ij}^{k=1} \mathbb{1}_{\left(S_{ij}^{k=1}>0\right)}}{d_i + 12} \right), \tag{3.1}$$

where $\mathbb{1}_{(\cdot)}$ takes a value 1 when the underlying condition $(\cdot)$ is hold and 0 otherwise. Notice that the first element D_{i1} measures negative deviation and the second D_{i2} measures positive deviation.

The interpretation of this bivariate index D_i is that average values in both dimensions represent a mother's exposure to normal rainfall. Cases where the positive exposure D_{i2} is high and the negative exposure D_{i1} is close to zero represent mothers exposed to higher rainfall than expected. Conversely, high values of negative exposure D_{i1} and positive exposure D_{i2} values close to zero represent mothers experiencing lower values of rainfall than expected. This indicator does not necessarily measure floods and droughts, but it is probably associated.

3.3.1.2 Exposure to Floods and Droughts

While in the previous index D_i, we were trying to measure deviations from seasonality, in our second index FD_i, we try to measure exposure to floods and droughts. We used the definition of drought (or flood) proposed by Mckee et al. (1993) as a period of time in which the SPI is continuously negative (or positive, for floods) reaching at least one value lower (or higher for floods) or equal to -1 (1). However, we use the MBSI instead of the SPI and allow the threshold of 1 to take other values in order to capture more extreme floods and droughts. For the computation of the MBSI, a time-scale equal to 8 weeks ($k = 8$) is used because this is related with agricultural floods and droughts, and it has performed adequately in Chacón-Montalván et al. (2018).

After identifying floods and droughts using the MBSI with the criteria explained above, the sum of standardized precipitation values corresponding to droughts (or floods) during pre-pregnancy and pregnancy period divided by the number of weeks are the elements of our bivariate indicator FD_i to measure exposure to floods and droughts such as

$$FD_i = \left(\sum_{j=-12}^{d_i} \frac{S_{ij}^{k=8} \mathbb{1}_{(j\ \in\ \text{drought event})}}{d_i + 12}, \sum_{j=-12}^{d_i} \frac{S_{ij}^{k=8} \mathbb{1}_{(j\ \in\ \text{flood event})}}{d_i + 12} \right), \tag{3.2}$$

where $\mathbb{1}_{(j\ \in\ \text{drought event})}$ takes a value 1 when the MBSI value $S_{ij}^{k=8}$ at week of pregnancy j for mother i belongs to a period where a drought has occurred and 0 otherwise; similarly, for $\mathbb{1}_{(j\ \in\ \text{flood event})}$. Notice that the first element FD_{i1} measures exposure to droughts and the second FD_{i2} measures exposure to floods and that the interpretation of this index is similar to D_i.

3.3.1.3 Exposure to Extreme Floods and Droughts

Our third bivariate index E_i is similar to the index for exposure to floods and droughts FD_i, but it tries to capture more extreme floods and droughts. Then, we have computed exposure to extreme floods and droughts E_i similarly to exposure to floods and droughts FD_i, but only the 8-week MBSI values greater that 2 or lower than -2 were considered for the computation of the bivariate index. The limit of 2 and -2 has been chosen because they are usually used to characterize extreme floods and droughts respectively (Mckee et al., 1993). Therefore, we define the bivariate index of exposure to extreme floods and droughts for mother i as

$$E_i = \left(\sum_{j=-12}^{d_i} \frac{S_{ij}^{k=8} \mathbb{1}_{\left(S_{ij}^{k=8} < -2\right)} \, \mathbb{1}_{(j \,\in\, \text{drought event})}}{d_i + 12}, \sum_{j=-12}^{d_i} \frac{S_{ij}^{k=8} \mathbb{1}_{\left(S_{ij}^{k=8} > 2\right)} \, \mathbb{1}_{(j \,\in\, \text{flood event})}}{d_i + 12} \right), \tag{3.3}$$

where $\mathbb{1}_{(\cdot)}$ takes a value 1 when the underlying condition is hold and 0 otherwise. Notice that the first element E_{i1} measures exposure to extreme droughts and the second E_{i2} measures exposure to extreme floods.

3.3.2 Statistical Modelling

In order to evaluate the effects of extreme events on birth-weight it is important to control for socio-economic status, sex and race, and include seasonal, temporal and spatial effects. Therefore, using Bayesian additive models for location, scale and shape (BAMLSS; see Umlauf et al., 2018), three models were proposed to include these effects and one was selected based on model adequacy.

3.3.2.1 Proposed Models

For the first two models, a Gaussian and Students-t distribution was assumed for birth-weight, but only the mean parameter was modelled with respect to the covariates and the scale parameter was assumed to be constant. In both cases, the mean was modelled as

$$\begin{aligned}\mu_{ij} = \eta_{ij} =&\beta_0 + h_1(\texttt{sex}_{ij}) + h_2(\texttt{marital status}_{ij}) + h_3(\texttt{study years}_{ij})+\\ &h_4(\texttt{birth place}_{ij}) + h_5(\texttt{ethnic race}_{ij}) + h_6(\texttt{consultations number}_{ij})+\\ &f_1(\texttt{age}_{ij}) + f_2(\texttt{remoteness}_{i}) + f_3(\texttt{malaria exposure}_{i})+\\ &f_4(\texttt{rural proportion}_{i}) + f_5(\texttt{tap toilet proportion}_{i})+\\ &s_1(\texttt{river level week}_{ij}) + f_6(\texttt{pregnancy date}_{ij})+\\ &f_7(\texttt{longitude}_{i}, \texttt{latitude}_{i})+\\ &f_8(\texttt{rain negative deviation}_{ij}, \texttt{rain positive deviation}_{ij})+\\ &f_9(\texttt{drought exposure}_{ij}, \texttt{flood exposure}_{ij})+\\ &f_{10}(\texttt{extreme drought exposure}_{ij}, \texttt{extreme flood exposure}_{ij}),\end{aligned} \tag{3.4}$$

where μ_{ij} is the mean birth-weight for mother j in municipality i and β_0 is the intercept. The functions $h_1(.), h_2(.), \ldots, h_6(.)$ represent the effects of categorical variables that are transformed to dummy variables. The functions $f_1(.), f_2(.), \ldots, f_6(.)$ represent thin plate regression splines for one variable, while $f_7(.), f_8(.), f_9(.), f_{10}(.)$ are bivariate. Lastly, $s(.)$ represents cyclic cubic regression splines to take into account seasonality.

The third model is an extension of the model with t-student distribution, where, in addition to the linear predictor in 3.4, the scale parameter is also modelled as

$$\begin{aligned}\log(\sigma_{ij}) = \eta^*_{ij} =&\beta^*_0 + h^*_1(\texttt{sex}_{ij}) + h^*_2(\texttt{study years}_{ij}) + h^*_3(\texttt{birth place}_{ij})+\\ &h^*_4(\texttt{ethnic race}_{ij}) + h^*_5(\texttt{consultations number}_{ij})+\\ &f^*_1(\texttt{age}_{ij}) + f^*_2(\texttt{pregnancy date}_{ij})+\\ &f^*_3(\texttt{longitude}_{i}, \texttt{latitude}_{i}),\end{aligned} \tag{3.5}$$

where σ^2_{ij} is the variance for mother j in municipality i and β^*_0 is the intercept.

The functions $h_1^*(\cdot), h_2^*(\cdot), \ldots, h_5^*(\cdot)$ represent the effects of categorical variables, $f_1^*(\cdot), f_2^*(\cdot)$ represent univariate thin plate regression splines, and $f_3^*(\cdot, \cdot)$ represents a bivariate thin plate regression spline.

3.3.2.2 Model Selection

Comparing the two first models, where only the mean was modelled, it is clear that the distribution assumption was not supported by the data in the Gaussian model (see left-side quantile plot in Figure 3.2). The points far from the straight line indicate that the residuals have heavier tails than a Gaussian distribution. On the other hand, the quantile plot for the t-distribution model looks better in terms of proximity to the straight line. Although the empirical quantiles lie out of the 95% confidence interval, this model is more adequate than the Gaussian model.

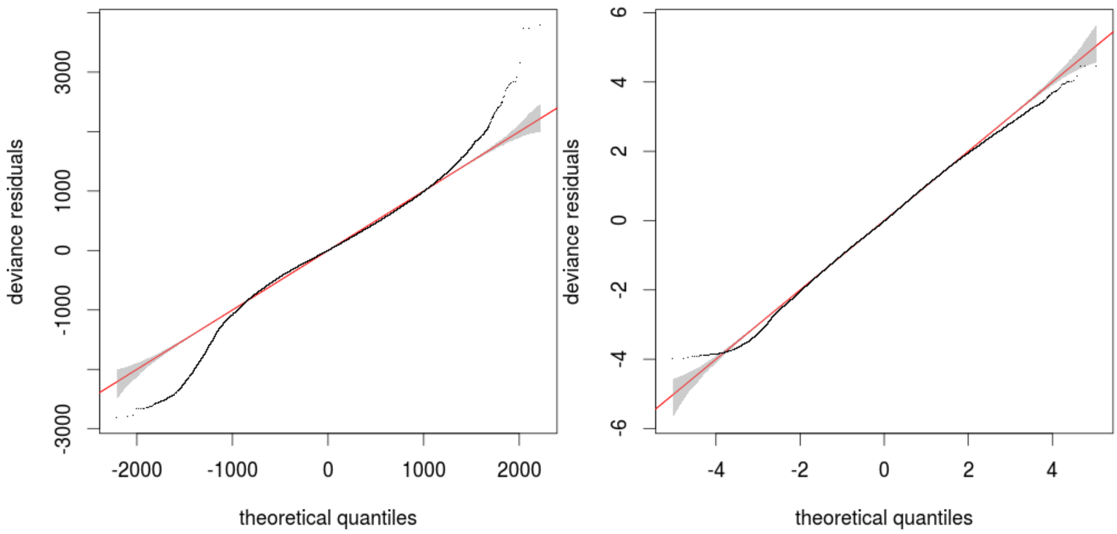

Figure 3.2: Quantile Plot of Residuals for a Generalized Additive Model (GAM) of Birth-Weight with Gaussian and t-Student Distribution from Left to Right

The quantile plot of the third model, which also models the scale parameter, shows a slight improvement on the tails (Fig. 3.3). Furthermore, as outlined in the next section, the effects of the covariates on the scale parameter were significant. Comparison using the deviance information criterion (DIC, see Spiegelhalter et al., 2002) leads to the same conclusion; this statistic was 2857483, 2848852 and 2847784 for the three models respectively. Hence, the third model was chosen as

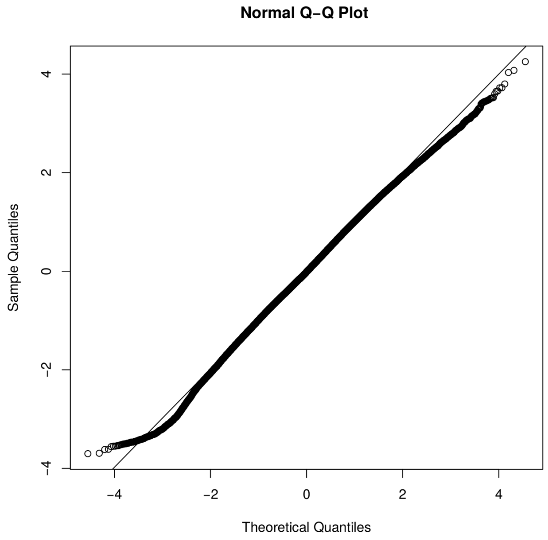

Figure 3.3: Quantile Plot Residuals for a BAMLSS Model with t-Student distribution

the definitive in this study because the assumptions are better held than the other two models and it has a better goodness of fit. Similar models to the third one were also tried by including 1, 2 or the 3 bivariate indicators to measure exposure to extremes hydro-climatic events in the mean parameter, but our third model was also better than them when comparing the DIC.

3.3.3 Results

This section presents effects of socio-economical and environmental predictors affecting birth-weight, through either *intrauterine growth rate* or *gestational duration*, obtained using the selected model which is a t-student Bayesian additive model for location, scale and shape, where the linear predictors are modelled as shown in equations 3.4 and 3.5.

3.3.3.1 Fixed effects

In order to understand the effects of the factors in Table 3.2, note that the intercept term represents the mean value for the group of mothers that: gave birth to a male offspring; was not married; had no formal education; gave birth in a hospital; was not indigenous (i.e. not a tribal Amerindian); received no formal antenatal care during her pregnancy. This group of mothers gave birth to offspring with an average weight of 3145 grams.

Table 3.2: Fixed Effects of Socio-Economic factors, Antenatal Care and Place of Birth on Mean Birth-weight: Mean value $\mathbb{E}[\beta]$ and quantiles $Q_{0.025}$, $Q_{0.5}$, $Q_{0.975}$ of the posterior distribution.

Terms	$\mathbb{E}[\beta]$	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$
(Intercept)	3144.62	3133.08	3144.62	3156.82
sex: female	-102.24	-106.21	-102.22	-98.28
marital status: married	29.80	23.30	29.76	36.33
marital status: widow	-10.88	-75.74	-11.02	55.78
marital status: divorced	-43.89	-101.99	-43.71	17.50
marital status: NA	22.38	15.94	22.38	28.98
study years: 1 - 3	55.36	45.33	55.50	64.58
study years: 4 - 7	75.95	66.34	76.07	84.49
study years: 8 - 11	73.57	63.64	73.83	82.88
study years: ≥ 12	73.21	60.31	73.44	85.23
birth place: another health center	-9.38	-47.42	-9.38	26.84
birth place: home	-83.42	-90.12	-83.36	-77.11
birth place: other	-116.87	-159.45	-115.83	-76.70
birth place: NA	-119.28	-340.51	-123.66	137.91
born race: indigenous	-58.70	-65.77	-58.73	-51.98
born race: NA	-12.31	-45.30	-12.76	20.54
consultations: 1 - 3	48.19	38.79	48.11	57.40
consultations: 4 - 6	85.61	77.20	85.48	94.56
consultations: ≥ 7	135.52	126.42	135.39	145.21
consultations: NA	19.12	-3.02	19.03	42.38

The following effects are conditional effects after accounting for the other covariates.

Female newborns had a mean weight 102 grams lower than males, consistent with the findings of Kramer (1987); Makhija et al. (1989). Offspring with a married mother had significant effects of around 30 grams greater birth-weight than single mothers.

Our analysis shows that maternal education plays an important role in determining birth-weight (and hence, health etc. in later life) because mothers with at least some formal education are expected to have offspring heavier than those without education. These differences are 55 grams (1-3 years education) and around 75 grams for greater than 3 years of education. Newborns born at home or other (non-hospital) locations, weighed less than those being born at hospital or another health centre. This difference was around 83 grams and 117 grams for those being born at home and other location, respectively. On average, indigenous offspring were 50 grams lighter than non-indigenous offspring. Lastly, the number of antenatal consultations had a positive effect on birth-weight, up 135 grams heavier for those borne to mothers with 7 or more consultations in comparison to those no consultations.

Regarding the variability parameter, $445 = \exp(6.1)$ was the estimated standard deviation for the group of mothers that gave birth to a male, were single (unmarried), had no formal education, gave birth in a hospital, were not indigenous and had no antenatal consultation during pregnancy (see Table 3.3). Around $4\% \simeq 100(1 - \exp(-0.0447))\%$ lower standard deviation has been estimated for females offspring in comparison to males. Similarly, we estimated around 4% higher standard deviation for mothers with relatively more education and higher variability for offspring born at home or other non-hospital location. On the other hand, the variability of birth-weight was lower for indigenous people and mothers with higher number of antenatal consultations.

Table 3.3: Fixed Effects of Covariates on Variance of Birth-weight: Mean value $\mathbb{E}[\beta]$ and quantiles $Q_{0.025}$, $Q_{0.5}$, $Q_{0.975}$ of the posterior distribution.

Terms	$\mathbb{E}[\beta]$	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$
(Intercept)	6.10111	6.07927	6.10094	6.12287
sex: female	-0.04447	-0.05236	-0.04461	-0.03697
study years: 1 - 3	0.00398	-0.01362	0.00395	0.02140
study years: 4 - 7	0.01020	-0.00756	0.01020	0.02708
study years: 8 - 11	0.02580	0.00653	0.02574	0.04440
study years: > 12	0.04314	0.02166	0.04317	0.06622
birth place: another health center	0.04112	-0.02374	0.04110	0.10545
birth place: home	0.02880	0.01821	0.02868	0.04012
birth place: other	0.07229	0.00013	0.07305	0.14125
birth place: NA	-0.14159	-0.67450	-0.13970	0.35574
born race: indigenous	-0.07126	-0.08427	-0.07118	-0.05836
born race: NA	-0.01814	-0.07776	-0.01798	0.04078
consultations: 1 - 3	-0.04751	-0.06460	-0.04716	-0.03128
consultations: 4 - 6	-0.08940	-0.10488	-0.08932	-0.07451
consultations: > 7	-0.11956	-0.13616	-0.11986	-0.10292
consultations: NA	-0.01112	-0.05203	-0.01130	0.03032

3.3.3.2 Non-linear effects on birth-weight at the individual level

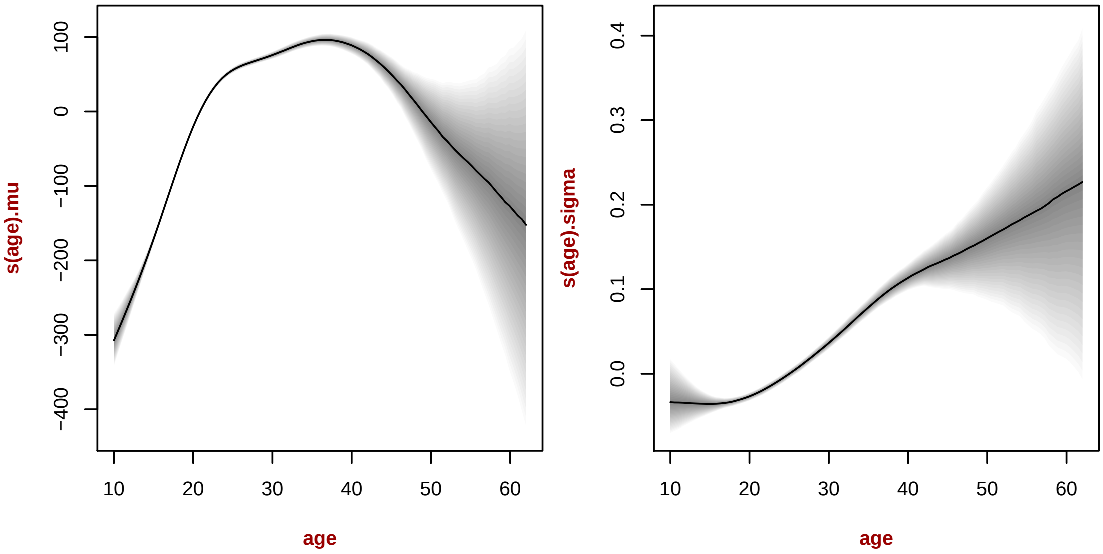

Figure 3.4: Non-linear Effects of Covariates at Individual Level on Birth-weight with 95% Credible Intervals (shaded areas). The left panel shows the effects on the mean parameter and right panel on the log-scale parameter.

The only socio-economic or constitutional variable with non-linear effect included in the selected model was age. Overall, the left panel in Figure 3.4 highlights the negative effect of pregnancy for very young women; this effect can exceed 200 grams (or more for teenage mothers) in comparison to the ideal age of 25-40 years old. This is congruent with the findings of previous studies (Kramer, 1987; Makhija et al., 1989). On the right panel in Figure 3.4, it can also be seen that the variability of birth-weight increases with maternal age indicating more uncertainty on the health of newborns when mothers are older. In part, this could be related to the health status of mothers given that older women are more prone to complications during pregnancy.

3.3.3.3 Non-linear effects on birth-weight at municipality level

The effects of municipality-scale covariates on birth-weight are shown in Figure 3.5. Remoteness from other cities in the region's hierarchical urban network had significant negative effects on birth-weight for values higher than 0.6. These more remote road-less cities include cases where boat-travel to the state capital of Manaus can take weeks and even the closest neighbouring small town may be several days away (Parry et al., 2017). The proportion of people with access to a toilet and piped water within the home - a well-recognized measure of development as seen in Brooks et al. (2005)- had a positive effect for municipalities where these basic services were accessible to more than 30% of the population. Nevertheless, it should be noted that remoteness and the proportion of people with toilet and water-on-tap are associated (Parry et al. (2017) also demonstrated that more remote places in Amazonia are less-developed), indicating that the effects shown in Figure 3.5 are conditional on each other.

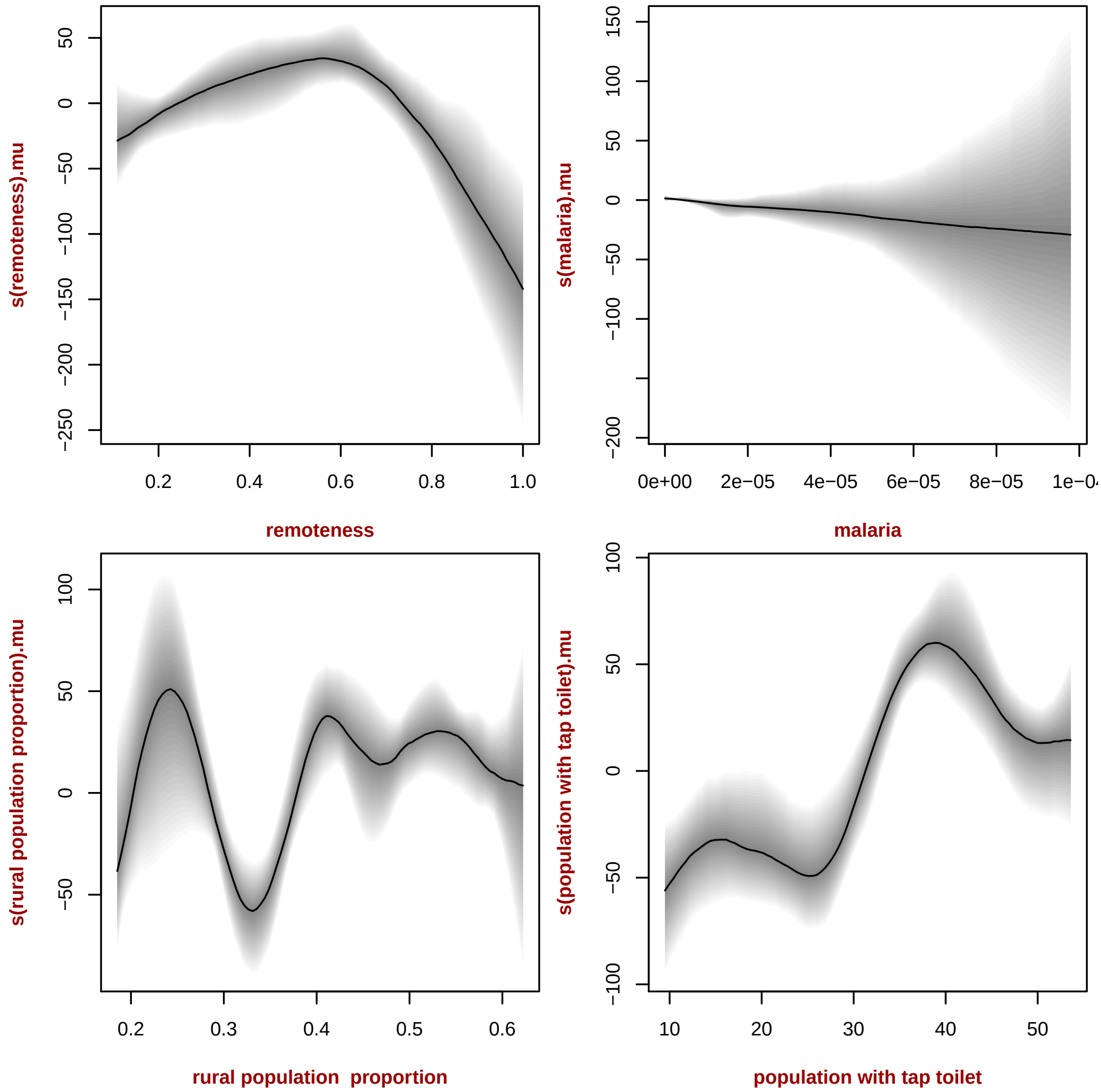

Figure 3.5: Non-linear Effects of Covariates at Municipality Level on the Mean Parameter of Birth-weight with 95% Credible Intervals (shaded areas).

The municipal-scale effect of exposure to malaria was negative, although the effect size was very small for values where the uncertainty of the effects was narrow. Lastly, the effect of 'rurality', defined as the proportion of a municipality's population that lived within the rural area, was somewhat unclear. The effect was negative for rurality values from 0.3-0.4, but this effect seems to over-fit the data. The inclusion of a higher penalty or the use of a hierarchical model could improve our models in terms of avoiding this over-fitting.

3.3.3.4 Temporal effects

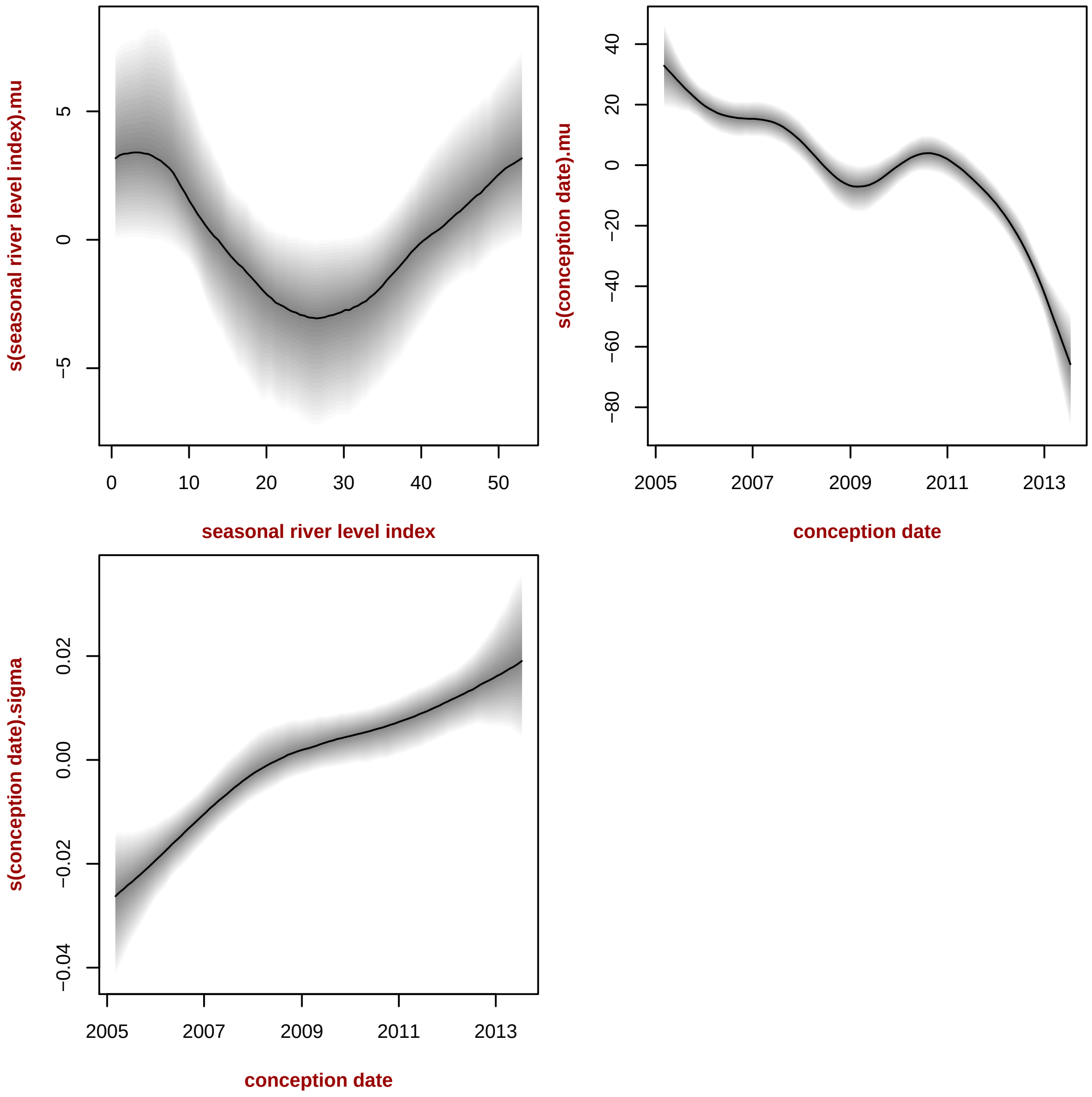

Figure 3.6: Temporal and Seasonal Effects on Birth-weight with 95% Credible Intervals (shaded areas). The top panels show the effects on the mean parameter and the bottom panel on the log-scale parameter.

Importantly, a seasonal river-level effect and a temporal effect, Figure 3.6, were included in the mean linear predictor of Equation (3.4) and one temporal effect on the standard deviation linear predictor of Equation (3.5). The seasonal effect (left panel) on the mean birth-weight parameter was relatively modest - approximately 5 grams difference between the peaks of the wet and dry seasons. In contrast, a surprising finding was that we found an unexplained temporal effect (right panel) of a global decrease in birth-weight and a main reduction around 2009. This may be associated with the major flood event in 2009, in which record-high water levels caused widespread disruption in Amazonas state (Chen et al., 2010; Filizola et al., 2014).

The tails of the conception date effects (right panel on Figure 3.6) around 2005 and 2013 years are less reliable because they are associated with mothers with higher and lower gestational duration for the left and right tail respectively. This happened because the mothers on the available dataset were selected based on birth date and not conception date.

Additionally, we found an increased temporal effect on the standard deviation of birth-weight (Figure 3.6), indicating that the uncertainty of birth-weight have increased during the period of study.

3.3.3.5 Spatial effects on birth-weight

The spatial effects for the mean parameter, $f_7(\texttt{longitude}_i, \texttt{latitude}_i)$, and scale parameter, $f_3^*(\texttt{longitude}_i, \texttt{latitude}_i)$, of birth-weight that is unexplained for the four municipal-scale predictors are shown in Figure 3.7. These effects appear to be significant, for both mean and standard deviation. The credible interval for the effects on birth-weight mean are negative (with mean greater than -50 grams) for two particular areas in the north and west of Amazonas state, and positive for an area in south-east Amazonas (with mean around 50). Furthermore, the effects on the standard deviation are positive in southern Amazonas and negative in western Amazonas. Their corresponding credible intervals suggest significant effects given

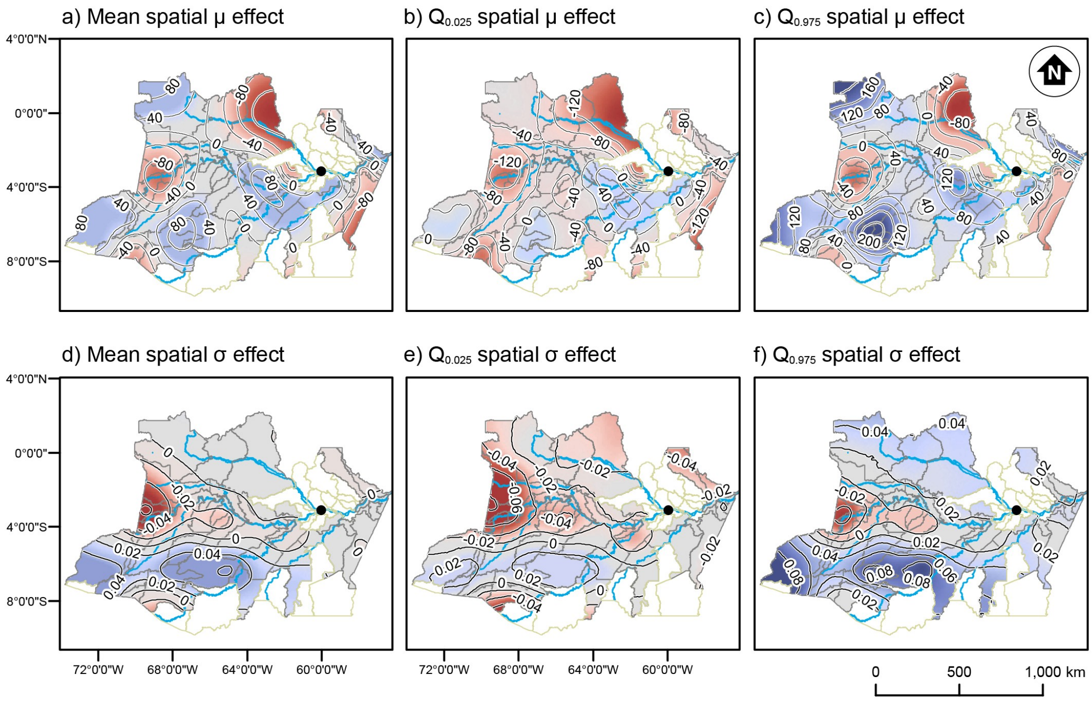

Figure 3.7: Spatial effects on the linear predictors of the birth-weight distribution: the effect on the mean $f_7(\texttt{longitude}_i, \texttt{latitude}_i)$ and on the standard deviation $f_3^*(\texttt{longitude}_i, \texttt{latitude}_i)$ from top to bottom. Red (blue) indicates a negative (positive) change in the relevant parameter. Mean, lower 95% bound and upper 95% bound from left to right.

that they do not include 0.

3.3.3.6 Effects of extreme hydro-climatic events

We used three pairs of indices to evaluate the effects of exposure to extreme events during pregnancy; exposure to: (1) rainfall which deviated from seasonal averages, (2) floods and droughts, and (3) extreme floods and droughts. These variables are, not surprisingly, co-linear, especially between exposure to rainfall deviations and exposure to floods and droughts.

The conditional effect of exposure to rainfall deviations indicates that high positive deviations negatively impact birth-weight (Fig. 3.8). The credible interval seems significant for levels of positive rainfall deviation around 0.6 and negative deviation around -0.3. It also suggests that birth-weight is greater when a pregnancy occurs during periods of rainfall that closely mirror long-term seasonal averages. In other words, when a mother is not exposed to either high positive or negative rainfall deviations from the mean. The positive effect looks significant around the position -0.2 and 0.3, where the credible interval is positive.

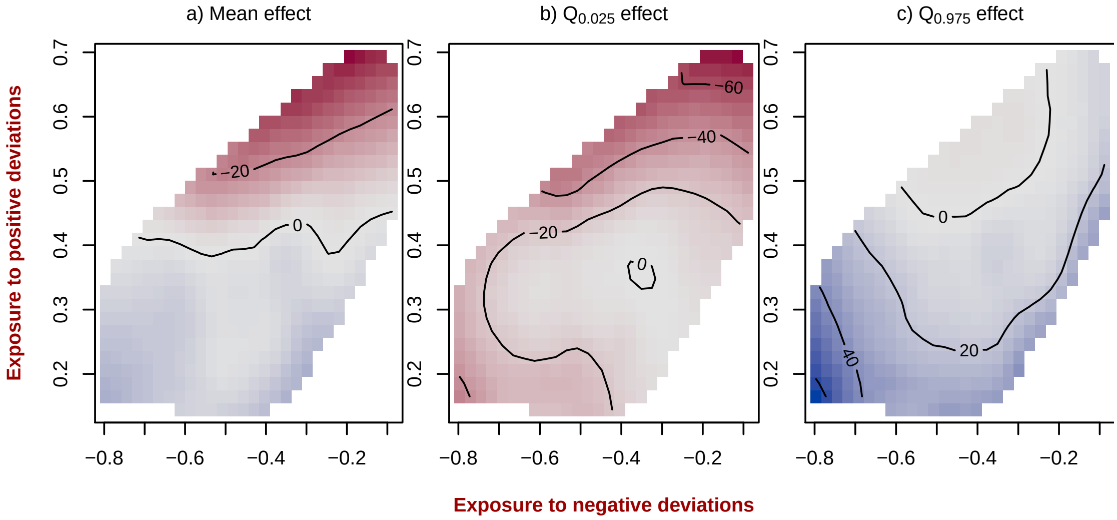

Figure 3.8: Effects on Birth-weight of Above or Below the Mean Rainfall Exposure during Pregnancy. Red (blue) indicates a negative (positive) change in mean birthweight. Mean, lower 95% bound and upper 95% bound from left to right.

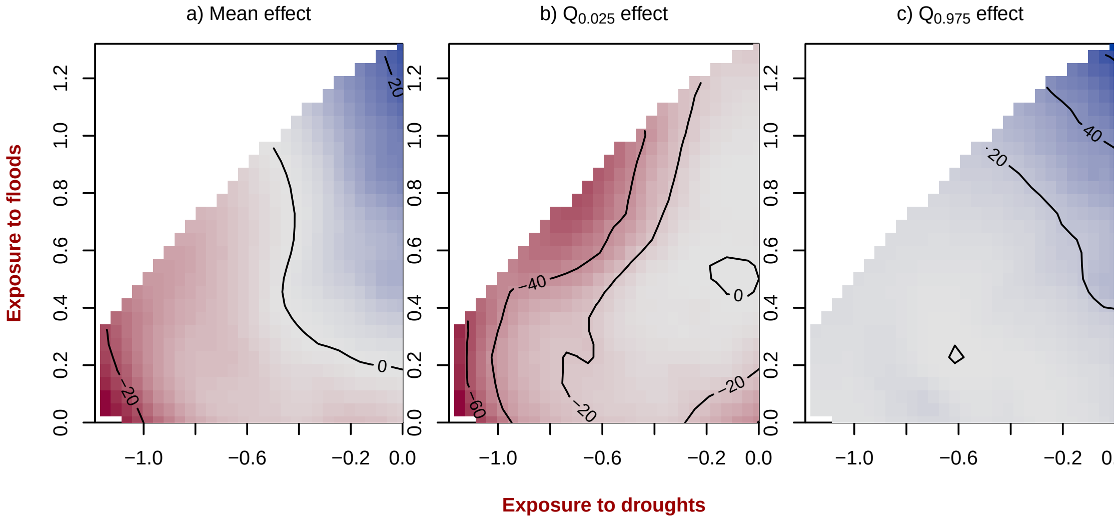

Figure 3.9: Effects on Birth-weight of Floods and Droughts as defined by Mckee et al. (1993). Red (blue) indicates a negative (positive) change in mean birthweight. Mean, lower 95% bound and upper 95% bound from left to right.

Our findings show a significant negative effect of floods (around 20 grams) but also a positive effect of droughts during pregnancy (Figure 3.9). However, it is important to note that our indices of droughts and floods and exposure to deviations are associated with one another and thus the marginal effects can lead to misleading conclusions. A clearer picture is obtained for a similar model without including exposure to floods and droughts. The effect of positive and negative deviations is shown in Figure 3.10, where negative effects are observed for higher exposure to positive and negative deviations. Hence, it is clear that exposure to extremes of precipitation affects birth-weight.

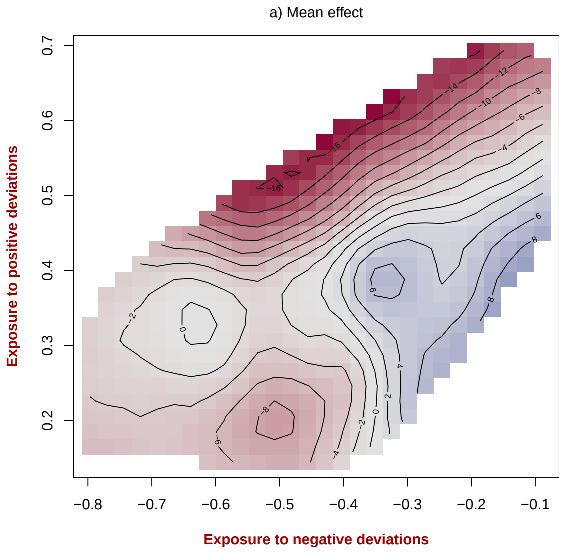

Figure 3.10: Mean Effects on Birth-weight of Above or Below the Mean Rainfall Exposure during Pregnancy: For an alternative model without including exposure to floods and droughts. Red (blue) indicates a negative (positive) change in mean birthweight.

Lastly, the mean of the effects of extreme floods and droughts in Figure 3.11 suggest a negative impact of droughts (exposure around -0.5) and negative impact for certain floods (exposure around 0.5), but also a positive impact around 0.6 exposure. Note that a mean decrease of 200 grams was observed around $(0, 0.4)$ and that the credible intervals of these effects are negative. This major result shows that extreme floods have strong negative impacts on birth-weight.

3.4 Discussion

In the Brazilian Amazonia, over the period of January 2006 - December 2013, the mean birth-weight was 3220 grams and the proportion of low birth-weight $(< 2500g)$ was 0.06. Our findings provide strong evidence that extreme climatic events - especially floods - negatively affect health, as suggested in Watts et al.

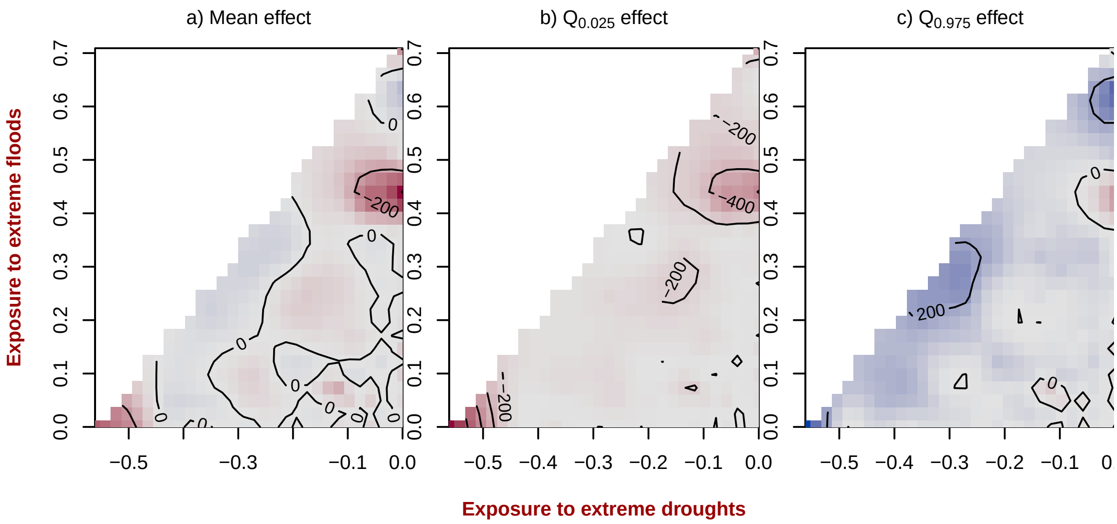

Figure 3.11: Effects on Birth-weight of Exposure to Extreme Floods and Droughts. Red (blue) indicates a negative (positive) change in mean birthweight. Mean, lower 95% bound and upper 95% bound from left to right.

(2015); Hales et al. (2003), including in road-less, vulnerable areas of the Brazilian Amazon (Parry et al., 2017; Hummell et al., 2016). We have shown how climatic extremes can deepen health inequities by causing lower birth-weight and exacerbating existing social vulnerabilities. Indeed, weight at birth is a strong indicator of maternal health and low birth-weight can confer life-long and inter-generational disadvantage (Kramer, 1987; Aizer and Currie, 2014).

In our study there were striking differences in birth-weight among children borne to mothers that were indigenous; young (< 25 years old); had no formal education and had little access to the formal healthcare system during pregnancy or birth (Moser et al., 2003; Reime et al., 2006; Nobile et al., 2007). This paper therefore makes an important and novel contribution to the literature on health and climate change, because few studies have explored the effects of climatic variation on birth-weight (but see Grace et al. (2015); Murray et al. (2000)), especially in relation to social inequities. From a regional perspective, this research also engages with the recent urgent call for more research on the social and health dimensions of climate change in Amazonia, which has been woefully neglected (Brondízio et al., 2016). Despite growing exposure to extreme floods and droughts in Amazonia (Marengo et al., 2013), to our knowledge this is the first study to systematically asses health impacts of floods.

3.4.1 Floods and droughts affect birth-weight in Amazonia

Using the MBSI index to identify extreme hydro-climatic events, a major finding of our study was that exposure to Amazonian flood events during pregnancy reduces mean birth-weight by around 200 grams. This effect is probably associated with the occurrence of severe flood(s) which mainly affected particular Amazonian sub-watersheds (e.g. along either the River Purus, River Jurua or main River Solimões) of the 43 municipalities we studied because a global reduction of 200 grams was not observed. Because extreme events tend to be infrequent, unique values of our three proposed bivariate indices were not very well distributed. This problem could be overcome by extending the study period backwards in time (or broaden the study area) in order to (i) improve the detection of extreme events, and also to (ii) include more people affected by those events.

Although our results provide clear evidence that extreme hydro-climatic events can negatively impact birth-weight (Section 3.3), our study was not intended (or able) to identify which causal mechanism(s) link(s) climate extreme and birth-weight. Nonetheless, identifying the relative importance of different causal pathways is clearly important for developing effective public policy to mitigate the health impacts of climate change. For the context of our study system - road-less areas in the Brazilian Amazon - principal candidate pathways for lower birth-weight include: deficiencies in maternal nutritional intake linked to food insecurity and disruption of the local food system (Sherman et al., 2015; Maru et al., 2014); maternal stress and anxiety (Mansour and Rees, 2012; Berry et al., 2010); restrictions on health-care access (including antenatal care) due to transportation difficulties or stresses on public services (De Onis et al., 2007; Haines et al., 2006); maternal morbidity, linked to insect-, water-borne or parasitic disease (Steketee, 2003). In addition to determining whether extreme event effects were caused by nutrition or disease, for example, there is an urgent need for further social-science research to understand the 'causes of the causes' of these health impacts. In other words, the social, economic and political processes through which floods and droughts affect population health in Amazonia and other vulnerable contexts (Watts et al., 2015). Building this evidence base is essential for eventually reducing the impact of extreme events through appropriate adaptation; the negative effects on birth-weight become somewhat inexorable if they lead to low birth-weight. Furthermore, the hazards posed by extreme climatic events will continue to grow if the frequency, intensity and duration of these events increase, as predicted (Field, 2012).

3.4.2 Seasonal trends on birth-weight in road-less, river-dependent places

River levels play a vital role in the lives of Amazonian people, around a million of whom live in urban centres that lack any access to Brazil's road network (Parry et al., 2017). Our study has shown that, despite traditional livelihoods that are adapted to the annual flood-pulse (Harris, 2000), seasonal changes in river levels impact birth-weight in road-less municipalities with a statistically significant drop in the mean weight of around 5 grams for mothers becoming pregnant in the dry season. Although this seasonality effect seems relatively small, consider that the estimated effect is the average for all 43 municipalities. It is likely that seasonality is much more important for some cities and not relevant for others, as we observed in our exploratory analysis. At this stage it is unclear whether the seasonal effect is related to temporal variation in food insecurity, disease prevalence (Olson et al., 2009) or access to public services (Parry et al., 2010).

In order to develop interventions that improve public health and mitigate seasonally-lower birth-weight, we require deeper insights into the seasonality of birth-weight at the municipality-scale and an understanding of the causes of inter-municipal variation in these differences. For example, whether a reduction birth-weight is due to fluctuation in the price of imported foodstuffs or reduced household access to local foods including açaí (*Euterpe spp.*), or access to bush-meat. In either of these cases, efforts could be made to provide nutritional substitutes for those products whose accessibility is reduced during the dry season. Our results may be explained by the argument posited by Vaitla et al. (2009) - that seasonal hunger or food insecurity is a neglected yet important development challenge in the Global South. The precise causes of seasonal differences in birth-weight are unclear, yet improving antenatal care both in general and during the dry season, including outside of urban centres, would be a 'no regret' strategy, yielding benefits even in the absence of climate change (Hallegatte, 2009; Watts et al., 2015).

3.4.3 Vulnerable groups of mothers give birth to smaller offspring

Our study clearly supports theoretical and empirical evidence that vulnerability to climate change is, at least partly, socially-determined (Birkmann, 2006). We identified characteristics of particularly vulnerable pregnant women; for instance, newborns of indigenous mothers weighed 50 grams less when considering only this variable. However, ethnicity - through structural discrimination and oppression - has complex links to social inequalities in Brazil and elsewhere (Guzmán, 2013; Young, 2009).

For example, when including low levels of maternal education and very low antenatal care received by many indigenous mothers, the negative effect on birth-weight can be exacerbated between 163 to 271 grams when comparing with more advantaged groups. Single mothers also had a negative effect on birth-weight, perhaps because married mothers are more economically advantaged Aizer and Currie (2014), and that they could exposure to more stressful scenarios. We found that low education and inadequate antenatal care probably had the most important impacts on birth-weight thus improving access to good quality education and healthcare are obvious contenders for improving the resilience of Amazonian societies in the context of climatic change.

3.4.4 Negative trend around 2009 flood and long-term worsening

An unexpected finding was that, after accounting for socio-economic and environmental predictors, birth-weight experienced a mean decline of around 10 grams around 2009 across our study region, which we have suggested is linked to a major flood in the Brazilian Amazon in that year in which the main River Solimões-Amazonas channel reached record levels (Chen et al., 2010; Filizola et al., 2014). Given the broad spatial extent of this flood, it is plausible that our indices do not fully capture the municipal-scale effects of this event.

In addition, a global decline on this unexplained temporal variation was observed during the period of study. It could represent serious problems for the population health; however, further analysis with more recent data per municipality are necessary for more concluding results about this temporal negative trend because it is probably the case that this temporal trend varies per municipality due to regional inequalities.

In summary, we have assessed the effects of socio-economical and environmental factors on birth-weight in the Brazilian Amazonia. Through our analyses, we have discovered significant negative effects of extreme hydro-climatic events, especially floods; seasonal trends on birth-weight related to river levels; vulnerable groups of mothers; negative effects around 2009 flood and long-term worsening trend.

Bibliography

Aizer, A. and Currie, J. (2014). The intergenerational transmission of inequality: Maternal disadvantage and health at birth. *Science*, 344(6186):856–861.

Berry, H. L., Bowen, K., and Kjellstrom, T. (2010). Climate change and mental health: A causal pathways framework. *International Journal of Public Health*, 55(2):123–132.

Birkmann, J. (2006). *Measuring vulnerability to natural hazards: towards disaster resilient societies.*

Blanc, A. K. and Wardlaw, T. (2005). Monitoring low birth weight: An evaluation of international estimates and an updated estimation procedure. *Bulletin of the World Health Organization*, 83(3):178–185.

Blanka, V., Ladányi, Z., Szilassi, P., Sipos, G., Rácz, A., and Szatmári, J. (2017). Public Perception on Hydro-Climatic Extremes and Water Management Related to Environmental Exposure, SE Hungary. *Water Resources Management*, 31(5):1619–1634.

Blumenshine, P., Egerter, S., Barclay, C. J., Cubbin, C., and Braveman, P. A. (2010). Socioeconomic disparities in adverse birth outcomes: A systematic review. *American Journal of Preventive Medicine*, 39(3):263–272.

Brondízio, E. S., de Lima, A. C., Schramski, S., and Adams, C. (2016). Social and health dimensions of climate change in the Amazon. *Annals of Human Biology*, 43(4):405–414.

Brooke, O. G., Anderson, H. R., Bland, J. M., Peacock, J. L., and Stewart, C. M. (1989). Effects on birth weight of smoking, alcohol, caffeine, socioeconomic factors, and psychosocial stress. *BMJ (Clinical research ed.)*, 298(6676):795–801.

Brooks, N., Adger, W. N., and Kelly, P. M. (2005). The determinants of vulnerability and adaptive capacity at the national level and the implications for adaptation. *Global Environmental Change*, 15(2):151–163.

Butler, N. R., Goldstein, H., and Ross, E. M. (1972). Cigarette smoking in pregnancy: its influence on birth weight and perinatal mortality. *British medical journal*, 2(5806):127–30.

Camacho, A. (2008). Stress and Birth Weight: Evidence from Terrorist Attacks. *The American Economic Review*, 98(571):0–10.

Campbell-lendrum, D., Manga, L., Bagayoko, M., and Sommerfeld, J. (2015). Climate change and vector-borne diseases : what are the implications for public health research and policy ? *Phil. Trans. R. Soc.*, B(370):20130552.

Chacón-Montalván, E. A., Parry, L., Davies, G., and Taylor, B. M. (2018). A Model-Based General Alternative to the Standardised Precipitation Index.

Chen, J. L., Wilson, C. R., and Tapley, B. D. (2010). The 2009 exceptional Amazon flood and interannual terrestrial water storage change observed by GRACE. *Water Resources Research*, 46(12):1–10.

Currie, J. and Moretti, E. (2007). Biology as Destiny? Short- and Long-Run Determinants of Intergenerational Transmission of Birth Weight. *Journal of Labor Economics*, 25(2):231–264.

Dadvand, P., Parker, J., Bell, M. L., Bonzini, M., Brauer, M., Darrow, L. A., Gehring, U., Glinianaia, S. V., Gouveia, N., Ha, E. H., Leem, J. H., van den Hooven, E. H., Jalaludin, B., Jesdale, B. M., Lepeule, J., Morello-Frosch, R., Morgan, G. G., Pesatori, A. C., Pierik, F. H., Pless-Mulloli, T., Rich, D. Q., Sathyanarayana, S., Seo, J., Slama, R., Strickland, M., Tamburic, L., Wartenberg, D., Nieuwenhuijsen, M. J., and Woodruff, T. J. (2013). Maternal exposure to particulate air pollution and term birth weight: A multi-country evaluation of effect and heterogeneity. *Environmental Health Perspectives*, 121(3):367–373.

Danielzik, S., Czerwinski-Mast, M., Langnäse, K., Dilba, B., and Müller, M. J. (2004). Parental overweight, socioeconomic status and high birth weight are the major determinants of overweight and obesity in 5-7 y-old children: baseline data of the Kiel Obesity Prevention Study (KOPS). *International Journal of Obesity*, 28(11):1494–1502.

De Onis, M., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., and Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organisation*, 85(10):812 – 819.

Field, C. B. (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

Filizola, N., Latrubesse, E. M., Fraizy, P., Souza, R., Guimarães, V., and Guyot, J. L. (2014). Was the 2009 flood the most hazardous or the largest ever recorded in the Amazon? *Geomorphology*, 215:99–105.

Foster, H. W., Wu, L., Bracken, M. B., Semenya, K., and Thomas, J. (2000). Intergenerational effects of high socioeconomic status on low birthweight and preterm birth in African Americans. *Journal of the National Medical Association*, 92(5):213–21.

Grace, K., Davenport, F., Hanson, H., Funk, C., and Shukla, S. (2015). Linking climate change and health outcomes: Examining the relationship between temperature, precipitation and birth weight in Africa. *Global Environmental Change*, 35:125–137.

Guzmán, T. D. (2013). *Native and national in Brazil: indigeneity after independence*. UNC Press Books.

Haines, A., Kovats, R. S., Campbell-Lendrum, D., and Corvalan, C. (2006). Climate change and human health: Impacts, vulnerability and public health. *Public Health*, 120(7):585–596.

Hales, S., Edwards, S. J., and Kovats, R. S. (2003). Impacts on health of climate extremes. *Climate change and human health: Risks and responses*, pages 79–102.

Hallegatte, S. (2009). Strategies to adapt to an uncertain climate change. *Global Environmental Change*, 19(2):240–247.

Harris, M. (2000). *Life on the Amazon: the anthropology of a Brazilian peasant village*. Oxford University Press.

Horikoshi, M., Yaghootkar, H., Mook-Kanamori, D. O., Sovio, U., Taal, H. R., Hennig, B. J., Bradfield, J. P., St Pourcain, B., Evans, D. M., Charoen, P., Kaakinen, M., Cousminer, D. L., Lehtimäki, T., Kreiner-Møller, E., Warrington, N. M., Bustamante, M., Feenstra, B., Berry, D. J., Thiering, E., Pfab, T., Barton, S. J., Shields, B. M., Kerkhof, M., van Leeuwen, E. M., Fulford, A. J., Kutalik, Z., Zhao, J. H., den Hoed, M., Mahajan, A., Lindi, V., Goh, L.-K., Hottenga, J.-J., Wu, Y., Raitakari, O. T., Harder, M. N., Meirhaeghe, A., Ntalla, I., Salem, R. M., Jameson, K. A., Zhou, K., Monies, D. M., Lagou, V., Kirin, M., Heikkinen, J., Adair, L. S., Alkuraya, F. S., Al-Odaib, A., Amouyel, P., Andersson, E. A., Bennett, A. J., Blakemore, A. I. F., Buxton, J. L., Dallongeville, J., Das, S., de Geus, E. J. C., Estivill, X., Flexeder, C., Froguel, P., Geller, F., Godfrey, K. M., Gottrand, F., Groves, C. J., Hansen, T., Hirschhorn, J. N., Hofman, A., Hollegaard, M. V., Hougaard, D. M., Hyppönen, E., Inskip, H. M., Isaacs, A., Jørgensen, T., Kanaka-Gantenbein, C., Kemp, J. P., Kiess, W., Kilpeläinen, T. O., Klopp, N., Knight, B. A., Kuzawa, C. W., McMahon, G., Newnham, J. P., Niinikoski, H., Oostra, B. A., Pedersen, L., Postma, D. S., Ring, S. M., Rivadeneira, F., Robertson, N. R., Sebert, S., Simell, O., Slowinski, T., Tiesler, C. M. T., Tönjes, A., Vaag, A., Viikari, J. S., Vink, J. M., Vissing, N. H., Wareham, N. J., Willemsen, G., Witte, D. R., Zhang, H., Zhao, J., Wilson, J. F., Stumvoll, M., Prentice, A. M., Meyer, B. F., Pearson, E. R., Boreham, C. A. G., Cooper, C., Gillman, M. W., Dedoussis, G. V., Moreno, L. A., Pedersen, O., Saarinen, M., Mohlke, K. L., Boomsma, D. I., Saw, S.-M., Lakka, T. A., Körner, A., Loos, R. J. F., Ong, K. K., Vollenweider, P., van Duijn, C. M., Koppelman, G. H., Hattersley, A. T., Holloway, J. W., Hocher, B., Heinrich, J., Power, C., Melbye, M., Guxens, M., Pennell, C. E., Bønnelykke, K., Bisgaard, H., Eriksson, J. G., Widén, E., Hakonarson, H., Uitterlinden, A. G., Pouta, A., Lawlor, D. A., Smith, G. D., Frayling, T. M., McCarthy, M. I., Grant, S. F. A., Jaddoe, V. W. V., Jarvelin, M.-R., Timpson, N. J., Prokopenko, I., and Freathy, R. M. (2012). New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature Genetics*, 45(1):76–82.

Hummell, B. M. d. L., Cutter, S. L., and Emrich, C. T. (2016). Social Vulnerability to Natural Hazards in Brazil. *International Journal of Disaster Risk Science*, 7(2):111–122.

Instituto Brasileiro de Geografia e Estatística (2010). Censo Demográfico da População [Demographic census of the population].

Kramer, M. S. (1987). Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization*, 65(5):663–737.

Little, R. E. (1977). Moderate alcohol use during pregnancy and decreased infant birth weight. *American Journal of Public Health*, 67(12):1154–1156.

Ludwig, D. S. and Currie, J. (2010). The association between pregnancy weight gain and birthweight: A within-family comparison. *The Lancet*, 376(9745):984–990.

Makhija, K., Murthy, G. V., Kapoor, S. K., and Lobo, J. (1989). Socio-biological determinants of birth weight. *Indian Journal of Pediatrics*, 56(5):639–43.

Mansour, H. and Rees, D. I. (2012). Armed conflict and birth weight: Evidence from the al-Aqsa Intifada. *Journal of Development Economics*, 99(1):190–199.

Marengo, J. A., Borma, L. S., Rodriguez, D. A., Pinho, P., Soares, W. R., and Alves, L. M. (2013). Recent Extremes of Drought and Flooding in Amazonia: Vulnerabilities and Human Adaptation. *American Journal of Climate Change*, 2(June):87–96.

Maru, Y. T., Stafford Smith, M., Sparrow, A., Pinho, P. F., and Dube, O. P. (2014). A linked vulnerability and resilience framework for adaptation pathways in remote disadvantaged communities. *Global Environmental Change*, 28:337–350.

McCormick, M. C. (1985). The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity. *New England Journal of Medicine*, 312(2):82–90.

McGuigan, C., Reynolds, R., and Wiedmer, D. (2002). Poverty and climate change: Assessing impacts in developing countries and the initiatives of the international community. *London School of Economics Consultancy Project for the Overseas Development Institute*, pages 1–40.

McIntire, D. D., Bloom, S. L., Casey, B. M., and Leveno, K. J. (1999). Birth Weight in Relation to Morbidity and Mortality among Newborn Infants. *New England Journal of Medicine*, 340(16):1234–1238.

Mckee, T. B., Doesken, N. J., and Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *AMS 8th Conference on Applied Climatology*, (January):179–184.

McMichael, A. J., Woodruff, R. E., and Hales, S. (2006). Climate change and human health: Present and future risks. *Lancet*, 367(9513):859–869.

Menendez, C., Ordi, J., Ismail, M., Ventura, P., Aponte, J., Kahigwa, E., Font, F., and Alonso, P. (2000). The Impact of Placental Malaria on Gestational Age and Birth Weight. *The Journal of Infectious Diseases*, 181(5):1740–1745.

Moser, K., Li, L., and Power, C. (2003). Social inequalities in low birth weight in England and Wales: trends and implications for future population health. *J Epidemiol Community Health*, 57(9):687–691.

Murray, L. J., O'Reilly, D. P. J., Betts, N., Patterson, C. C., Davey Smith, G., and Evans, A. E. (2000). Season and outdoor ambient temperature: Effects on birth weight. *Obstetrics and Gynecology*, 96(5):689–695.

Nobile, C. G., Raffaele, G., Altomare, C., and Pavia, M. (2007). Influence of maternal and social factors as predictors of low birth weight in Italy. *BMC Public Health*, 7(1):192.

Olson, S. H., Gangnon, R., Elguero, E., Durieux, L., Guégan, J. F., Foley, J. a., and Patz, J. a. (2009). Links between climate, malaria, and wetlands in the amazon basin. *Emerging Infectious Diseases*, 15(4):659–662.

Parry, L., Davies, G., Almeida, O., Frausin, G., de Moraés, A., Rivero, S., Filizola, N., and Torres, P. (2017). Social Vulnerability to Climatic Shocks Is Shaped by Urban Accessibility. *Annals of the American Association of Geographers*, 4452(October):1–19.

Parry, L., Day, B., Amaral, S., and Peres, C. A. (2010). Drivers of rural exodus from Amazonian headwaters. *Population and Environment*, 32(2):137–176.

Porporato, A., Vico, G., and Fay, P. A. (2006). Superstatistics of hydro-climatic fluctuations and interannual ecosystem productivity. *Geophysical Research Letters*, 33(15):2–5.

Reime, B., Ratner, P. A., Tomaselli-Reime, S. N., Kelly, A., Schuecking, B. A., and Wenzlaff, P. (2006). The role of mediating factors in the association between social deprivation and low birth weight in Germany. *Social Science and Medicine*, 62(7):1731–1744.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Risnes, K. R., Vatten, L. J., Baker, J. L., Jameson, K., Sovio, U., Kajantie, E., Osler, M., Morley, R., Jokela, M., Painter, R. C., Sundh, V., Jacobsen, G. W., Eriksson, J. G., Sørensen, T. I., and Bracken, M. B. (2011). Birthweight and mortality in adulthood: A systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3):647–661.

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., and Chivian, E. (2001). Climate change and extreme weather events. *Global change & human health*, 2(2):90–104.

Sherman, M., Ford, J., Llanos-Cuentas, A., Valdivia, M. J., and Bussalleu, A. (2015). Vulnerability and adaptive capacity of community food systems in the Peruvian Amazon: a case study from Panaillo. *Natural Hazards*, 77(3):2049–2079.

Smith, L. T., Aragão, L. E. O. C., Sabel, C. E., and Nakaya, T. (2014). Drought impacts on children's respiratory health in the Brazilian Amazon. *Scientific reports*, 4:3726.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):583–616.

Steketee, R. W. (2003). Pregnancy, nutrition and parasitic diseases. *The Journal of nutrition*, 133(5 Suppl 2):1661S–1667S.

Stephenson, T. (2002). Maternal nutrition as a determinant of birth weight. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 86(1):4F–6.

Stieb, D. M., Chen, L., Eshoul, M., and Judek, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environmental Research*, 117:100–111.

Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627.

Vaitla, B., Devereux, S., and Swan, S. H. (2009). Seasonal hunger: A neglected problem with proven solutions. *PLoS Medicine*, 6(6).

Watts, N., Adger, W. N., Agnolucci, P., Blackstock, J., Byass, P., Cai, W., Chaytor, S., Colbourn, T., Collins, M., Cooper, A., Cox, P. M., Depledge, J., Drummond, P., Ekins, P., Galaz, V., Grace, D., Graham, H., Grubb, M., Haines, A., Hamilton, I., Hunter, A., Jiang, X., Li, M., Kelman, I., Liang, L., Lott, M., Lowe, R., Luo, Y., Mace, G., Maslin, M., Nilsson, M., Oreszczyn, T., Pye, S., Quinn, T., Svensdotter, M., Venevsky, S., Warner, K., Xu, B., Yang, J., Yin, Y., Yu, C., Zhang, Q., Gong, P., Montgomery, H., and Costello, A. (2015). Health and climate change: Policy responses to protect public health. *The Lancet*, 386(10006):1861–1914.

Young, I. (2009). Five faces of oppression. Geographic Thought. *A Praxis Perspective*, pages 55–71.

Chapter 4

In Chapter 3, we have found significant effects of extreme hydro-climatic events affecting birthweight and detected certain characteristics of disadvantaged groups. However, further studies are required to understand the causal mechanism that links hydro-climatic extremes and birthweight. An attempt to better understand this link is presented in this chapter by studying *food insecurity*, which is a latent construct to represent a situation in which individual or household access to sufficient, safe and nutritious food is not a guaranteed (National Research Council, 2006). We present a novel approach to model latent constructs with spatial structure and apply it to the modelling and prediction of food insecurity; we obtain areas with high food insecurity and evaluate if they are prone to extreme hydro-climatic events.

Spatial Item Factor Analysis With Application to Mapping Food Insecurity

Erick A. Chacón-Montalván[1], Luke Parry[2,5], Emanuele Giorgi[1], Patricia Torres[3], Jesem Orellana[4], Benjamin M. Taylor[1]

[1]Centre for Health Informatics, Computing, and Statistics (CHICAS), Lancaster Medical School, Lancaster University, United Kingdom.
[2]Lancaster Environment Centre, Lancaster University, United Kingdom.
[3]Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
[4]Instituto Leônidas e Maria Deane, Fundação Oswaldo Cruz, Manaus, Brazil
[5]Núcleo de Altos Estudos Amazônicos, Universidade Federal do Pará, Belém, Brazil

Abstract

Item factor analysis is widely used for studying the relationship between a latent construct and a set of observed variables. One of the main assumptions of this method is that the latent construct or factor is independent between subjects, which might not be adequate in certain contexts. In the study of food insecurity, for example, this is likely not true due to a close relationship with socio-economic characteristics, that are spatially structured. In order to capture these effects, we propose an extension of item factor analysis to the spatial domain that is able to predict the latent factors at unobserved spatial locations. We develop a Bayesian sampling scheme for providing inference and illustrate the explanatory strength of our model by application to a study of the latent construct 'food insecurity' in a remote urban centre in the Brazilian Amazon. We use our method to map the dimensions of food insecurity in this area and identify the most severely affected areas. Our methods are implemented in an R package, `spifa`, available from Github.

Keywords: Continuous spatial variation, Factor analysis, Gaussian processes, Item Factor Analysis, Kriging, Model-based geostatistics, Multivariate regression, Spatial prediction.

4.1 Introduction

This paper concerns the analysis of geo-referenced survey data in which there is interest in understanding a set of spatially-varying *latent constructs*. A latent construct is a complex attribute or property that can be described by a number of characteristics, sometimes elicited through responses to survey questions for example. They are not rigidly defined, rather the characteristics suggest the construct and may be debated and revised as time progresses. Latent constructs are very widely used across many areas of scientific research; in psychological research for instance, an example of a latent construct would be extroversion. This characteristic is not directly measurable for an individual (unlike age for example), but it can be measured through questionnaires such as the Keirsey Temperament Sorter (Briggs Myers and Myers, 1980; Keirsey, 1998). The idea is that the construct, extroversion, can be indirectly measured through responses to a subset of questions designed to elicit social behaviour and preferences. The collective response to these questions, created for example using a summative operation (in the case of binary data), is used to infer the degree of extroversion, as opposed to introversion, in a person.

Using the language of *Item Response Theory* (IRT), the individual questions in a survey (or test) are referred to as *items*, see Hambleton and Swaminathan (1989) for a detailed review. The responses to these items measure different concrete characteristics, known as *observable variables*. To continue the extroversion example above, item 15 from the Kiersey Temperament Sorter is "At a party, do you (a) interact with many, even strangers or (b) interact with a few friends?" and the observable variable in this case might be 'interaction preferences in social situations'. Item response theory is a family of statistical models used to relate responses to items to the latent construct(s). These models assume the latent construct or *ability* (degree of extroversion in this case) is defined on a continuum. This allows us, for instance to score each individual's ability; to identify which items have the greatest capacity to *discriminate* between individuals of differing abilities (i.e. how well each item identifies the trait of extroversion in individuals); or to identify the *difficulty* associated to each item – more 'difficult' items in this context would tend to be endorsed by more extroverted individuals, but less often by less extroverted individuals (De Ayala, 2013).

Item response theory has been widely applied in many areas of research. In psychometrics, for example, it has been used to measure the theory of mind ability (Shryane et al., 2008), emotional intelligence (Fiori et al., 2014), self-esteem (Gray-Little et al., 1997). In health and medicine, it is used to determine the health status of patients using self-reported outcomes (Edelen and Reeve, 2007), to measure individual scores of child developmental status (Drachler et al., 2007) and to asses achievement and evaluation of clinical performance (Downing, 2003). In mental health research, it has been used to study disorders like psychopathy

(Laurens et al., 2012), alcohol use (Saha et al., 2006) and depression (Sharp et al., 2006). In e-learning, item response theory has been used to develop personalized intelligent tutoring systems that match learner ability and difficulty level (Chen and Duh, 2008). In computerized adaptive testing, it is used in tests like GMAT, GRE or TOEFL to dynamically select the most appropriate items for examinees according to individual abilities (Chen et al., 2006). In marketing, it has been used to measure customer relationship satisfaction (Funk and Rogge, 2007) and to measure extreme response styles (ERS) (de Jong et al., 2008). In criminology, it is applied to the analysis of the causes of crime and deviance using self-reporting measures of delinquency (Osgood et al., 2002) and to measure self-control (Piquero et al., 2000).

Our motivating application concerns the assessment of household food insecurity which is mediated through a family's ability to access food and also through the supply of food potentially available. Both factors are relevant in the context of our study located in Ipixuna, a remote urban centre in the Brazilian Amazon. Food insecurity was measured using responses to a modified version of the questionnaire proposed by the United States Department of Agriculture (Carlson et al., 1999; National Research Council, 2006). Food insecurity in these remote and roadless urban centres, accessible only by boat or plane, is partly affected by seasonal variation in river levels. During particularly dry months it may be difficult for cargo boats to access the city and in very wet months there are risks of large-scale flooding - disease, loss of home and income. But there are other factors at play too: community, governmental and non-governmental support can bolster a family's food resources in difficult times (Garrett and Ruel, 1999; Battersby, 2011). As is the case with cities in the West, neighbourhoods with certain characteristics tend to cluster together: it is for exactly this reason that in this paper we propose to extend traditional IRT models to accommodate spatial structure, among other attributes detailed below.

One of the main limitations of classical IRT models is that they assume that the latent construct is unidimensional: this assumption may not be adequate for more complex latent constructs. For example, the items developed to study food insecurity capture a number of different concepts including: (i) the perception of reduction in the quality or quantity of food, (ii) an actual reduction in quality of food, (iii) an actual reduction in quantity of food, and (iv) a reduction in the quantity or quality of food for children in the household. Hence, the construct food insecurity has more than one dimension, and might also depend on characteristics of the population under study, Froelich and Jensen (2002) for example found a further dimension associated with the protection of children from hunger.

In this context, where unidimensional models are not appropriate, researchers have developed *Multidimensional Item Response Theory* (MIRT) or *Item Factor Analysis* (IFA), both approaches being conceptually similar (Bock et al., 1988; Chalmers, 2012). These models extend the concept of standard multivariate factor analysis so it can be applied to binary or ordinal data and allows us to study the interaction between multiple items and a multi-dimensional latent construct. Although item factor analysis addresses the problem of uni-dimensionality, there are other limitations of this approach that we seek to address in the present paper.

Firstly, IFA assumes the latent construct of a particular subject to be independent of any other subject. In our subsequent example of food insecurity, this seems inadequate given that households near to each other are more likely to share similar socio-economic conditions and environmental exposures and thus a similar risk of food insecurity. This observation also applies to the analysis of latent constructs in other disciplines where spatial correlation is naturally expected, an example would be socio-economic status itself. Connected to this, an item factor analysis model incorporating spatial random effects would allow us to map the latent factors at unobserved locations, which can be (and is in our case) of scientific interest. With respect to our own and other similar application(s), a complete map of the latent factors over the area under study will improve our understanding of the construct and help to better inform the decision-making process.

Secondly, IFA only relates items to the latent construct, but not to possible covariates that could help explain why certain individuals might have particularly high or low values of the latent construct. For example, our previous research in this area suggests socio-economic and environmental variables play an important role in determining food insecurity (Parry et al., 2017). In our case, therefore, understanding the relationship between the items, the latent construct and the covariates is highly desirable.

The above summarises our motivation for developing an extension to IFA which we here denominate *spatial item factor analysis*. Our hierarchical framework allows the latent construct to be split into multiple latent factors, the number and composition of which are determined by initial exploratory analyses. These latent factors are explained by observed covariates and also by spatially-correlated random effects. The relationship between the latent factors and the item responses, in the case of binary outcomes, is mediated through a set of auxiliary variables which handle the conversion between continuous to discrete data forms. We present an efficient Metropolis-within-Gibbs sampling strategy for Bayesian inference with our model.

The structure of the paper is as follows. Details of our proposed model for spatial item factor analysis is presented in Section 4.2. This model is implemented in our R package described in Section 4.3. Bayesian inference for our model through Markov chain Monte Carlo methods is explained in Section 4.4. Spatial prediction for the latent construct is developed in Section 4.5. Then we detail application of the model to predicting food insecurity in Section 4.6. Finally, this paper concludes with a discussion of the advantages, disadvantages and possible extensions to our model in Section 4.7.

4.2 Spatial Item Factor Analysis

In this section we develop a modelling framework for spatial item factor analysis. We first introduce classical item factor analysis in Section 4.2.1, then in

Section 4.2.3 we introduce our new methods. Solutions to identifiability issues in our model are discussed in Section 4.2.4. We then introduce the matrix form of the auxiliary variables of our model in Section 4.2.7. We conclude this section with the specification of the likelihood function in Section 4.2.8.

4.2.1 Item Factor Analysis

Item factor analysis can be seen as an extension of factor analysis for binary or ordinal data. In the present article, we concentrate on binary outcomes and discuss extensions of the proposed framework to a mix of continuous, binary and ordinal items in the Discussion (Section 4.7) and in Appendix E.

We begin by considering the response variable Y_{ij} for item $j = 1, 2, \ldots, q$ from subject $i = 1, 2, \ldots, n$ as a binarization around zero of a continuous but unobservable process Z_{ij}, explained by m latent factors (also called latent abilities) $\theta_{1i}, \ldots, \theta_{mi}$,

$$Z_{ij} = c_j + \sum_{k=1}^{m} a_{jk}\theta_{ki} + \epsilon_{ij}, \tag{4.1}$$

where $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ and $\{c_j\}$ are intercept parameters that take into model the difficulty of items. High positive (negative) values for c_j increase (reduce) the probability of endorsing the j-th item, which is why they are also referred to as *easiness* parameters (Chalmers, 2015). The slopes $\{a_{jk}\}$, commonly called *discrimination* parameters, indicate how well the j-th item can discriminate the k-th ability between the subjects under study. If $a_{jk} = 0$, the k-th latent factor does not explain the variability of the j-th response item, in other words this item does not help to discriminate the k-th latent ability between the subjects. In our paper, we also use this parameterisation i.e. using intercepts and slopes. Further details on this model including inference via the expectation-maximization algorithm can be found in Bock et al. (1988).

As well as estimating the easiness and discrimination parameters, interest may also lie in making inferences for the latent factors θ_{ki}, this allows us to differentiate individuals with high or low levels of the construct under study. A practical application of this is in the area of *ideal point estimates*, where the objective is to estimate the ideological position of a political legislator in order to predict whether they will vote in favour of a particular motion, see Bafumi et al. (2005) for example.

4.2.2 Exploratory and Confirmatory Item Factor analysis

The model defined by Equation 4.1 is not identifiable due to different types of aliasing, as explained in Section 4.2.4. We can make the model identifiable by placing restrictions on some parameters. The way this is done yields two different approaches.

We obtain an exploratory item factor analysis if the restrictions are imposed only to make the inference possible, i.e. the restrictions are not related to the construct and data under analysis. In this case, estimates can be rotated under the preference of the researcher.

We obtain a confirmatory item factor analysis if the restrictions are established in a semi-formal manner: the researcher uses their own (or expert) knowledge about the latent construct to establish the structure of an appropriate model (Cai, 2010b). In a confirmatory item factor analysis, the restrictions are designed with a particular study and context in mind, while in exploratory item factor analysis, the restrictions are generally imposed and are not problem-specific. Where experts cannot agree on a particular structure for the model, the option to use measures of model fit (e.g. WAIC or DIC for a Bayesian analysis) is still possible, as is model averaging.

4.2.3 Extension to the Spatial Domain

In our application, we are interested in estimating the easiness and discrimination parameters in order to understand the relationship between the underlying latent factors with the response variables. We also want to be able to predict the latent factors not only in places where the observations were taken, but also in locations where we have no observations. Our data were costly, difficult and time-consuming to collect, thus our method for predicting food insecurity at new locations is an important step for identifying particularly vulnerable areas that could be targeted for intervention. Since our method can also be used to map and predict the different dimensions of food insecurity, this information could be used to tailor specific interventions to specific regions. This is our motivation for the development of spatial item factor analysis.

The extension of item factor analysis to the spatial domain can be achieved by including a spatial process in the predictor in Equation 4.1. Frichot et al. (2012), for example, proposed such a model by including a spatially correlated error term ϵ_{ij}. This extension tries to correct the principal components by modelling the residual spatial variation. Our proposed method, spatial item factor analysis allows the latent factors θ_{ki} to be spatially correlated because the nature of the particular construct we are studying suggests they should be treated in this way. For example, we expect there to be spatial patterns in food insecurity scores across a municipality due to the relationship with socio-economic and environmental variables.

We model the binary response variables as a discrete-state stochastic processes $\{Y_j(s) : s \in D\}$ where $D \subset \mathbb{R}^2$ and the notation $Y_j(s)$ is the response to item j at spatial location s. The response variables take values 0 or 1 according to the value assumed by an auxiliary spatial stochastic process $\{Z_j(s) : s \in D\}$:

$$Y_j(s) = \begin{cases} 1 & \text{if } Z_j(s) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Conditional on $Z_j(s)$, the values assumed by $Y_j(s)$ are deterministic. We model the auxiliary process as follows:

$$Z_j(s) = c_j + \boldsymbol{a}_j^{\intercal}\boldsymbol{\theta}(s) + \epsilon_j(s), \quad \epsilon_j(s) \sim \mathcal{N}(0, 1), \tag{4.3}$$

where c_j and $\boldsymbol{a}_j$ are respectively the easiness and discrimination parameters. The latent factors, $\boldsymbol{\theta}(s)$, are defined as a function of covariates, continuous-space m-dimensional stochastic process and a non-spatially correlated error term as in Equation 4.4. Note that this process is the only source of spatial correlation in $Z_j(s)$ and $Y_j(s)$: if the spatial variation is removed from $\theta_j(s)$, then the model reduces to a simple item factor analysis.

The different assumptions that one can make with respect to $\boldsymbol{a}_j$ and $\boldsymbol{\theta}(s)$ generate different types of models. For example, under the assumption that $\boldsymbol{\theta}(s)$ and $\boldsymbol{\theta}(s')$ are uncorrelated with the further assumption that $\boldsymbol{\theta}(s) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$, this generates an exploratory item factor analysis (Cai, 2010a). Alternatively, restrictions on $\boldsymbol{a}_j$ lead to a confirmatory item factor analysis (Cai, 2010b). The reasons why we include these assumptions and restrictions will be explained in Section 4.2.4: the concern is identifiability and our spatial item factor analysis model requires specific choices here.

In a similar manner, we can impose a particular structure on the latent factor $\boldsymbol{\theta}(s)$ in order to create our spatial item factor model. Since one of our interests is predicting the latent factors at unobserved locations s^*; we define the structure of $\boldsymbol{\theta}(s)$ through a set of spatial covariates $\boldsymbol{x}(s) = (x_1(s), \ldots, x_p(s))$ that preferably are also available at unobserved locations. This way the model allows us to understand *why* certain individuals have high or low scores. The inclusion of covariates in factor analyses leads to *multiple indicators, multiple causes models* (MIMIC) in the literature on structural equation modelling (SEM), see Tekwe et al. (2014) for example. We include a latent spatial stochastic process $\{\boldsymbol{w}(s) : s \in D\}$ into our model for $\boldsymbol{\theta}(s)$, defining the m-dimensional latent factor as:

$$\boldsymbol{\theta}(s) = \boldsymbol{B}^\intercal \boldsymbol{x}(s) + \boldsymbol{w}(s) + \boldsymbol{v}(s), \tag{4.4}$$

where $\boldsymbol{B}$ is an $p \times m$ matrix of slopes associating a set of covariates $\boldsymbol{x}(s)$ with the latent factor $\boldsymbol{\theta}(s)$ and $\boldsymbol{v}(s)$ as defined below. Note that we will eventually assume that the covariates have been standardised, see Section 4.2.4 for further details.

We define $\boldsymbol{w}(s) = \{w_k(s)\}_{k=1}^{m}$ to be a set of zero-mean, independent, stationary and isotropic Gaussian processes with variance σ_k^2 and correlation function $\rho_k(u)$ at distance u,

$$w_k(s) \sim \mathrm{GP}(0, \sigma_k^2, \rho_k(u)), \qquad k = 1, \dots, m. \tag{4.5}$$

This definition might seem restrictive, but the independence assumption of these spatial processes is adequate when the latent factors $\theta_k(s)$ are independent and it could still be adequate when the latent factors are not independent. We use vector notation to denote $\boldsymbol{w}(s)$ because later, in Section 4.2.5, we discuss further extensions to the structure of $\boldsymbol{w}(s)$, such as allowing correlation between the $w_k(s)$ and thus at the outset we wish to think of this as a multivariate Gaussian process (MGP).

Finally, the m-dimensional random vector $\boldsymbol{v}(s)$ is the remaining uncertainty in the latent factors that is neither explained by the covariates nor by $\boldsymbol{w}(s)$. We assume $\boldsymbol{v}(s)$ is a zero-mean multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}_v$,

$$\boldsymbol{v}(s) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_v). \tag{4.6}$$

Equation 4.4 has the same structure as a multivariate geostatistical model. However, in our case, the dependent variable, $\boldsymbol{\theta}(s)$, is a low-dimensional latent process instead of a high-dimensional observed process as in Gelfand et al. (2004). A similar structure including fixed and random effects is also discussed in Chalmers (2015), but the author does not attempt to model unexplained spatial variation. In addition, the author mainly focuses on including covariates at the item level, whereas our emphasis is on the inclusion of covariates at the subject level which will then allow us to make predictions about individuals at unobserved locations.

Substituting the structure of the latent factors $\boldsymbol{\theta}(s)$ into Equation 4.3 results in

$$Z_j(s) = c_j + \boldsymbol{a}_j^{\intercal}[\boldsymbol{B}^{\intercal}\boldsymbol{x}(s) + \boldsymbol{w}(s) + \boldsymbol{v}(s)] + \epsilon_j(s). \tag{4.7}$$

We note that if $\boldsymbol{a}_j$ were known, then Equation 4.7 would be a multivariate geostatistical model. The main challenges in our proposed model come from the inclusion of the interaction between the latent variables with the (unknown) slopes, $\boldsymbol{a}_j$.

In theory, the proposed model could be used in both exploratory and confirmatory factor analysis. However, we suggest using the model for confirmatory factor analysis in which there is no rotation of the latent factors - in this way, the correlation parameters are directly interpretable. If on the other hand, the latent factors have been rotated, as in exploratory analysis, interpreting the correlation parameters is then more difficult.

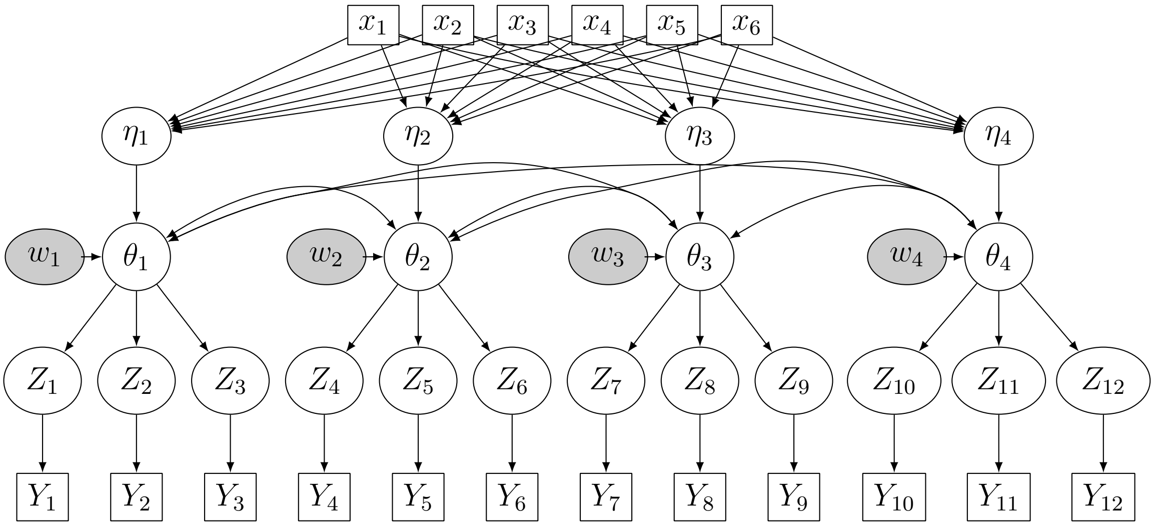

Figure 4.1: Directed Graph for the Spatial Item Factor Model: This example has twelve response items (Y_j), twelve auxiliary variables (Z_j), four latent factors (θ_k), four Gaussian processes (w_k), four linear predictors (η_k) and six covariates (x_l).

The relationship between covariates $\boldsymbol{x}(s)$, latent factors $\boldsymbol{\theta}(s)$, auxiliary latent variables $\boldsymbol{Z}(s)$ and response variables $\boldsymbol{Y}(s)$ can be seen more clearly through an example of spatial confirmatory factor analysis, as shown in Figure 4.1. This figure shows a directed graph with twelve items $Y_j(s)$, or response variables, four latent factors $\theta_k(s)$, four Gaussian processes $w_k(s)$ and six covariates $x_l(s)$. We have introduced η_k as a linear combination of the covariates in order to have a more clear visualization of the model. In this example some of the coefficients $\boldsymbol{a}_j$ are set to zero so that each factor is only explained by a subset of items; this is usually decided using an exploratory item factor analysis. It can be seen that the 12-dimensional response vector $\boldsymbol{Y}(s)$ is reduced to a 4-dimensional space of factors $\boldsymbol{\theta}(s)$. These factors are allowed to be correlated with each other and also spatial correlation is permitted within factors. The top row in the figure shows how covariates $\boldsymbol{x}(s)$ are used to predict the latent factors $\boldsymbol{\theta}(s)$.

4.2.4 Identifiability and restrictions

The model presented above is subject to the same *identifiability* problems as those found in factor analysis and structural equation modelling. Identifiability issues arise when different sets of parameters lead to the same likelihood in a structured way - this leads to symmetry in the posterior distribution in a Bayesian framework, (or objective function in a classical approach), i.e. there are multiple modes. In our model, these identifiability issues could be due to *additive*, *scaling*, *rotational* or *reflection* aliasing, which will be discussed in detail below.

Additive aliasing occurs when the item difficulties c_j and the product $\boldsymbol{a}_j^\intercal \boldsymbol{\theta}_j$ have free means. Under this situation a constant value could be added and subtracted to each term respectively and the probability density function will be unchanged. Similarly, if $\boldsymbol{a}_j^\intercal$ is multiplied by a constant and $\boldsymbol{\theta}_j$ divided by the same constant, then the probability density function is constant leading to *scaling* aliasing.

In order to address the issues of scaling and additive aliasing in classical item factor analysis as in Equation 4.1 it is common to assume that $\boldsymbol{\theta}_{ik} \sim \mathcal{N}(0, 1)$ (Bafumi et al., 2005). A generalisation of this would be to assume $\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$, whence the previous solution is obtained by setting that $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}$ for an exploratory factor analysis or by setting $\text{diag}(\boldsymbol{\Sigma}_\theta) = \mathbf{1}$ for a confirmatory factor analysis.

The spatial item factor model presented in Section 4.2.3 does not suffer from additive aliasing because we are already assuming the processes $\boldsymbol{w}(s)$ and $\boldsymbol{v}(s)$ are zero-mean. As mentioned above, we assume that the covariates included in Equation 4.4 are standardised, which leads to the latent factor $\boldsymbol{\theta}(s)$ having mean zero.

However, our model does suffer from scaling aliasing, so we are required to restrict the variances of the latent factors $\theta_k(s)$, this is complicated by the presence of covariates because we cannot directly ensure that $\mathbb{V}\left[\boldsymbol{b}_k^\intercal \boldsymbol{x}(s) + w_k(s) + v_k(s)\right] = 1$. One simple way of achieving the required restriction is by fixing the variance of one of the terms inside the structure of the latent factors in Equation 4.4, see Appendix A.1 for details. This is usually applied to the multivariate error term, $\boldsymbol{v}(s)$, as in Tekwe et al. (2014). It is sufficient to fix a diagonal matrix $\boldsymbol{D}$, which contains the marginal standard deviations of $\boldsymbol{v}(s)$, $\text{diag}(\boldsymbol{D}) = (\sigma_{v_1}, \ldots, \sigma_{v_m})^\intercal$, such that

$$\boldsymbol{\Sigma}_v = \boldsymbol{D}\boldsymbol{R}_v\boldsymbol{D}, \tag{4.8}$$

where $\boldsymbol{R}_v$ is a correlation matrix. The usual restrictions applied in exploratory or confirmatory item factor analysis are equivalent to setting $\boldsymbol{D} = \boldsymbol{I}$. If the model includes both covariates and Gaussian processes and we are conducting an exploratory item factor analysis, then this method does not work well because the marginal variances of the latent factors $\theta_k(s)$ might become big (greater than 1) and consequently, the discrimination parameters would have to be close to zero and become unidentifiable in practice. This happens because the modes of the posterior distribution will not be very well separated and the MCMC chains will be jumping between modes that are equivalent solutions. For confirmatory item factor analysis the condition $\boldsymbol{D} = \boldsymbol{I}$ is sufficient to eliminate issues of scaling aliasing, see Appendix A.1.

Although the restrictions imposed modify the interpretation of the discrimination parameters $\boldsymbol{a}_j$ because the latent factors are on different scales, they are only necessary in order to make the inference possible. Therefore, a scaling transformation can applied post-estimation in order to recover the correct interpretation of the discrimination parameters:

$$Z_j(s) = c_j + \boldsymbol{a}^\intercal \boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{\theta}(s) + \epsilon_j(s), \tag{4.9}$$

where $\boldsymbol{Q}$ is a diagonal matrix of the standard deviations for $\boldsymbol{\theta}(s)$. This transformation leads to a new vector of latent abilities $\boldsymbol{Q}^{-1}\boldsymbol{\theta}(s)$ with unit variances and discrimination parameters $\boldsymbol{a}^\intercal\boldsymbol{Q}$ with the usual interpretation as in item factor analysis, see Section 4.4.4 for further details.

Returning to classical item factor analysis, the other two types of aliasing, *rotational* and *reflection*, are due to the fact that linear transformations of the slope parameters $\boldsymbol{a}_j^* = \boldsymbol{a}_j^\intercal \boldsymbol{\Lambda}^{-1}$ and of the latent factors $\boldsymbol{\theta}_i^* = \boldsymbol{\Lambda}\boldsymbol{\theta}_i$ lead to the same probability density function of the original parameters $\boldsymbol{a}_j$ and $\boldsymbol{\theta}_j$ given that $\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda} = \boldsymbol{I}$ (Erosheva and Curtis, 2011). In exploratory factor analysis $\boldsymbol{\Lambda}$ is an orthogonal matrix because it is assumed $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}$; it can be shown that this implies $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\intercal = \boldsymbol{I}$. In the case of rotational aliasing the matrix of the linear transformation has $m(m-1)/2$ degrees of freedom. Hence, $m(m-1)/2$ restrictions can be applied to eliminate this type of aliasing. The usual criteria is to set $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q)^\intercal$ to be a lower triangular matrix (Geweke and Zhou, 1996). For reflection aliasing, there are 2^m orthogonal matrices $\boldsymbol{\Lambda}$ obtained by simultaneously changing the signs of $\boldsymbol{a}_j$ and $\boldsymbol{\theta}_i$. In this case, identifiability can be ensured by setting the diagonal elements of $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J)^\intercal$ to be positive (Geweke and Zhou, 1996).

For spatial exploratory item factor analysis the above restrictions on the discrimination parameters, or similar, are necessary. For confirmatory factor analysis it is sufficient to fix $m(m-1)/2$ entries (usually the value chosen is zero) of $\boldsymbol{A}$ and also set as positive (or negative) one element from each column of $\boldsymbol{A}$; the former addresses rotation aliasing, and the latter reflection aliasing. More generally, a set of restrictions can be induced through a linear association between the constrained parameters $\boldsymbol{a}_j^*$ and the free parameters $\boldsymbol{a}_j$,

$$\boldsymbol{a}_j^* = \boldsymbol{u}_j + \boldsymbol{L}_j \boldsymbol{a}_j, \tag{4.10}$$

where the vector $\boldsymbol{u}_j$ are the values that are to be fixed, while the matrix $\boldsymbol{L}_j$ indicates which elements of the free-parameter $\boldsymbol{a}_j$ are to be activated (Cai, 2010b). In the example below, the third and fourth elements of the parameter vector are set to 0 and 1 respectively.

$$\begin{pmatrix} a_{j1}^* \\ a_{j2}^* \\ a_{j3}^* \\ a_{j4}^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{j1} \\ a_{j2} \\ a_{j3} \\ a_{j4} \end{pmatrix} = \begin{pmatrix} a_{j1} \\ a_{j2} \\ 0 \\ 1 \end{pmatrix} \tag{4.11}$$

In practice, achieving the required positivity (or negativity) constraints is accomplished through the appropriate specification of the marginal prior distributions, see Section 4.4.2 for details.

4.2.5 Allowing Further Flexibility on the Multivariate Spatial Structure

In the discussion above, we proposed using a set of independent Gaussian processes in the structure of the latent factors $\boldsymbol{\theta}(s)$. It is adequate when each latent factor $\theta_k(s)$ has a spatial structure and they are independent; the second condition might be held when the restrictions on the discrimination parameters are imposed based on an exploratory item factor analysis with varimax rotation. However, in more general situations, it can be the case that some of the factors $\theta_k(s)$ are not spatially correlated, or that some of the unexplained variation in two or more factors may have a common (spatially-correlated) component. In this situation it will be desirable, respectively, to include spatial structure on only a subset of the factors, or to share the spatial structure across several factors.

In a similar way to how restrictions were imposed on the discrimination parameters, we can use an $m \times g$ transformation matrix $\boldsymbol{T}$ to convert g independent standard Gaussian processes in $\boldsymbol{w}(s)$ into an m-dimensional multivariate Gaussian process, $\boldsymbol{w}^*(s)$:

$$\boldsymbol{w}^*(s) = \boldsymbol{T}\boldsymbol{w}(s). \tag{4.12}$$

An example is given in Equation 4.13, where after transforming, $w_1(s)$ is common to the first and second factor and the second factor has an additional spatial structure, namely $w_2(s)$; $w_3(s)$ features in the third factor, and the last factor does not include any Gaussian process i.e. it is not spatially structured.

$$\begin{pmatrix} w_1^*(s) \\ w_2^*(s) \\ w_3^*(s) \\ w_4^*(s) \end{pmatrix} = \begin{pmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ 0 & 0 & t_{33} \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} w_1(s) \\ w_2(s) \\ w_3(s) \end{pmatrix} = \begin{pmatrix} t_{11}w_1(s) \\ t_{21}w_1(s) + t_{22}w_2(s) \\ t_{33}w_3(s) \\ 0 \end{pmatrix} \tag{4.13}$$

Notice that the variance of $\boldsymbol{w}^*(s)$ is controlled by $\boldsymbol{T}$. Using this stochastic process in Equation 4.4, we re-define the m-dimensional latent factor of our model as:

$$\boldsymbol{\theta}(s) = \boldsymbol{B}^{\intercal}\boldsymbol{x}(s) + \boldsymbol{w}^*(s) + \boldsymbol{v}(s). \tag{4.14}$$

The methods described in the section are closely connected to multivariate geostatistical models of coregionalization (Gelfand et al., 2004; Fanshawe and Diggle, 2012). The main difference is that, due to the sparse structure of $\boldsymbol{T}$, it is not ensured that $\boldsymbol{w}^*(s)$ has a positive definite covariance matrix, and that we are using this structure as a way for the user to control the nature of interrelationships between factors (which would obviously change according to the problem and data under study), rather than allowing free reign estimating all the elements of $\boldsymbol{T}$. Although we propose $\boldsymbol{w}^*(s)$ because it is simple and easy to interpret, our model is not limited to this choice, it could be replaced with a more attractive multivariate stochastic process (e.g. see Gneiting et al., 2010).

There is a sense in which the restrictions imposed can be thought of as prior specification. Provided the 'correct' overall sparse structure of $\boldsymbol{T}$ has been chosen, such restrictions are also beneficial; in particular if $m > g$ then inference becomes more tractable – both in terms of computation, and subsequently interpretation. In the absence of expert opinion (but preferably in the presence of it), we suggest using an exploratory item factor analysis before applying our model in order to evaluate these characteristics and decide on the structure of the multivariate spatial correlation defined through $\boldsymbol{T}$.

4.2.6 Auxiliary Variables in the Identifiable Spatial Item Factor Analysis

Using the restricted discrimination parameters $\boldsymbol{a}^*$ defined in Equation 4.10 and the new definition of the latent factor $\boldsymbol{\theta}(s)$ in Equation 4.14, we obtain an identifiable and flexible model for spatial item factor analysis where the auxiliary variables $Z_j(s)$ have the following structure

$$Z_j(s) = c_j + \boldsymbol{a}_j^{*\intercal}\boldsymbol{\theta}(s) + \epsilon_j(s) = c_j + \boldsymbol{a}_j^{*\intercal}[\boldsymbol{B}^\intercal \boldsymbol{x}(s) + \boldsymbol{w}^*(s) + \boldsymbol{v}(s)] + \epsilon_j(s). \quad (4.15)$$

We are assuming that the structure of the restricted discrimination parameters $\boldsymbol{a}_j^*$ and also the multivariate Gaussian process $\boldsymbol{w}^*(s)$ will be informed by expert opinion through direct involvement of researchers in the area of application and/or through consulting the academic literature in that area.

Doing this not only allows our model to be identifiable, but it also allows us to obtain interpretable latent factors which are practically useful to researchers in the field under consideration.

4.2.7 Matrix Form of the Auxiliary Variables

Expressing the terms in our model at the individual level as above (and in Equation 4.15) is convenient for understanding the various components; however, in the sequel, we will use the matrix form of our model in order to define the likelihood function (Section 4.2.8) and later derive the conditional distributions of the posterior (Section 4.4).

Before proceeding with the matrix form of our model, we introduce some further notational conventions. Let $\boldsymbol{\alpha}(s)$ be a q-variate random variable at spatial location s. Then if $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)^\intercal$ is a set of locations, we will define the q-vector $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}(s_i) = (\alpha_1(s_i), \ldots, \alpha_q(s_i))^\intercal$ and the n-vector $\boldsymbol{\alpha}_{[j]} = \boldsymbol{\alpha}_j(\boldsymbol{s}) = (\alpha_j(s_1), \ldots, \alpha_j(s_n))^\intercal$.

With the above conventions, the collection of auxiliary random variables $\boldsymbol{Z} = (\boldsymbol{Z}_{[1]}^\intercal, \ldots, \boldsymbol{Z}_{[q]}^\intercal)^\intercal$ for q items at n locations can be expressed as

$$\boldsymbol{Z} = (\boldsymbol{I}_q \otimes \mathbf{1}_n)\boldsymbol{c} + (\boldsymbol{A}^* \otimes \boldsymbol{I}_n)\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{4.16}$$

where $\boldsymbol{I}_q$ and $\boldsymbol{I}_n$ are identity matrices of dimension q and n respectively, $\mathbf{1}_n$ is a n-dimensional vector with all elements equals to one, $\boldsymbol{c} = (c_1, \ldots, c_q)^\intercal$ is a vector arrangement of the easiness parameters, $\boldsymbol{A}^*_{q\times m} = (\boldsymbol{a}^*_1, \ldots, \boldsymbol{a}^*_q)^\intercal$ is a matrix arrangement of the restricted discrimination parameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}_{[1]}^\intercal, \ldots, \boldsymbol{\theta}_{[m]}^\intercal)^\intercal$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{[1]}^\intercal, \ldots, \boldsymbol{\epsilon}_{[q]}^\intercal)^\intercal$ is a nq-vector of residual terms.

The vector of latent abilities $\boldsymbol{\theta}$ with respect to Equation 4.14 can be expressed as

$$\boldsymbol{\theta} = (\boldsymbol{I}_m \otimes \boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{w} + \boldsymbol{v}, \tag{4.17}$$

where $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B})$ is a column-vectorization of the multivariate fixed effects, $\boldsymbol{X}_{n\times p} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal$ is the design matrix of the covariates, $\boldsymbol{w} = (\boldsymbol{w}_{[1]}^\intercal, \ldots, \boldsymbol{w}_{[m]}^\intercal)^\intercal$ is the collection of the multivariate Gaussian process and $\boldsymbol{v} = (\boldsymbol{v}_{[1]}^\intercal, \ldots, \boldsymbol{v}_{[m]}^\intercal)^\intercal$ is the collection of the multivariate residual terms. Substituting Equation 4.17 into Equation 4.16, we obtain:

$$\boldsymbol{Z} = (\boldsymbol{I}_q \otimes \mathbf{1}_n)\boldsymbol{c} + (\boldsymbol{A}^* \otimes \boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{w} + (\boldsymbol{A}^* \otimes \boldsymbol{I}_n)\boldsymbol{v} + \boldsymbol{\epsilon}. \tag{4.18}$$

This matrix representation is useful for deriving the multivariate marginal and conditional distributions of $\boldsymbol{Z}$ in the following sections.

Alternatively, the collection of auxiliary variables $\boldsymbol{Z}$ can also be expressed as

$$\begin{aligned}\boldsymbol{Z} &= (\boldsymbol{I}_q \otimes \boldsymbol{1}_n)\boldsymbol{c} + (\boldsymbol{I}_q \otimes \boldsymbol{\Theta})\boldsymbol{a}^* + \boldsymbol{\epsilon} \\ &= (\boldsymbol{I}_q \otimes \boldsymbol{1}_n)\boldsymbol{c} + (\boldsymbol{I}_q \otimes \boldsymbol{\Theta})\boldsymbol{u} + (\boldsymbol{I}_q \otimes \boldsymbol{\Theta})\boldsymbol{L}\boldsymbol{a} + \boldsymbol{\epsilon}, \end{aligned} \tag{4.19}$$

where $\boldsymbol{\Theta}_{n\times m} = (\boldsymbol{\theta}_{[1]}, \ldots, \boldsymbol{\theta}_{[m]})$ is the matrix of latent abilities, $\boldsymbol{u} = (\boldsymbol{u}_1^\intercal, \ldots, \boldsymbol{u}_q^\intercal)^\intercal$ are the restrictions defined in Equation 4.10, $\boldsymbol{a} = (\boldsymbol{a}_1^\intercal, \ldots, \boldsymbol{a}_q^\intercal)^\intercal$ are the free discrimination parameters and $\boldsymbol{L} = \oplus_{j=1}^q \boldsymbol{L}_j$ is the direct sum of the activation matrices defined in Equation 4.10 (recall these link the free discrimination parameters $\boldsymbol{a}$ with the constrained discrimination parameters $\boldsymbol{a}^*$). We later use Equation 4.19 in the derivation of the conditional posterior distribution of the discrimination parameters $\boldsymbol{a}$.

4.2.8 Likelihood Function

A challenging aspect of some motivating applications is the fact that not all items are observed for all subjects. More generally, it is common to have to deal with missing data (in this case item responses) in statistics, therefore in the present section we begin to introduce notation for observed and missing data; this will be revisited several times in Section 4.4 and is also connected with prediction.

Let $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)^\intercal$ be a set of locations at which data from q items has been collected. Let the random variable $Y_{ij} = Y_j(s_i)$ be the j-th item response at location s_i. Using notation introduced in Section 4.2.7, let $\boldsymbol{Y} = (\boldsymbol{Y}_{[1]}^\intercal, \ldots, \boldsymbol{Y}_{[q]}^\intercal)^\intercal$ be the collection of responses to all items. These can be divided into two groups; the set of observed variables $\boldsymbol{Y}_{obs}$ and the set of variables that were missing $\boldsymbol{Y}_{mis}$.

The marginal likelihood function for our spatial item factor analysis model is obtained by integrating the joint density of the observed variables $\boldsymbol{Y}_{obs}$, the associated auxiliary variables $\boldsymbol{Z}_{obs}$ and the collection of latent abilities $\boldsymbol{\theta} = (\boldsymbol{\theta}_{[1]}, \ldots, \boldsymbol{\theta}_{[m]})^\intercal$;

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) = \int\int \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right) \Pr\left(\boldsymbol{z}_{obs}, \boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right) \, d\boldsymbol{z}_{obs} \, d\boldsymbol{\theta}, \tag{4.20}$$

where $\boldsymbol{a} = (\boldsymbol{a}_1^\intercal, \ldots, \boldsymbol{a}_q^\intercal)^\intercal$ is vector arrangement of all the discrimination parameters and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_g)^\intercal$ is the vector of scale parameters of the g-dimensional Gaussian process $\boldsymbol{w}(s)$. Note that the main computational cost inside the integral, $\mathcal{O}(n^3m^3)$, comes from evaluating the distribution associated with $\boldsymbol{\theta}$ which has a $mn \times mn$ covariance matrix.

In Equation 4.20, the structure of our model implies

$$\Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right) = \prod_{o_{ij}=1} \Pr\left(y_{ij} \mid z_{ij}\right), \tag{4.21}$$

where o_{ij} is an indicator variable with value equals to one when the variable Y_{ij} has been observed (i.e. is not missing) and zero otherwise. We further have:

$$\Pr\left(\boldsymbol{z}_{obs}, \boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right) = \prod_{o_{ij}=1} \Pr\left(z_{ij} \mid \boldsymbol{\theta}_i, c_j, \boldsymbol{a}_j, \boldsymbol{B}\right) \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right), \tag{4.22}$$

and variables on the right hand side are normally distributed.

Note that the definition of the likelihood function through Equation 4.20, 4.21 and 4.22 does not depend on the missing observations. Therefore, if some items were not observed in some of the locations, inference will still be possible provided the missing data are *missing at random* (Merkle, 2011). Using this likelihood, inference from the model can proceed in a number of ways. Maximum likelihood estimation can be achieved by approximating the likelihood function in Equation 4.20 using a variety of Monte Carlo methods or via stochastic approximation (Cai, 2010b). However in the present article, we focus on a Bayesian approach as shown in Section 4.4 because of the simplicity and reliability of uncertainty computation.

Our likelihood function can also be written using the auxiliary variables associated with both the observed and missing responses:

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) = \int \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right) \Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right) \, d\boldsymbol{z}. \tag{4.23}$$

The advantage of this representation is that the joint density of the auxiliary variables $\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right)$ can be obtained in a straightforward manner using Equation 4.18. It is normally distributed with mean

$$\boldsymbol{\mu}_z = (\boldsymbol{I}_q \otimes \boldsymbol{1}_n)\boldsymbol{c} + (\boldsymbol{A}^* \otimes \boldsymbol{X})\boldsymbol{\beta} \tag{4.24}$$

and covariance matrix

$$\boldsymbol{\Sigma}_z = (\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n) + (\boldsymbol{A}^* \otimes \boldsymbol{I}_n)\boldsymbol{D}\boldsymbol{R}_v\boldsymbol{D}(\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n) \tag{4.25}$$

where $\boldsymbol{\Sigma}_w = \oplus_{k=1}^{g} \boldsymbol{\Sigma}_{w_k}$ is the direct sum of the covariance matrices of the independent Gaussian processes. We prefer this last definition of the likelihood function as it allows us to handle the missing data using data augmentation, see Section 4.4.3.

4.3 R Package

Our model is implemented in an open-source `R` package, `spifa`, available from Github, `https://github.com/ErickChacon/spifa`. This package implements the Bayesian inferential method outlined below in full, allowing the user to specify the structure of the multivariate Gaussian processes and prior hyperparameters; model selection is also available through the DIC. The package has functions for summarising model output, for MCMC diagnostics and for the production of predictive maps via `sf` methods (Pebesma, 2018). The inferential code is written using `C++`, `Rcpp`, `RcppArmadillo` and `OpenBLAS` to make efficient use of multi-CPU hardware architectures.

4.4 Bayesian Inference Using Markov Chain Monte Carlo

In this section we describe a Metropolis-within-Gibbs algorithm for Bayesian inference with the spatial item factor analysis model proposed in Section 4.2. We first present the Bayesian formulation of our model in Section 4.4.1; then in Section 4.4.2, we provide details of the prior specifications; lastly, we conclude by explaining the sampling scheme for the parameters and auxiliary variables in Section 4.4.3.

4.4.1 Bayesian Spatial Item Factor Analysis Model

As illustrated in Figure 4.1, we factorised the joint likelihood in an natural way into four levels. The first three levels are: the data level $\Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right)$, the auxiliary variable level $\Pr\left(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{a}\right)$, and the latent factor level $\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right)$. For our Bayesian model, we add an additional level for the prior distribution of the parameters $\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}$ and $\boldsymbol{R}_v$. The posterior distribution of the model is

$$\begin{aligned}\Pr\left(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v \mid \boldsymbol{y}_{obs}\right) \propto \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right) \Pr\left(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{a}\right) \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right) \\ \Pr\left(\boldsymbol{c}\right) \Pr\left(\boldsymbol{a}\right) \Pr\left(\boldsymbol{\beta}\right) \Pr\left(\boldsymbol{T}\right) \Pr\left(\boldsymbol{\phi}\right) \Pr\left(\boldsymbol{R}_v\right).\end{aligned} \tag{4.26}$$

This choice of factorisation allows us to take advantage of conjugacy for some parameters and also marginalise terms that may lead to slow convergence/mixing e.g. the multivariate Gaussian process $\boldsymbol{w}$ and the multivariate residual term $\boldsymbol{v}$.

4.4.2 Priors

We assume Gaussian distributions for the easiness, discrimination and fixed effects parameters:

$$\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_c), \qquad \boldsymbol{a}_j \sim \mathcal{N}(\boldsymbol{\mu}_{a_j}, \Sigma_{a_j}), \qquad \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_\beta), \tag{4.27}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\mu}_{a_j}$ are the mean parameters, and $\boldsymbol{\Sigma}_c$, $\boldsymbol{\Sigma}_{a_j}$ and $\boldsymbol{\Sigma}_\beta$ are diagonal covariance matrices.

With respect to the m-dimensional Gaussian process $\boldsymbol{w}^*(s)$, we assume that the associated parameters have a log-normal distribution,

$$\text{vec}^*(\boldsymbol{T}) \sim \mathcal{LN}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T), \qquad\qquad \boldsymbol{\phi} \sim \mathcal{LN}(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi), \tag{4.28}$$

where $\text{vec}^*(\boldsymbol{T})$ is a vector of the non-zero values of $\boldsymbol{T}$, $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_\phi$ are the mean parameters and $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_\phi$ are diagonal covariance matrices of the log-transformation of the parameters.

Finally, as proposed in Lewandowski et al. (2009, Section 3) we use an LKJ distribution for the correlation matrix $\boldsymbol{R}_v$ of the multivariate residual term, which is defined as:

$$\Pr(\boldsymbol{R}_v) \propto \det(\boldsymbol{R}_v)^{\eta-1}. \tag{4.29}$$

Here, η is the shape parameter of the LKJ distribution. If $\eta = 1$, the density is uniform; for bigger values $\eta > 1$, the mode is a identity matrix; and band diagonal matrices are more likely when $0 < \eta < 1$.

Bayesian inference can be sensitive to the choice of hyperparameters for small sample sizes on the prior distributions described above; however, this is less highlighted in factor models due to the high number of observations nq. In our experience, inference does not vary drastically for the prior distribution of the easiness, discrimination and fixed parameters as long as reasonable hyperparameters are defined. More careful specification is needed for the scale parameters of the Gaussian processes. This can be achieved by using the maximum spatial distance between observations to define more informative prior distributions for these parameters.

4.4.3 Sampling Scheme

Samples from the posterior distribution (Equation 4.26) are drawn using blocked Gibbs sampling where possible. In cases where the conditional posterior distribution is not available analytically, we use Metropolis Hastings to update parameters, details below.

4.4.3.1 Auxiliary variables

Recall from above that we introduced a distinction between the observed variables $\boldsymbol{Y}_{obs}$ and the set that could not been observed $\boldsymbol{Y}_{mis}$. In a similar way, we divide the associated auxiliary variables into two groups, $\boldsymbol{Z}_{obs}$ and $\boldsymbol{Z}_{mis}$. From Equation 4.18, the joint vector of auxiliary variables $\boldsymbol{Z}$ is normally distributed given the easiness parameters $\boldsymbol{c}$, the discrimination parameters $\boldsymbol{a}$ and the latent factors $\boldsymbol{\theta}$:

$$\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) = \mathcal{N}(\boldsymbol{z} \mid (\boldsymbol{I}_q \otimes \mathbf{1}_n)\boldsymbol{c} + (\boldsymbol{A}^* \otimes \boldsymbol{I}_n)\boldsymbol{\theta}, \boldsymbol{I}_{nq}). \tag{4.30}$$

In the equation above it can be seen that any two elements of $\boldsymbol{Z}$ are conditionally independent given $\boldsymbol{c}$, $\boldsymbol{a}$ and $\boldsymbol{\theta}$ because the covariance is the identity matrix. Using the fact that this joint density can also be written as the product of two marginal densities and that $\boldsymbol{Y}_{obs}$ is conditionally independent of $\boldsymbol{Z}_{mis}$ given $\boldsymbol{Z}_{mis}$, as shown in Appendix B.1, the conditional posterior distribution for the auxiliary variables $\Pr\left(\boldsymbol{z} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right)$ is

$$\Pr\left(\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \propto \Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{z}_{mis} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right). \tag{4.31}$$

Hence, using Equation 4.21, the conditional posterior distribution for $\boldsymbol{Z}_{obs}$ is

$$\Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \propto \Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \prod_{o_{ij}=1} \Pr\left(y_{ij} \mid z_{ij}\right), \tag{4.32}$$

which is a marginal truncated normal distribution obtained from Equation 4.30. Note that $\Pr(y_{ij} \mid z_{ij}) = \mathbb{1}_{(z_{ij}>0)}^{y_{ij}} \mathbb{1}_{(z_{ij}\leq 0)}^{1-y_{ij}}$, where $\mathbb{1}_{(.)}$ is the indicator function. In a similar way, we obtain that the conditional posterior distribution of the auxiliary variables related to the missing data $\boldsymbol{Z}_{mis}$ as

$$\Pr(\boldsymbol{z}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}) \propto \Pr(\boldsymbol{z}_{mis} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}), \tag{4.33}$$

which is a marginal distribution of Equation 4.30. Hence, the only difference between the posterior of both sets of variables is that it is truncated for the $\boldsymbol{Z}_{obs}$ and unrestricted for $\boldsymbol{Z}_{mis}$.

4.4.3.2 Latent Factors

The conditional posterior distribution of the latent abilities is

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\beta}) \propto \Pr(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v), \tag{4.34}$$

where the joint density of the auxiliary variables $\Pr(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta})$ is a Gaussian distribution, given in Equation 4.30, and the density of the latent factors, as defined in Equation 4.17, is also a Gaussian distribution,

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) = \mathcal{N}(\boldsymbol{\theta} \mid (\boldsymbol{I}_m \otimes \boldsymbol{X})\boldsymbol{\beta}, (\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \boldsymbol{D}\boldsymbol{R}_v\boldsymbol{D} \otimes \boldsymbol{I}_n), \tag{4.35}$$

where $\boldsymbol{\Sigma}_w = \oplus_{k=1}^{g} \boldsymbol{\Sigma}_{w_k}$. Hence, the conditional posterior $\Pr(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\beta})$ is defined by the product of two normal densities that leads to a normal density with covariance matrix

$$\boldsymbol{\Sigma}_{\theta|\cdot} = \left((\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n)(\boldsymbol{A}^{*} \otimes \boldsymbol{I}_n) + ((\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \boldsymbol{D}\boldsymbol{R}_v\boldsymbol{D} \otimes \boldsymbol{I}_n)^{-1}\right)^{-1}, \tag{4.36}$$

and mean

$$\begin{aligned}\boldsymbol{\mu}_{\theta|\cdot} = {} & \boldsymbol{\Sigma}_{\theta|\cdot}(\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n)(\boldsymbol{z} - (\boldsymbol{I}_q \otimes \mathbf{1}_n)\boldsymbol{c}) + \\ & \boldsymbol{\Sigma}_{\theta|\cdot}\left((\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \boldsymbol{D}\boldsymbol{R}_v\boldsymbol{D} \otimes \boldsymbol{I}_n\right)^{-1}(\boldsymbol{I}_m \otimes \boldsymbol{X})\boldsymbol{\beta}. \end{aligned} \tag{4.37}$$

4.4.3.3 Fixed effects

For the multivariate fixed effects $\boldsymbol{\beta}$, the conditional posterior

$$\Pr(\boldsymbol{\beta} \mid \boldsymbol{y}_{obs}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{a}) \propto \Pr(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v)\Pr(\boldsymbol{\beta}) \tag{4.38}$$

is given by the product of two normal densities obtained from Equation 4.17 and 4.27,

$$\mathcal{N}(\boldsymbol{\theta} \mid (\boldsymbol{I}_m \otimes \boldsymbol{X})\boldsymbol{\beta}, (\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \boldsymbol{R} \otimes \boldsymbol{I}_n)\mathcal{N}(\boldsymbol{\beta} \mid \mathbf{0}, \boldsymbol{\Sigma}_\beta), \tag{4.39}$$

that also leads to a multivariate normal distribution with covariance matrix

$$\boldsymbol{\Sigma}_{\beta|\cdot} = \left((\boldsymbol{I}_m \otimes \boldsymbol{X})^\intercal\left((\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \boldsymbol{R} \otimes \boldsymbol{I}_n\right)^{-1}(\boldsymbol{I}_n \otimes \boldsymbol{X}) + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1}, \tag{4.40}$$

and mean

$$\boldsymbol{\mu}_{\beta|\cdot} = \boldsymbol{\Sigma}_{\beta|\cdot}(\boldsymbol{I}_m \otimes \boldsymbol{X}^\intercal)\left((\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \boldsymbol{R} \otimes \boldsymbol{I}_n\right)^{-1}\boldsymbol{\theta}. \tag{4.41}$$

4.4.3.4 Easiness parameters

The conditional posterior distribution of the easiness parameters $\boldsymbol{c}$,

$$\Pr(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{\theta}) \propto \Pr(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{a})\Pr(\boldsymbol{c}), \tag{4.42}$$

is also the product of two normal densities obtained from Equation 4.30 and 4.27,

$$\Pr\left(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{\theta}\right) \propto \mathcal{N}(\boldsymbol{z} \mid (\boldsymbol{I}_q \otimes \boldsymbol{1}_n)\boldsymbol{c} + (\boldsymbol{A}^* \otimes \boldsymbol{I}_n)\boldsymbol{\theta}, \boldsymbol{I}_{nq})\mathcal{N}(\boldsymbol{c} \mid \boldsymbol{0}, \boldsymbol{\Sigma}_c), \tag{4.43}$$

leading to a multivariate normal density with covariance matrix

$$\boldsymbol{\Sigma}_{c|\cdot} = ((\boldsymbol{I}_q \otimes \boldsymbol{1}_n)^\intercal(\boldsymbol{I}_q \otimes \boldsymbol{1}_n) + \boldsymbol{\Sigma}_c^{-1})^{-1} = (\mathrm{diag}(\boldsymbol{\Sigma}_c)^{-1} + n)^{-1}, \tag{4.44}$$

and mean

$$\boldsymbol{\mu}_{c|\cdot} = \boldsymbol{\Sigma}_{c|\cdot}(\boldsymbol{I}_q \otimes \boldsymbol{1}_n^\intercal)(\boldsymbol{z} - (\boldsymbol{A}^* \otimes \boldsymbol{I}_n)\boldsymbol{\theta}). \tag{4.45}$$

4.4.3.5 Discrimination parameters

Due to the structure of our hierarchical model in Section 4.4.1, the conditional posterior distribution of the discrimination parameters,

$$\Pr\left(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\theta}\right) \propto \Pr\left(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{a}\right) \Pr\left(\boldsymbol{a}\right), \tag{4.46}$$

is determined by the product of two Gaussian densities obtained from Equation 4.30 and 4.27,

$$\mathcal{N}(\boldsymbol{z} \mid (\boldsymbol{I}_q \otimes \boldsymbol{1}_n)\boldsymbol{c} + (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}^*)\boldsymbol{u} + (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}^*)\boldsymbol{L}\boldsymbol{a}, \boldsymbol{I}_n)\mathcal{N}(\boldsymbol{a} \mid \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \tag{4.47}$$

which, similar to previous parameters, leads to a Gaussian density with covariance matrix

$$\boldsymbol{\Sigma}_{a|\cdot} = (\boldsymbol{L}^\intercal(\boldsymbol{I}_q \otimes \boldsymbol{\Theta}^{*\intercal}\boldsymbol{\Theta}^*)\boldsymbol{L} + \boldsymbol{\Sigma}_a^{-1})^{-1}, \tag{4.48}$$

and mean

$$\boldsymbol{\mu}_{a|\cdot} = \boldsymbol{\Sigma}_{a|\cdot}\boldsymbol{L}^\intercal(\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_j^{*\intercal})(\boldsymbol{z} - (\boldsymbol{I}_q \otimes \boldsymbol{1}_n)\boldsymbol{c} - (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}^*)\boldsymbol{u}) + \boldsymbol{\Sigma}_{a|\cdot}\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a. \tag{4.49}$$

4.4.3.6 Covariance parameters

Unlike the previous parameters, the parameters of the multivariate Gaussian process $\boldsymbol{w}^*(s)$ and the multivariate residual term $\boldsymbol{v}(s)$ can not be directly sampled from their conditional posterior density as they are not available analytically. However, this density can be defined up to a constant of proportionality,

$$\Pr\left(\text{vec}^*(\boldsymbol{T}), \boldsymbol{\phi}, \boldsymbol{R}_v \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) \propto \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right) \Pr\left(\boldsymbol{T}\right) \Pr\left(\boldsymbol{\phi}\right) \Pr\left(\boldsymbol{R}_v\right). \tag{4.50}$$

In order to obtain an MCMC chain that mixes over the real line, we work with $\log(\boldsymbol{\phi})$ instead of $\boldsymbol{\phi}$ and $\log(\text{vec}^*(\boldsymbol{T}))$ instead of $\text{vec}^*(\boldsymbol{T})$. For the correlation $\boldsymbol{R}_v$, we use canonical partial correlation, transforming to a set of free parameters $\boldsymbol{\nu} \in \mathbb{R}^{m(m-1)/2}$, see Lewandowski et al. (2009) for further details.

We use an adaptive random-walk Metropolis Hastings algorithm to sample from this part of the posterior distribution. The covariance matrix of the proposal, is adapted to reach a fixed acceptance probability (e.g. 0.234). More specifically, we implemented algorithm 4 proposed in Andrieu and Thoms (2008) using a deterministic adaptive sequence $\gamma_i = C/i^{\alpha}$ for $\alpha \in ((1+\lambda)^{-1}, 1]$, where $\lambda > 0$. In the tests we have run and in our food insecurity application, this algorithm and choice of parameters performs well (see details below for our choice of C and α).

4.4.4 Scaling Samples for Interpretation

In Section 4.2.4, we saw how restricting the standard deviations of the multivariate residual term $\boldsymbol{v}(s)$ is necessary to make our model identifiable (Equation 4.8). However, we can not ensure that the latent factors will be on the same scale, which leads to a loss of interpretation of the discrimination parameters $\boldsymbol{a}_j$. As proposed in the same section, after the samples of the MCMC have obtained, we can transform the parameters in order to obtain latent factors with expected variance equal to 1 to solve this problem. We can then obtain the matrix $\boldsymbol{Q}$ of Equation 4.8 by filling the diagonal elements with the expected variances of the samples of the latent factors $\boldsymbol{\theta}(s)$. We then make the following transformations

$$\boldsymbol{a}_j \leftarrow \boldsymbol{Q}\boldsymbol{a}_j, \quad \boldsymbol{\theta}_i \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{\theta}_i, \quad \boldsymbol{B} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{B}, \quad \boldsymbol{T} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{T}, \quad \boldsymbol{D} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{D}; \quad (4.51)$$

the correct interpretation of the parameters is then recovered.

4.4.5 Model Selection Using the Deviance Information Criterion

Bayesian model selection for the spatial item factor analysis can be done by using any of the information criteria normally applied in Bayesian modelling; here we focus on the *deviance information criterion* (DIC) proposed by Spiegelhalter et al. (2002), though note compteting alternatives such as the Watanabe-Akaike Information Criterion (WAIC). A Bayesian version of the *Akaike information criterion*, the DIC encapsulates the trade-off between goodness of fit and model complexity. This complexity, measured through the effective number of parameters, is determined by the difference between the mean of the deviance and the deviance of the mean,

$$p_D = \overline{D(\boldsymbol{\alpha})} - D(\bar{\boldsymbol{\alpha}}). \quad (4.52)$$

The deviance in our case is given by

$$D(\boldsymbol{\alpha}) = -2\log\{\Pr(\boldsymbol{y} \mid \boldsymbol{\alpha})\} + 2\log\{\Pr(\boldsymbol{y} \mid \mu(\boldsymbol{\alpha}) = \boldsymbol{y})\}, \quad (4.53)$$

where $\Pr(\boldsymbol{y} \mid \mu(\boldsymbol{\alpha}) = \boldsymbol{y})$ is the likelihood associated with a saturated model. The DIC can then be calculated as:

$$DIC = \overline{D(\boldsymbol{\alpha})} + p_D, \quad (4.54)$$

where models with a lower DIC are preferred.

In order to be able to calculate this quantity for our model, we require the density function of the responses $\boldsymbol{Y}$ given all the parameters of the model, which is expressed as

$$\log(\Pr(\boldsymbol{y} \mid \boldsymbol{\alpha})) = \sum_{o_{ij}=1} \left(y_{ij} \log(\Phi(c_j + \boldsymbol{a}_j^\intercal \boldsymbol{\theta}_j)) + (1 - y_{ij}) \log(1 - \Phi(c_j + \boldsymbol{a}_j^\intercal \boldsymbol{\theta}_j))\right), \tag{4.55}$$

where o_{ij} is a binary variable taking value equal to one when the variable Y_{ij} has been observed and zero otherwise.

4.5 Prediction of Latent Factors

In this section, our interest is on the spatial prediction of the latent factors $\tilde{\boldsymbol{\theta}}$ at a set of locations that we have not observed data, $\tilde{\boldsymbol{s}}$. As is customary, we obtain the predictive distribution by integrating out the parameters of the model from the joint density of $\tilde{\boldsymbol{\theta}}$ and the parameters,

$$\Pr\left(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right) =$$
$$\int \Pr\left(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{R}_v, \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right) \Pr\left(\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{R}_v \mid \boldsymbol{y}, \boldsymbol{X}\right) d\boldsymbol{\theta} d\boldsymbol{B} d\boldsymbol{\sigma}^2 d\boldsymbol{\phi} d\boldsymbol{R}_v.$$

Note that a vectorized version of Equation 4.4 can be expressed as

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{I}_m \otimes \tilde{\boldsymbol{X}})\boldsymbol{\beta} + (\boldsymbol{T} \otimes \boldsymbol{I}_{\tilde{n}})\tilde{\boldsymbol{w}} + \tilde{\boldsymbol{v}}. \tag{4.56}$$

Under these expressions, it can be shown that $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ are normally distributed with parameters

$$\mathbb{E}[\boldsymbol{\theta}] = (\boldsymbol{I}_m \otimes \boldsymbol{X})\boldsymbol{\beta}, \qquad \mathbb{V}[\boldsymbol{\theta}] = (\boldsymbol{T} \otimes \boldsymbol{I}_n)\mathbb{V}[\boldsymbol{w}](\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_n) + \mathbb{V}[\boldsymbol{v}] \tag{4.57}$$

$$\mathbb{E}\left[\tilde{\boldsymbol{\theta}}\right] = (\boldsymbol{I}_m \otimes \tilde{\boldsymbol{X}})\boldsymbol{\beta}, \qquad \mathbb{V}\left[\tilde{\boldsymbol{\theta}}\right] = (\boldsymbol{T} \otimes \boldsymbol{I}_{\tilde{n}})\mathbb{V}[\tilde{\boldsymbol{w}}](\boldsymbol{T}^\intercal \otimes \boldsymbol{I}_{\tilde{n}}) + \mathbb{V}[\tilde{\boldsymbol{v}}]. \tag{4.58}$$

Furthermore, the cross-covariance can be obtained as

$$\begin{aligned}\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\right] &= \mathrm{Cov}\left[(\boldsymbol{T} \otimes \boldsymbol{I}_{\tilde{n}})\tilde{\boldsymbol{w}} + \tilde{\boldsymbol{v}}, (\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{w} + \boldsymbol{v}\right] \\ &= \mathrm{Cov}\left[(\boldsymbol{T} \otimes \boldsymbol{I}_{\tilde{n}})\tilde{\boldsymbol{w}}, (\boldsymbol{T} \otimes \boldsymbol{I}_n)\boldsymbol{w}\right] \\ &= (\boldsymbol{T} \otimes \boldsymbol{I}_{\tilde{n}})\mathrm{Cov}\left[\tilde{\boldsymbol{w}}, \boldsymbol{w}\right](\boldsymbol{T}^{\intercal} \otimes \boldsymbol{I}_n), \qquad (4.59)\end{aligned}$$

where $\mathrm{Cov}\left[\tilde{\boldsymbol{w}}, \boldsymbol{w}\right]$ is a block diagonal matrix as both $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$ are multivariate independent Gaussian process, see Section 4.2.3. Hence, the conditional distribution of $\tilde{\boldsymbol{\theta}}$ is $\Pr\left(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{R}_v, \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right)$, a normal distribution with mean and variance

$$\mathbb{E}\left[\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}\right] = \mathbb{E}\left[\tilde{\boldsymbol{\theta}}\right] + \mathrm{Cov}\left[\tilde{\theta}, \theta\right] \mathbb{V}\left[\theta\right]^{-1} (\boldsymbol{\theta} - \mathbb{E}\left[\boldsymbol{\theta}\right])) \qquad (4.60)$$

$$\mathbb{V}\left[\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}\right] = \mathbb{V}\left[\tilde{\boldsymbol{\theta}}\right] - \mathrm{Cov}\left[\tilde{\theta}, \theta\right] \mathbb{V}\left[\theta\right]^{-1} \mathrm{Cov}\left[\theta, \tilde{\theta}\right]. \qquad (4.61)$$

Predictions are obtained by generating $\tilde{\boldsymbol{\theta}}$ from this conditional distribution for a set of samples $\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{R}_v$ obtained from the joint posterior via MCMC.

4.6 Case of Study: Predicting Food Insecurity in an Urban Centre in Brazilian Amazonia

In this section, we detail results from our motivating application: modelling and prediction of food insecurity in a remote urban centre, Ipixuna, in the Amazonas state, Brazil.

"*Food security* [is] a situation that exists when all people, at all times, have physical, social and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life" (FAO, 2003, pp 313). *Food insecurity* describes the opposite situation, in which individual or household access to sufficient, safe and nutritious food is not a guarantee (National Research Council, 2006). For policy makers, understanding the level of food insecurity in a region is crucial in the planning of interventions designed to foster development and improve the quality of life for these populations. Therefore, being able to understand the spatial structure of food insecurity and to be able to map (i.e. predict) it is highly relevant for both fundamental science and policy makers alike.

Ipixuna, shown in Figure 4.2, is a 'jungle town' located on the banks of the River Juruá; it is unconnected to the Brazilian road network, and is several thousand kilometers of upstream boat travel from the Amazonas state capital, Manaus. Being remote and 'roadless', Ipixuna exhibits very high social vulnerability and it is also prone to extreme hydro-climatic events such as floods and droughts, which pose a serious risk of harm to the local population (Parry et al., 2018).

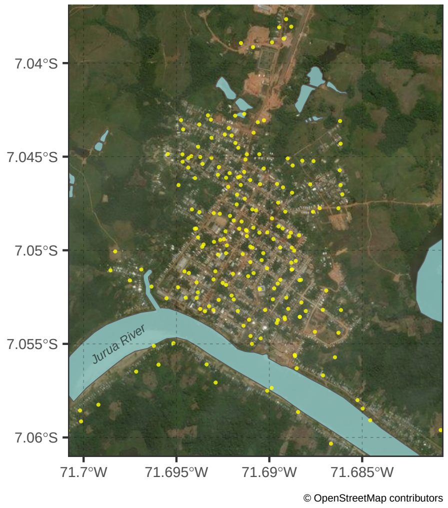

Figure 4.2: Spatial distribution of the sampled households (yellow points) in the urban area of Ipixuna. Note the points have been jittered.

4.6.1 Data description

Our data were collected in August 2015 (low-water dry season) and March 2016 (high water rainy season) with 200 randomly sampled households in total. The spatial distribution of these samples can be seen in Figure 4.2; these points have been jittered for privacy reasons: they just give a general sense of where samples were taken from. Following our analysis, we presented and discussed our results with local authorities and citizens to explain how food insecurity mapping can be done and to encourage the development of a food insecurity early warning system; we also conducted site visits to neighbourhoods identified as particularly vulnerable to observe their characteristics. These interactions were highly beneficial in interpreting the results of our spatial models below.

The questionnaire contained items initially validated by the United States Department of Agriculture and additional items that are relevant in the context of Brazilian Amazonìa, the full questionnaire is available at: `https://www.lancaster.ac.uk/staff/taylorb1/food-insecurity-questions.html`. In total the questionnaire contains 18 items relating to food insecurity. Items in Section A of our questionnaire referred to the household as a whole, those in Section B referred to adults only, Section C concerned children and Section D included items related to the regional context of our study. The regionally-specific questions in Section D were designed to measure similar aspects as contained in the general scale, but measured through common coping strategies employed in this locality.

The items with higher endorsement probability were numbers 15, 3, 1, 18, and 2, see Table 4.1. In the present context, endorsement simply means 'answering with an affirmative'. This indicates that it is common that Ipixuna citizens obtain credit for eating, eat few food types, are worried that food will end, reduce meat or fish consumption, or run out of food. Of the 200 surveyed households, 25 did not have children and this led to missing data on the 6 items of associated with food insecurity in children, see Section D in Table 4.1. This is treated as missing because it is desirable to obtain a joint model for all the population.

4.6.2 Confirmatory item factor analysis

Before undertaking a confirmatory item factor analysis (CIFA), we performed an exploratory item factor analysis (EIFA) in order to choose the number of dimensions and identify which items should be related to each factor. We compared models whose dimensions ranged from one to six, and selected a model with 3

Table 4.1: Summary of the food insecurity items: i) the number of missing values (#NA) and the proportion of endorsement (π) are shown for the descriptive analysis, ii) the posterior median of the discrimination parameters $\{\hat{\boldsymbol{A}}_{\cdot 1}, \hat{\boldsymbol{A}}_{\cdot 2}, \hat{\boldsymbol{A}}_{\cdot 3}\}$ are shown for the confirmatory factor analysis (CIFA), and iii) the posterior median of the discrimination and easiness parameters $\{\hat{\boldsymbol{A}}_{\cdot 1}, \hat{\boldsymbol{A}}_{\cdot 2}, \hat{\boldsymbol{A}}_{\cdot 3}, \hat{\boldsymbol{c}}\}$ are shown for the spatial item factor analysis (SPIFA).

Item	**Section**	**Question**	**Descriptive**		**CIFA**			**SPIFA**			
			#NA	π	$\hat{\boldsymbol{A}}_{\cdot 1}$	$\hat{\boldsymbol{A}}_{\cdot 2}$	$\hat{\boldsymbol{A}}_{\cdot 3}$	$\hat{\boldsymbol{A}}_{\cdot 1}$	$\hat{\boldsymbol{A}}_{\cdot 2}$	$\hat{\boldsymbol{A}}_{\cdot 3}$	$\hat{\boldsymbol{c}}$
1	A	worry that food ends	0	0.56	·	1.62	·	·	1.79	·	0.44
2		run out of food	0	0.52	·	·	1.49	·	·	1.87	0.21
3		ate few food types	0	0.64	1.68	·	·	1.83	·	·	1.47
4	B	skip a meal	0	0.30	1.48	·	1.01	1.91	·	1.00	-0.62
5		ate less than required	0	0.41	0.88	·	1.77	1.39	·	1.50	-0.07
6		hungry but did not eat	0	0.24	1.26	·	1.52	1.83	·	1.51	-1.44
7		one meal per day	0	0.26	1.57	·	·	1.82	·	·	-0.53
8	C	ate few food types	25	0.49	1.69	·	·	1.90	·	·	0.60
9		ate less than required	25	0.31	1.89	·	·	2.24	·	·	-0.34
10		decreases food quantity	25	0.36	2.16	·	·	2.51	·	·	-0.03
11		skip a meal	25	0.23	2.01	·	·	2.54	·	·	-1.06
12		hungry but did not eat	25	0.20	2.11	·	·	2.56	·	·	-1.32
13		one meal per day	25	0.18	1.95	·	·	2.45	·	·	-1.52
14	D	food just with farinha	0	0.17	0.34	·	1.24	0.63	·	1.28	-1.60
15		credit for eating	0	0.68	·	0.72	·	·	0.79	·	0.62
16		borrowed food	0	0.14	·	1.42	·	·	1.61	·	-1.89
17		meal at neighbors	0	0.17	·	0.97	·	·	1.01	·	-1.24
18		reduced meat or fish	0	0.54	1.28	·	·	1.43	·	·	0.76

dimensions because a likelihood ratio test indicated no significant improvement for higher dimensions (p-value 0.594). We applied a varimax rotation to try to obtain independent factors.

For the structure of the CIFA model, we decided to include only those items with a discrimination parameter greater than 0.5 in the EIFA model. In our CIFA this leads to: the first factor being explained by items 3–14 and 18; the second, by items 1 and 15–17; and the third by items 2, 4–6 and 14. To perform Bayesian inference, we used standard normal priors for the easiness parameters c_j; standard normal priors for the discrimination parameters with exception of $\{A_{11,1}, A_{13,1} A_{16,2}, A_{14,3}\}$ for which we used normal priors with mean $\mu = 1$ and standard deviation $\sigma = 0.45$; and an LKJ prior distribution with hyper-parameter $\eta = 1.5$ for the correlation matrix of the latent factors. The adaptive MCMC scheme had parameters $C = 0.7$ and $\alpha = 0.8$ with target acceptance probability of 0.234. We ran the Metropolis-within-Gibbs algorithm for 100,000 iterations discarding the first 50,000 iterations and storing 1 in 10 of the remaining iterations. Convergence was assessed visually; stationarity was observed from around iteration

10,000 of the burn-in period.

The posterior median of the discrimination parameters $\{\hat{\boldsymbol{A}}_{\cdot 1}, \hat{\boldsymbol{A}}_{\cdot 2}, \hat{\boldsymbol{A}}_{\cdot 3}\}$ of the CIFA model is shown in Table 4.1. These values show that questions related to reduction of quality and quantity of food in the diet of children, items 10–12, are the top three most important items for the first factor. The second factor includes three items relating to Amazonian coping strategies (15–17) and one concerning anxiety (1). Note that using credit (15), borrowing food (16) or relying on neighbours for meals (17) are likely sources of anxiety in their own right. Finally, the third factor is related mainly to the reduction in quantity of food (2 and 4–6) and one item associated with substitution of normal foods with only toasted manioc flour, a staple carbohydrate in low-income households (14).

In order to evaluate the spatial correlation in the obtained factors, we use the empirical variogram: see Figure 4.3. This exploratory tool for determining the extent and form of (spatial) correlation is defined as a function of the distance u,

$$\hat{\gamma}(u) = \frac{1}{2}\hat{\mathbb{E}}\left[(w(s) - w(s+u))^2\right].$$

The initial increasing behaviour of the variogram, mainly, observed in the first and third factors is evidence for spatial correlation in these dimensions of food insecurity.

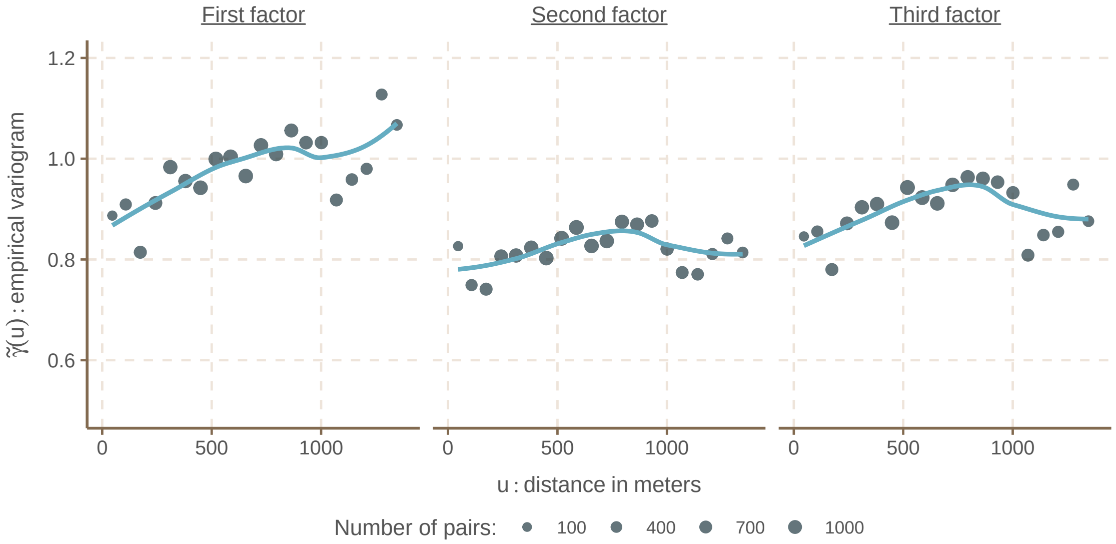

Figure 4.3: Empirical variogram $\tilde{\gamma}(u)$ for each latent factor: the points represent the empirical values and the lines the smoothed version of the empirical variogram.

4.6.3 Spatial confirmatory item factor analysis

We placed the same restrictions on the discrimination parameters for the spatial item factor as we did for the confirmatory item factor analysis. Based on the empirical variograms shown in Figure 4.3, we proposed three models; model 1 includes a Gaussian process in the first latent factor (SPIFA I), model 2 includes Gaussian processes in the first and third factor (SPIFA II), and model 3 includes Gaussian processes in the three factors (SPIFA III). We used the exponential correlation function to model the spatial structure of each of the Gaussian processes in our model. Spatial predictors were not included in our model because these are insufficiently finely resolved in our study area. For instance, there are only 8 census sectors (from the 2010 demographic census by the Brazilian Institute for Geography and Statistics (IBGE)) covering Ipixuna - in the future, we are planning a larger scale analysis in which spatial predictors *will* be included; our software package is already able to handle this case.

We used the same prior specifications as in the CIFA model for the easiness parameters c_j, discrimination parameters A_{jk} and correlation matrix $\boldsymbol{R}_v$. In addition, we used log-normal priors $\mathcal{LN}(\log(160), 0.3)$, $\mathcal{LN}(\log(80), 0.3)$ and

$\mathcal{LN}(\log(80), 0.3)$ for the scale parameters $\{\phi_1, \phi_2, \phi_3\}$ of the Gaussian processes in factor 1, 2 and 3 respectively; and the log-normal prior distribution $\mathcal{LN}(\log(0.4), 0.4)$ for all the free elements of $\boldsymbol{T}$. The adaptive MCMC scheme had parameters $C = 0.7$ and $\alpha = 0.8$ with target acceptance probability of 0.234. We ran the Metropolis-within-Gibbs algorithm for 300,000 iterations discarding the first 150,000 iterations and storing 1 in every 150 iterations. Convergence was again assessed visually with stationarity occurring around the iteration 40,000 of the burnin period. Usually, mixing is slower for the elements of $\boldsymbol{T}$ and the scale parameter $\boldsymbol{\phi}$ of the Gaussian processes.

We compared these three models using the Deviance Information Criterion (DIC), see Table 4.2. We can see that the classical confirmatory model (CIFA) has lowest effective number of parameters (328.14); this model has independent random effects only. In contrast, the spatial models include both independent and spatial random effects as explained in Section 4.2. The DIC for the three spatial models is lower than that for CIFA, hence by this measure, it is statistically advantageous in terms of model fit to allow the factors to be spatially correlated. Of the three spatial models, SPIFA III, the model including Gaussian processes in all three factors, has the best performance in terms of DIC (2195.529). Hence in the remainder of this section, we focus on the results from this model. The trace-plots for the elements of $\boldsymbol{T}$ and the scale parameters $\boldsymbol{\phi}$ of the Gaussian processes for our selected model can be seen on Figure 4.10 and 4.11 respectively. Additional (representative) trace-plots for random selected parameters can be seen in Appendix D.

Table 4.2: Deviance Information Criteria (DIC) for the Proposed Models: without spatial correlation (CIFA), with spatial correlation in factor 1 (SPIFA I), with spatial correlation in factor 1 and 3 (SPIFA II) and with spatial correlation in all factors (SPIFA III).

Model	**Diagnostics**		
	Posterior Mean Deviance	Effective Number of Parameters	DIC
CIFA	1894.228	328.1377	2222.365
SPIFA I	1865.85	334.228	2200.078
SPIFA II	1862.156	339.2756	2201.432
SPIFA III	1856.354	339.1752	2195.529

The posterior medians of the discrimination parameters $\{\hat{\boldsymbol{A}}_{\cdot 1}, \hat{\boldsymbol{A}}_{\cdot 2}, \hat{\boldsymbol{A}}_{\cdot 3}\}$ for the selected model (SPIFA III) are shown in Table 4.1 under the column of SPIFA. We can see that the median of the obtained parameters have a broadly a similar structure as for the CIFA model, so their interpretation is as discussed in the previous section; notice that most of the discrimination parameters are higher for the SPIFA model. The last column of Table 4.1 shows the posterior median of the easiness parameters $\hat{\boldsymbol{c}}$; note the items with high easiness are those most frequently answered with an affirmative ('endorsed'). This column shows that eating few food types (item 3), obtaining credit for eating (item 15) and worrying that food will end (item 1) are the most common behaviours in the population of Ipixuna. Borrowing food (item 16), eating food just with farinha (item 14), having children with one meal per day (item 13) and feeling hungry but do not eat (item 6 and 12) are less common.

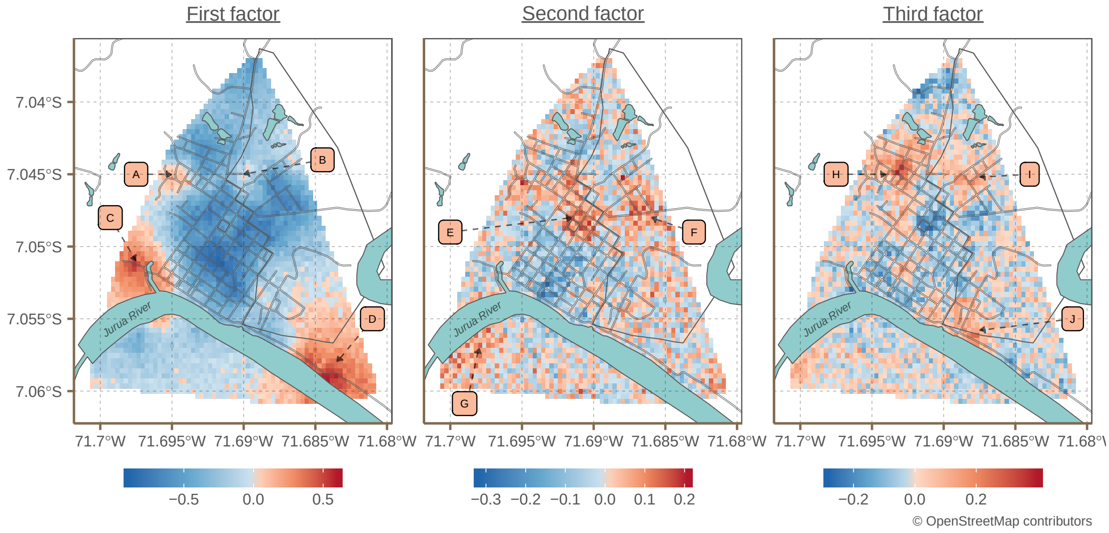

Figure 4.4: Median of the predicted latent factors of food insecurity.

Figure 4.4 shows the posterior median of each of the three factors over our study area. The left plot shows that the first factor has a strong spatial structure; the respective posterior median of the standard deviation and scale parameters of the associated Gaussian process are $\hat{T}_{1,1} = 0.465$ and $\hat{\phi}_1 = 214$ meters. Examining the middle plot, for the second factor, it can be seen that the spatial structure is not as strong as the first factor. The respective parameters of the associated Gaussian process have posterior medians $\hat{T}_{2,2} = 0.205$ and $\hat{\phi}_2 = 83.6$. The right hand plot, referring to the third factor, shows moderate spatial structure with similar median posterior estimates as the second factor: $\hat{T}_{3,3} = 0.287$ and $\hat{\phi}_3 = 78.8$.

Examining the obtained maps of food insecurity for the first factor, we can see there are lower levels of food insecurity around the center of the study area and more severe food insecurity around the locations A (71.695° W, 7.045° S), B (71.69° W, 7.045° S), C (71.698° W, 7.052° S) and D (71.685° W, 7.06° S). In this city, location C refers to the flood-prone, politically marginalized and poor neighbourhood of Turrufão. Housing is on stilts, and there is no sanitation and poor provision of public services. Point A refers to the flood-prone, poorest part of another marginalized neighbourhood, Multirão Novo. These households are also located at the edge of a large stream called Igarapé Turrufão. The area B is a relatively new neighbourhood, Morro dos Encanados, which is poor and prone to surface flooding from rainfall. Area D is at the edge of the River Juruá and is highly flood-prone. It is also a relatively new and very poor neighbourhood called Bairro da Várzea. Hence, the common characteristic among these locations is that they are poor, marginalized and mostly flood-prone neighbourhoods on the peri-urban fringe. Most of the heads of households in these neighbourhoods are rural-urban migrants (often relatively recent), and many of their livelihoods are still based in rural areas. These relatively large areas are capturing indications of relatively severe food insecurity, yet without apparent anxiety and a distinct absence of some coping strategies: borrowing food, eating in other households or accessing credit.

With respect to the map of the second factor, we can identify higher levels of food insecurity around location E (71.69° W, 7.048° S), F (71.686° W, 7.048° S) and G (71.698° W, 7.058° S). Location E covers a large and older area of the town, covering proportions of two neighbourhoods: Bairro do Cemetério and Multirão Velho. They are not flood prone and not so marginalized and poor, though certainly not wealthy. It is plausible that this factor captures more moderate food insecurity and coping strategies associated with higher levels of horizontal social capital and access to credit. Area F is the larger part of Morro dos Encanados (see above). Area G is another flood-prone peri-urban neighbourhood on the other side of the River Jurua, by the name of Bairro da Ressaca.

In the map of the third factor of food insecurity, we can see areas of severe food insecurity around H (71.693° W, 7.045° S), I (71.687° W, 7.057° S) and J (71.688° W, 7.056° S). Area H covers the border between two poor, peri-urban neighbourhoods: Bairro da Liberdade and Multirão Novo. Point I is an area of Morro dos Encanados. Area J is the beginning of the peri-urban, flood-prone region and the poor area, Bairro da Várzea.

While spatial plots of the posterior median tell us where food insecurity is high and low on average, we ideally also need to take into account the spatial sampling design, since we will be better able to estimate food insecurity where we have more data points. One such measure are exceedance probabilities: the posterior probability that the factor exceeds a given threshold; this takes into account both the mean and the variance of the factor at each location. In figure Figure 4.5, we show the probability that the latent factor is greater than zero in order to identify areas over and below average. It so happens that in the present case, the pattern of high and low food insecure areas remain similar with respect to the maps of the median for each factor.

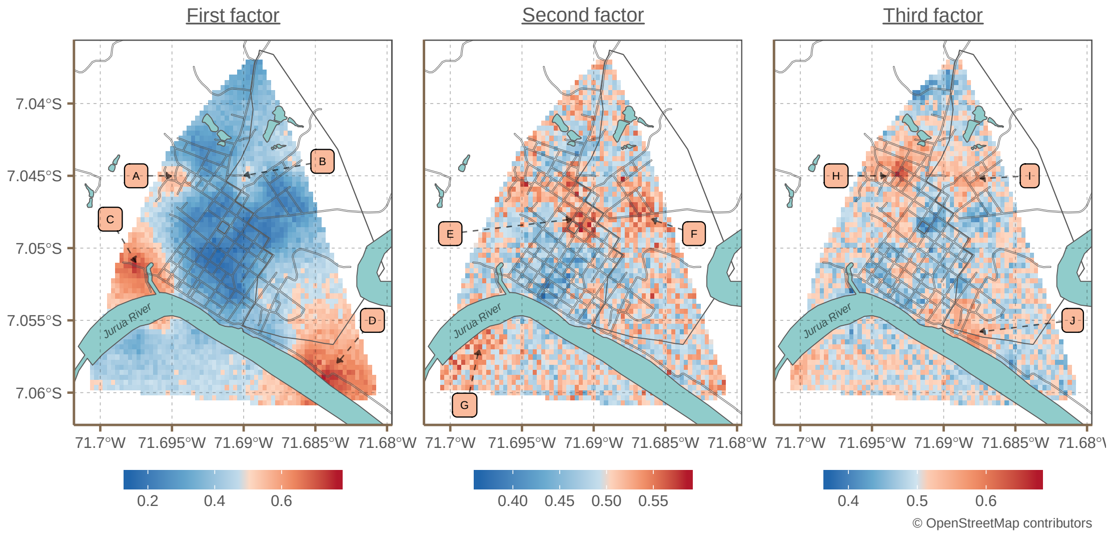

Figure 4.5: Exceedance probabilities $\Pr(\theta_k(s) > 0)$ of the latent factors of food insecurity.

Identifying these areas of high (and also low) food insecurity is of relevance for future research in this area, for example: exploring the social and environmental (e.g. household flood risk due to elevation) determinants of vulnerability to food insecurity. Understanding the spatial-variation of food insecurity at local (e.g. neighbourhood or street) scales will also allow us to continue our dialogue with local government and other stakeholders around which are the priority areas for intervention and what type(s) of intervention should be deployed in order to reduce the risk of food insecurity.

4.7 Discussion

In this work we have developed a new extension of item factor analysis to the spatial domain, where the latent factors are allowed to be spatially correlated. Our model allows for the inclusion of predictors to help explain the variability of the factors. These developments allow us to make prediction of the latent factors at unobserved locations as shown in our case of study of food insecurity in the Brazilian Amazon. We solved the issues of identifiability and interpretability by employing a similar strategy as for confirmatory item factor analysis in order to obtain an identifiable model, and by standardizing the resulting factors after inference. Our model has been successfully implemented in an open source R package.

Since item factor analysis is used across such a wide range of scientific disciplines, we believe that our model and method of inference will be important for generating and investigating many new hypotheses. For instance, it could be used to model socio-economic status.

Computationally, our model is more efficient compared to a model where the spatial structure is used at the level of the response variables. By including spatial structure at the level of the factors, we reduce the computational cost from $\mathcal{O}(q^3n^3)$ to $\mathcal{O}(m^3n^3)$ where the number of items (q) is usually much greater than the number of latent factors (m). For larger datasets, we can reduce the computational burden by using alternatives to the Gaussian process. For example, we could use spatial basis functions (Fahrmeir et al., 2004), nearest neighbour Gaussian processes (Datta et al., 2016) or stochastic partial differential equations (Lindgren et al., 2011) to reduce the cost. This is not so obvious because some of the nice properties of these processes can be lost when working with multivariate models.

Our model can be extended to the spatio-temporal domain, though again with increased computational expense, depending on the chosen parameterisation of the spatio-temporal correlation. A more complex extension of our model would allow the use of binary, ordinal and continuous items and would also allow predictors to be related in a non-linear way to the latent factors. These extensions would allow us to answer more complex research questions and would also improve prediction of the latent factors, see Appendix E. Extensions to other distributional assumptions (e.g. heavier tailed densities) are also possible if one desires to trade the convenience conjugacy for realism; the Gaussian model fitted our particular dataset well.

Acknowledgements

This research was funded by a Future Research Leaders Fellowship to LP (ES/K010018/1), the Newton Fund/FAPEAM (ES/M011542/1), Brazil's CNPq (CsF PVE 313742/2013-8) and CAPES-ProAmazonia (Projeto 3322-2013), and the European Commission Horizon 2020 RISE programme (Project 691053 - ODYSSEA). Erick Chacon's studentship was funded by the Faculty of Health and Medicine, Lancaster University.

Bibliography

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187.

Battersby, J. (2011). Urban food insecurity in Cape Town, South Africa: An alternative approach to food access. *Development Southern Africa*, 28(4):545–561.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement*, 12(3):261–280.

Briggs Myers, I. and Myers, P. B. (1980). *Gifts Differing: Understanding Personality Type.* Davies-Black Pub.

Cai, L. (2010a). High-dimensional Exploratory Item Factor Analysis by A Metropolis-Hastings Robbins-Monro Algorithm. *Psychometrika*, 75(1):33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis. *Journal of Educational and Behavioral Statistics*, 35(3):307–335.

Carlson, S. J., Andrews, M. S., and Bickel, G. W. (1999). Measuring food insecurity and hunger in the United States: development of a national benchmark measure and prevalence estimates. *The Journal of Nutrition*, 129(2S Suppl):510S–516S.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6).

Chalmers, R. P. (2015). Extended Mixed-Effects Item Response Models With the MH-RM Algorithm. *Journal of Educational Measurement*, 52(2):200–222.

Chen, C. M. and Duh, L. J. (2008). Personalized web-based tutoring system based on fuzzy item response theory. *Expert Systems with Applications*, 34(4):2298–2315.

Chen, C. M., Liu, C. Y., and Chang, M. H. (2006). Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications*, 30(2):378–396.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514):800–812.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

de Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., and Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research (JMR)*, 45(1):104–115.

Downing, S. M. (2003). Item response theory: applications of modern test theory in medical education. *Medical Education*, 37(8):739–745.

Drachler, M. L., Marshall, T., Carlos, J., and Leite, D. C. (2007). A continuous-scale measure of child development for population- based epidemiological surveys : a preliminary study using Item Response Theory for the Denver Test. *Pediatric and Perinatal Epidemiology*, 21:138–53.

Edelen, M. O. and Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(SUPPL. 1):5–18.

Erosheva, E. A. and Curtis, S. M. (2011). Dealing with rotational invariance in bayesian confirmatory factor analysis. *Department of Statistics, University of Washington, Seattle, Washington, USA*.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective. *Statistica Sinica*, 14:731–761.

Fanshawe, T. R. and Diggle, P. J. (2012). Bivariate geostatistical modelling: A review and an application to spatial variation in radon concentrations. *Environmental and Ecological Statistics*, 19(2):139–160.

FAO (2003). Food and agriculture organization of the united nations. trade reforms and food security: Conceptualizing the linkages. Technical report, Food and Agriculture Organization of the United Nations.

Fiori, M., Antonietti, J.-P., Mikolajczak, M., Luminet, O., Hansenne, M., and Rossier, J. (2014). What Is the Ability Emotional Intelligence Test (MSCEIT) Good for? An Evaluation Using Item Response Theory. *PLoS ONE*, 9(6):e98827.

Frichot, E., Schoville, S., Bouchard, G., and François, O. (2012). Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in Genetics*, 3(NOV):1–9.

Froelich, A. G. and Jensen, H. H. (2002). Dimensionality of the usda food security index. *Unpublished Manuscript*.

Funk, J. L. and Rogge, R. D. (2007). Testing the Ruler With Item Response Theory: Increasing Precision of Measurement for Relationship Satisfaction With the Couples Satisfaction Index. *Journal of Family Psychology*, 21(4):572–583.

Garrett, J. L. and Ruel, M. T. (1999). Are determinants of rural and urban food security and nutritional status different? some insights from Mozambique. *World Development*, 27(11):1955–1975.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2):263–312.

Geweke, J. and Zhou, G. (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. *Review of Financial Studies*, 9(2):557–587.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Gray-Little, B., Williams, V. S. L., and Hancock, T. D. (1997). An Item Response Theory Analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 23(5):443–451.

Hambleton, R. and Swaminathan, H. (1989). *Item Response Theory: Principles and Applications*. Evaluation in education and human services. Kluwer-Nijhoff Pub.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1 A):145–168.

Keirsey, D. (1998). *Please Understand Me 2*. Prometheus Nemesis.

Laurens, K. R., Hobbs, M. J., Sunderland, M., Green, M. J., and Mould, G. L. (2012). Psychotic-like experiences in a community sample of 8000 children aged 9 to 11 years: An item response theory analysis. *Psychological Medicine*, 42(7):1495–1506.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(4):423–498.

Merkle, E. C. (2011). A Comparison of Imputation Methods for Bayesian Factor Analysis Models. *Journal of Educational and Behavioral Statistics*, 36(2):257–276.

National Research Council (2006). Item Response Theory and Food Insecurity. In Wunderlich, G. S. and Norwood, J. L., editors, *Food Insecurity and Hunger in the United States: An Assessment of the Measure*, chapter 5. The National Academies Press, Washington, DC.

Osgood, D. W., McMorris, B. J., and Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology*, 18(3):267–296.

Parry, L., Davies, G., Almeida, O., Frausin, G., de Moraés, A., Rivero, S., Filizola, N., and Torres, P. (2017). Social Vulnerability to Climatic Shocks Is Shaped by Urban Accessibility. *Annals of the American Association of Geographers*, 4452(October):1–19.

Parry, L. T. W., Davies, G., Almeida, O., Frausin Bustamante, G. G., de Moraés, A., Rivero, S., Filizola, N., and Torres, P. (2018). Social vulnerability to climatic shocks is shaped by urban accessibility. *Annals of the Association of American Geographers*, 108(1):125–143.

Pebesma, E. (2018). *sf: Simple Features for R.* R package version 0.6-3.

Piquero, A. R., MacIntosh, R., and Hickman, M. (2000). Does Self-Control Affect Survey Response? Applying Exploratory, Confirmatory, and Item Response Theory Analysis To Grasmick Et Al.'S Self-Control Scale. *Criminology*, 38(3):897–930.

Saha, T. D., Chou, S. P., and Grant, B. F. (2006). Toward an alcohol use disorder continuum using item response theory: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychological Medicine*, 36(7):931–941.

Sharp, C., Goodyer, I. M., and Croudace, T. J. (2006). The Short Mood and Feelings Questionnaire (SMFQ): A unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of Abnormal Child Psychology*, 34(3):379–391.

Shryane, N. M., Corcoran, R., Rowse, G., Moore, R., Cummins, S., Blackwood, N., Howard, R., and Bentall, R. P. (2008). Deception and false belief in paranoia: Modelling Theory of Mind stories. *Cognitive Neuropsychiatry*, 13(1):8–32.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):583–616.

Tekwe, C. D., Carter, R. L., Cullings, H. M., and Carroll, R. J. (2014). Multiple indicators, multiple causes measurement error models. *Statistics in Medicine*, 33(25):4469–4481.

Appendix A Spatial item factor analysis

A.1 Scaling aliasing

Restricting the variances of the latent abilities to one, $\mathrm{diag}(\boldsymbol{\Sigma}_\theta) = \mathbf{1}$, is the same as restricting the variances of the residual term $\boldsymbol{v}(s)$ because

$$\mathbb{V}\left[\boldsymbol{b}_k^\intercal \boldsymbol{x}(s) + w_k(s) + v_k(s)\right] = 1 \tag{4.62}$$

for $k = 1, \ldots, m$ implies that

$$\mathbb{V}\left[v_k(s)\right] = 1 - \mathbb{V}\left[\boldsymbol{b}_k^\intercal \boldsymbol{x}(s) + w_k(s)\right] = \sigma^2_{v_k}. \tag{4.63}$$

More generally, the constrain $\mathrm{diag}(\boldsymbol{\Sigma}_\theta)$ is equivalent to set the covariance matrix of $\boldsymbol{v}(s)$ as $\boldsymbol{\Sigma}_v = \boldsymbol{D}_1 \boldsymbol{R}_v \boldsymbol{D}_1$, where $\boldsymbol{D}_1$ is a diagonal matrix with elements σ_{v_k}. Then the covariance matrix of the latent abilities $\boldsymbol{\theta}(s)$ is expressed as

$$\mathbb{V}\left[\boldsymbol{\theta}(s)\right] = \mathbb{V}\left[\boldsymbol{B}^\intercal \boldsymbol{x}(s)\right] + \mathbb{V}\left[\boldsymbol{w}(s)\right] + \boldsymbol{D}_1 \boldsymbol{R}_v \boldsymbol{D}_1, \tag{4.64}$$

the problem with this restriction is that σ_{v_k} need to be known.

Inference can be attained by introducing arbitrary values. Consider the transformation $\boldsymbol{D}_2 = \boldsymbol{D}\boldsymbol{D}_1^{-1}$, where $\boldsymbol{D}$ is a diagonal matrix with arbitrary values, then we can define

$$\hat{\boldsymbol{a}}_j^\intercal \hat{\boldsymbol{\theta}}(s) = \boldsymbol{a}_j^\intercal \boldsymbol{D}_2^{-1} \boldsymbol{D}_2 \boldsymbol{\theta}(s) = \boldsymbol{a}_j^\intercal \boldsymbol{\theta}(s). \tag{4.65}$$

Note that under this transformation, the variance of the new latent variable $\hat{\boldsymbol{\theta}}(s) = \boldsymbol{D}_2 \boldsymbol{\theta}(s)$ is defined as

$$\mathbb{V}\left[\hat{\boldsymbol{\theta}}(s)\right] = \mathbb{V}\left[\hat{\boldsymbol{B}}^\intercal \boldsymbol{x}(s)\right] + \mathbb{V}\left[\hat{\boldsymbol{w}}(s)\right] + \boldsymbol{D}\boldsymbol{R}_v\boldsymbol{D}, \tag{4.66}$$

where $\hat{\boldsymbol{B}}^{\intercal} = \boldsymbol{D}_2\boldsymbol{B}^{\intercal}$ and $\hat{\boldsymbol{w}}(s) = \boldsymbol{D}_2\boldsymbol{w}(s)$. It can be seen that defining an arbitrary diagonal matrix $\boldsymbol{D}$ still allows us to make inference given that the marginal variances of $\hat{\boldsymbol{\theta}}(s)$ are still restricted. In this case, the variances are equal to the squared values of the diagonal matrix $\boldsymbol{D}_2$, $\text{diag}(\boldsymbol{\Sigma}_{\hat{\theta}}) = \text{diag}(\boldsymbol{D}_2^2)$.

If we choose $\boldsymbol{D} = \boldsymbol{I}$; then $\boldsymbol{D}_2 = \boldsymbol{D}_1^{-1}$, $\hat{\boldsymbol{\theta}}(s) = \boldsymbol{D}_1^{-1}\boldsymbol{\theta}(s)$, $\text{diag}(\boldsymbol{\Sigma}_{\hat{\theta}}) = \text{diag}(\boldsymbol{D}_1^{-2})$ and

$$\mathbb{V}\left[\hat{\boldsymbol{\theta}}(s)\right] = \mathbb{V}\left[\hat{\boldsymbol{B}}^{\intercal}\boldsymbol{x}(s)\right] + \mathbb{V}\left[\hat{\boldsymbol{w}}(s)\right] + \boldsymbol{R}_v. \tag{4.67}$$

This transformation allows us to make inference, but the interpretation of the transformed parameters $\hat{\boldsymbol{a}}_j$ are not the same as in the classical item factor analysis because the marginal variances of $\hat{\boldsymbol{\theta}}(s)$ are not equal to 1, $\text{diag}(\Sigma_{\hat{\theta}}) \neq \boldsymbol{1}$. To recover the interpretation of the discrimination parameters, we simply compute the standard deviations of $\hat{\boldsymbol{\theta}}(s)$ after sampling to obtain the estimated $\boldsymbol{Q} = \hat{\boldsymbol{D}}_1^{-1}$, and back-transformed $\boldsymbol{a}_j = \boldsymbol{Q}\hat{\boldsymbol{a}}_j$ and $\boldsymbol{\theta}(s) = \boldsymbol{Q}^{-1}\hat{\boldsymbol{\theta}}(s)$ as explained in Section 4.4.4.

Appendix B Markov chain Monte Carlo scheme sampling

B.1 Posterior of auxiliary variables

We show the details of how to obtain Equation 4.31 to specify the posterior distribution of the auxiliary variables of our model.

$$\begin{aligned}
\Pr\left(\boldsymbol{z} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) &= \Pr\left(\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \\
&\propto \Pr\left(\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}\right) \\
&\propto \Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}\right) \\
&\propto \Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{z}_{mis} \mid \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right).
\end{aligned} \tag{4.68}$$

The last line is obtained because $\boldsymbol{Z}_{mis}$ and $\boldsymbol{Z}_{obs}$ are conditionally independent given $\{\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\theta}\}$ and because $\boldsymbol{Y}_{obs}$ is conditionally independent of $\boldsymbol{Z}_{mis}$ given $\boldsymbol{Z}_{obs}$.

Appendix C Alternative Sampling Schemes

C.1 Alternative sampling scheme using marginalization

In Section 4.4.1, we defined the Bayesian model such us the conditional probability $\Pr\left(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{a}\right)$ plays a main role to derive the posterior conditional distributions of the associated parameters. This was convenient to obtain the analytical expression of the conditional posterior distributions; however, convergence can be slow due to nested relationship in the updates of the Gibbs sampling. An Alternative approach to achieve faster convergence, in terms of iterations, is to marginalize some parameters such as the nested relationship is reduced.

Considering the definition of the auxiliary variables in Equation 4.18, we can see that any element from the set $\{\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{v}\}$ can be marginalized due to conjugacy Gaussian properties. Let $\boldsymbol{\alpha}$ be the subset of parameters that we wish to marginalize and $\boldsymbol{\gamma}$ the subset of remaining parameters which will not be marginalized. Additionally, let $\boldsymbol{X}_{\alpha}$ and $\boldsymbol{X}_{\gamma}$ be the associated design matrix, and let $\boldsymbol{\Sigma}_{\alpha}$ and $\boldsymbol{\Sigma}_{\gamma}$ be the associated covariance matrices. Then the auxiliary variables can be expressed as

$$\boldsymbol{Z} = \boldsymbol{X}_{\gamma}\boldsymbol{\gamma} + \boldsymbol{X}_{\alpha}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{4.69}$$

where at least one of the design matrices $\boldsymbol{X}_{\beta}$ and $\boldsymbol{X}_{\gamma}$ will depend of the restricted discrimination parameters $\boldsymbol{A}^{*}$. Then, *composition* sampling, as shown in Holmes and Held (2006), can be used to sample from the posterior distribution of the model using the following equivalence

$$\Pr\left(\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_{v} \mid \boldsymbol{y}\right) = \Pr\left(\boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_{v} \mid \boldsymbol{y}\right) \Pr\left(\boldsymbol{\alpha} \mid \boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_{v}\right), \tag{4.70}$$

such as the first term of the right hand side does not depend of the set of parameters $\boldsymbol{\alpha}$. This way convergence is expected to be faster and the parameters included in $\boldsymbol{\alpha}$ can be simulated by *composition* sampling once the convergence of the remaining parameters is ensured.

Sampling from the marginalized parameters $\boldsymbol{\alpha}$ can be done straight away because the conditional distribution given $\boldsymbol{\gamma}$ is a Gaussian density,

$$\Pr(\boldsymbol{\alpha} \mid \boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) \propto \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{X}_\gamma \boldsymbol{\gamma} + \boldsymbol{X}_\alpha \boldsymbol{\alpha}, \boldsymbol{I}_{nq}) \mathcal{N}(\boldsymbol{\alpha} \mid \boldsymbol{0}, \boldsymbol{\Sigma}_\alpha), \tag{4.71}$$

with mean and covariance:

$$\boldsymbol{\Sigma}_{\alpha|z} = (\boldsymbol{X}_\alpha^\intercal \boldsymbol{X}_\alpha + \boldsymbol{\Sigma}_\alpha^{-1})^{-1} \tag{4.72}$$

$$\boldsymbol{\mu}_{\alpha|z} = \boldsymbol{\Sigma}_{\alpha|z} \boldsymbol{X}_\alpha^\intercal (\boldsymbol{z} - \boldsymbol{X}_\gamma \boldsymbol{\gamma}). \tag{4.73}$$

In some cases, computational advantage can be gained considering that

$$(\boldsymbol{X}_\alpha^\intercal \boldsymbol{X}_\alpha + \boldsymbol{\Sigma}_\alpha^{-1})^{-1} = \boldsymbol{\Sigma}_\alpha - \boldsymbol{\Sigma}_\alpha \boldsymbol{X}_\alpha^\intercal (\boldsymbol{X}_\alpha \Sigma_\alpha \boldsymbol{X}_\alpha^\intercal + \boldsymbol{I})^{-1} \boldsymbol{X}_\alpha \boldsymbol{\Sigma}_\alpha. \tag{4.74}$$

Obtaining posterior samples for $\{\boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\}$ is more complicated, but can be achieved using Metropolis within Gibbs sampling. For this, we should notice that

$$\begin{aligned}\Pr(\boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v \mid \boldsymbol{y}) \propto \Pr(\boldsymbol{y} \mid \boldsymbol{z}) \Pr(\boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) \Pr(\boldsymbol{\gamma}) \Pr(\boldsymbol{a}) \\ \Pr(\boldsymbol{T}) \Pr(\boldsymbol{\phi}) \Pr(\boldsymbol{R}).\end{aligned} \tag{4.75}$$

Hence, using Equation 4.69, the conditional posterior for $\boldsymbol{Z}$ is

$$\Pr(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) \propto \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{X}_\gamma \boldsymbol{\gamma}, \boldsymbol{X}_\alpha \boldsymbol{\Sigma}_\alpha \boldsymbol{X}_\alpha^\intercal + \boldsymbol{I}_{nq}) \prod_{i,j} \Pr(y_{ij} \mid z_{ij}) \tag{4.76}$$

which is a truncated multivariate normal distribution. Unfortunately, sampling can not be done directly, but Gibbs sampling can be used taking into advantage that the conditional posterior of the marginalized parameters $\boldsymbol{\alpha}$ given all the auxiliary variables except Z_k, $\Pr(\boldsymbol{\alpha} \mid \boldsymbol{z}_{-k}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v)$, is a Normal distribution with covariance and mean:

$$\begin{aligned} \boldsymbol{\Sigma}_{\alpha|z_{-k}} &= \boldsymbol{\Sigma}_{\alpha|z} + \frac{\boldsymbol{\Sigma}_{\alpha|z}\boldsymbol{x}_{\alpha_k}\boldsymbol{x}_{\alpha_k}^{\intercal}\boldsymbol{\Sigma}_{\alpha|z}}{1-h_{kk}} && (4.77) \\ \boldsymbol{\mu}_{\alpha|z_{-k}} &= \boldsymbol{\mu}_{\alpha|z} - \frac{\boldsymbol{\Sigma}_{\alpha|z}\boldsymbol{x}_{\alpha_k}}{1-h_{kk}}(z_k - \boldsymbol{x}_{\beta_{ck}}^{\intercal}\boldsymbol{\beta}_c - \boldsymbol{x}_{\alpha_k}^{\intercal}\boldsymbol{\mu}_{\alpha|z}), && (4.78) \end{aligned}$$

where $h_{kk} = \boldsymbol{x}_{\alpha_k}^{\intercal}\boldsymbol{\Sigma}_{\alpha|y}\boldsymbol{x}_{\alpha_k}$. Then, we can sample from the leave-one-out marginal predictive densities,

$$\Pr(z_k \mid \boldsymbol{z}_{-k}, y_k, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v) = \int \Pr(z_k \mid \boldsymbol{\alpha}, y_k, \boldsymbol{\gamma}, \boldsymbol{a}) \Pr(\boldsymbol{\alpha} \mid \boldsymbol{z}_{-k}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v)\, d\boldsymbol{\alpha},$$

being proportional to

$$\mathbb{1}_{(z_k>0)}^{y_k}\mathbb{1}_{(z_k\leq 0)}^{1-y_k} \int \mathcal{N}(z_k \mid \boldsymbol{x}_{\gamma_k}^{\intercal}\boldsymbol{\gamma} + \boldsymbol{x}_{\alpha_k}^{\intercal}\boldsymbol{\alpha}, 1) \Pr(\boldsymbol{\alpha} \mid \boldsymbol{z}_{-k}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v)\, d\boldsymbol{\alpha} \qquad (4.79)$$

which are univariate Normal truncated densities,

$$\propto \mathcal{N}\left(\boldsymbol{x}_{\gamma_k}^{\intercal}\boldsymbol{\gamma} + \boldsymbol{x}_{\alpha_k}^{\intercal}\boldsymbol{\mu}_{\alpha|z} - w_k(z_k - \boldsymbol{x}_{\gamma_k}^{\intercal}\boldsymbol{\gamma} - \boldsymbol{x}_{\alpha_k}^{\intercal}\boldsymbol{\mu}_{\alpha|z}), 1 + w_k\right) \mathbb{1}_{(z_k>0)}^{y_k}\mathbb{1}_{(z_k\leq 0)}^{1-y_k}, \qquad (4.80)$$

where $w_k = h_{kk}/(1-h_{kk})$. As explained in Holmes and Held (2006), each time a sample z_k is drawn, the conditional mean $\boldsymbol{\mu}_{\alpha|z}$ must be updated. Denoting $\boldsymbol{S} = \boldsymbol{\Sigma}_{\alpha|\cdot}\boldsymbol{X}_{\alpha}^{\intercal}$, the conditional mean can be expressed as $\boldsymbol{\mu}_{\alpha|z} = \boldsymbol{S}_{-i}\boldsymbol{z}_{-i} + \boldsymbol{S}_i z_i - \boldsymbol{S}\boldsymbol{X}_{\gamma}\boldsymbol{\gamma}$, and it can efficiently be updated as

$$\begin{aligned} \boldsymbol{\mu}_{\alpha|z}^{\text{new}} &= \boldsymbol{S}_i z_i^{\text{new}} + \boldsymbol{S}_{-i}\boldsymbol{z}_{-i} - \boldsymbol{S}\boldsymbol{X}_{\beta_c}\boldsymbol{\beta}_c && (4.81) \\ &= \boldsymbol{S}_i z_i^{\text{new}} + \boldsymbol{S}_{-i}\boldsymbol{z}_{-i} - \boldsymbol{S}\boldsymbol{X}_{\beta_c}\boldsymbol{\beta}_c && (4.82) \\ &= \boldsymbol{S}_i z_i^{\text{new}} + \boldsymbol{S}\boldsymbol{z} - \boldsymbol{S}_i z_i^{\text{old}} - \boldsymbol{S}\boldsymbol{X}_{\beta_c}\boldsymbol{\beta}_c && (4.83) \\ &= \boldsymbol{\mu}_{\alpha|z}^{\text{old}} + \boldsymbol{S}_i(z_i^{\text{new}} - z_i^{\text{old}}). && (4.84) \end{aligned}$$

Finally, because we do not get analytically expressions for the conditional distributions of remaining parameters $\{\boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\}$, we can use Metropolis-Hasting or others samplers like Hamiltonian Monte Carlo to obtain draws from them. The posterior is only defined up to a constant of proportionality

$$\Pr\left(\boldsymbol{a}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v \mid \boldsymbol{\gamma}, \boldsymbol{z}\right) \propto \Pr\left(\boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{R}_v\right) \Pr\left(\boldsymbol{a}\right) \Pr\left(\boldsymbol{T}\right) \Pr\left(\boldsymbol{\phi}\right) \Pr\left(\boldsymbol{R}_v\right). \quad (4.85)$$

Note that an adequate transformation will be required to sample these parameters as explained in Section 4.4.3.6.

C.2 Marginalizing the Gaussian process and individual random effect

In the spatial item factor analysis, it seems reasonable to desired to marginalized the more high-dimensional terms like the multivariate Gaussian process $\boldsymbol{w}$ and the multivariate residual term $\boldsymbol{v}$. In this case, the marginalized parameters is defined as $\boldsymbol{\alpha} = (\boldsymbol{w}^\intercal, \boldsymbol{v}^\intercal)^\intercal$ with associated design matrix $\boldsymbol{X}_\alpha = (\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n, \boldsymbol{A}^* \otimes \boldsymbol{I}_n)$. The remaining parameters would be $\boldsymbol{\gamma} = (\boldsymbol{c}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$ with design matrix $\boldsymbol{X}_\gamma = (\boldsymbol{I}_q \otimes \mathbf{1}_n, \boldsymbol{A}^* \otimes \boldsymbol{X})$. The covariance matrix of these collections of parameters are obtained as $\boldsymbol{\Sigma}_\alpha = \operatorname{diag}(\boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_v)$ and $\boldsymbol{\Sigma}_\gamma = \operatorname{diag}(\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_\beta)$. Given these definitions, it can be noticed that the some of the terms required for the sampling are

$$\begin{aligned}
\boldsymbol{X}_\gamma\boldsymbol{\gamma} &= (\boldsymbol{I}_q \otimes \mathbf{1}_n)\boldsymbol{c} + (\boldsymbol{A}^* \otimes \boldsymbol{X})\boldsymbol{\beta} \\
\boldsymbol{X}_\alpha\boldsymbol{\Sigma}_\alpha\boldsymbol{X}_\alpha^\intercal &= (\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n)(\oplus_{k=1}^m \boldsymbol{\Sigma}_{w_k})(\boldsymbol{T}^\intercal\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n) + (\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n), \quad (4.86)
\end{aligned}$$

$$\boldsymbol{X}_\alpha^\intercal\boldsymbol{X}_\alpha = \begin{pmatrix} \boldsymbol{T}^\intercal\boldsymbol{A}^{*\intercal}\boldsymbol{A}^*\boldsymbol{T} & \boldsymbol{T}^\intercal\boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \\ \boldsymbol{A}^{*\intercal}\boldsymbol{A}^*\boldsymbol{T} & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \end{pmatrix} \otimes \boldsymbol{I}_n, \quad (4.87)$$

and also

$$\begin{aligned}\boldsymbol{X}_\alpha^\top(\boldsymbol{z}-\boldsymbol{X}_\gamma\boldsymbol{\gamma}) &= \begin{pmatrix}(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)(\boldsymbol{z}-\boldsymbol{X}_\gamma\boldsymbol{\gamma})\\ (\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)(\boldsymbol{z}-\boldsymbol{X}_\gamma\boldsymbol{\gamma})\end{pmatrix} && (4.88)\\ &= \begin{pmatrix}\operatorname{vec}((\boldsymbol{Z}-\boldsymbol{1}_n\boldsymbol{c}^\top-\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{*\top})\boldsymbol{A}^*\boldsymbol{T})\\ \operatorname{vec}((\boldsymbol{Z}-\boldsymbol{1}_n\boldsymbol{c}^\top-\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{*\top})\boldsymbol{A}^*)\end{pmatrix}. && (4.89)\end{aligned}$$

As mentioned before, we can take advantage of Equation 4.74 and additionally reduce the dimension of the computational cost considering that

$$\begin{aligned}\boldsymbol{X}_\alpha\boldsymbol{\Sigma}_\alpha\boldsymbol{X}_\alpha^\top+\boldsymbol{I}_{nq} &= (\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+\boldsymbol{I}_{nq}\\ &= (\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)\otimes\boldsymbol{I}_n),\end{aligned} \tag{4.90}$$

that the inverse of this is

$$\begin{aligned}&(\boldsymbol{X}_\alpha\boldsymbol{\Sigma}_\alpha\boldsymbol{X}_\alpha^\top+\boldsymbol{I}_{nq})^{-1} && (4.91)\\ &\quad= ((\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+\boldsymbol{I}_{nq})^{-1}\\ &\quad= ((\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)\otimes\boldsymbol{I}_n)^{-1}\\ &\quad= (\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\otimes\boldsymbol{I}_n-((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\\ &\qquad(\boldsymbol{\Sigma}_w^{-1}+\boldsymbol{T}^\top\boldsymbol{A}^{*\top}(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)^{-1}(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\otimes\boldsymbol{I}_n)\\ &\quad= (\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\otimes\boldsymbol{I}_n-((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w\\ &\qquad(\boldsymbol{I}_{mn}+\big(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n\big)\,\boldsymbol{\Sigma}_w)^{-1}(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\otimes\boldsymbol{I}_n)\end{aligned}$$
$$(4.92)$$

and that the determinant is

$$\begin{aligned}&\det(\boldsymbol{X}_\alpha\boldsymbol{\Sigma}_\alpha\boldsymbol{X}_\alpha^\top+\boldsymbol{I}_{nq})\\ &\quad= \det\left((\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)\otimes\boldsymbol{I}_n\right)\\ &\quad= \det((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)\otimes\boldsymbol{I}_n)\det(((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\otimes\boldsymbol{I}_n)(\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w(\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)+\boldsymbol{I}_{nq})\\ &\quad= \det((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)\otimes\boldsymbol{I}_n)\det((\boldsymbol{T}^\top\boldsymbol{A}^{*\top}\otimes\boldsymbol{I}_n)((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\otimes\boldsymbol{I}_n)(\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w+\boldsymbol{I}_{nm})\\ &\quad= \det((\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)\otimes\boldsymbol{I}_n)\det((\boldsymbol{T}^\top\boldsymbol{A}^{*\top}(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w+\boldsymbol{I}_{nm})\\ &\quad= \det(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^n\det((\boldsymbol{T}^\top\boldsymbol{A}^{*\top}(\boldsymbol{A}^*\boldsymbol{R}_v\boldsymbol{A}^{*\top}+\boldsymbol{I}_q)^{-1}\boldsymbol{A}^*\boldsymbol{T}\otimes\boldsymbol{I}_n)\boldsymbol{\Sigma}_w+\boldsymbol{I}_{nm}). && (4.93)\end{aligned}$$

C.3 Marginalizing all the posibble set of parameters

Let $\boldsymbol{\alpha} = (\boldsymbol{c}^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{w}^\intercal, \boldsymbol{v}^\intercal)^\intercal$ denote the collection of model terms that will be marginalized with associated design matrix $\boldsymbol{X}_\alpha = (\boldsymbol{I}_q \otimes \mathbf{1}_n, \boldsymbol{A}^* \otimes \boldsymbol{X}, \boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n, \boldsymbol{A}^* \otimes \boldsymbol{I}_n)$. The covariance matrix of this collection of parameters is obtained as $\boldsymbol{\Sigma}_\alpha = \text{diag}(\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_\beta, \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_v)$. Then $\boldsymbol{\gamma}$ would be an empty set and will simply be removed from the expressions shown in Section C.1.

The sampling follows the explanation presented in Section C.1, but it is worth to notice that

$$\boldsymbol{X}_\alpha^\intercal \boldsymbol{X}_\alpha = \begin{pmatrix} \boldsymbol{I}_q \otimes \mathbf{1}_{\boldsymbol{n}}^\intercal \mathbf{1}_{\boldsymbol{n}} & \boldsymbol{A}^* \otimes \mathbf{1}_{\boldsymbol{n}}^\intercal \boldsymbol{X} & \boldsymbol{A}^*\boldsymbol{T} \otimes \mathbf{1}_{\boldsymbol{n}}^\intercal & \boldsymbol{A}^* \otimes \mathbf{1}_{\boldsymbol{n}}^\intercal \\ \boldsymbol{A}^{*\intercal} \otimes \boldsymbol{X}^\intercal \mathbf{1}_{\boldsymbol{n}} & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \otimes \boldsymbol{X}^\intercal \boldsymbol{X} & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{X}^\intercal & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \otimes \boldsymbol{X}^\intercal \\ \boldsymbol{T}^\intercal \boldsymbol{A}^{*\intercal} \otimes \mathbf{1}_{\boldsymbol{n}} & \boldsymbol{T}^\intercal \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \otimes \boldsymbol{X} & \boldsymbol{T}^\intercal \boldsymbol{A}^{*\intercal}\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n & \boldsymbol{T}^\intercal \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \otimes \boldsymbol{I}_n \\ \boldsymbol{A}^{*\intercal} \otimes \mathbf{1}_{\boldsymbol{n}} & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \otimes \boldsymbol{X} & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^*\boldsymbol{T} \otimes \boldsymbol{I}_n & \boldsymbol{A}^{*\intercal}\boldsymbol{A}^* \otimes \boldsymbol{I}_n \end{pmatrix}, \tag{4.94}$$

$$\boldsymbol{X}_\alpha^\intercal \boldsymbol{z} = \begin{pmatrix} (\boldsymbol{I}_q \otimes \mathbf{1}_{\boldsymbol{n}}^\intercal)\boldsymbol{z} \\ (\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{X}^\intercal)\boldsymbol{z} \\ (\boldsymbol{T}^\intercal \boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n)\boldsymbol{z} \\ (\boldsymbol{A}^{*\intercal} \otimes \boldsymbol{I}_n)\boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \text{vec}(\mathbf{1}_{\boldsymbol{n}}^\intercal \boldsymbol{Z}) \\ \text{vec}(\boldsymbol{X}^\intercal \boldsymbol{Z}\boldsymbol{A}^*) \\ \text{vec}(\boldsymbol{Z}\boldsymbol{A}^*\boldsymbol{T}) \\ \text{vec}(\boldsymbol{Z}\boldsymbol{A}^*) \end{pmatrix}. \tag{4.95}$$

Appendix D Traceplots of the Case of Study

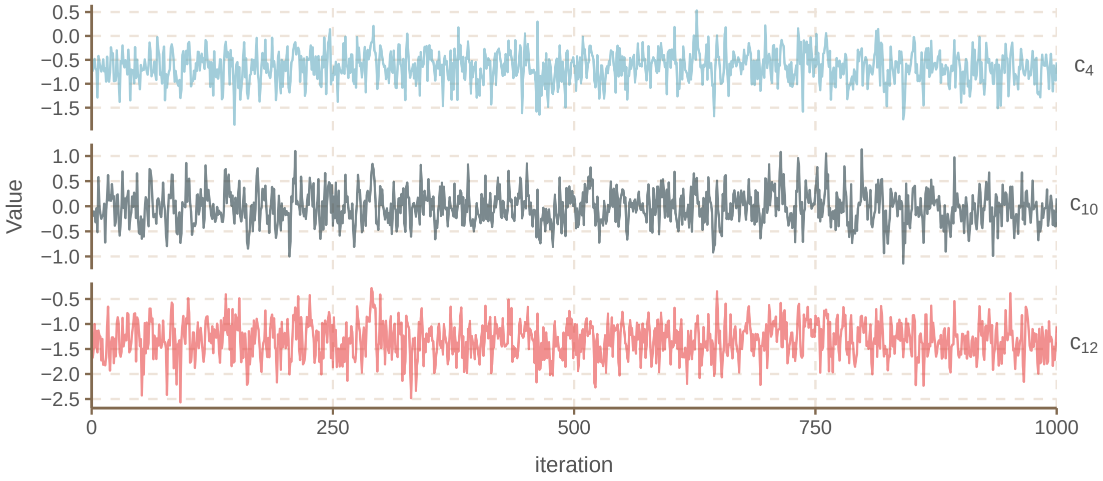

Figure 4.6: Traceplots of difficulty parameters: only 3 out of 18 were randomly selected to be shown.

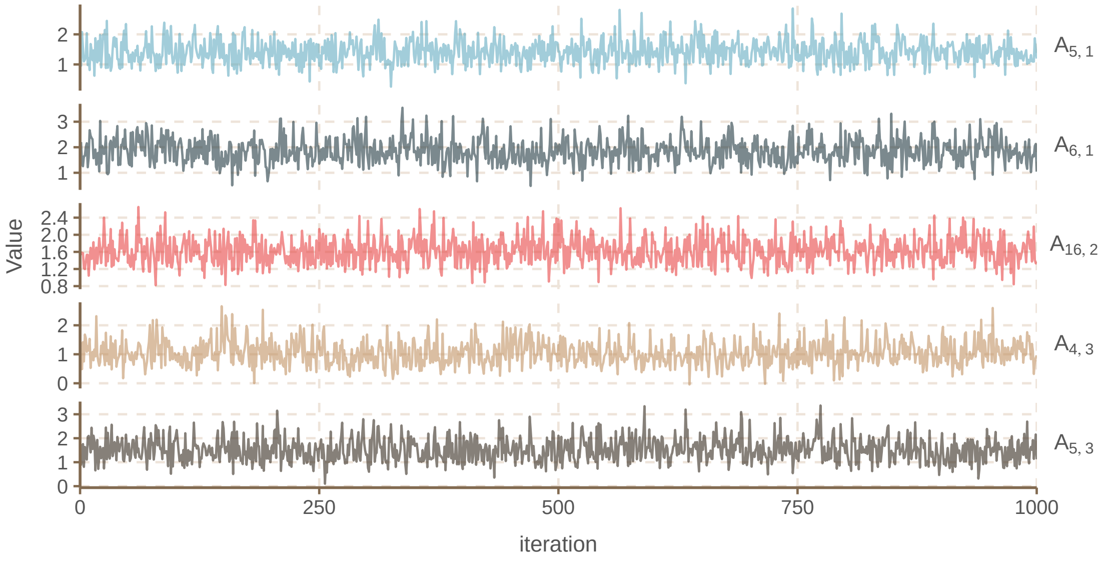

Figure 4.7: Traceplots of discrimination parameters: only 5 out of 22 were randomly selected to be shown.

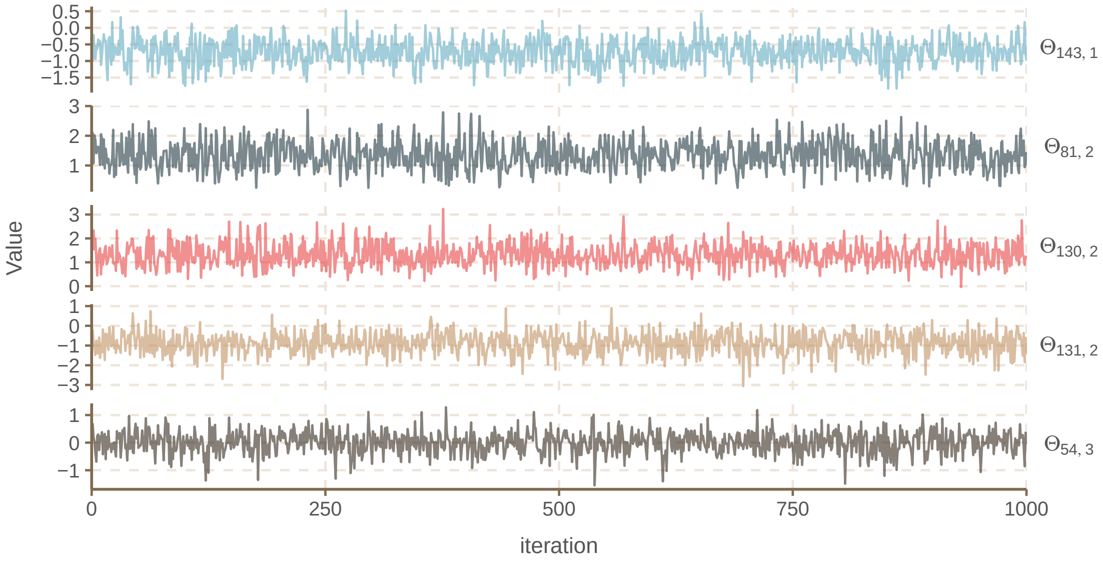

Figure 4.8: Traceplots of discrimination parameters: only 5 out of 600 were randomly selected to be shown.

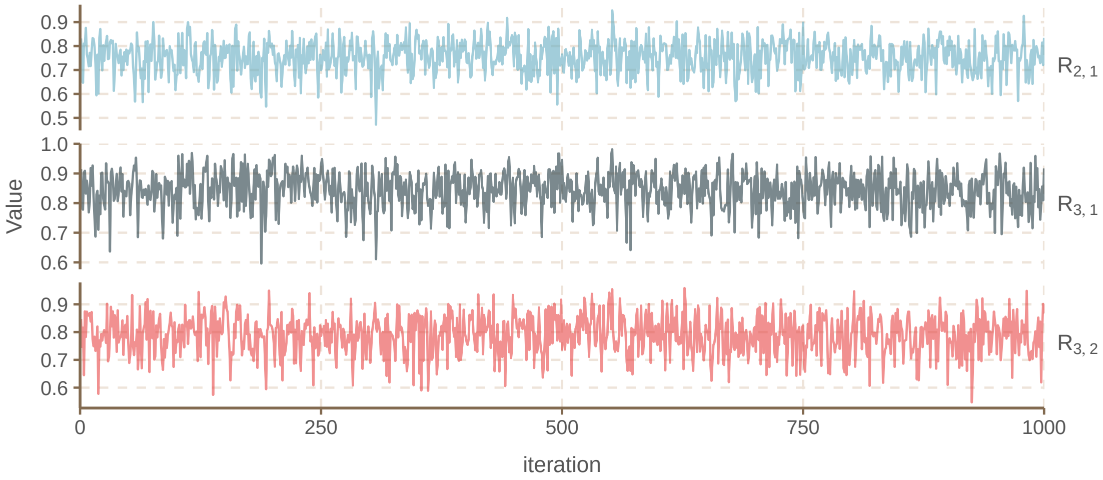

Figure 4.9: Traceplots of correlation parameters.

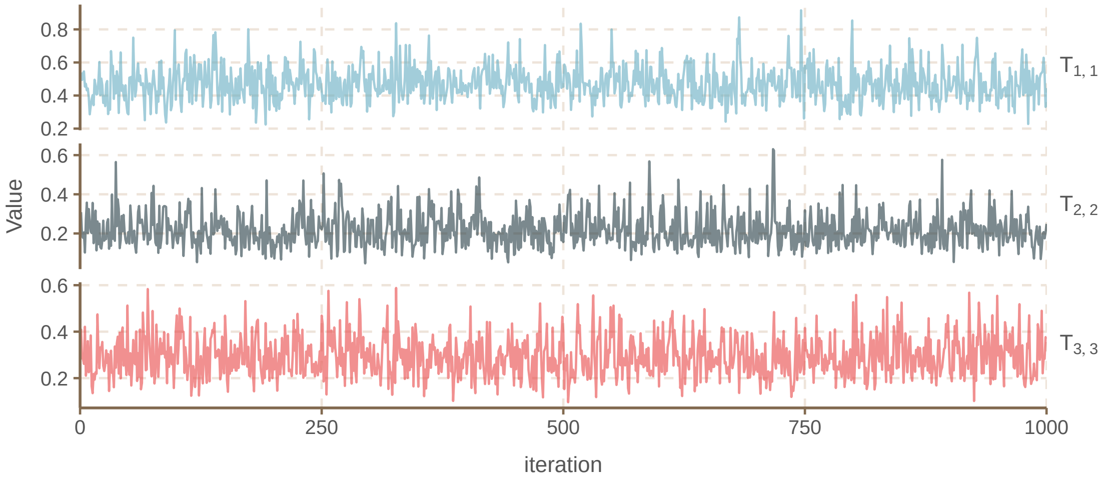

Figure 4.10: Traceplots of unrestricted standard deviations parameters for the multivariate Gaussian process.

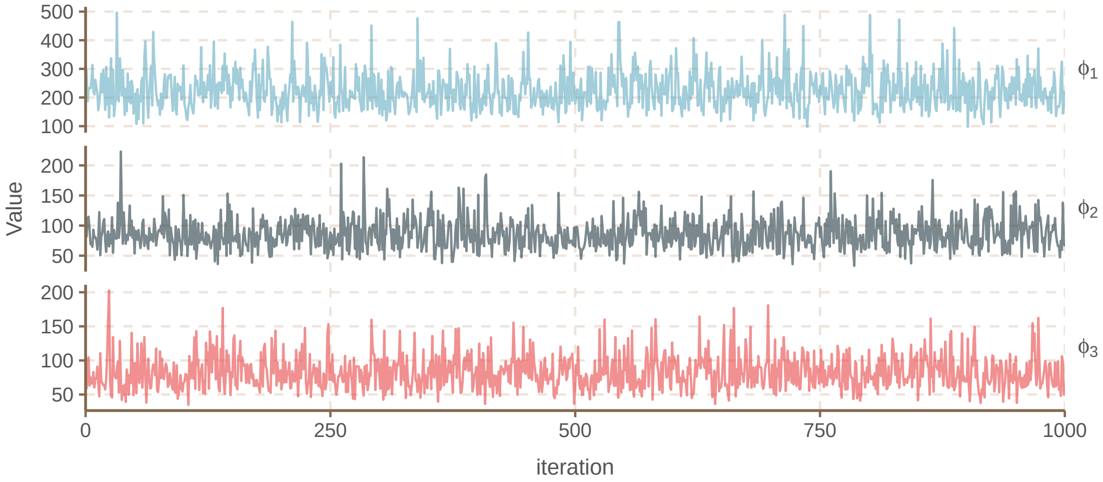

Figure 4.11: Traceplots of the scale parameters of the multivariate Gaussian process.

Appendix E Extension to Mixed Outcome Types

In order to deal with binary, ordinal or continuous items, we can extend the spatial item factor analysis by considering q_1 ordinal items and q_2 continuous items. We do not need to differentiate another set of binary items given the they are simply ordinal items with two categories. The q_1 ordinal times can be modelled as spatial discrete-valued stochastic processes $\{Y_j(s) : s \in D\}$, where $D \subset \mathbb{R}^2$ and the random variable $Y_j(s)$ can take values $\{0, 1, \ldots, K_j - 1\}$. Notice that

K_j represents the number of categories for ordinal item $j = 1, \ldots, q_1$. We assume that the values of the q_1 discrete-valued stochastic processes are determined by an auxiliary real-valued stochastic processes $\{Z_j^o(s) : s \in D\}$ and thresholds $\boldsymbol{\gamma}_j = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{j(K_j-1)})^\intercal$ such as

$$Y_j(s) = k \iff -\gamma_{jk} \leq Z_j^o(s) < -\gamma_{j(k+1)}, \text{ for } k = 0, 1, \ldots, (K_j - 1),$$

where $\gamma_{j0} = -\infty$ and $\gamma_{j(K_j)} = \infty$. The q_2 continuous items can be modelled as real-valued stochastic processes $\{Z_j^c(s) : s \in D\}$ for $j = 1, \ldots, q_2$. Then, we can defined the spatial random vector $\boldsymbol{Z}(s) = (Z_1^o(s), \ldots, Z_{q_1}^o(s), Z_1^c(s), \ldots, Z_{q_2}^c(s))^\intercal$, a collection of the auxiliary random variables $Z_j^o(s)$ associated to the ordinal items and the observable random variables $Z_j^c(s)$ associated to the continuous items, and define the factor model at this level such as

$$Z_j(s) = c_j + \boldsymbol{a}_j^{*\intercal}\boldsymbol{\theta}(s) + \epsilon_j(s), \text{ for } j = 1, \ldots, q_1 + q_2$$

where, due to identifiability, $c_j = 0$ for $j = 1, \ldots, q_1$ and where the m-dimensional latent factors are modelled including multivariate non-linear effects, $\boldsymbol{f}(x_j(s)) : \mathbb{R} \to \mathbb{R}^m$,

$$\boldsymbol{\theta}(s) = \sum_{i=1}^{p} \boldsymbol{f}(x_i(s)) + \boldsymbol{w}^*(s) + \boldsymbol{v}(s).$$

Finally, to make the model identifiable, the error term is defined as

$$\epsilon_j(s) \sim \begin{cases} \mathcal{N}(0, 1) & \text{for } j = 1, \ldots, q_1 \\ \mathcal{N}(0, \sigma_j) & \text{for } j = q_1 + 1, \ldots, q_1 + q_2 \end{cases}.$$

Chapter 5
Conclusions

In this thesis, we have presented three studies that aim to understand effects of extreme hydro-climatic events on the health of vulnerable populations in Brazilian Amazonia. In our first study, we proposed a *model-based standardized index* (MBSI) to quantify and identify floods and droughts (Chacón-Montalván et al., 2018a). In the second, we studied the effects of floods and droughts on birthweight (Chacón-Montalván et al., 2018b). In the third, we proposed and developed inferential methods for *spatial item factor analysis* (SPIFA) to model and predict food insecurity (Chacon-Montalvan et al., 2018). In this section, we summarise the contributions of these studies to (i) our collective scientific understanding of the impacts of extreme hydro-climatic events on population health and (ii) with respect to advances in statistical modelling.

5.1 Linking Extreme Hydro-climatic Events with Population Health

In our studies, we have found consistently that floods, which are increasing in frequency and magnitude, can have severe effects on population health. We have also found identified spatial regions and popultaion characteristics of disadvantaged groups in roadless cities of Brazilian Amazonia, in which the impacts of extreme events are likely to be exacerbated due to the fact that these populations are vulnerable. From the aid perspective, our results can help in the development of policies to protect these groups. In the future, we would like to see the development of an early warning system for food security and health indicators.

The effects of floods on population health was investigated in Chapter 3 by analysing birthweight and by modelling food insecurity in Chapter 4. We studied these two variables because we hypothesized that one mechanism by which extreme hydro-climatic events could affect human health would be nutrition, which is highly connected with both birthweight and food insecurity.

In the study of birthweight, it was found that exposure to extreme floods during pregnancy reduces birth-weight by around 200 grams (see Section 3.4.1); while in the study of food security, we found that hotspot areas of food insecurity, in different dimensions, are related with flood-prone neighborhoods (see Section 4.6.3). This strongly indicates that there is a need for improving policies for prevention of extreme hydro-climatic events to reduce the impact on the remote populations of Brazilian Amazonia.

Disadvantaged groups have also been found our studies of birthweight in Chapter 3 and food insecurity in Chapter 4. The effects of extreme hydro-climatic events can be of major impact on these groups. For example, in Section 3.4.3, it was found that babies from disadvantaged groups were between 163 and 271 grams lighter compared to advantaged groups. The characteristics of these vulnerable groups when studying birthweight was indigenous mothers with low level of education and low ante-natal care. The additional effect of experiencing extreme hydro-climatic events for these groups could lead to newborns with low weight (< 2500 grams). With respect to food insecurity, high areas of food insecurity were also associated with poor and marginalized neighborhoods (see Section 4.6.3). This indicates spatial structure in disadvantaged groups (which is a common demographic phenomenon, even in the UK). When proposing policies for monitoring and preventing floods to reduce the impact on populations, the characteristics and location of the vulnerable groups should be used in order to prioritize and select intervention efforts and resources given that they are more likely to be affected because of their low resilience.

We would like to see our work forming a basis for the development of an early warning system (EWS) to reduce the impacts of floods and droughts on human health. We could focus on ensuring an adequate level of food security in this EWS. The system could first use the MBSI presented in Chapter 2 to predict droughts and floods by relating the obtained MBSI of precipitation with river levels and we could use additional climatic and environmental variables like topography to improve this predictions. Later, similar to Chapter 4, an SPIFA model could be used to predict areas of high food insecurity by including the data of predicted floods and droughts as covariates on the SPIFA model. Finally, prioritization inside these areas could be done by identifying the vulnerable groups as in our study of birthweight in Chapter 3.

5.2 Statistical Modelling

By understanding the effects of extreme hydro-climatic events on population health, we developed novel methodologies that contribute to the area of statistical modelling. These contributions are based on known theory of generalized additive models for location, scale and shape (GAMLSS), item factor analysis (IFA) and spatial statistics. Using these approaches, and mixing them when necessary, allowed us to propose very flexible approaches like the model-based standardised index (MBSI) in Chapter 2, the spatial item factor analysis (SPIFA) in Chapter 4 and the mixed spatial item factor analysis (MSPIFA) in Section 5.2.3.

5.2.1 Model-based Standardised Index

The model-based standardised index (MBSI) selected in Chapter 2 is an index to detect extreme events of an discrete-time stochastic process. It has been applied to precipitation data, but it can be used in more general cases where data collected through time is available. This index turned out to be more theoretically attractive than the standardised precipitation index (SPI), but we also showed the advantages of it in practice (see Section 2.5). Lastly, given that advocate the use of a model-based approach, this index can be easily extended, at least theoretically, to consider temporal trends, to include covariates or to the spatio-temporal domain.

5.2.2 Spatial Item Factor Analysis

Spatial item factor analysis (SPIFA), presented in Chapter 4, is a novel approach to model and predict spatially structured multi-dimensional latent factors when the observed items are dichotomous. These models use concepts of item factor analysis and spatial statistics. In our application study, our method and inferential techniques enabled us to detect areas with high food insecurity in each dimension of this construct. Our method will allow scientists to obtain a deep understanding of the spatial, environmental and socioeconomic aspects of food insecurity and hopefully to inform and make better decisions in order to reduce it. The SPIFA model can be applied in more general examples, as highlighted in that chapter, which could lead to new scientific findings on the spatial structure of other latent constructs.

Our model can be easily extended depending of the interest of the study. For example, we propose an extension in the next subsections that could allow prediction of food insecurity in cities where the associated survey was not observed due to financial and logistic constraints.

5.2.3 Mixed Spatial Item Factor Analysis with Aggregated Covariates Effects

In this section we describe an extension to spatial item factor analysis (SPIFA), proposed in Chacon-Montalvan et al. (2018), by introducing random effects at the levels of items and also covariates that have been spatially aggregated at the level of latent factors.

This extension is motivated by the desire to predict food insecurity in several remote municipalities of Brazilian Amazonia, where resources only permitted us to conduct food insecurity surveys in only a few cities. Knowing that food insecurity is linked to socio-economic characteristics, we could use secondary data, like the census data, to link it to food insecurity thus enabling us to be able to predict this latent construct in other municipalities. Our extension will also handle the case that covariates from the secondary data might only be available at an aggregated level.

5.2.3.1 Model at Item Level

Let $Y_{kj}(s)$ be a binary random variable to item $j = 1, 2, \ldots, q$ in group (e.g. city) $k = 1, 2, \ldots, K$ at location s. These binary responses are modelled as discrete-state stochastic processes $\{Y_{kj}(s) : s \in D\}$, where $D \subset \mathbb{R}^2$, that take values 0 or 1 according to an auxiliary stochastic process $\{Z_{kj}(s) : s \in D\}$:

$$Y_{kj}(s) = \begin{cases} 1 & \text{if } Z_{kj}(s) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

In the SPIFA model, the structure of the auxiliary variables, $Z_{kj}(s) = c_j + \boldsymbol{a}_j^{*\intercal}\boldsymbol{\theta}(s) + \epsilon_j(s)$, includes the easiness parameters $\{c_j\}$ to account for the difference in the chance of endorsing each item, an interaction between the m-dimensional restricted discrimination parameter $\boldsymbol{a}_j^*$ and the m-dimensional latent factor $\boldsymbol{\theta}(s)$ to account for the influence of the latent factor to endorse item j, and an error term $\epsilon_j(s) \sim \mathcal{N}(0, 1)$ (Chacon-Montalvan et al., 2018). The discrimination parameters are restricted in the sense that some elements are set to zero and can be defined as $\boldsymbol{a}_j^* = \boldsymbol{L}_j\boldsymbol{a}_j$, where $\boldsymbol{a}_j$ is the vector of free discrimination parameters and $\boldsymbol{L}_j$ is a $m \times m$ matrix of ones and zeros that defines the structure imposed by the researcher. The l-element a_{jl}^* of the restricted discrimination parameters can be interpreted as the capacity of of item j to discriminate the dimension l of the latent factor $\boldsymbol{\theta}(s)$; as long as a_{jl}^* is far apart of zero, the capacity to discriminate increases.

We can modify the model imposed on the auxiliary variables $Z_{kj}(s)$ to account for other sources of variability. In our case, to consider the variability between groups (i.e. cities), we extend the structure of the auxiliary variables as

$$Z_{kj}(s) = c_j + \gamma_{kj} + [\boldsymbol{a}_j^* + \boldsymbol{\alpha}_{kj}^*]^\intercal \boldsymbol{\theta}_k(s) + \epsilon_{kj}(s), \quad \epsilon_{kj}(s) \sim \mathcal{N}(0, 1), \tag{5.2}$$

where c_j and γ_{kj} are the fixed and random easiness parameters respectively; the first indicates the global easiness for item j and the second, which is assumed to be normally distributed $\gamma_{kj} \sim \mathcal{N}(0, \sigma_\gamma^2)$, provides an additional increment on easiness in item j for belonging to group k. The m-dimensional continuous-space stochastic process $\{\boldsymbol{\theta}_k(s) : s \in D\}$ represent the multi-dimensional latent construct under interest in location s at group k. The latent factor has an effect on the auxiliary variable $Z_{kj}(s)$ through the sum of the fixed $\boldsymbol{a}_j^*$ and random $\boldsymbol{\alpha}_{kj}^*$ discrimination parameters. In a similar way to the easiness parameters, $\boldsymbol{a}_j^*$ represent the global discrimination parameter, while $\boldsymbol{\alpha}_{kj}^*$, assumed to be normally distributed as $\mathcal{N}(\boldsymbol{0}, \sigma_\alpha^2 \boldsymbol{I})$, is the additional increment on the discrimination for belonging to group k. The random discrimination parameters also need to be restricted such as $\boldsymbol{\alpha}_{kj}^* = \boldsymbol{L}_j \boldsymbol{\alpha}_{kj}$, where $\boldsymbol{\alpha}_{kj}$ is the free random discrimination parameters.

5.2.3.2 Model at Latent Factor Level

The structure imposed on the m-dimensional latent factor $\boldsymbol{\theta}_k(s)$ is the same model proposed in Chacon-Montalvan et al. (2018); the difference is that the parameters are going to vary with respect to each city. This structure considers three sources of variability; the effects of covariates, multivariate continuous spatial variation and multivariate residual variation. Then, the model for the latent factor in group k is defined as

$$\boldsymbol{\theta}_k(s) = \boldsymbol{B}^\intercal \boldsymbol{x}(s) + \boldsymbol{T}_k \boldsymbol{w}_k(s) + \boldsymbol{v}_k(s), \tag{5.3}$$

where $\boldsymbol{B}$ is an $p \times m$ matrix of slopes associating a set of standardized covariates $\boldsymbol{x}(s)$, non-necessarily observed, with the latent factor $\boldsymbol{\theta}_k(s)$. The spatial component of the latent factor is explained by a $m \times g$ linear transformation $\boldsymbol{T}_k$, whose sparsity is defined by the researcher, of a set $\boldsymbol{w}_k(s) = \{w_{kl}(s)\}_{l=1}^{g}$ of standardized, independent, stationary and isotropic Gaussian processes with correlation function $\rho_{kl}(u)$,

$$w_{kl}(s) \sim \mathrm{GP}(0, 1, \rho_{kl}(u)), \qquad k = 1, \ldots, K, \qquad l = 1, \ldots, g. \tag{5.4}$$

Finally, the m-dimensional random vector $\boldsymbol{v}_k(s)$ accounts for the remaining uncertainty in the latent factors that is neither explained by the covariates nor by $\boldsymbol{w}_k(s)$. It is assumed $\boldsymbol{v}_k(s)$ is a zero-mean multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}_{v_k}$,

$$\boldsymbol{v}_k(s) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{v_k}), \qquad k = 1, \ldots, K. \tag{5.5}$$

It is useful to decompose the covariance matrix $\boldsymbol{\Sigma}_{v_k} = \boldsymbol{D}_k \boldsymbol{R}_{v_k} \boldsymbol{D}_k$ in terms of a correlation matrix $\boldsymbol{R}_{v_k}$ and a diagonal matrix of standard deviations $\boldsymbol{D}_k$.

5.2.3.3 Model on Covariates at Aggregated Level

Notice that the covariates $\boldsymbol{x}(s) = (x_1(s), \ldots, x_p(s))^{\intercal}$ included in Equation 5.3 need to be on the continuous spatial scale at the locations where the items $Y_{kj}(s)$ have been observed. However, the data available could be at aggregated level (e.g. census sectors, post-codes, etc) or at individual level measured at different locations to the items. One possibility would be to model the spatial covariates and use this to predict these variables at any required (and reasonable) location s. Different types of models could be used for this, but we prefer to use models based on basis functions because they computationally cheaper when compared to other alternatives like Gaussian processes.

Let $X(s)$ denote the random variable associated with the covariate, or realization, $x_i(s)$ for $i = 1, \ldots, p$ and spatial location $s = (s_1, s_2)$ that can be modelled separately as an additive model

$$X(s) = f(s) + \epsilon(s) = \sum_{j=1}^{q_1}\sum_{l=1}^{q_2} \delta_{jl} b_{1j}(s_1) b_{2l}(s_2) + \epsilon(s), \tag{5.6}$$

where ϵs represents the perturbation term and the flexible function $f(s)$, defined on the spatial domain $s \in \mathcal{D}$, is expressed as a linear combination of the interaction $b_{jl}(s) = b_{1j}(s_1)b_{2l}(s_2)$ of the basis functions b_{1j} and b_{2j} defined for the axis s_1 and s_2 respectively. Notice that δ_{jl} defines the importance of the interaction of the basis functions in the spatial structure and are usually modelled using intrinsic Gaussian Markov random fields as a prior distribution such as

$$\Pr(\boldsymbol{\delta}) \propto \exp(-\frac{1}{2\tau^2}\boldsymbol{\delta}^\intercal \boldsymbol{P}\boldsymbol{\delta}), \tag{5.7}$$

where $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{q_1 q_2})^\intercal$, $\boldsymbol{P}$ is the penalty matrix and τ is the smoothing parameter (Rue and Held, 2005).

In case $X(s)$ has been observed individually, we can easily fit our model in Equation 5.6 and make prediction at locations where the items were observed. However, if the data has been observed at aggregated level we can consider the random variable $X(A)$ for the area A such as $X(A) = |A|^{-1} f(A) + \epsilon(s)$, where $f(A) = \int_{s\in A} f(s)ds$, leading to the model

$$X(A) = \sum_{j=1}^{q_1}\sum_{l=1}^{q_2} \delta_{jl} \left(|A|^{-1} \int_{s\in A} b_{1j}(s_1)b_{2l}(s_2)ds \right) + \epsilon(A), \tag{5.8}$$

considering $b^*_{jl}(A) = |A|^{-1} \int_{s\in A} b_{1j}(s_1)b_{2l}(s_2)ds$, the above equation has the same structure as an additive model where the basis functions are build based on the areas under study, so common approached for inference of additive models can be used. Evaluating $b^*_{jl}(s)$ is not straightforward, but Monte Carlo methods can be used to easily evaluate this function. Notice that δ_{jl} in remains unmodified with respect to the individual model in Equation 5.6, so they can be used to make inference of $X(s)$ at any location $s \in \mathcal{D}$.

Bibliography

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Chacon-Montalvan, E., Parry, L., Giorgi, E., Torres, P., Orellana, J., and Taylor, B. M. (2018). Spatial Item Factor Analysis With Application to Mapping Food Insecurity. *Arxiv*, pages 1–44.

Chacón-Montalván, E. A., Parry, L., Davies, G., and Taylor, B. M. (2018a). A Model-Based General Alternative to the Standardised Precipitation Index.

Chacón-Montalván, E. A., Parry, L., Torres, P., Orellana, J., Davies, G., and Taylor, B. M. (2018b). Evaluating the Effects of Extreme Hydro-climatic Events on Birth-weight in Amazonia.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.

Appendix A Matrix Form of the Mixed Model

In this section, we present the matrix form of the model at item and latent factor level. This is a key aspect to be able to make inference and prediction for our proposed mixed model.

Let $\boldsymbol{s}_k = \{s_{k1}, \ldots, s_{kn_k}\}$ be the set of n_k locations in group k, and $\theta_{kji} = \theta_{kj}(s_{ki})$ the j-dimension of the latent factor in group k at location s_{ki}. We denote the m-vector $\boldsymbol{\theta}_{k\cdot i} = \boldsymbol{\theta}_k(s_{ki}) = (\theta_{k1i}, \ldots, \theta_{kmi})^\intercal$, the n_k-vector $\boldsymbol{\theta}_{kj\cdot} = \boldsymbol{\theta}_{kj}(\boldsymbol{s}_k) = (\theta_{kj1}, \ldots, \theta_{kjn_k})^\intercal$, the mn_k-vector $\boldsymbol{\theta}_{k\cdot\cdot} = (\boldsymbol{\theta}_{k1\cdot}^\intercal, \ldots, \boldsymbol{\theta}_{km\cdot}^\intercal)^\intercal$, and the mn-vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_{1\cdot\cdot}^\intercal, \ldots, \boldsymbol{\theta}_{K\cdot\cdot}^\intercal)^\intercal$ with $n = \sum_{k=1}^{K} n_k$. These notations will be used for the multivariate processes considered in our mixed SPIFA model such as the response item $\boldsymbol{Y}_k(s)$, the auxiliary variables $\boldsymbol{Z}_k(s)$, the multivariate Gaussian process $\boldsymbol{w}_k(s)$, and the multivariate residual term $\boldsymbol{v}_k(s)$.

With the above conventions, the collection of auxiliary random variables $\boldsymbol{Z} = (\boldsymbol{Z}_{1\cdot\cdot}^\intercal, \ldots, \boldsymbol{Z}_{K\cdot\cdot}^\intercal)^\intercal$ for q items at n_1 locations in group 1, n_2 locations in group 2, so on, can be expressed as

$$\boldsymbol{Z} = \boldsymbol{\mathcal{I}}_c \boldsymbol{c} + \boldsymbol{\mathcal{I}}_\gamma \boldsymbol{\gamma} + \boldsymbol{\mathcal{A}} \boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{5.9}$$

where $\boldsymbol{c} = (c_1, \ldots, c_q)^\intercal$ is a vector arrangement of the fixed easiness parameters, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\intercal, \ldots \boldsymbol{\gamma}_K^\intercal)^\intercal$ is a vector arrangement of random easiness parameters with $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots \gamma_{kq})^\intercal$, $\boldsymbol{\theta}$ the vector of the latent factors as defined above, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{1\cdot\cdot}^\intercal, \ldots, \boldsymbol{\epsilon}_{K\cdot\cdot}^\intercal)^\intercal$ is a nq-vector of residual terms. The vectors $\boldsymbol{c}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are related with $\boldsymbol{Z}$ through the matrices $\boldsymbol{\mathcal{I}}_c$, $\boldsymbol{\mathcal{I}}_\gamma$ and $\boldsymbol{\mathcal{A}}$ respectively, which are defined as

$$\boldsymbol{\mathcal{I}}_c = (\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_1}^\intercal, \ldots, \boldsymbol{I}_q \otimes \boldsymbol{1}_{n_K}^\intercal)^\intercal,\ \boldsymbol{\mathcal{I}}_\gamma = \oplus_{k=1}^{K} (\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}),\ \boldsymbol{\mathcal{A}} = \oplus_{k=1}^{K} ([\boldsymbol{A}^* + \boldsymbol{\alpha}_k^*] \otimes \boldsymbol{I}_{n_k}), \tag{5.10}$$

where $\boldsymbol{I}_q$ and $\boldsymbol{I}_{n_k}$ are identity matrices of dimension q and n_k respectively, $\boldsymbol{1}_{n_k}$ is a n_k-dimensional vector with all elements equals to one, $\oplus$ represents a direct sum of matrices, $\boldsymbol{A}^* = (\boldsymbol{a}_1^*, \ldots, \boldsymbol{a}_q^*)^\intercal$ is a $q \times m$ matrix arrangement of the restricted fixed discrimination parameters, and $\boldsymbol{\alpha}_k^* = (\boldsymbol{\alpha}_{k1}^*, \ldots, \boldsymbol{\alpha}_{kq}^*)^\intercal$ is a $q \times m$ matrix arrangement of random discrimination parameters for city k.

The vector of latent factors $\boldsymbol{\theta}$ for m dimensions at n_1 locations in group 1, n_2 locations in group 2, so on, can be expressed as

$$\boldsymbol{\theta} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\mathcal{T}}\boldsymbol{w} + \boldsymbol{v}, \tag{5.11}$$

where $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B})$ is a column-vectorization of the multivariate fixed effects, $\boldsymbol{w} = (\boldsymbol{w}_{1\cdot\cdot}^\intercal, \ldots, \boldsymbol{w}_{K\cdot\cdot}^\intercal)^\intercal$ is the collection of the multivariate Gaussian processes and $\boldsymbol{v} = (\boldsymbol{v}_{1\cdot\cdot}^\intercal, \ldots, \boldsymbol{v}_{K\cdot\cdot}^\intercal)^\intercal$ is the collection of the multivariate residual terms. The terms $\boldsymbol{\beta}$ and $\boldsymbol{w}$ influence $\boldsymbol{\theta}$ through $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{T}}$ respectively, which are defined as

$$\boldsymbol{\mathcal{X}} = (\boldsymbol{I}_m \otimes \boldsymbol{X}_1^\intercal, \ldots, \boldsymbol{I}_m \otimes \boldsymbol{X}_K^\intercal)^\intercal, \qquad \boldsymbol{\mathcal{T}} = \oplus_{k=1}^K (\boldsymbol{T}_k \otimes \boldsymbol{I}_{n_k}), \tag{5.12}$$

where $\boldsymbol{X}_k = (\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kn})^\intercal$ is the $n_k \times p$ design matrix of the covariates for group k, and $\boldsymbol{T}_k$ as defined in Equation 5.3.

Equation 5.9 and 5.11 are very useful to obtained conditional posterior parameters, but not for the discrimination parameters. For this reason, it is also convenient to express the collection of auxiliary variables $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \boldsymbol{\mathcal{I}}_c \boldsymbol{c} + \boldsymbol{\mathcal{I}}_\gamma \boldsymbol{\gamma} + \boldsymbol{\vartheta}_a \boldsymbol{L} \boldsymbol{a} + \boldsymbol{\vartheta}_\alpha (\boldsymbol{I}_K \otimes \boldsymbol{L}) \boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{5.13}$$

where $\boldsymbol{a} = (\boldsymbol{a}_1^\intercal, \ldots, \boldsymbol{a}_q^\intercal)^\intercal$ is an mq-vector of the free fixed discrimination parameters and $\boldsymbol{\alpha} = (\mathrm{vec}(\boldsymbol{\alpha}_1), \ldots, \mathrm{vec}(\boldsymbol{\alpha}_K))^\intercal$ is an mqK-vector of the random discrimination parameters. $\boldsymbol{L} = \oplus_{j=1}^q \boldsymbol{L}_j$ is the direct sum of the activation matrices that constrain the fixed discrimination parameters. $\boldsymbol{a}$ and $\boldsymbol{\alpha}$ are related to Z through the matrices $\boldsymbol{\vartheta}_a$ and $\boldsymbol{\vartheta}_\alpha$, which are defined as

$$\boldsymbol{\vartheta}_a = (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_1^\intercal, \ldots, \boldsymbol{I}_q \otimes \boldsymbol{\Theta}_K^\intercal)^\intercal, \qquad \boldsymbol{\vartheta}_\alpha = \oplus_{k=1}^K (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_k) \tag{5.14}$$

where $\boldsymbol{\Theta}_k = (\boldsymbol{\theta}_{k1\cdot}, \ldots, \boldsymbol{\theta}_{km\cdot})$ is a $n_k \times m$ matrix of latent abilities for group k.

Appendix B Bayesian Inference

The model proposed in this paper assumes that n_k locations have been sample in group $k = 1, \ldots, K$ and that q items have been observed at each location. Using notation presented in Section A, we denote the collection of response variables for item j in city k as $\boldsymbol{Y}_{kj\cdot} = (Y_{kj1}, \ldots, Y_{kjn_k})^\intercal$, the collection of response variables in city k as $\boldsymbol{Y}_{k\cdot\cdot} = (\boldsymbol{Y}_{k1\cdot}^\intercal, \ldots, \boldsymbol{Y}_{kq\cdot}^\intercal)^\intercal$, and the collection of all response variables as $\boldsymbol{Y} = (\boldsymbol{Y}_{1\cdot\cdot}^\intercal, \ldots, \boldsymbol{Y}_{K\cdot\cdot}^\intercal)^\intercal$. Given that it is possible some items are missing at certain sample locations, we denote $\boldsymbol{Y}_{obs}$ as the set of response variables where the data has been observed and $\boldsymbol{Y}_{mis}$ the set of response variables where the data is missing.

B.1 Hierarchical Model

We factored the joint likelihood into three model hierarchies: the data level $\Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right)$, at the level of the auxiliary variables $\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$, and at the level of the latent factors $\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{R}_{v_k}\}\right)$, easiness random parameter $\Pr\left(\boldsymbol{\gamma} \mid \sigma_\gamma^2\right)$ and discrimination random parameter $\Pr\left(\boldsymbol{\alpha} \mid \sigma_\alpha^2\right)$. Our hierarchical model includes an additional level for the prior distributions of the parameters $\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\beta}, \boldsymbol{T}_k, \boldsymbol{\phi}_k, \boldsymbol{R}_{v_k}, \sigma_\gamma^2$ and σ_α^2. Hence, the posterior distribution of the model is

$$\begin{aligned}
&\Pr\left(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{R}_{vk}\}, \sigma_\gamma^2, \sigma_\alpha^2 \mid \boldsymbol{y}_{obs}\right) \propto \\
&\quad \Pr\left(\boldsymbol{y}_{obs} \mid \boldsymbol{z}_{obs}\right) \Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{R}_{v_k}\}\right) \Pr\left(\boldsymbol{\gamma} \mid \sigma_\gamma^2\right) \\
&\quad \Pr\left(\boldsymbol{\alpha} \mid \sigma_\alpha^2\right) \Pr\left(\boldsymbol{c}\right) \Pr\left(\boldsymbol{a}\right) \Pr\left(\boldsymbol{\beta}\right) \Pr\left(\{\boldsymbol{T}_k\}\right) \Pr\left(\{\boldsymbol{\phi}_k\}\right) \Pr\left(\{\boldsymbol{R}_{v_k}\}\right) \Pr\left(\sigma_\gamma^2\right) \Pr\left(\sigma_\alpha^2\right).
\end{aligned} \tag{5.15}$$

The advantage of this factorisation is that we marginalise the multivariate Gaussian process $\boldsymbol{w}$ and the multivariate residual term $\boldsymbol{v}$; in addition, the posterior conditional distribution of some parameters are analytically obtained.

B.2 Prior Distributions

The prior distributions of the parameters for our extension of the SPIFA model are defined as in Chacon-Montalvan et al. (2018), but the prior of some parameters is defined for each group in our extension.

The fixed easiness parameter $\boldsymbol{c}$, the fixed discrimination parameter $\boldsymbol{a}$ and the regression coefficients $\boldsymbol{\beta}$ are normally distributed with diagonal covariance matrices, where the prior mean of $\boldsymbol{c}$ and $\boldsymbol{\beta}$ is set to zero. A prior log-normal distribution is used for each free element of the sparse linear transformation $\boldsymbol{T}_k$ and for the scale parameters $\{\phi_{kl}\}$ of the correlation functions $\{\rho_{kl}(u)\}$. The correlation matrix $\boldsymbol{R}_{v_k}$ are assumed to have a prior LKJ distribution as defined in Lewandowski et al. (2009). Finally, we impose an inverse-gamma distribution $\mathcal{IG}(\cdot, \cdot)$ for the additional parameters of our extended model, σ^2_γ and σ^2_α.

B.3 Sampling from the Posterior Distributions

In the same way as in Chacon-Montalvan et al. (2018), we use a blocked Gibbs sampling to obtain samples from the posterior defined in Equation 5.15. The parameters that are not conjugate are sample together using adaptive Metropolis-Hastings (Andrieu and Thoms, 2008). Details of the conditional posteriors of our scheme can be found in the Appendix C.

B.4 Scaling Samples for Interpretation

In Chacon-Montalvan et al. (2018), restricting the standard deviations of the multivariate residual term $\boldsymbol{v}(s)$ is necessary to make the SPIFA model identifiable. However, it can not be ensured that the latent factors will be on the same scale leading to a loss of interpretation of the discrimination parameters $\boldsymbol{a}_j$. As proposed in Chacon-Montalvan et al. (2018), after the samples of the MCMC have been obtained, we can transform the parameters in order to obtain latent factors with expected variance equal to 1. We can then a diagonal matrix $\boldsymbol{Q}$ by filling the diagonal with the expected variances of the samples of the latent factors $\boldsymbol{\theta}(s)$. We then make the following transformations

$$\boldsymbol{a}_j \leftarrow \boldsymbol{Q}\boldsymbol{a}_j, \quad \boldsymbol{\alpha}_j \leftarrow \boldsymbol{Q}\boldsymbol{\alpha}_j, \quad \boldsymbol{\theta}_i \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{\theta}_i, \quad \boldsymbol{B} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{B}, \quad \boldsymbol{T} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{T}, \quad \boldsymbol{D} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{D}; \tag{5.16}$$

the correct interpretation of the parameters is then recovered.

Appendix C Markov chain Monte Carlo scheme sampling

As mentioned in Section B.3, we use Metropolis-within-Gibbs algorithm to obtain samples from the posterior distribution of our mixed SPIFA model with aggregated covariates. Our scheme sampling is just an extension to the scheme proposed in Chacón-Montalván et al. (2018b), where the parameters are updated by blocks. Details of the conditional posterior distribution for each block of parameters are shown below.

C.1 Auxiliary Variables

Using Equation 5.9, we can see that the joint density of the vector of auxiliary variables $\boldsymbol{Z}$ given given the fixed easiness parameters $\boldsymbol{c}$, the random easiness parameters $\boldsymbol{\gamma}$, the fixed discrimination parameters $\boldsymbol{a}$, the random discrimination parameters $\boldsymbol{\alpha}$ and the latent factors $\boldsymbol{\theta}$ is normally distributed with mean $\boldsymbol{\mathcal{I}}_c\boldsymbol{c}+\boldsymbol{\mathcal{I}}_\gamma\boldsymbol{\gamma}+\boldsymbol{\mathcal{A}}\boldsymbol{\theta}$ and identity covariance matrix,

$$\Pr(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mathcal{I}}_c\boldsymbol{c} + \boldsymbol{\mathcal{I}}_\gamma\boldsymbol{\gamma} + \boldsymbol{\mathcal{A}}\boldsymbol{\theta}, \boldsymbol{I}_{nq}). \tag{5.17}$$

For sampling, it is required to differentiate between the auxiliary variables where the response items were observed $\boldsymbol{Z}_{obs}$ and where the response items could not been observed $\boldsymbol{Z}_{mis}$. Given that these two random vectors are conditionally independent in Equation 5.17, we can sample $\boldsymbol{Z}_{obs}$ from

$$\Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) \propto \Pr\left(\boldsymbol{z}_{obs} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) \prod_{o_{kij}=1} \Pr\left(y_{kji} \mid z_{kji}\right), \quad (5.18)$$

which is a marginal truncated normal distribution obtained from Equation 5.17. The truncation direction is defined by $\Pr\left(y_{kji} \mid z_{kji}\right) = \mathbb{1}^{y_{kji}}_{\left(z_{kji}>0\right)} \mathbb{1}^{1-y_{kji}}_{\left(z_{kji}\leq 0\right)}$, where $\mathbb{1}_{(.)}$ is the indicator function and o_{kij} take value 1 when the item j at location i of group k has been observed. The conditional posterior distribution of the auxiliary variables when the item are not observed, $o_{kji} = 0$, is

$$\Pr\left(\boldsymbol{z}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) \propto \Pr\left(\boldsymbol{z}_{mis} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) \quad (5.19)$$

which is simply a marginal distribution of Equation 5.17.

C.2 Latent Factors

From Equation 5.11, it can be seen that the vector of latent factors is normally distributed with a block diagonal covariance matrix,

$$\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{R}_{v_k}\}\right) = \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{\mathcal{X}}\boldsymbol{\beta}, \oplus_{k=1}^{K} \boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}}\right), \quad (5.20)$$

where $\oplus$ represent the direct sum of matrices and the covariance matrix of $\boldsymbol{\theta}_{k\cdot\cdot}$ is

$$\boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}} = (\boldsymbol{T}_k \otimes \boldsymbol{I}_{n_k})\boldsymbol{\Sigma}_{w_{k\cdot\cdot}}(\boldsymbol{T}_k^{\intercal} \otimes \boldsymbol{I}_{n_k}) + \boldsymbol{D}_k \boldsymbol{R}_{v_k} \boldsymbol{D}_k \otimes \boldsymbol{I}_{n_k}, \quad (5.21)$$

with $\boldsymbol{\Sigma}_{w_{k\cdot\cdot}} = \oplus_{l=1}^{g} \boldsymbol{\Sigma}_{w_{kl\cdot}}$.

The conditional posterior distribution of the latent factors $\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ is proportional to the product of two Gaussian densities, $\Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{R}_{vk}\}\right)$ and $\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$, resulting in another Gaussian density with block diagonal covariance matrix $\boldsymbol{\Sigma}_{\theta|\cdot} = \oplus_{k=1}^{K}\boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}|\cdot}$, where

$$\boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}|\cdot} = \left((\boldsymbol{A}^{*\intercal} + \boldsymbol{\alpha}_k^{*\intercal})(\boldsymbol{A}^{*} + \boldsymbol{\alpha}_k^{*}) \otimes \boldsymbol{I}_{n_k} + \boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}}^{-1}\right)^{-1},$$

and mean $\boldsymbol{\mu}_{\theta|\cdot} = (\boldsymbol{\mu}_{\theta_{1\cdot\cdot}|\cdot}^{\intercal}, \ldots, \boldsymbol{\mu}_{\theta_{K\cdot\cdot}|\cdot}^{\intercal})^{\intercal}$, where

$$\boldsymbol{\mu}_{\theta_{k\cdot\cdot}|\cdot} = \boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}|\cdot}\left[([\boldsymbol{A}^{*\intercal} + \boldsymbol{\alpha}_k^{*\intercal}] \otimes \boldsymbol{I}_{n_k})(\boldsymbol{z}_{k\cdot\cdot} - (\boldsymbol{c} + \boldsymbol{\gamma}_k) \otimes \boldsymbol{1}_{n_k}) + \boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}}^{-1}(\boldsymbol{I}_m \otimes \boldsymbol{X}_k)\boldsymbol{\beta}\right].$$

C.3 Fixed Regression Effects

The conditional posterior distribution of the multivariate fixed effects $\boldsymbol{\beta}$ given $\boldsymbol{\theta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}$ and $\{\boldsymbol{R}_{v_k}\}$ is proportional to the product of two Gaussian densities, $\Pr(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \{\boldsymbol{T}_k\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{R}_{v_k}\})$ and $\Pr(\boldsymbol{\beta})$, leading to another Gaussian density with covariance matrix and mean:

$$\boldsymbol{\Sigma}_{\beta|\cdot} = \left(\sum_{k=1}^{K}(\boldsymbol{I}_m \otimes \boldsymbol{X}_k^{\intercal})\boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}}^{-1}(\boldsymbol{I}_m \otimes \boldsymbol{X}_k) + \boldsymbol{\Sigma}_{\beta}^{-1}\right)^{-1}, \quad \boldsymbol{\mu}_{\beta|\cdot} = \boldsymbol{\Sigma}_{\beta|\cdot}\left(\sum_{k=1}^{K}(\boldsymbol{I}_m \otimes \boldsymbol{X}_k^{\intercal})\boldsymbol{\Sigma}_{\theta_{k\cdot\cdot}}^{-1}\boldsymbol{\theta}_{k\cdot\cdot}\right).$$

C.4 Fixed Easiness parameters

The conditional posterior distribution of the fixed easiness parameters $\Pr(\boldsymbol{c} \mid \boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is also proportional to the product of two Gaussian densities, $\Pr(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ and $\Pr(\boldsymbol{c})$, that leads to another Gaussian density with covariance matrix

$$\boldsymbol{\Sigma}_{c|\cdot} = \left(\sum_{k=1}^{K}(\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}^{\intercal})(\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}) + \boldsymbol{\Sigma}_c^{-1}\right)^{-1} = \left(n + \operatorname{diag}(\boldsymbol{\Sigma}_c)^{-1}\right)^{-1}, \tag{5.22}$$

and mean

$$\boldsymbol{\mu}_{c|\cdot} = \boldsymbol{\Sigma}_{c|\cdot}\left(\sum_{k=1}^{K}(\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}^{\intercal})(\boldsymbol{z}_{k\cdot\cdot} - \boldsymbol{\gamma}_k \otimes \boldsymbol{1}_{n_k} - ([\boldsymbol{A}^{*} + \boldsymbol{\alpha}_k^{*}] \otimes \boldsymbol{I}_{n_k})\boldsymbol{\theta}_{k\cdot\cdot})\right). \tag{5.23}$$

C.5 Random Easiness parameters

The conditional posterior distribution of the random easiness parameters $\boldsymbol{\gamma}$ given $\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}$ and σ_γ^2 is also proportional to the product of two Gaussian densities, $\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$ and $\Pr\left(\boldsymbol{\gamma} \mid \sigma_\gamma^2\right)$, that leads to another Gaussian density with covariance matrix $\boldsymbol{\Sigma}_{\gamma|\cdot} = \oplus_{k=1}^{K} \boldsymbol{\Sigma}_{\gamma_k|\cdot}$, where

$$\boldsymbol{\Sigma}_{\gamma_k|\cdot} = \left((\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}^\intercal)(\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}) + \sigma_\gamma^{-2} \boldsymbol{I}_q\right)^{-1} = \left(n_k + \sigma_\gamma^{-2}\right)^{-1} \boldsymbol{I}_q, \tag{5.24}$$

and mean $\boldsymbol{\mu}_{\gamma|\cdot} = (\boldsymbol{\mu}_{\gamma_1|\cdot}^\intercal, \ldots, \boldsymbol{\mu}_{\gamma_K|\cdot}^\intercal)^\intercal$, where

$$\boldsymbol{\mu}_{\gamma_k|\cdot} = \boldsymbol{\Sigma}_{\gamma_k|\cdot}(\boldsymbol{I}_q \otimes \boldsymbol{1}_{n_k}^\intercal)(\boldsymbol{z}_{k\cdot\cdot} - \boldsymbol{c} \otimes \boldsymbol{1}_{n_k} - ([\boldsymbol{A}^* + \boldsymbol{\alpha}_k^*] \otimes \boldsymbol{I}_{n_k})\boldsymbol{\theta}_{k\cdot\cdot}). \tag{5.25}$$

C.6 Fixed Discrimination parameters

Similar to previous parameters, the conditional posterior distribution of the discrimination parameters $\Pr\left(\boldsymbol{a} \mid \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$ is proportional to the product between $\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$ and $\Pr\left(\boldsymbol{a}\right)$. This is a Gaussian density with covariance matrix

$$\boldsymbol{\Sigma}_{a|\cdot} = \left(\boldsymbol{L}^\intercal \left(\boldsymbol{I}_q \otimes \sum_{k=1}^{K} \boldsymbol{\Theta}_k^\intercal \boldsymbol{\Theta}_k\right) \boldsymbol{L} + \boldsymbol{\Sigma}_a^{-1}\right)^{-1},$$

and mean

$$\boldsymbol{\mu}_{a|\cdot} = \boldsymbol{\Sigma}_{a|\cdot}\boldsymbol{L}^\intercal \left(\sum_{k=1}^{K} (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_k^\intercal)(\boldsymbol{z}_{k\cdot\cdot} - (\boldsymbol{c} + \boldsymbol{\gamma}_k) \otimes \boldsymbol{1}_{n_k} - (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_k)\boldsymbol{\alpha}_k)\right) + \boldsymbol{\Sigma}_{a|\cdot}\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a.$$

C.7 Random Discrimination parameters

Similar to previous parameters, the conditional posterior distribution of the discrimination parameters $\Pr\left(\boldsymbol{\alpha} \mid \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$ is proportional to the product between $\Pr\left(\boldsymbol{z} \mid \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$ and $\Pr\left(\boldsymbol{\alpha} \mid \sigma_\alpha^2\right)$. This is a Gaussian density with covariance matrix $\boldsymbol{\Sigma}_{\alpha|\cdot} = \oplus_{k=1}^{K} \boldsymbol{\Sigma}_{\alpha_k|\cdot}$, where

$$\boldsymbol{\Sigma}_{\alpha_k|\cdot} = \left(\boldsymbol{L}^{\intercal}\left(\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_k^{\intercal}\boldsymbol{\Theta}_k\right)\boldsymbol{L} + \sigma_{\alpha}^{-2}\boldsymbol{I}_{qm}\right)^{-1},$$

and mean $\boldsymbol{\mu}_{\alpha|\cdot} = (\boldsymbol{\mu}_{\alpha_1|\cdot}^{\intercal}, \ldots, \boldsymbol{\mu}_{\alpha_K|\cdot}^{\intercal})^{\intercal}$, where

$$\boldsymbol{\mu}_{\alpha_k|\cdot} = \boldsymbol{\Sigma}_{\alpha_k|\cdot}\boldsymbol{L}^{\intercal}(\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_k^{\intercal})(\boldsymbol{z}_{k\cdot\cdot} - (\boldsymbol{c} + \boldsymbol{\gamma}_k) \otimes \boldsymbol{1}_{n_k} - (\boldsymbol{I}_q \otimes \boldsymbol{\Theta}_k)\boldsymbol{a}).$$

C.8 Variances of the Random Effects

The posterior conditional distribution of the variance of the random easiness parameters σ_{γ}^2 and the random discrimination parameters σ_{α}^2 are

$$\begin{aligned}\Pr\left(\sigma_{\gamma}^2 \mid \boldsymbol{\gamma}\right) &= \mathcal{IG}(a_{0_{\gamma}} + qK/2, b_{0_{\gamma}} + \boldsymbol{\gamma}^{\intercal}\boldsymbol{\gamma}/2),\\ \Pr\left(\sigma_{\alpha}^2 \mid \boldsymbol{\alpha}\right) &= \mathcal{IG}(a_{0_{\alpha}} + qmK/2, b_{0_{\alpha}} + \boldsymbol{\alpha}^{\intercal}\boldsymbol{\alpha}/2),\end{aligned}$$

where $a_{0_{\gamma}}$, $b_{0_{\gamma}}$, $a_{0_{\alpha}}$ and $b_{0_{\alpha}}$ are the hyper-parameters of the prior $\mathcal{IG}$ distributions.

C.9 Covariance parameters

Let vec$^*(.)$ denote a vector of the free elements in (.). The conditional posterior distribution of the parameters vec$^*(\boldsymbol{T}_k)$, $\boldsymbol{\phi}_k$ and $\boldsymbol{R}_{v_k}$,

$$\Pr\left(\text{vec}^*(\boldsymbol{T}_k), \boldsymbol{\phi}_k, \boldsymbol{R}_{v_k} \mid \boldsymbol{\theta}_{k\cdot\cdot}, \boldsymbol{\beta}\right) \propto \Pr\left(\boldsymbol{\theta}_{k\cdot\cdot} \mid \boldsymbol{\beta}, \boldsymbol{T}_k, \boldsymbol{\phi}_k, \boldsymbol{R}_{v_k}\right)\Pr\left(\boldsymbol{T}_k\right)\Pr\left(\boldsymbol{\phi}_k\right)\Pr\left(\boldsymbol{R}_{v_k}\right),$$

can not be solved analytically. We obtain samples from this posterior using an adaptive random-walk Metropolis Hastings algorithm, algorithm 4 of Andrieu and Thoms (2008), on transformed variables, log(vec$^*(\boldsymbol{T}_k)$), log($\boldsymbol{\phi}_k$) and $\boldsymbol{R}_{v_k}^*$, whose each dimensional domain is on the real line $\mathbb{R}$. The transformation $\boldsymbol{R}_{v_k}^*$ is done using canonical partial correlations (Lewandowski et al., 2009).