UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXI

# Bayesian hierarchical modelling for population size estimation: application to Italian data

**Coordinatore del Corso:** Prof. Nicola Sartori

**Supervisore:** Prof. Brunero Liseo

**Dottoranda:** Charlotte Taglioni

$1^{st}$ October 2018

# Abstract

Bayesian demography developments, global trends for substituting traditional censuses with cheaper methods able to use available information, and new technologies require investigating and providing new models to answer new requirements. In Italy in particular, during the last years Istat worked for launching in October 2018 the "permanent census of population and housing". After a first discussion on censuses, changes recommended by organisations such as the UN and the European Union to the National Statistical Institutes, and on new demographic models for population size estimation, the model proposed by Bryant and Graham (2013) is analysed. The model allows for integration of different data sources, for demographic series estimation, and it is very flexible and complex at the same time. Applications of this model to the Italian population are performed, highlighting its advantages and limits. Data for the period considered (2006-2015) and metadata come from Istat. Data are not always consistent, confirming the need of statistical methods able to integrate sources and reconstruct demographic series. As expected, census data and migration flows estimation caused most of the problems. The method still needs further experimentations, therefore applications aim to compare results when varying initial assumptions and to identify their pros and cons rather than provide actual results on the Italian population. Eventually a model extension, along with the first results of its application, is proposed using the Conway-Maxwell Poisson distribution (Conway and Maxwell, 1962), a flexible two parameters version of the Poisson distribution.

# Sommario

Con i recenti sviluppi della demografia Bayesiana, la tendenza globale a sostituire i tradizionali censimenti della popolazione con metodi più economici, capaci di sfruttare la grande quantità di informazioni disponibile e le nuove tecnologie, si rende necessario esaminare e fornire nuovi metodi in grado di rispondere a queste nuove esigenze. In Italia in particolare, negli ultimi anni l'Istat ha lavorato per il lancio ad ottobre 2018 del "censimento permanente della popolazione e delle abitazioni". Dopo una prima discussione sui censimenti, i cambiamenti raccomandati agli Istituti nazionali di statistica da organizzazioni internazionali come l'ONU e dall'Unione Europea, e gli sviluppi dei modelli demografici per la stima della popolazione, si approfondisce il metodo proposto da (Bryant and Graham, 2013). Il modello consente l'integrazione di varie fonti per la stima delle serie demografiche e si caratterizza per notevole flessibilità unita ad un'elevata complessità. Di questo modello si presentano varie applicazioni alla popolazione italiana con lo scopo evidenziarne vantaggi e limiti. I dati per il periodo di riferimento (2006-2015) e i metadati provengono direttamente dall'Istat. I dati non sempre sono coerenti tra loro, confermando la necessità di servirsi di metodi statistici per integrare fonti e ricostruire le serie demografiche. In linea con quanto previsto, l'inclusione dei dati censuari e la stima dei flussi migratori si sono rivelati particolarmente problematici. Il metodo è ancora in fase di sperimentazione, perciò le applicazioni si propongono di confrontare i risultati al variare delle assunzioni e di evidenziarne pro e contro piuttosto che fornire risultati quantitativi sulla popolazione italiana. Infine si propone un'estensione del modello con l'uso della distribuzione di Conway-Maxwell Poisson (Conway and Maxwell, 1962), dotata di elevata flessibilità, e se ne presentano i primi risultati.

"The Lord says: Look at the new thing I am going to do.
It is already happening. Don't you see it? I will make
a road in the desert and rivers in the dry land."

Isaiah 43:19

"Defenceless under the night
Our world in stupor lies;
Yet, dotted everywhere,
Ironic points of light
Flash out wherever the Just
Exchange their messages:
May I, composed like them
Of Eros and of dust,
Beleaguered by the same
Negation and despair,
Show an affirming flame."

September 1, 1939
W. H. Auden

*To my points of light*

# Acknowledgements

I would like to thank my supervisors, Prof. Brunero Liseo for giving me the chance to do all this and for his support and advice. Prof. John Bryant for the time he spent helping me, his support and enthusiasm and for our productive discussions. Prof. Patrick Graham for his advice. The whole Department of Statistical Sciences of the University of Padua and PhD colleagues I shared the journey with. The University of Rome "La Sapienza" and PhD students I met there, it is always good to "come back home". Professors and students of the European Doctorate School of Demography 2016-2017 for the passion they have for this subject. The School of Mathematics and Statistics of the University of Canterbury in New Zealand for welcoming me, and in particular Prof. Marco Reale for all his help and generosity. The Istat research group I have been in contact with for their help with data.

I thank my family and friends for all their support throughout the whole journey, when I was far, closer and further, they are my bedrock. I thank God for all this.

# Contents

# List of Figures

# List of Tables

# Introduction

## Overview

During the last decades Bayesian methods have been applied to population size estimation mainly addressing problems of population projection and reconstruction, poor data coverage and missing data, and data source integration. After a review on the main developed methods and progress in official statistics, the thesis focuses on the model proposed by Bryant and Graham (2013), studying its strengths and limitations and applying the method to Italian data to estimate the Italian population.

The area of Bayesian demography is currently growing fast as its framework suits well to the uncertainty underlying population estimation and forecasting problems. Traditional demographic models mostly oversimplify the complexity of demographic phenomena such as the dynamic nature of real demographic systems, bias due to integration of sources with different levels of completeness and reliability, and "uncertainty arising from incomplete knowledge of historical trends or causal mechanisms, or from random variation in disaggregated counts." (Bijak and Bryant, 2016). Demography has long been considered as a subject mostly relying on deterministic data, especially when talking about official statistics and census data which, in theory, should have the best coverage given the wide access to information and public offices cooperation all over the country.

The notion of uncertainty related to demographic quantities is rather new or, more precisely, it has been explicitly introduced and included in demographic studies only recently.

Reasons for joining probabilistic statements to demographic data comes from different factors. First of all, data problems like bias, under- or over-coverage and sparseness are more and more evident and are increasingly addressed by the scientific community. It is important to consistently deal with such issues and to explicitly point out what is their impact on the results. Secondly, data sources and their accessibility are increasing and the ability to combine, integrate and organise them has become a key point. A third aspect concerns outputs typically given in demographic estimations and forecasts from official sources. Their interpretation and the underlying hypotheses are not always clear and can be questionable. For example, UN projections used to assume convergence of mortality, fertility and migration but this assumption is itself not certain and this was not taken into account in projections (World Population Prospectus `data.un.org`). The UN have now changed their approach and use Bayesian methods. Eventually, *a priori* information not directly coming from data (e.g. expert opinions) could help the inference process if included in the model, especially when data are sparse, incomplete or biased. The Bayesian framework is therefore particularly appropriate to deal with these issues as it allows for inclusion of prior information and provides uncertainty estimates.

The main critique to methods typically used by organisations making projections on population, like United Nations or National statistical institutes (NSIs), is that they provide results for different scenarios but the probability of each scenario is not always clear and projections do not clearly reflect the corresponding uncertainty. Furthermore, the quality of data used for projections or estimations is only indirectly accounted for (Daponte *et al.* (1997), Abel *et al.* (2010)). Non-Bayesian examples of progress in this sense can be found in Lee and Carter (1992) where confidence intervals are added to projections; stochastic approaches of Lee and Tuljapurkar (1994), Tuljapurkar and Lee (1997) , Tuljapurkar and Boe (1999), where some quantities are considered as known, ignoring confidence intervals; or in Pflaumer (1988) where the importance of prior knowledge in choosing the demographic distribution is stated even if not incorporated in the model.

Bayesian methods naturally account for uncertainty and give probabilistic results; they also incorporate prior information and several data sources. Examples of population reconstructions and projections are Raftery *et al.* (2012); Rendall *et al.* (2009); Wheldon *et al.* (2012, 2016).

The method proposed in Bryant and Graham (2013) and discussed more in details in Bryant and Zhang (2018) makes possible to analyse the population according to structural parameters as region, sex, age and year, along with the possibility to add covariates or other dimensions. It also applies both to single demographic series (like births, deaths or migration counts) and to the whole demographic account.

The demographic account is defined as a “complete description of the demographic stocks and flows of interest, subject to accounting identities that relate stocks to flows” (Bryant and Graham, 2013) and it is a tool initially used in economics. Classical discussions on population accounting models are the Nobel Memorial Lecture 1984: “The Accounts of Society” reported in Stone (1986) and Rees (1979) where a simple account model is firstly introduced, followed by related issues, limitations and assumptions, along with alternatives and improvements. A case study on regions of East Anglia, the South East, the Rest of Britain and the Rest of the World is also included.

The demographic account model proposed by Bryant and Graham (2013, 2015) is a complex population size estimation model. Taking advantages on the reliability and abundance of data in New Zealand, they use multiple data sources and combine administrative data and official statistics to estimate the population of six different regions in the country.

The demographic account used as example is a four-dimensional array where dimensions are age, region, sex and time. Bryant and Graham divide their model in two parts: the "system model" and the "data model". The first one is based on the demographic account itself treating as sub-models counts of population, births, deaths, internal and external in and out migration; it is meant to catch regularities in the demographic series, and, potentially, to link them to external characteristics (covariates). Parameters of each sub-model are assumed *a priori* independent. The data model explains the relationship between the datasets and the demographic account and is chosen according to data accuracy; tasks previously accomplished by experts, such as data accuracy evaluation or systematic biases, are embedded in the model as *a priori* information; missing data do not represent an issue as "the model *predicts* the contents of each dataset from the contents of the demographic account and the corresponding data model" (Bijak and Bryant, 2016). Model complexity comes from (i) the diversity of data sources used; (ii) the application of a probabilistic model to both data sources for coverage errors and to counts themselves; (iii) the high adaptability to the specific data despite the common form of the model (a hierarchical Poisson-Log-normal model); (iv) the constraints required for demographic account consistency. An R package (`demest`) for Bryant and Graham (2013) demographic account model has been implemented but is still under revision and, according to the Authors, it still needs one or two years of testing and tuning before being ready for a broad use.

# Main contributions of the thesis

The Italian NSI (Istat) showed interest in Bryant and Graham (2013) method and started investigating it (Toti *et al.*, 2017). The model satisfies the requirements for population size estimation aiming to implement the project of "permanent census of population and housing" Istat has been planning during the last years. Therefore, with the support of an Istat research group, the thesis provides an application of Bryant and Graham (2013) method to Italian data. The model is very flexible and adapts to different data sources. The application to the Italian case adds to and extends the two examples of demographic account estimation presented in Bryant and Zhang (2018) for New Zealand and China. These two latter applications mainly differ for data quality, if New Zealand has very high standards because of its relatively small population and the control on immigration, in China registers show major internal contradictions that heavily affect the estimation (Bryant and Zhang, 2018). In Italy, Istat keeps high data quality standards according to international organisation recommendations. Main problems affecting data quality are restrictions from privacy laws and migration estimation. Privacy is a delicate topic, on the one hand European and Italian Institutions constantly work to control, enhance and safeguard citizens' rights; on the other hand this laws affect and limit not only illicit procedures but also official statistics tasks. Application to the Italian data allowed to extensively test the model and highlighted both its advantages and limitations. Moreover, the necessity to satisfy the population balance equations showed some Italian data inconsistencies. In this sense, the thesis contributed to the study of the model and potentially to the implementation of permanent census. The results obtained show trend and main dimensions driving the Italian population change during the period 2006-2015 and they also give an idea of the coverage of data sources.

An extension of the model is also proposed so that the model can account for heterogeneity of the observed population. The Poisson distribution is a natural choice for counting people as it is discrete and has good properties but it assumes equi-dispersion of data, implying from a demographic point of view that the population modelled is homogeneous. To overcome this assumption, the Conway-Maxwell Poisson (CMP) distribution has been introduced as a possible data model. This distribution had a "revival" in Shmueli *et al.* (2005) after it was initially proposed in Conway and Maxwell (1962). The CMP distribution allows for modelling both under-dispersion and over-dispersion. In one hand the extension adds flexibility to the model and gives a measure of population data homogeneity, on the other hand this extension further complicates the Bryant and Graham (2013) model due to higher number of parameter to estimate when considering the CMP.

Chapter 1 discusses censuses origin and methods, new requirements and recommendations, the challenges National statistical institutes (NSIs) have to face and how they are changing. A focus on Istat is also presented. Eventually a review on Bayesian methods applied to demography is included. Chapter 2 introduces, explains, comments and proposes possible developments for the model initially proposed by Bryant and Graham (2013). Chapter 3 contains applications to Italian data, comments and results for single demographic series and for the demographic account estimation. Chapter 4 introduces the CMP distribution, its inclusion in the demographic account model (DAM) and problems linked to it, and first results from the application to birth and death counts. Discussion and future directions of research form the Conclusion chapter.

# Chapter 1

# Bayesian demography and Official Statistics challenges

## 1.1 Bayesian statistics applied to demography [1]

The first application of Bayesian statistics to demography dates back to the end of the XVIII century when Laplace (1781) applied Bayesian inference to estimate sex ratios at birth in Paris and London, not log after the publication of Bayes theorem (Bayes, 1763). After its introduction, Bayesian statistics and, consequently, its applications to demography have hardly been used until technology and new theory (e.g. Markov Chains Monte Carlo methods) allowed to overcome computational and theoretical issues related to Bayesian techniques. The interest in Bayesian demography rose again in the late 1990s (Daponte *et al.*, 1997) and it is growing fast, especially after the UN adopted a Bayesian probabilistic approach for its projections (Gerland *et al.*, 2014).

The Bayesian and the classical (frequentist) approaches present differences from many points of view. First of all, from a philosophical perspective, the definition of probability changes. Bayesians relate probability to subjective beliefs whether for Frequentists an event's probability is the limit of its relative frequency in a large number of trials. Bayesian theory is based on Bayes theorem whereas classical inference is mainly based on the likelihood; the outcome of classical statistics is a point estimate with standard error, whereas in Bayesian statistics a whole probability distribution is available increasing the amount of information available, e.g. multiple modes, probability mass concentration, and, in general, all quantities derivable from a full probability distribution; the interpretation of credible intervals in Bayesian statistics is different from classical confidence intervals. Considering the usual 95% interval, when considering a 95% credible

[1] This introductory section is widely based on Bijak and Bryant (2016)

intervals it is correct to say that the interval contains the true value with a probability of 95%; with confidence intervals this interpretation is not right. In classical statistics the 95% "refers to the proportion of hypothetical confidence intervals that would contain the true value if the study were replicated many times" (Bijak and Bryant, 2016). Typical of Bayesian statistics is the use of prior distributions meant to embed *a priori* information not contained in the data. Prior distributions can be informative, weakly informative or even non-informative depending on the data and the approach. When there are little data, or they are sparse, it is common to use informative priors as very little information is contained in the data. If data provide a satisfying level of information usually a weakly informative prior is used mostly providing qualitative information about data as trend or neighbouring similarities (as following age groups, similar incomes or neighbouring regions). When the prior is non-informative, the whole analysis is mostly likelihood driven and Bayesian results are very similar to Frequentist ones. Very important for the model presented in chapter 2 (Bryant and Graham, 2013) and often used in Bayesian models are the hierarchical priors, i.e. prior parameters are given hyper-prior distributions with hyper-prior parameters. Equivalent Frequentist versions of Bayesian hierarchical models are multilevel models or random effects models. Nevertheless, the Bayesian framework seems to be a natural way to approach problems hierarchically as parameters are treated as random variables and always have their own probability distributions whereas they are point estimates in the classical framework. Hierarchical models are introduced in section 2.1.1.

The main critic Frequentists have towards the Bayesian method is the subjectivity introduced in the model through the prior distribution. The prior is chosen by the user according to prior beliefs on the data and it can influence results.
From a Bayesian point of view the prior is a transparent way to introduce *a priori* choices that are unavoidable in any statistics. Also using the classical approach, there are choice that can be considered as subjective, e.g. the choice of error distributions or what method to use to deal with missing values. Moreover, models are often robust to the prior choice and the influence of the prior decreases as the amount and the quality of data increase. The prior distribution allows for an explicit introduction of additional information in the model, if available. The level of informativeness can be chosen and robustness to this choice can be tested.

Many of the typical aspects of Bayesian statistics are suitable for application to demography. The availability of a whole probability distribution instead of a single point estimate and the suitability of the Bayesian framework for hierarchical modelling is useful for demographic phenomena as these type of model are popular both in demography and ecology. With hierarchical modelling complex phenomena as social ones can be described, it is possible to include information at individual level, group level or population level and account for all of them in the same model. In many cases, and particularly in demography, the influence of prior distributions can be tuned. On the one hand it can be reduced either by the choice of the distribution itself or letting its influence on the posterior calculation naturally fading increasing data information and quantity. This last option is a viable choice in demography as population studies typically rely on empirical data and, especially since census-based information are available, this abundance of data can be used to reduce the impact of the prior. On the other hand, it is not rare in demography to use either different sources, or *a priori* knowledge or to "borrow" data from different populations than the one analysed to improve the model. This is possible as there are demographic processes presenting similarities across different populations and because demographers have knowledge that, despite not being available from the data, can be included in the prior and help the estimation process when data are sparse or poor. Also, with Bayesian methods, uncertainty measures are provided along with the parameter estimations, turning typical deterministic projections into probabilistic ones. The inclusions of uncertainty in social sciences is of major interest as source of uncertainty are multiple and models have to account for it in order to provide a more complete picture of reality and not be limited only to qualitative indications of reliability but to rely on quantitative ones.

Bijak and Bryant (2016) also highlight the advantages Bayesian statistics can obtain from demographical applications. Demography has a "strong empirical orientation" (Bijak and Bryant, 2016), its applications have important political consequences in terms of social and economic policies, and it has developed methods and solutions that might be used for other fields. For Bayesian statistics, it represents a good opportunity for new applications, and Bayesians can take advantage of these peculiarities and develop new methods starting from demographical models.

The exchange of knowledge and methods between Bayesian statistics and demography benefits both fields and provides new challenges and opportunities. Application of Bayesian methods to demography provides new tools to support traditional demographic methods allowing for constant updating and incorporation of different information (prior beliefs, expert opinions, quantitative and qualitative data). Bayesian methods role is not to compete with older methods but rather to support and complement them.

Bijak and Bryant (2016) point out three main areas of application of Bayesian method to demography and provide examples for each of them: (i) forecasting, (ii) limited data; (iii) structured and complex models. In the following section a review of models for population estimation or reconstruction is presented particularly focusing on studies related to official statistics.

## 1.2 A review of population estimation through Bayesian methods

Application of Bayesian methods to demography is a convenient partnership for both areas. Demographers can quantify via probability distribution the uncertainty of their results, combine different data sources, and have a suitable tool to deal with multilevel analysis. Especially Bayesian hierarchical models allow for information exchange on unknown quantities across regions, "based on the assumption that these quantities are drawn from a common probability distribution" (Alkema *et al.*, 2015). For Bayesian statisticians, demography is a subject providing a good opportunity to test Bayesian methods and also a policy relevant area of application; demographic models can also enrich the Bayesian landscape and be a starting point for new developments (Bijak and Bryant, 2016). In addition to these reasons pointed out in the last section, the application of Bayesian methods to statistics has been a solution to a critique to the usual way of making projections. One of the main critiques to methods typically used by organisations making projection on population and for population size estimation (e.g. United Nations or Census offices) is that they make assumptions and give projections according to scenarios that do not clearly reflect uncertainty, and they only indirectly take into account the quality of data they use (Daponte *et al.*, 1997; Abel *et al.*, 2010). Non-Bayesian examples of progress in this sense can be found in Lee and Carter (1992) where confidence intervals are added to the projections; in stochastic approaches (Lee and Tuljapurkar, 1994; Tuljapurkar and Lee, 1997; Tuljapurkar and Boe, 1999), where still some estimated quantities are considered as known, ignoring confidence intervals; in Pflaumer (1988) where the importance of prior knowledge in choosing the demographic distribution is stated even if not incorporated in the model. The advantage of Bayesian methods is that they naturally account for uncertainty, give probabilistic results and incorporate prior information. Applications of Bayesian methods to official statistics and population reconstruction or projections are diverse and literature is growing fast.

In official statistics it has been easier to introduce Bayesian methods in Small area Estimation (SAE) than to use them for census adjustments, as both scientific (Freedman and Navidi, 1986) and political communities appear to be reluctant to this change Fienberg (2011). For examples and theories of small area estimation the book from Rao (2003) provide an overview whereas Ballin *et al.* (2005) and Trevisani and Torelli (2004) focus more on Bayesian hierarchical formulation. Bias problems for missing data imputation are addressed in King and P. (2001). Authors combine and extend the Markov chain Monte Carlo model composition (Madigan and York, 1997) applying the method to incomplete data and coming from different sources also optimising the best data source combination. Dellaportas and J. (1999) instead propose hierarchical log-linear model using reversible jump Markov chain Monte Carlo simulation techniques, discussing the prior specification and presenting *ad hoc* loss functions approaches to select the appropriate data sources. Comparing frequentist and Bayesian approach applied in presence of multiple and incomplete data sources, simulation methods based on the Bayesian approach seem to perform better when the number of source becomes high, i.e. when estimator for missing cells are not "highly sensitive to the choice of model" (King and P., 2001), and to obtain good model-average estimates. Frequentist approaches for assessing the accuracy of register-based household statistics is presented in Zhang (2011) using a unit-error theory with application to Norwegian registers whereas Yildiz and Smith (2015) developed a model to correct the over-coverage of the Patient Register of UK and Wales with more accurate auxiliary data sources when census data are not available.

One of the first time a Bayesian approach has been used in demography was in Daponte *et al.* (1997) to project and reconstruct the Iraqi Kurdish population from 1977 to 1990. In the article Authors point out how with Bayesian analysis it is possible to explicitly include personal beliefs or uncertainty about parameters, to get probabilistic results. The task was particularly hard since data on Kurdish minority where severely under-reported and biased due to social, political and geographical problems of the area. With a cohort-component method they adjust the last census data available and use it as baseline populations. Then they assume levels and patterns of fertility, mortality and net migration for the projection period. Opinions, judgements, experience, outlooks and data quality assessments are also included in the process. Another historical reconstruction with Bayesian method is the one in Bertino and Sonnino (2003) whereas Bryant and Graham (2013) estimate New Zealand regional population.

An important approach adopted by the UN is the one firstly proposed in Wheldon *et al.* (2010) and then applied to different cases in Wheldon *et al.* (2012, 2013, 2015, 2016). This method simultaneously estimates population counts, vital rates and net international migration at the country level, by age, together with uncertainty. One of the first applications was on female population, then Wheldon *et al.* (2013) reconstruct both sex populations for India, Thailand and Laos studying the sex-ratios at birth and the sex-ratios of mortality. From data quality and availability point of view, these are challenging countries, good for experimenting Bayesian methods performance. A comparison among results from countries with three different data quality is presented in Wheldon *et al.* (2016). Authors reconstruct female population from Laos, Sri Lanka and New Zealand, with Laos having the poorest data quality and New Zealand the best. Wheldon *et al.* (2016) integrate the two main model categories in demographic literature of population reconstruction: (1) the cohort component method of population projection (CCMPP) (Lewis (1942), Leslie (1945), Leslie (1948)), a deterministic method typically used for reconstruction of distant past population and population dynamics after extreme crises; (2) Bayesian hierarchical modelling allowing for uncertainty of measurement errors. Moreover, despite being similar to Daponte *et al.* (1997), this approach does not assume any specific age pattern through the period of reconstruction, and it overcomes the requirement of regular census data of Wheldon *et al.* (2013). Prior distributions embed information and expert opinions available for the countries of interest and treat bias and variance in measurement errors separately. This method requires at least two data sources, e.g. baseline population estimates based on bias-adjusted census counts, and fertility and mortality estimates from surveys. Results from the three different countries have uncertainty levels inversely proportional to the quality of data. Wheldon *et al.* (2013) also implemented an R package for population reconstruction `popReconstruct`.

## 1.3 Changes in demography and Official statistics

During the last years both demography and official statistics are facing a period of change. Considering demography, paradigms, topics, approaches are changing so that the nature itself of the subject has been questioned (Billari, 2015). From a mainly descriptive subject demography is now more about population studies increasing the importance of statistical methods. The way to name demographical studies (demography, population studies, political arithmetic), what components of population change is more prominent (fertility, mortality, migration), what kind of analysis is more used (cohort analysis, cross-sectional analysis, longitudinal analysis, event history, biographical analysis, multilevel analysis) has changed and evolved over time (Bijak *et al.*, 2014). A controversial and deep change is the shift from a macro and descriptive approach to a micro and dynamic (or “life-course”) approach based on agent-based models (Rettaroli, 2011; Billari, 2015), and suspected to cause demography abandoning its “core” (Lee, 2001) by some, and considered a “seminal idea”, already adopted in other sciences like economics, by others (discussion in Billari (2015)). Also, Bijak and Bryant

(2016) highlight how traditional demographic models often oversimplify the complexity of demographic phenomena such as the dynamic nature of real demographic systems, bias due to integration of sources with different levels of completeness and reliability, and "uncertainty arising from incomplete knowledge of historical trends or causal mechanisms, or from random variation in disaggregated counts." Bijak and Bryant (2016). Dealing with uncertainty, the change of perspective, and the increasing technology give statistical models a prominent place in demography, and Bayesian methods, that typically perform well in high uncertain situations, have started being studied and applied also to demography. The notion of uncertainty related to demographic quantities is rather new, or better, it has been explicitly introduced and included in demographic studies only recently. Reasons for interpreting and providing results in a probabilistic way come from different factors: first of all problems with data such as bias, under- or over-coverage and sparseness need to be addressed properly, techniques accounting for these characteristics and explicitly pointing out what impact they have on results are required. Secondly, data sources and their accessibility are increasing, and the ability to combine, integrate and organise them is a key point for efficient and reliable official statistics and research in general. A third aspect concerns outputs typically given in demographic estimations and forecasts from official sources. For instance, UN projections used to hypothesise convergence of mortality, fertility and migration but this assumption is itself not certain and this was not taken into account in projections (World Population Prospectus data.un.org). Also, the habit to make projections according to scenarios (typically low, medium and high) without probabilistic statements was not always clear, and therefore, made results questionable. The UN have now switched to probabilistic projections (Gerland *et al.*, 2014). Eventually, knowledge and information not directly coming from data (e.g. expert opinions) could help the inference process especially when data are sparse, incomplete or biased. The Bayesian framework is particularly appropriate to deal with these issues as it allows for prior information inclusion and provides uncertainty estimates.

Changes in demography have affected Official statistics which is closely related to it. On its side, Official statistics is also experiencing a period of deep transformations. New technology, new requirements, new data source and higher data quantity, globalisation are only few of the driving force of the changes National Statistical Institutes (NSIs) have undertaken. NSIs are facing new challenges and, in many cases, they have started modernisation and industrialisation processes. Commissions and groups of research have been established in international organisations to study current situations, propose new solutions and standards, and harmonise definitions and procedures as far as possible.

Examples of these groups are: the UNECE High-Level Group of Modernisation of Official Statistics (HLG-MOS) with all their committees and activities (Generic Statistical Business Process Model (GSBPM), the Generic Statistical Information Model (GSIM), and the Common Statistical Production Architecture (CSPA), Generic Activity Model for Statistical Organizations (GAMSO), the sets of Generic Statistical Data Editing Models (GSDEMs) and the Big Data project). Complementary initiatives within the European Statistical System (ESS) are, for instance, the recommended practices for editing and imputation in cross-sectional business surveys (EDIMBUS manual), the European Statistics Code of Practice with the related Quality Assurance Framework of the ESS (ESS QAF), the Euro-SDMX Metadata Structure (ESMS) for the dissemination of reference metadata, and the Validation and Transformation Language (VTL) as a standard language to express data editing validation rules, to name a few (Salgado, 2016).

With technology NSIs are now able to perform better and faster. These changes require new methods, ways of proceeding, suitable knowledge and tools (software, accessing data, internal and external communication and organisation), and a substantial revision of organisations processes and structures. NSIs have started adapting to the new requirements and in many European countries novelty have been introduced already during the 2011 census round. In addition to "formal" innovations concerning the number of institutions and/or commissions and new rules, also substantial changes occurred in the procedures and in the theoretical framework used. If, on the one hand, NSIs might be slow adapting to new solutions due to bureaucracy, regulation and country-specific issues, on the other hand, other organisations and research institutes can drive innovation faster. For example, the UN switched to probabilistic population projections for all countries in July 2014 using a Bayesian framework (Alkema *et al.*, 2015).

## 1.4 Censuses: history and methods

A key task of NSIs and, more generally, a fundamental point of Official statistics and a precious demographic data source is the census. Following changes in demography and Official statistics, also census is undergoing a lot of changes in many countries.

Census is "a count for official purposes, especially one to count the number of people living in a country and to collect information about them" (`http://dictionary.cambridge.org`). There are evidences of censuses since the dawn of civilisation (Egyptians, Mesopotamian societies, Chinese, Romans are only few examples). Reasons for conducting an enumeration of population are diverse and changed over time, they go from military and fiscal to welfare or descriptive reasons. As the importance and the requirements of censuses have increased over time, the limitations of the traditional way of conducting censuses have arisen more and more clearly. A review of these limitations as well as alternatives and issues can be found in Coleman (2013).

Traditionally it is said that the first modern census date back to the XVII century even though it is difficult to be precise about what was the first modern census (Egidi and Ferruzza, 2009). Nowadays it is possible to identify five essential characteristics of a census: (i) the individual enumeration, i.e. separate recording of personal characteristics allowing for cross-classification; (ii) simultaneity, so that information are all collected at the same time or adjustments are made to the data to have the same reference period; (iii) universality, as the census is meant to enumerate every person/household/housing residing and/or present in the country; (iv) small area data, i.e. to produce data related to the smallest geographic area; (v) defined periodicity, each country decide its own but general recommendations are for at least a census every ten years (UNECE report, 2006).

Census regulations and recommendations come from different institutions and there are different levels. The most general organisation, encompassing 193 countries is the United Nations Organisation. UN give general recommendations, individuates the main issues, possible and needed innovation and provides reports about countries. Then, in Europe, the European Commission regulates some aspects for member countries and, eventually, each country has its own laws and rules for conducting censuses.

The main phases of censuses are: (a) involvement of stakeholders, (b) preparatory work (including legislation, testing and outsourcing), (c) enumeration, (d) data processing, (e) quality assurance of data prior to its dissemination, (f) dissemination of the results, (g) evaluation of the coverage and data quality, and (h) analysis of the results. In each phase, needs and constraints of who conducts the census must be taken into account. On the one hand, National Statistical Institutes (NSIs) would like to have detailed, reliable and accurate data with previous censuses; on the other hand, a census should respect people's privacy, not be too long or ask sensitive or complex questions, it should adapt to changes in the society and avoid inconveniences or ambiguous goals. More formally, "a topic should NOT be included in a census if: (a) it is sensitive or potentially intrusive, or requires lengthy explanations or instructions to collect; (b) it imposes an excessive burden on the population, or seeks information not readily known; (c) its inclusion is likely to have a detrimental impact on coverage or the quality of the information collected; (d) it enquires about opinions or attitudes; or (e) it is likely to present major coding problems or extensive processing or significantly add to the overall cost of the census. In addition to these factors, the census should be considered as an exercise carried out purely for statistical purposes, and should not, therefore, be used to collect data that will deliberately promote political or sectarian groups, or sponsor particular causes" (UNECE report, 2006). One of the main UN commission for census recommendations is the United Nations Economic Commission for Europe (UNECE), set up in 1947. "UNECE's major aim is to promote pan-European economic integration. UNECE includes 56 member States in Europe, North America and Asia. However, all interested United Nations member States may participate in the work of UNECE". UNECE mission is to "focus on raising UNECE countries' capacity in official statistics by helping national statistical offices and other stakeholders to coordinate their work and fill statistical gaps. Our work aims to address the increasing demand for high quality and comparable data among countries." (`https://www.unece.org/stats/stats_h.html`). In general, but in particular in the UNECE region, there are three main ways to conduct census:

1. the traditional methods of full enumeration, conducted either once in a regular basis, or a rolling census where information is collected by a continuous cumulative survey covering the whole country over an extended period (rather than on a particular day or short period of enumeration). Traditional censuses give a specific and detailed picture of a country's population but it is also an elaborate, complex and costly activity. Because it can be carried out only once in a while, collected data go through a long revision process before being released, and, therefore, data can never be considered up-to-date. Furthermore, cooperation from the whole population is necessary to have reliable data and it has not always been the case for different reasons (mainly religious, racial, and political). Traditional census can be carried out either with interviews or with questionnaires, depending on the literacy of the population, and on countries resources (economic and technological).

2. administrative and register-based census, possibly supported by sample surveys for selected variables. This system is more and more popular and it is replacing traditional censuses. Transition though is a long process because it requires reliable and up-to-date registers and the ability to correctly link people/units in different registers and combine them (e.g. people to household, dwelling, buildings and places; employer to employee). Mainly Nordic countries (Denmark, Finland, Iceland, Norway and Sweden) and few others like Austria, Belgium, Slovenia and Turkey use this system. It reduces costs and allows for increased frequency of outputs, but information is limited to what administrative registers provide which

| **Type of census** | **Number of countries** |
|---|---|
| Traditional censuses | 35 |
| Combination between register-based censuses and traditional censuses | 3 (Czech Republic, Latvia, Lithuania) |
| Combination between register-based censuses and sampling surveys | 6 (Spain, Germany, The Netherlands, Poland, Switzerland, Israel) |
| Register-based censuses | 9 (Austria, Belgium, Denmark, Finland, Iceland, Norway, Slovenia, Sweden, Turkey) |
| Appropriate surveys with rotating samples (continuous censuses) | 1 (France) |

Table 1.1: Type of censuses in 2011 for UNECE region countries (INE Spain, 2011).

is supposed to have administrative and not statistical purpose. Moreover, differences in contents limit comparability with other countries and it is less flexible than a traditional census questionnaire that can be adapted each time according to current necessities.

3. combined approach: data come from administrative registers and, for other variables not contained in registers, data are collected by full enumeration or by sample. This method violates the principle stating that census information is used only for statistical purposes and assumes NSIs have access to administrative data and are able to collect required information and link it to registers. This option is less costly than the traditional census and it is perceived as much less intrusive.

Table 1.1 "shows the type of Census that, according to a survey conducted by the United Nations in June 2009, it was planned to conduct in 2011 by the 50 Member States of the UNECE, and an additional 4 countries (Australia, Japan, Mexico and Kosovo)" (INE Spain, 2011)

Regardless countries' census choice, the UN advice countries to "take into account a wide range of issues such as: (a) users needs, (b) quality of the data, (c) completeness of the count, (d) data protection and security, (e) comparability of the results between countries and over time, (f) burden on the respondents, (g) timeliness of outputs, (h) costs, (i) political and legislative implications, and (j) public understanding and acceptance" (UNECE report, 2006) in addition to confidentiality requirements. UN also monitor what technological innovations might have a good impact on efficiency, quality and costs, give directions on data collection and processing, and define quality standards for relevance, accuracy, timeliness, accessibility, interpretability and coherence.

In Europe, the European Statistical System (ESS) and Eurostat coordinate and work on Census programmes giving main guidelines, collecting and comparing results at European level. After the 2021 European Census round, the next is scheduled for 2031 embedded in the world-wide UN Census round. In addition to international recommendation, two important topics the European Union is focusing on are migration and the change on the geographical boundaries used for data collection. During the last years, European and national political institutions are particularly interested in migration, mobility, migrant populations and, more generally, in collecting migration related information and investigating dynamics of the phenomenon which has become a major topic of research. Another European countries main goal is to implement grid-based data collection method rather than the usual administrative based one. In this way data collection would be always geographically consistent avoiding inconvenience deriving from administrative boundary changes.

### 1.4.1 The case of Istat and the "permanent census"

The Italian NSI, Istat, is adapting to the new requirements and directions of international and European institutions. Changes in such big institutions require time and cooperation but there are visible progresses and results, especially during the last few years.

Since the first Italian population census in 1861, it is possible to individuate differences and novelties in each round, but in 2011 for the $15^{th}$ census, Istat has deeply changed its procedures in order to satisfy the increasing need of better longitudinal and spatio-temporal data, to release census information with higher frequency than once every ten years, to decrease the cost for conducting censuses [3], and to start integrating different sources. Innovations of the 2011 census concern, for instance: the use of administrative and territorial data, even though just as a support to other data; some activities have been carried out through the web; there was the introduction of sampling techniques; a Post Enumeration Survey (PES) was conducted according to European Commission Regulation n. 1151/2010; new registers and systems to harmonised municipalities; surveys, usually considered independently of one another, have been looked at in an integrated way in order to get more precise and richer information.

The idea of a system where institutions communicate and work together to integrate different sources of information is not new. In the 1960s, Professor De Finetti wrote

[3] 2011 census had an overall cost of 604 million, including personnel costs, `http://ec.europa.eu/eurostat/cache/metadata/EN/cens_11r_esmscs_it.htm`

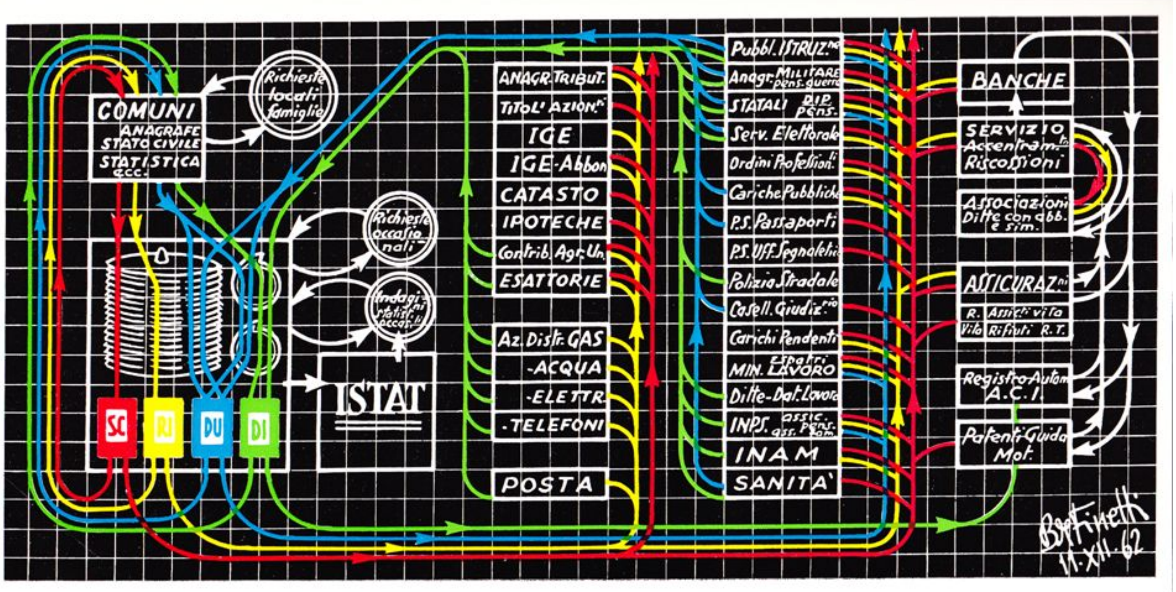


Figure 1.1: A "futuristic" blackboard, Prof. De Finetti drew and signed it in 1962 during a seminar at the Demographic Institute of the University of Rome. It represents the information flow of different kind of data among public and private institutions. Information flows regard: marital status in red (SC = *stato civile*), residence and address in yellow (RI = *residenza e indirizzo*), official data in blue (DU = *dati ufficiali*) and informational data in green (DI =*dati informativi*).

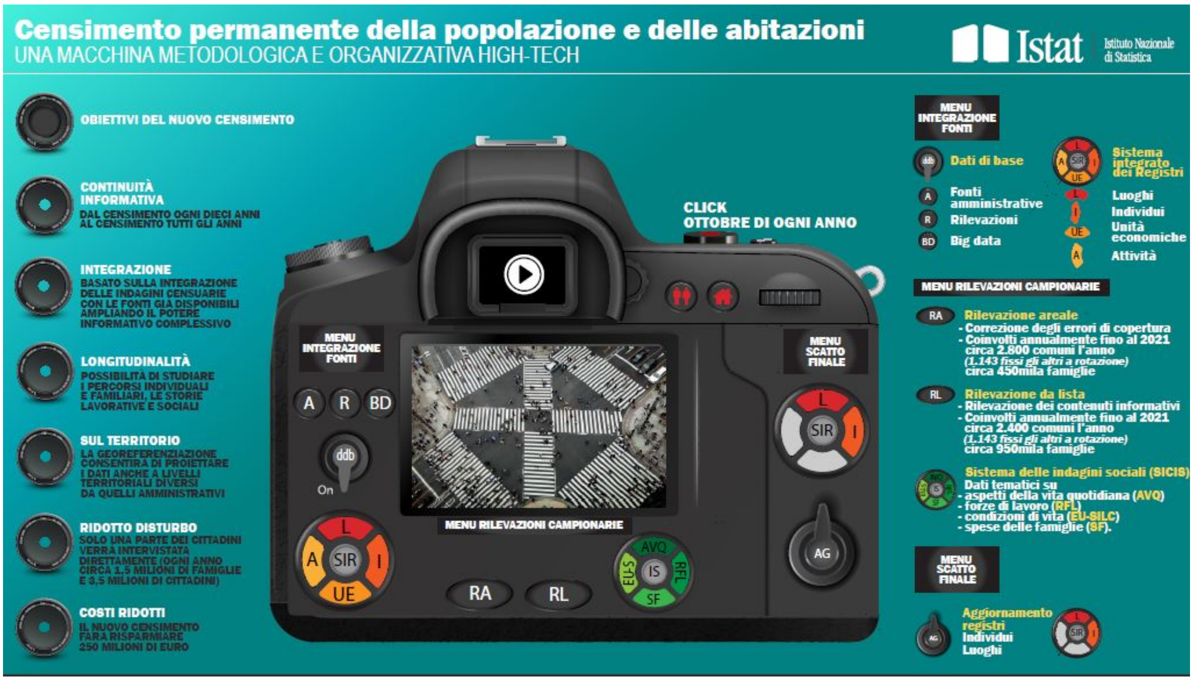
about how to improve and extend official statistics production through a network linking many institutions, both public and private, and draw the (still) "futuristic" system showed in figure 1.1. In his essay De Finetti, B. (1965) proposes a system at national and possibly European level, where a unique code associated to each person would be used everywhere allowing all information to be linked to the right person, and where institutions public and private communicate and cooperate creating an efficient and integrated system. He also highlights obstacles and problems for such a transformation, and procedures to implement for such a system to work (updating, storing data, communication among institutions).

If the idea proposed in De Finetti, B. (1965) is still "futuristic", it is true that Istat and many NSIs are taking steps towards an always more integrated system. In particular, since the 2011 census, Istat has been working to eventually switch from a traditional census to a "permanent census". Permanent census is regulated at European level since 2008 (CE n. 73/2008) with three other regulations released in 2017 for its practical implementation. As Istat explains, permanent census does "not involve all citizens, enterprises and institutions, but parts of them from time to time, that is representative samples. However, the data disseminated to the Country will be census data, and therefore referable to the entire field of the survey." (`https://www.istat.`

`it/en/permanent-censuses`). Hence, with permanent census, it is possible to provide census data every year at a much lower cost. Istat will integrate survey results and data from registers belonging to the Integrated System of Registers (*Sistema Integrato di Registri*, SIR). In 2017 there have been experimental surveys, and the actual project starts in October 2018. Figure 1.2, produced by Istat, summarises characteristics of permanent census:

- Objectives: (i) Yearly information, at October every year (ii) integration of existing administrative sources with surveys, (iii) production of longitudinal data, (iv) data by territorial grid different from administrative one, (v) less public inconvenience, (vi) reduced cost.
- Integration of sources: (i) Administrative data, (ii) surveys, (iii) big data.
- Integrated System of Registers including: (i) Places, (ii) People, (iii) Economic units, (iv) Activities.
- Sample surveys: (i) surveys by area to correct coverage errors involving 2800 municipalities every year, (ii) surveys from administrative lists for information, (iii) System of social surveys (Sistema delle Indagini Sociali, SICIS) for daily life aspects, workforce, life conditions, households expenses.
- Surveys help registers update.

Permanent census is a complex and new project whose first part is planned for the four years 2018-2021, ending when the next census round was planned. For such an important change research is essential to provide suitable tools able to analyse and integrate data, and to suggest potential adjustments, corrections and procedures to improve this new official statistics framework.



**Censimento permanente della popolazione e delle abitazioni**
UNA MACCHINA METODOLOGICA E ORGANIZZATIVA HIGH-TECH

Istat Istituto Nazionale di Statistica

OBIETTIVI DEL NUOVO CENSIMENTO

CONTINUITÀ INFORMATIVA
DAL CENSIMENTO OGNI DIECI ANNI AL CENSIMENTO TUTTI GLI ANNI

INTEGRAZIONE
BASATO SULLA INTEGRAZIONE DELLE INDAGINI CENSUARIE CON LE FONTI GIÀ DISPONIBILI AMPLIANDO IL POTERE INFORMATIVO COMPLESSIVO

LONGITUDINALITÀ
POSSIBILITÀ DI STUDIARE I PERCORSI INDIVIDUALI E FAMILIARI, LE STORIE LAVORATIVE E SOCIALI

SUL TERRITORIO
LA GEOREFERENZIAZIONE CONSENTIRÀ DI PROIETTARE I DATI ANCHE A LIVELLI TERRITORIALI DIVERSI DA QUELLI AMMINISTRATIVI

RIDOTTO DISTURBO
SOLO UNA PARTE DEI CITTADINI VERRÀ INTERVISTATA DIRETTAMENTE (OGNI ANNO CIRCA 1,5 MILIONI DI FAMIGLIE E 3,5 MILIONI DI CITTADINI)

COSTI RIDOTTI
IL NUOVO CENSIMENTO FARÀ RISPARMIARE 250 MILIONI DI EURO

CLICK
OTTOBRE DI OGNI ANNO

MENU INTEGRAZIONE FONTI

ddb Dati di base
A Fonti amministrative
R Rilevazioni
BD Big data

Sistema integrato dei Registri
L Luoghi
I Individui
UE Unità economiche
A Attività

MENU RILEVAZIONI CAMPIONARIE

RA Rilevazione areale
- Correzione degli errori di copertura
- Coinvolti annualmente fino al 2021 circa 2.800 comuni l'anno *(1.143 fissi gli altri a rotazione)* circa 450mila famiglie

RL Rilevazione da lista
- Rilevazione dei contenuti informativi
- Coinvolti annualmente fino al 2021 circa 2.400 comuni l'anno *(1.143 fissi gli altri a rotazione)* circa 950mila famiglie

Sistema delle indagini sociali (SICIS)
Dati tematici su
- aspetti della vita quotidiana (AVQ)
- forze di lavoro (RFL)
- condizioni di vita (EU-SILC)
- spese delle famiglie (SF).

MENU SCATTO FINALE

AG Aggiornamento registri
Individui
Luoghi

FIGURE 1.2: Istat permanent census representation with its objectives, sources, registers, sample surveys and periodicity. Source `https://www.istat.it/it/censimenti-permanenti`.

# Chapter 2

# Demographic Account Model

## 2.1 Preliminary concepts

The model described in this chapter has been firstly introduced by Bryant and Graham (2013). According to the classification proposed in Bijak and Bryant (2016), it is one of the "highly structured and complex models" in Bayesian demography. Before introducing the model itself, the next 2 subsections provide a short descriptions of the main elements the model combines for estimating the population size: hierarchical models and demographic account. As the concept of Bayesian demography implies, also in this model the two arguments are one typical of Bayesian statistical modelling (hierarchical models) and the other of demography (demographic account).

### 2.1.1 Hierarchical models

When parameters describing a statistical model present some similarities or are somehow connected this information should be included in the model. Hierarchical models are a useful way to perform this task and Bayesian framework naturally suits their structure. The basic structure of any hierarchical model includes the data model $y \sim f(y|\theta)$ conditioned on on parameters in the vector $\theta$ which has itself a distribution conditioned on additional parameters called hyper-parameters $g_1(\theta|\lambda)$. Hyper-parameters can also have their distribution $\lambda \sim g_2(\lambda)$, be considered as known, estimated through frequentist methods and then treated as known (empirical Bayes). This structure based on layers is flexible and presents many strengths. The structure just presented shows why hierarchical model fit well in the Bayesian framework. A parallel between hierarchical models structure and the typical Bayesian models one is immediate. The part modelling data $y \sim f(y|\theta)$ corresponds to the likelihood, whereas the parameter model $g_1(\theta|\lambda)$ is the prior distribution used in the Bayesian framework which can have hyper-priors depending on the structure. Then the resulting distribution is the posterior.

Hierarchical models can usually fit complex structure where other models fail. Especially with large datasets, non-hierarchical models risk either to include too few parameters, maybe ignoring important features, or to consider too many parameters, ending up over-fitting the data. Instead, because of their structure, hierarchical models can handle complicated problems. Combining common and simple model for the data, priors and hyper-priors, it is possible to obtain a complex posterior which would be difficult to model and work with directly. This also bring computational advantages. Also the reverse process is possible, i.e. starting from a complex posterior a decompose it in different parts. These models are called hidden Markov models, hidden mixtures or deconvolution.

Hierarchical models are usually robust to model misspecification, i.e. usually similar results are obtained with different priors and, especially if priors are relatively flat, results are often close to those obtained with empirical Bayes analysis. This property represents an advantage as it allows to concentrate less on the exact prior or hyper-prior specification, which can be left flat non-informative. Lehman and Casella (1998) write that “hierarchical models allow easier modelling of prior with flatter tails which can lead to Bayes estimators with more desirable frequentist properties”. And also “ordering in the hierarchy allows to order the importance of the parameters and to incorporate some of our uncertainty about the prior specification”.

A key concept for hierarchical model is the concept of exchangeability. Real-valued random variables $Y_1, ..., Y_n$ are said to be finitely exchangeable if $(Y_1, ..., Y_n) \overset{D}{=} (Y_{i_1}, \ldots, Y_{i_n})$ [1] for any permutation $(i_1, i_2, \ldots, i_n)$ of $(1, 2, \ldots, n)$, $1 < n < \infty$. If $Y_1, Y_2, ...$ is an infinite sequence then it is called infinitely exchangeable if every finite subset of it is finitely exchangeable. An important theoretical contribution is De Finetti's representation theorem. It proves that if observations are judged to be exchangeable, then they must indeed be a random sample from some model and there must exist a prior probability distribution over the parameter of the model (Bernardo, 1996).

[1] $\overset{D}{=}$ represents equality in distribution.

**Theorem 2.1.** *A set of binary variables $Y_1, Y_2, ...$ is infinitely exchangeable if and only if there is a random variable $\Theta : \Omega \rightarrow [0,1]$, with distribution $F(\theta)$ on $[0,1]$, such that:*

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n) &= \int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} dF(\theta) \\ &= \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \underbrace{f(\theta)}_{\textit{prior}} d\theta \end{aligned} \tag{2.1}$$

*And $\overline{Y_n} = \frac{1}{n}\sum Yi \rightarrow \Theta$ almost surely as $n \rightarrow \infty$, for the strong law of large numbers. So if a sequence of observations is judged to be exchangeable, then, any finite subset of them is a random sample of some model $f(y_i|\theta)$, and there exists a prior distribution describing the initial information about the parameter.*

A more general version of equation 2.1 exists for real random variables. And Davison (2003) explains how "that certain quantities are exchangeable implies that they may be represented as a random sample conditional on a variable that itself has a distribution. This provides the basis of a case in favour of Bayesian inference, because it implies that the conditional density $\Pr(Y_{n+1}|Y_1, ..., Y_n)$ for a future variable $Y_{n+1}$ given the outcomes of $Y_1, ..., Y_n$, may be represented as a ratio of two integrals of form" of equation 2.1, "and this is formally equivalent to Bayesian prediction using a prior density on $\Theta$". And continues highlighting that "the essence of hierarchical modelling is to treat not data but particular sets of parameters as exchangeable. For if our model contains parameters $\Theta_1, ..., \Theta_n$ and if we believe a priori that these are to be treated completely symmetrically, then they are exchangeable and may be thought of as a random sample from a distribution that is itself unknown" (Davison, 2003). Depending on the characteristics of the random variables, the sample can be fully, partially or conditionally exchangeable, and, whereas random variables independence always implies exchangeability, the contrary is not always true. A set of exchangeable random variables might not be independent. The differences in the exchangeability properties of random samples reflect on the structure of hierarchical models and on the pooling options, i.e. on how differences between parameters are considered and how they are grouped. Usually in hierarchical models parameters are considered as partially exchangeable, i.e. parameters are grouped and each group has its own sub-model whose properties need to be estimated. With partial exchangeability, in each group variables are considered exchangeable and have a common prior distribution. There are therefore as many prior as groups. Grouping parameters according to certain characteristics and assigning a different distribution to each group is called partial pooling. Partial pooling is a compromise between "complete pooling, in which differences between groups are ignored, and no pooling, in which data from different sources are analysed separately" (Gelman and Hill, 2007). Sometimes random variables $Y_1, ...Y_n$ are not exchangeable on their own, but providing additional information $X_1, ..., X_n$ it is possible to obtain a joint model for $(Y_i, X_i)$ or a conditional model for $Y_i|X_i,\ i = 1, ..., n$, where the couples $(Y_i, X_i)$ are still exchangeable. In real applications exchangeability almost never holds, considering variables as exchangeable is rather a simplification to model ignorance about random variables, as when assuming a sample from a common population is independent and identically distributed. As argued in Gelman (2006) "the valid concern is not about exchangeability, but about encoding relevant knowledge as explanatory variables where possible". Assuming exchangeability at the beginning of the analysis corresponds to admitting ignorance about the random variables characteristics, additional information can then be incorporated in the model only as the analysis goes on. In hierarchical models, the prior belief of parameters exchangeability influences the choice of the prior distribution. A sensible way of proceeding is to start with simple priors and then test sensitivity to prior changes and check the fitted model with the predictive distribution.

As briefly described, the properties and structure of hierarchical models are convenient for modelling demographical data in a Bayesian framework. The information available from demographical data is usually enough to group data according to available characteristics, e.g. age, sex and region, which naturally leads to a hierarchical structure as the one proposed by Bryant and Graham (2013) for population size estimation and described in the next sections.

### 2.1.2 The demographic account

In a population study, one of the first points to clarify is the demographic system the study focuses on. A demographic system describes how a population grows and changes by defining:

1. the *membership criteria*: what is common to all the people belonging to the population;

2. *classification system*: from the population as a whole it is necessary to individuate sub-populations, i.e. to give a structure to the population by identifying characteristics of interest that help the study. These characteristics can be attributes as age, sex, region of residence, education or income and they can be changeable or fixed;

3. *ways of entering, exiting, or moving within the system*: typically one enters a population by birth, immigration or enrolment, exits by death, emigration or

cancellation from a register or list and moves by migrating internally from one place of the region of interest to another.

A way to summarise information about the demographic system is to cross-tabulate data in a demographic array, usually according to the dimensions of the classification system. A demographic array collects data in aggregate form and data are typically either population counts or life events occurred during the considered time frame. In tables 2.1 and 2.2 there are two examples of demographic array. In the first one each cell contains the number of deaths by sex occurred in 2015 in the regions of Southern Italy, and the second one gives the total number of birth from 2012 to 2015 by mothers age groups. The two arrays have different dimensions stressing different aspects: the first one focuses on a geographic and sex division, whereas the second highlights age and time.

| | *Sex* | |
|---|---|---|
| *Region* | **Female** | **Male** |
| **Abruzzo** | 7998 | 7367 |
| **Molise** | 1967 | 1917 |
| **Campania** | 29277 | 27519 |
| **Puglia** | 20294 | 19231 |
| **Basilicata** | 3240 | 3174 |
| **Calabria** | 10182 | 10129 |
| **Sicilia** | 27484 | 25633 |
| **Sardegna** | 8172 | 8356 |

TABLE 2.1: Demographic array by region and sex for deaths in the Southern regions of Italy in 2015.

Source: Istat, `www.istat.it`

| | *Year* | | | |
|---|---|---|---|---|
| *Age* | **2012** | **2013** | **2014** | **2015** |
| **15-19** | 8706 | 7989 | 7862 | 7240 |
| **20-24** | 50095 | 47084 | 46260 | 43335 |
| **25-29** | 118936 | 113590 | 113361 | 109864 |
| **30-34** | 178141 | 169240 | 168654 | 164610 |
| **35-39** | 137531 | 131121 | 128793 | 124827 |
| **40-44** | 36168 | 36546 | 36806 | 37368 |
| **45-49** | 2847 | 2993 | 3188 | 3220 |

TABLE 2.2: Demographic array by mothers' age and year for births in Italy from 2012 to 2015.

Source: Istat, `www.istat.it`.

| **Population** | *Women* | *Men* |
|---|---|---|
| *1/01/2015* | 31294022 | 29501590 |
| *31/12/2015* | 31209230 | 29456321 |

Table 2.3: First part of Italian demographic account considering population by sex in 2015.

| 2015 | *Women* | *Men* |
|---|---|---|
| **Births** | 235830 | 249950 |
| **Deaths** | 339607 | 307964 |
| **Immigration** | 129076 | 151002 |
| **Emigration** | 68633 | 78322 |

Table 2.4: Second part of Italian demographic account considering life events by sex in 2015.

When arrays describing population counts, entries, exits and movements within the demographic system are consistently organised together they form a demographic account. In this case the dimensions considered for all the arrays must be the same. A simple example of demographic account is given in tables 2.3, 2.4. Along with the initial and final population (table 2.3), it considers the total number of births, deaths, international (or external) immigration and emigration in 2015 in Italy divided by sex (table 2.4). Adding dimensions the structure becomes more detailed and also more complicated.

A theoretical requirement of the demographic account is internal consistency, which means that it has to be possible to retrieve the initial population by combining all life events occurred in the period with the final population. This concept is summarised in the demographic balance equation (2.2) and must be true for every population system if data are right.

$$\text{Pop}_{t_1} = \text{Pop}_{t_0} + \text{Entries}_{t_0} - \text{Exits}_{t_0} \tag{2.2}$$

The equation is very general and states that a population at the end of a period ($t_1$) is equal to the population at the beginning of the period ($t_0$), plus entries and minus exits occurred during the period, in the demographic account in table 2.3 entries are births and immigration and exits are deaths and emigrations. The specific demographic equation in this case is then

$$\text{Pop}_{31/12/15} = \text{Pop}_{1/1/2015} + \text{Births}_{15} - \text{Deaths}_{15} + \text{Immigration}_{15} - \text{Emigration}_{15} \tag{2.3}$$

Despite being a very simple and intuitive equation in practice it is very difficult to have perfectly matching data and retrieve the truth about the evolution of the population. Reconstruct a consistent demographic account in a probabilistic way is the goal of the model presented in the chapter.

Initially, the account structure was applied to economics and, as the same concepts of “stock” and “flow” can apply to money and populations, it was then adapted to demographic use. Classical examples of discussions, projections and models for the demographic account are Rees (1979) and the Nobel Memorial Lecture 1984: "The Accounts of Society" reported in Stone (1986). The demographic account is a practical tool to represent a population. It is flexible, as it allows for any choice of dimensions and it relates all population movements in a consistent and unique framework. These good characteristics meet the needs for demographers to experiment new tools and new way of estimating population. The interest in new demographic models has risen both at academic and political level and a shift from traditional census methods to register based system or to probabilistic population estimations is now a general trend in many countries encouraged by governments and international organisations. An example of population estimation model using the demographic account is Bryant and Graham (2013) (initial model) and, later, Bryant and Graham (2015) and Bryant and Zhang (2018). The Bayesian hierarchical model for population size estimation proposed is among the most complex and recent models in Bayesian demography.

## 2.2 Super and finite population

Bryant and Graham (2013, 2015) propose a Bayesian hierarchical model for population size estimation and forecasting integrating traditional demography and more recent Bayesian demographic models enhancing the estimation process by controlling assumptions and embedding extra-model information.

The main assumption for population estimation models is that the true value of the population is unknown because values given by the datasets have accuracy and reliability varying from source to source and, usually, they cannot be considered as perfect. They can suffer from bias, incompleteness, under- or over-coverage. For this reason the model distinguishes between hypotheses on the *true* population (*system model*) and hypotheses on datasets (*data model*); this distinction is addressed in section 2.3. Another distinction is between the "underlying risks or propensities (*super-population quantities*), and the random events governed by these risks or propensities (*finite-population quantities*)" (Bryant and Zhang, 2018). Difference between *super-population* and *finite-population quantities* starts from the assumption that the population of interest is a random variable assumed to follow a suitable probability distribution. It is then necessary to distinguish between a realisation of the distribution and the hypotheses on the distribution itself. Super-population quantities are values like rates, probabilities, percentages, means, growth rates or other quantities representing theoretical values, underlying risks or propensities the population is subject to. They provide trends or general features of a population which are quantities more of interest for researchers and demographers. Finite-population quantities are population direct estimates. All quantities are directly calculated from actual data that can be affected by random variations. They aim to recreate the actual population of interest, in this sense, they are very specific values. These quantities concern more politicians or institutes in need to have the actual number of people rather than theoretical features or trends. For instance, if no death occurred during the time considered, the "mortality rate" is zero when computed from actual data (finite-population quantity) but, the underlying risk of dying, even if very small, would be different from zero (super-population quantity). Finite- and super-population quantities are hardly ever equal and their difference tends to be higher when counts are small but, if quantities are large enough, estimates for one can be used as proxy for the other (Bryant and Zhang, 2018).

Bryant and Graham (2013)'s model considers all these different aspects and it is graphically represented in Figure 2.1. In the central box "Account" there are the final arrays of interest that are estimated in the model (population, births, deaths and migrations). They are all connected to ensure the account consistency and they represent the estimation of the true population counts which are finite-population quantities. The account is estimated combining information from the *system model* (upper part) and the *data model* (lower part). The system model has a prior model which helps to define the arrays of rates or means that are super-population quantities. The data models are models *a priori* defined that create, along with raw data, the arrays of components or counts which are finite-population quantities. The only observed quantity of the whole model are these finite-population quantities in red in figure 2.1. Interactions between these three parts through the Demographic account model (DAM) lead to the estimation of all the unknown quantities. Except from raw data, in figure 2.1 all white filled boxes are unknown and need to be inferred, only red filled rectangles are partly or fully observed. Each series in the demographic account have one and only one corresponding system model. For data models, as each dataset has its own data model and there can be one or more datasets referring to each series (or no dataset at all), there can be more than one array of population count referring to each demographic account series.

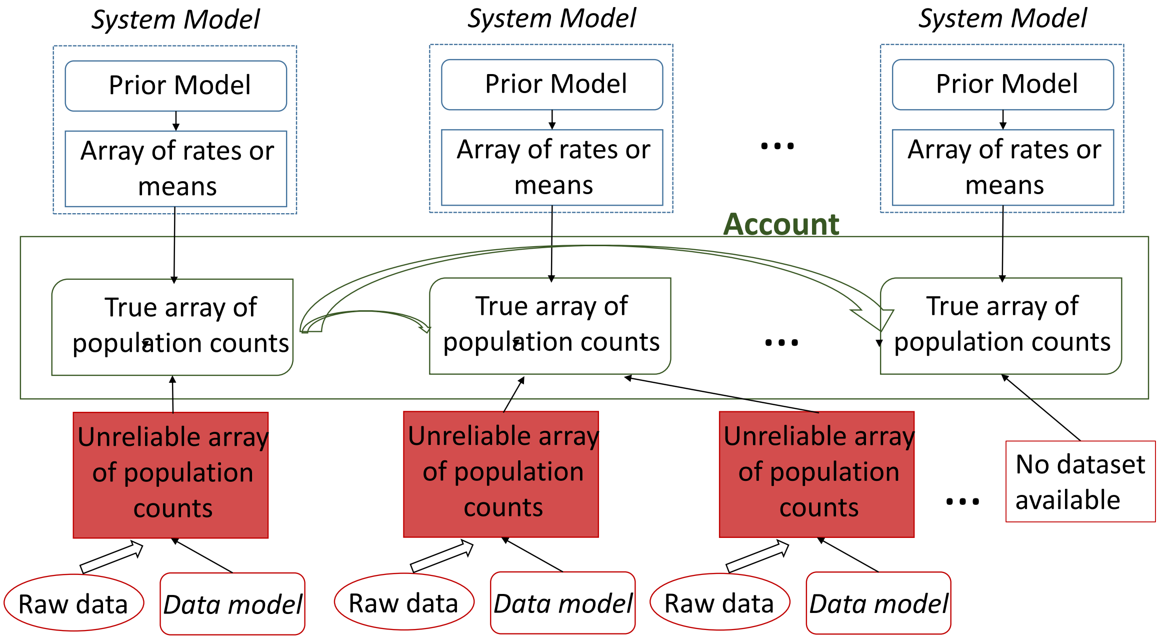


FIGURE 2.1: Inferring a demographic account. Rectangles in the upper part are the system models formed by prior models and super-population quantities arrays, each system model correspond to only one array in the demographic account (middle part boxes). All the arrays in the accounts are interconnected to ensure the account consistency. The lower part is formed by datasets. Each dataset has its own data model and each demographic account array can have one, several or no dataset referring to it.

## 2.3 Model framework

As shown in the examples in section 2.1, arrays and consequently the demographic account is composed of cells. Each cell is a population count for the specific dimensions it corresponds to. For example in Table 2.2 the cell corresponding to the number of births occurred in 2013 from mothers aged between 25 and 29 years contains number 113590. Let now $Y^{births}$ denote the array of births, each cell $y^{births}$ can be identified either by a number, $i = 1, \cdots, N$ with $N$ being the total number of cells, or by numbers or names corresponding to the dimensions of the array, e.g. $a = 1, ..., A$ for age dimension, $r = 1, ..., R$ for the region, $t = 1, ..., T$ for time. The selected cell in Table 2.2 is then indicated in one of the three following equivalent ways

$$y_{10}^{births} = y_{3,2}^{births} = y_{24-29,2013}^{births} = 113,590$$

According to the degree of specification needed the more or less compact notation is used. The array of births is only one of the arrays of the demographic account, the whole demographic account is denoted by $Y$ and contains all the arrays in the demographic system, in our example from table 2.3 there are six arrays, initial population, final population, births, deaths, immigrations, emigrations: $Y = \{Y^{in.pop}, Y^{fin.pop}, Y^{bir}, Y^{dea}, Y^{imm}, Y^{emi}\}$. Each array has sex dimension with categories "women" and "men".

In general, at the first level of the hierarchical model, each cell of a demographic array is considered separately from the others and has its own parameters. Specifically, each number in the cell $y_i$ is assumed to be a realisation from a random variable with Poisson distribution with its own mean $\gamma_i$

$$y_i \sim Pois(\gamma_i) \tag{2.4}$$

Or, if exposure term ($\omega_i$) is considered

$$y_i \sim Pois(\gamma_i \omega_i) \tag{2.5}$$

It is important to consider this second version including exposure because demographers usually work with rates rather than counts. Rates are comparable between demographic systems and give a clearer idea about the magnitude of phenomena than a stock number which is always relative to the population size it refers to. This is why, despite the aim of the DAM is to estimate counts, the models linked to life events estimation (births, deaths, migration) usually have the form of equation (2.5), whereas for population arrays, which have no exposure to refer to (but population itself), the model is obviously the first one (equation (2.4)). The exposure is internally calculated either through the approximation

$$\omega_i = 1/2(y_{i-1} + y_i) + \epsilon \tag{2.6}$$

or, if more information is available from the data, it is possible to calculate the exposure using "person-years" values resulting in a more accurate exposure term. The accuracy of life tables reconstruction depends on data quality. Parameter $\gamma_i$ was initially assumed to be Gamma distributed (Bryant and Graham, 2013) but, for reasons that will be clarified later, a Lognormal distribution has been eventually preferred (Bryant and Graham, 2015) so that

$$\log(\gamma_i) \sim N(\mu_i, \sigma^2) \tag{2.7}$$

The Poisson-Lognormal structure is quite common for hierarchical models as it allows to easily assume a regression model on the mean of the Normal distribution ($\mu_i$). Whereas the mean parameter $\mu_i$ is specific for each cell, the variance $\sigma^2$ is unique for each the array. What motivates this choice is the difficulty to make specific assumptions for variance on each cell as there is no particular reason to expect different variances across the array, and it also substantially reduces the number of parameters to estimate. Mean parameter $\mu_i$ is itself a sum of parameters ($\beta^{(k)}$, single element or a vector) that can have different distributions according to what they refer to, and $\sigma$ is assumed to follow a Half-t distribution for its good properties (Gelman, 2006). The specification of both parameters is further discussed in section 2.4:

$$\mu_i = \sum_{k=0}^{K} \beta^{(k)} \tag{2.8}$$

$$\sigma \sim t^*_\nu(A) \tag{2.9}$$

where $A$ is the rate parameter of the Half-t distribution and $\nu$ denotes the degree of freedom.

### 2.3.1 System model and the demographic account

The system model refers to a part of the DAM which is not directly observed. The aim is to include prior knowledge and regularities in the true population through prior distributions on dimensions like age, regional differences or evolution through time. Each pattern coming from *a priori* information have to be plausible and motivated. The structure contains *a priori* independent models of parameter denoted by $\Theta$, each one corresponding to an array of the demographic account. So each model applies directly to its associated series. Therefore, if we have an account as the one in the previous section $Y = \{Y^{in\_pop}, Y^{fin\_pop}, Y^{bir}, Y^{dea}, Y^{imm}, Y^{emi}\}$ there will be six independent parameter models, one for each array $\Theta = \{\Theta_{in\_pop}, \Theta_{fin\_pop}, \Theta_{bir}, \Theta_{dea}, \Theta_{imm}, \Theta_{emi}\}$. This way of proceeding is different from the usual one. Most of the time there is one series which is not modelled but only derived combining the others through the demographic balance equations, in this way the equation satisfaction and parsimony are guaranteed. Nevertheless, the advantage of modelling *all* series directly as in the DAM is that they are treated the same way, each series has its own independent prior model, and it prevents from implausible results for the series derived from a mere application of the demographic balance equations. It also results in a relatively simple structure because each model is assumed *a priori* independent so the conditional joint posterior distribution is simply obtained by the product of the models distributions (equation (2.10)). On the other hand, despite each series is independently modelled, consistency still has to be respected within the account, so every change in one series has to be balanced by a change in, at least, one other series. In this sense series are not completely independent but bounded by this constraint. Furthermore, cells share parameters in the hierarchy so, in order to capture possible dependencies, there are: i) explicit conditioning on population counts, ii) covariates, iii) demographic balance equations.

Before introducing the demographic account conditional posterior distribution, there is another distinction to point out between demographic arrays. The distinction follows from the nature of the count the array contains. If the array refers to a population *at a specific time* it is a *point* array, this is the case for $Y^{in.pop}$ and $Y^{fin.pop}$ in section 2.3. Arrays of this kind are denoted from now on with letter $N$, and they typically refer to population counts. Instead, if the array contains counts of life events occurred *during a time interval* then, they are *interval* arrays, like $Y^{bir}$, $Y^{dea}$, $Y^{imm}$ and $Y^{emi}$ and they are now denoted with $C_l$, $l = 1, \cdots, L$. So the demographic account is a collection of arrays with general form

$$Y = \{N, C_1, \cdots, C_L\}$$

This distinction is particularly important during the updating process described in section 2.5 and Appendix. Therefore, the conditional prior distribution for the unknown demographic account is

$$p(Y|\Theta_Y, Z) \propto p(N|\Theta_N, Z)p(C_1|N, \Theta_1)p(C_2|N, \Theta_2)...p(C_L|N, \Theta_L)I(Y) \tag{2.10}$$

where $Z$ is the set of covariates, $\Theta$s are the parameter sets of the arrays and $I(Y)$ is an indicator function. The indicator function ensure the respect of the constraints taking value 0 when values are not consistent with the balance equations or impossible.

### 2.3.2 Data model

Whereas the system model catches regularities in the demographic account and relates to an unobserved part of the model, the data model relates datasets, the only observed part of the model, to the demographic account. The system model includes all demographic series to be estimated with all the dimensions and no missing data, whereas series in the data model coming from datasets can be incomplete, have fewer dimensions or cover only a part of the series of the system model. Each system model corresponds to one and only one series in the demographic account whereas it is possible that more than one dataset link to the same series in the demographic account. For instance, in the system model there is only one array for births but there can be more than one dataset in the data model containing birth counts. This is coherent with the fact that there is only one series of super-population quantities relating to the *true* value of births occurred in the period considered by the model, but birth registration datasets can come from different sources (e.g. hospital, city council, surveys). Also, the dataset quality can vary and have different degree of detail (e.g. with or without parents age, sex, citizenship, missing data, coverage).

Each dataset is different and has its own characteristics and level of accuracy, therefore each data model has to consider it and adapt to the dataset it refers to. Furthermore, reliable sources, like census data, are not always available or recent enough to make inference so the possibility to have multiple datasets referring to the same demographic account array can help to improve estimates quality. Because datasets can have missing data or lesser dimensions than their corresponding demographic account arrays, in theory all dataset providing information about the population can be used, from official registers covering the whole population to surveys. In practice, a sensible selection of the datasets used is necessary and if a reliable, detailed enough and accurate source is already available there is no need to add a worse one.

In Figure 2.1 the only parts filled in red are the arrays of population counts but the data model boxes themselves are white, meaning that they need to be estimated. One could wonder why there is need for a data model in addition to the system model. The necessity to estimate the data model could appear somehow counter-intuitive, it is normal to estimate the true value of the population given the data, it is less obvious to do the opposite, i.e. estimate data given the true unknown value, but it is sensible from a demographical perspective. As datasets are not perfect, and they are usually not consistent when compared and combined, each datum can be considered as a random variable generated from a distribution depending on the *true* value of the population. In demography the true value of the population exists for sure even if it is unknown, whereas datasets can contain errors, missing data and have different coverage. Therefore, it is sensible to consider datasets as random variables generated from the true population and to find what model originates the data to better reconstruct the unknown true value of the population. In this way, all the three parts of the DAM contribute and exchange information with one another during the estimation process. This theoretical explanation translates to a computational point of view introducing for each cell of the datasets a dependence on the corresponding true value of the demographic account (in addition to the dataset parameter set). Let each dataset be $X_m$, $m = 1, \cdots, M$, with cells $x_{jm}$, with $j = 1, \cdots, J$ number of observations in the $m$-th dataset, let $\Omega_m$ be the corresponding parameter set, and $y_{j[m]}$ the cell in the demographic account corresponding to $x_{jm}$ and $Y_{[m]}$ the corresponding array. Note that: $j \neq i$ as datasets can have a different number of cell than the corresponding array; $M$ is the number of datasets corresponding to the same array $Y_{[m]}$ therefore for all $X_m$, $m = 1, \cdots, M$ the corresponding demographic account array $Y_{[m]}$ is always the same; the data model does not share any parameter with the system model therefore they are denoted with two different letters $\Omega$ for the data models and $\Theta$ for the system model. Another assumption is that all datasets are conditionally independent, therefore for each data model the structure is

$$\begin{aligned} p(X|Y, \Omega_X) &= \prod_{m=1}^{M} p(X_m|Y_{[m]}, \Omega_{X_m}) \\ p(\Omega_X) &= \prod_{m=1}^{M} p(\Omega_{X_m}) \end{aligned} \tag{2.11}$$

Each observation $x_{jm}$ is conditionally independent on $y_{[m]}$ and $\Omega_{X_m}$ so that

$$p(X_m|Y_{[m]}, \Omega_{X_m}) = p(X_m|y_{[m]}, \Omega_{X_m}) = \prod_{j=0}^{J} p(x_{jm}|y_{j[m]}, \Omega_{jm}) \tag{2.12}$$

Note that from the first to the second term of equation (2.12), $Y_{[m]}$, which is the true unknown value of the population is replaced by $y_{[m]}$ which is only a realisation of the true value, in practice, it is the value available at the moment of the estimation, i.e. at the $z$-th iteration of the Markov Chain Monte Carlo (MCMC).

Selection of the right $y_{j[m]}$s from the demographic account corresponding to the cells in $X_m$ happens via an indicator function $I_i^{jm}$, and collapsing extra dimensions in $Y$ if $X_m$ has fewer dimensions than $Y_{[m]}$.

$$y_{j[m]} = \sum_i Y_i^{[m]} I_i^{jm} + \epsilon \tag{2.13}$$

The indicator function $I_i^{jm}$ takes value 1 if dimensions of cell $x_{jm}$ match the one of $y_i$, and 0 otherwise. The error term $\epsilon$ allows for $x_{jm}$ to be positive if $y_{[m]}$ cells happen to be all 0. Through prior distributions on the data model it is possible to include prior beliefs on the datasets and to allow for systematic biases. Usually data models do not need to be very complex as allowing for too much flexibility might affect estimation process and deviate from $Y$.

According to the accuracy or reliability of data, three main models can apply to $X$ with corresponding prior models and link functions ($g(\cdot)$) for parameter $\gamma_{jm}$. The link function $g(\cdot)$ changes according to the model in order to always have the transform $g(\gamma_{jm})$ such that

$$g(\gamma_{jm}) \sim N(\mu_{jm}, \sigma^2) \tag{2.14}$$

Parametrisation of $\mu_{jm}$ and $\sigma$ follow equations (2.8) and (2.9), as in the system model.

Models for $X$ are:

1. **Poisson**: same model as in equation (2.4) and (2.5), it is used when the source is not very reliable or accurate. Variance is restricted to be as large as the mean so, usually, the model provides output with larger credible intervals than with the other models and estimates can then be far from the data. The link function $g()$ is the logarithmic function as in equation (2.7).

$$\begin{aligned} x_{jm} &\sim Pois(\gamma_{jm}) \qquad \text{model without exposure} \\ x_{jm} &\sim Pois(\gamma_{jm}\omega_{jm}) \qquad \text{model with exposure} \\ \log(\gamma_{jm}) &\sim N(\mu_{jm}, \sigma^2) \end{aligned}$$

2. **Normal**: used for rather reliable data source, it allows for tuning variance parameter according to data quality. A Normal distribution might seem a questionable choice for counting people as it is continuous and defined on the whole real line. Despite this, if the estimate of $\gamma_{jm}$ and the of $\sigma$ are good, it performs quite well in practice. Moreover, mean values are usually far from zero and variances are quite low so values are seldom negative.

$$x_{jm} \sim N(\gamma_{jm}, \phi^2/w_{jm}) \tag{2.15}$$

where $w_{jm}$ is a weight term that can be introduced if variance has to vary across cells. If necessary, it is possible to use the integer-only version of the Normal distribution. The integer-only version is mainly used if, instead of separate arrays for immigration and emigration, there is only one array of net migration which can assume also negative values.

$$x_{jm} \sim \text{round}\big(N(\gamma_{jm}, \phi^2/\omega_{jm})\big) \tag{2.16}$$

For the Normal distribution the link function is the identity function therefore

$$\gamma_{jm} \sim N(\mu_{jm}, \sigma^2) \tag{2.17}$$

Variance parameter $\phi^2$ can take a fixed value or have a Half-t prior $\phi \sim t^+_{\nu_\phi}(0, A^2_\phi)$ as in equation (2.9).

3. **Poisson-Binomial**: this is a mixture model used for reliable and accurate data, each cell count is divided in a Binomial and a Poisson part. The higher the

Binomial probability parameter the lower the flexibility, meaning that trust in the data is high. The parameter of the Binomial "can be interpreted as the probability that a person or event is detected and appropriately enumerated by the dataset" (Bryant and Graham, 2013).

Each cell is then considered as the sum of a Binomial ($h_{jm}$) and a Poisson ($g_{jm}$) random variable.

$$\begin{aligned} x_{jm} &= h_{jm} + g_{jm} \quad \text{where} \\ h_{jm} &\sim Bin(\omega_{jm}, \gamma_{jm}) \\ g_{jm} &\sim Pois(\omega_{jm}(1-\gamma_{jm})) \end{aligned} \tag{2.18}$$

where $\omega_{jm}$ is, at the same time, the exposure for the Poisson part and the sample size for the Binomial. The Binomial part $h_{jm}$ is the number of people correctly included in cell $j$ and $g_{jm}$ is the "over-count", people that are counted twice or incorrectly included in cell $j$. Value of $g_{jm}$ is assumed proportional to the cell count. The expected value of the mixture is $E[x_{jm}] = \omega_{jm}$ and variance $\text{Var}[x_{jm}] = \omega_{jm}(1-\gamma_{jm}^2)$. The probability $\gamma_{jm}$ depends on the dataset but in general the model is quite robust to this choice, estimates are always very close to the original data. The link function for the Binomial distribution is the *logit* function so

$$\text{logit}(\gamma_{jm}) \sim N(\mu_{jm}, \sigma^2)$$

Parameters $\mu_{jm}$ and $\sigma$ follow equations (2.8) and (2.9) respectively.

## 2.4 Choice of prior distribution for hyper-parameters

If the structure for cells and $\gamma$s parameters is quite standard and choice is limited, the distribution choice for standard deviations $\sigma$ and, especially for components of parameter $\mu$s ($\beta$s coefficients as shown in equation (2.8)) is wider and is still a field of ongoing research for the model.

The choice of prior distributions is the strength but also the most discussed characteristic of the Bayesian framework. On the one hand prior distributions introduce a structure embedding informations not available or hidden in the data, and they strengthen the model when information from the data is weak or missing. On the other hand, they introduce a degree of subjectivity that could be misleading during the estimation process and which is absent, or hidden according to Bayesians, in Frequentist models. It is therefore important to carefully select prior distributions and test the sensitivity of the model to different choices.

The use of prior distributions can be particularly helpful in demography because demographic phenomena often present regularities and patterns demographers are aware of, and that can speed up estimation and forecast of population size, if embedded in prior distributions. For instance, mortality rates have now quite regular patterns, especially in countries with high life expectancy, and they are easier to estimate and to predict than migration. Migration is a complex phenomenon, even the definition itself of migrant changes according to different countries. Migration depends on economic, political and social factors that change over time so that a "sending" country can become a "receiving" country and vice-versa.

Fertility rates are an intermediate case, regularities can be identified and factors influencing it have been widely studied. For example mothers' age, the economic and social status, country healthcare, laws regulating maternity and paternity leaves, childhood related facilities, all these factors influence fertility. Nonetheless, these factors are partly influenced by national policies and migration which can both change over time making prediction on fertility a delicate point. The possibility to include *a priori* knowledge can help modelling any kind of phenomenon from regular to less stable ones, and the ability to choose suitable prior distributions helps to deal with complicated and realistic models untreatable otherwise.

In section 2.3 parameters $\mu$s and $\sigma$s are only quickly defined, but they actually play a central role in the model. Recalling equations (2.8) and (2.9), details are now further developed.

$$\mu_i = \sum_{k=0}^{K} \beta^{(k)}_{h^k_i} \tag{2.19}$$

$$\sigma \sim t^*_\nu(A) \tag{2.20}$$

Standard deviation $\sigma$ is unique for each array and it is assumed to be Half-t distributed (Gelman, 2006) with an updating process in the MCMC algorithm involving Slice Sampling (Radford, 2003) (see Appendix). Mean parameter $\mu$ is unique for each cell and it is a sum of $K$ coefficients that can be main effects or interactions among the array dimensions. Even if the combination of $\beta$s is unique for each $\mu_i$, i.e. there is a $\mu$ for each cell, the number of $\beta$s is much lower depending on the dimensions considered. For each cell "$i$" the value of $\mu_i$ depends on the sum of coefficients $\beta$s where each $\beta^{(k)}$ is a vector of length $h$ representing an effect included in the regression. The index $k$ refers to the variable the coefficient corresponds. The length $h$ depends on the number of categories each effect can take, e.g. $h = 1$ if it refers to the intercept, for a variable like "sex" $h = 2$ (values are "Female" and "Male"), or for Italian regions $h = 20$.

For example, let cell $i$ be the number of female (corresponding to sex number 1), in region number "16". Let the mean $\mu_i$ include intercept ($k = 0$), sex ($k = 1$) and region ($k = 2$) effects, then $\mu_i = \beta^0 + \beta^{(1)}_{1^1_i} + \beta^2_{16^2_i} = \sum_{k=0}^{2} \beta^{(k)}_{h^k_i}$. If an interaction sex:region is introduced then $\mu_i = \beta^0 + \beta^{(1)}_{1^1_i} + \beta^2_{16^2_i} + \beta^3_{(16\times 1)^3_i}$. Length of vector $\beta^3$ is the product of the number of region and sexes 20 × 2 and the element corresponding to the $16^{th}$ and first sex is 16 × 1.

Assumptions on vectors $\beta^{(k)}$s consider different distributions and research is still ongoing. At the moment the most common and experimented priors are the exchangeable prior (Normal distributions), dynamic linear models (DLMs) or Student's t distribution. These are analysed one by one in the following paragraphs. In very simple cases the mean $\mu_i$ can also be assumed to simply follow a standard Normal distribution, $\mu_i \sim N(0, 1)$. There are many possible prior models on $\mu_i$ allowing for more or less flexibility depending on the application. Complexity of prior distributions can vary but, as a general rule, when it comes to $\beta$ vectors, it is a good practice to keep complexity low when information is weak, especially for interactions as it is usually difficult to understand or identify clear patterns or their influence on the variables. Also, if the model includes interactions then also their corresponding marginal terms have to be included. For instance if an "age-time" interaction is included, then also "age" and "time" effects have to be included in the model; if a more complicated one as "age-time-region" is in the model, then also, "age-time", "age-region", "time-region" interactions and "age", "time" and "region" effect have to be included, like in most ANOVA models.

### 2.4.1 *Ad hoc* prior distributions, an example

Currently only standard prior distributions have been experimented but there is room for development in this field. An example of *ad hoc* prior distribution is provided in Wiśniowski *et al.* (2013) where the prior is built starting from expert opinions is. Authors' starting point is the difference on the registration of migration flows between origin and destination country. Ideally the number of people cancelling from country A to live in country B should match, or at least be very similar, to the number of people from country A registering in country B over the same period of time. It comes out that this is not the case. As an example they cite the case of Germany and Spain when, in the same year (2007), Germany registered 15, 515 immigrants from Spain and Spain only 3, 601 emigrants to Germany. In order to reconstruct the true value of migration flows starting from the country-specific registered values, Authors present an equation to relate these two quantities through four parameters: (i) accuracy of data collection; (ii) how much of the divergence is due to difference in the duration criteria used to qualify migrants, using the UN criterion as baseline (12 months); (iii) underestimation in capturing migration flows in different countries; (iv) a country-specific parameter of coverage in migration flows including subgroups of population like students or refugees. The estimation of these parameters is performed in a Bayesian framework where prior distributions are mixtures of distributions reflecting expert opinions collected through a two stages Delphi method. First of all this method requires a lot of time and work; secondly, reliability of results is arguable as, despite the research involved only experts, subjectivity and differences in responses were large. Therefore, the resulting mixtures from non-homogeneous responses are not very informative with respect to the effort required. Wiśniowski *et al.* (2013) were pioneer in this demographical prior distribution building process. The motivation of the research was the need of proper tools to take advantage of the amount of information and knowledge about demographic phenomena and what emerged was encouraging but also challenging for further research.

Another *ad hoc* prior distribution partially investigated but still needing experimentation is a mixed distribution proposed in Dunson and Xing (2009) and Kunihama and Dunson (2013). They investigate in a non-parametric way how to model trends among categorical variables and, in general, relationships among multivariate unordered categorical variables. They use Dirichlet process mixture of product of Multinomial distributions where weights change over time. If Kunihama and Dunson (2013) use Multinomials, Bryant and Graham (2015) have tested Normal distributions and the model have been proven to work as well. The model share some feature with intrinsic conditional autoregressive models (ICAR) and principal component. These models appear to be a parsimonious way to handle interactions if compared with other options, and it is also useful to model changes in patterns as the framework allows for breaks and for changes over time. It is still not very clear though how to set parameters and how they interact within the Demographic account model (DAM).

Keeping in mind that there is room for improvement and research in this field, the next sections describe the principal prior distributions used in the model. Of particular interest are distributions of the components of mean parameter $\mu_i = \sum_{k=1}^{K} \beta^{(k)}_{h_i^k}$ and priors for standard deviation terms. Instead, the prior on transformed parameter $g(\gamma_i)$ is stable as it is always a Normal distribution $N(\mu_i, \sigma^2)$, see equation (2.14). The priors described in the following sections are those implemented in the `demest R package`. So far, rather simple priors seem to be the best choice for this complex model, but there is space for experimenting and implementing new options.

### 2.4.2 Exchangeable prior

When elements $\beta_h^{(k)}$ seem not to have a specific pattern or reason to be sequentially linked to one another, the $h$ units are assumed to be exchangeable, i.e. their labels or ordering does not affect the distribution. This is the most general prior to use, but it is also quite flexible, as means and standard deviations can vary. It is also possible to control for covariates if they have a role in the exchangeability assumption, as sometimes there are factors not included in the model that prevent elements from being assumed exchangeable.

For vector $\beta^{(k)}$ with only one or two elements, such as intercept or effects with only two elements (e.g. "sex"), prior distributions are as follows

- For the intercept $\beta^{(0)}$ ($k = 0$ is always used for intercept as it does not represent any specific effect): $\beta^{(0)} \sim N(0, \tau_0^2)$
- For an independent two element vector $\beta^{(k)}$: $\beta_h^{(k)} \sim N(0, \tau_2^2)$, $h = 1, 2$, e.g. $1 =$ Female and $2 =$ Male

When vectors have length of one or two ($h < 3$) a simple prior is enough as it is not proper to mention exchangeability when the order of the vector cannot substantially change. Instead, when $\beta^{(k)}$ has three or more components and covariates can be introduced, there are four main model options for each element $h$ of $\beta^{(k)}$:

1. Normal without covariates: $\beta_h^{(k)} \sim N(0, \tau_k^2)$
2. Robust version without covariates: $\beta_h^{(k)} \sim t_{\nu\beta}(0, \tau_m^2)$
3. Normal with covariates: $\beta_h^{(k)} \sim N(z_h^{(k)}\eta^{(k)}, \tau_k^2)$
4. Robust version with covariates: $\beta_h^{(k)} \sim t_{\nu_\beta}(z_h^{(k)}\eta^{(k)}, \tau_k^2)$

In all cases the model is presented for one element ($\beta_h^{(k)}$) of vector $\beta^{(k)}$. In the robust version a Student's t distribution is used instead of a Normal. With a Student's t tails are heavier and further values from the mean have higher density values, i.e. there is less concentration of value on the mean. The value recommended in Bryant and Zhang (2018) for the degrees of freedom parameter is $\nu_\beta = 4$, implying thick tails, but any other valid value can be chosen.

In the version with covariates ($Z$), for each $\beta^{(k)}$ covariates are all standardised and stored in a $H_k \times P_k$ matrix $Z^{(k)}$ with elements denoted as $z_h^{(k)}$. Matrix dimension depends on the length of vector $\beta^{(k)}$ ($H_k$), and on the length of the covariate vector ($P_k$). As usually in regression model, the first element is the intercept and hence the first column of $Z^{(k)}$ is a column of 1s, whereas the vector of coefficients, $\eta^{(k)}$, has length $p = 1, \cdots, P_k$, where each element $\eta_p^{(k)}$ also has either Normal or Student's t prior.

$$\eta_1^{(k)} \sim N(0, A_0^2) \tag{2.21}$$

$$\eta_p^{(k)} \sim t_{\nu_\eta}(0, A_{\eta k}^2),\ p = 2, \cdots, P_k \tag{2.22}$$

Standard deviation $\tau$ can be either fixed (set to a pre-determined value, like 1 or 10, or equal to the standard deviation of data), or assumed Half-t distributed $\tau_k \sim t_{\nu_\tau}^{+}(A_{\tau k}^2)$ with degrees of freedom $\nu_\tau$ and scale parameter $A$ which can be fixed or depend on data standard deviation. Bryant and Zhang (2018) recommend $\nu_\tau = 7$ based on empirical results, but any other valid value for degrees of freedom can be chosen. Usually, the more uncertainty there is the lower is A, for example for interactions a good choice for scale value is half the one of main effect.

When vectors $\beta^{(k)}$ are large enough, it can be worth considering pooling options. If elements have similar value and can be imputed to the same probability distribution then a complete pooling is possible, i.e. all element have a unique prior distribution; if elements can be grouped according to a criterion then a partial pooling is performed, i.e. there are as many prior distribution as groups; if elements share no common feature and they all need their own distribution then there is no-pooling and there are $h$ prior, i.e. as many prior as elements of $\beta^{(k)}$. This means that elements $\beta_h^{(k)}$ can be normally distributed either all with the same mean, or with a mean common only to a subset of them, or with their own mean.

### 2.4.3 Dynamic Linear Model prior

The Dynamic Linear Model (DLM) prior is a convenient model for ordered parameters with higher correlation for neighbouring categories than for non-neighbouring ones. Typically, a DLM model suits variables such as time, age and, to some extent, also to education and income. Originally the DLM is a time series model suitable for time series with non-stationary components and Prado and West (2010) provide a good description. With respect to an exchangeable prior setting which is more general, if a DLM is suitable for the variable of interest, then this prior usually speeds up convergence. In its simplest version, a DLM prior is a local level model, and the $h$-th element of coefficient vector $\beta^{(k)}$ is assumed normally distributed with means $\alpha_h^{(k)}$ (level term), and linked to the previous term $h-1$ as in equation (2.24). Each $\beta_h^{(k)}$ has distribution

$$\beta_h^{(k)} \sim N(\alpha_h^{(k)}, \tau_k^2),\ h = 1, \cdots, H_k \tag{2.23}$$

with mean $\alpha_h^{(k)}$ that depend on the mean of the previous term $\alpha_{h-1}^{(k)}$:

$$\alpha_h^{(k)} \sim N(\alpha_{h-1}^{(k)}, \tau_{\alpha k}^2),\ h = 1, \cdots, H_k \tag{2.24}$$

when $h = 1$ then $\alpha_{h-1}^{(k)} = \alpha_0^{(k)} \sim N(0, A_0^2)$. A local level model works as a random walk where the level of the mean can be higher or lower than the previous one but has expected value equal to the previous one. Parameter $\alpha_h^{(k)}$ is the *level term.* If there is a trend in the pattern, i.e. the values $\alpha_h^{(k)}$ are expected to be always higher or lower than the previous one, then a *trend term* can be introduced. Let $\delta_h^{(k)}$ be the trend term for $\alpha_h^{(k)}$ then equation (2.24) becomes

$$\alpha_h^{(k)} \sim N(\alpha_{h-1}^{(k)} + \delta_{h-1}^{(k)}, \tau_{\alpha k}^2),\ h = 1, \cdots, H_k \tag{2.25}$$

and the general trend term $\delta_h^{(k)}$ has distribution

$$\delta_h^{(k)} \sim N(\delta_{h-1}^{(k)}, \tau_{\delta k}^2),\ h = 1, \cdots, H_k \tag{2.26}$$

$$\delta_0^{(k)} \sim N(0, A_{\delta k}^2) \text{ for } h = 0 \tag{2.27}$$

When a trend term is included then the DLM assumed is a local trend model.

Especially when considering trends in the long run, but sometimes also in other situations, it is not appropriate to expect upward or downward trends to always continue at the same pace. Sometimes trends tends to get weaker or reach a lower/upper bound so that an ordinary random walk as in equation (2.26) is not appropriate. A good option in those cases is a damped random walk where each step tends to be smaller than the one before it. In Bryant and Zhang (2018), Authors take the example of age-time interactions for mortality rates, pointing out that: “Human mortality rates have a characteristic age-profile, which recurs, with variations, across many populations. Damping prevents forecast age-profiles from departing too far from their observed historical average, which, arguably, increases their plausibility” (Bryant and Zhang, 2018). A damped random walk includes a *damping term* $\zeta \in \{0, 1\}$ that affects the value of the trend term mean so that equation (2.26) becomes

$$\delta_h^{(k)} \sim N(\zeta_k \delta_{h-1}^{(k)}, \tau_{\delta k}^2),\ h = 1, \cdots, H_k \tag{2.28}$$

$$\zeta_k \sim Unif(V_{min}, V_{max}) \tag{2.29}$$

As values get closer to 0, damping term effect on the random walk steps gets higher, whereas, if $\zeta = 1$ it goes back to the classic random walk. The range for $\zeta$, as shown in equation (2.28), can be chosen, but empirical results show the lower bound is hardly ever below $V_{min} \approx 0.8$. When a local trend model is assumed, damping is particularly useful for forecasts in the long run. As a general rule, when modelling interactions damping should not be used and variance should be fixed. Another option for damping term $\zeta$ is to assume a Beta prior possibly restricted to values inferred from the data.

In addition to level term, it is possible to include in the model for $\beta^{(k)}$ a *season effect* ($s^{(k)}$) and/or covariates ($Z^{(k)}$). Adding these terms to equation (2.23), distribution for term $\beta_h^{(k)}$ is

$$\beta_h^{(k)} \sim N(\alpha_h^{(k)} + s_h^{(k)} + z_h^{(k)}\eta^{(k)}, \tau_k^2),\ h = 1, \cdots, H_k \tag{2.30}$$

The covariates structure is the same as described in subsection 2.4.2 and equation (2.21). For seasons, let $S_k$ be the total number of seasons considered then, like for level and trend terms, the mean term depends on the value of the one before so that

$$s_h^{(k)} \sim N(0, A_{sk}^2),\ h = -S_k, \cdots, 0,\ S_k \text{ for the first season} \tag{2.31}$$

$$s_h^{(k)} \sim N(s_{h-S_k}^{(k)}, \tau_{sk}^2),\ h = 1, \cdots, H_k \text{ for the following seasons} \tag{2.32}$$

Standard deviations $\tau$s have Half-t distribution as in previous cases, with degrees of freedom $\nu$ and scale parameter $A$.

As in subsection 2.4.2, there is a robust version for $\beta^{(k)}$ prior (2.30):

$$\beta_h^{(k)} \sim t_{\nu_\beta}(\alpha_h^{(k)} + s_h^{(k)} + z_h^{(k)}\eta^{(k)}, \tau_m^2) \tag{2.33}$$

Informative prior assumptions on the terms just defined (local levels, local trends, damping terms, season effects or covariates) can improve the model fit and speed up computations. They can also improve convergence and lower auto-correlation functions. Nonetheless, prior distribution have to be handled with caution as they can distort the estimation.

During the estimation process of parameter $\beta^{(k)}$ a centring step has been introduced to mitigate identifiability problems often occurring during the estimation. Especially when level ($\alpha_h$) and season ($s_h$) parameters are included in the definition of $\beta_h^{(k)}$, there can be more than one combination of $\beta^{(k)}$, $k = 0, ...K$, providing a sensible and/or optimal $\mu_i$ (recall $\mu_i = \sum_{k=0}^{K} \beta_{h_i^k}^{(k)}$). This identifiability problem makes $\beta_h^{(k)}$ values determination hard. The problem has been reduced introducing a centring step during the process, i.e. at each iteration the estimate mean of each term is embedded in the intercept value.

### 2.4.4 Half-t or Folded-t distribution for standard deviation parameters

Half-t distribution is a special case of the Folded non-central t distribution introduced in Gelman (2006). Starting from a Folded non-central t distribution and restricting the prior mean to zero the result is the absolute value of a Student's t distribution centred at zero with two parameters: the scale term $A$ and degrees of freedom $\nu$, with probability density function

$$p(\sigma|\cdot) \propto \frac{1}{\sigma^n} \exp\left( - \frac{V_\sigma}{2\sigma^2}\right) \left(\sigma^2 + \nu_\sigma A_\sigma^2)\right)^{(\nu_\sigma+1)/2} \tag{2.34}$$

where $V_\sigma = \sum_{i=1}^n \left(g(\gamma_i) - \mu_i\right)^2$. Two special cases occur when $\nu = -1$ and when $\nu = 1$ giving respectively an improper Uniform density and a proper Half-Cauchy . Gelman (2006) praises the use of the Half-t with respect to a Uniform or an Inverse-Gamma in hierarchical models. The first has problem when the number of groups is small and the second one has problems for small values of standard deviation. The Half-t distribution is very flexible and it is generally weakly informative and suitable for both (i) restricting standard deviation from assuming very large values, unlike the Uniform and the sometimes recommended uninformative Inverse-Gamma($\epsilon, \epsilon$), where $\epsilon$ is a small number, and (ii) dealing with standard deviation values near zero, as it behaves better than the two others for very low values. If even the Half-t provides too large values in a weakly informative case, a truncated version can be used, but usually truncation takes place at very high quantiles (0.999 or 0.98) not heavily distorting estimation. This distribution works well in hierarchical models especially when the number of groups is small, it is a good option for a non-informative or weakly informative prior, it is conditionally conjugate with the Normal and, from a computational point of view, it is an easy distribution to sample from. For these properties it is a common choice for prior distribution for the standard deviation in the DAM.

## 2.5 Posterior calculations and account updating

The previous sections describe the assumptions made on data and parameters and the different options available for them. These assumptions will now be combined to calculate the posterior distribution the DAM aims to estimate. The number of parameters involved in the model is very high, especially if compared to the data. The only known parts are the datasets and the assumptions made *a priori* but then parameters of both the system model and data model plus the whole demographic account have to be estimated. The hierarchical structure with possibly three or four layers, the high number of parameters and the interactions among all the parts of the model create a complex structure whose full understanding will require further testing and time. Nevertheless, many of the assumptions ease the calculation of the posterior, and of the updating process.

The joint posterior distribution includes the demographic account $Y$ with the system model parameter set $\Theta$, the data model parameter set $\Omega$, the datasets $X$ and, possibly, the set of covariates $Z$. Its general form is

$$\begin{aligned} p(Y,\Theta,\Omega|X,Z) &\propto p(X|Y,\Theta,\Omega,Z)p(Y|\Theta,\Omega,Z)p(\Theta,\Omega|Z) \\ &= p(X|Y,\Omega)p(Y|\Theta)p(\Theta|Z)p(\Omega) \end{aligned} \tag{2.35}$$

with full conditionals:

$$\begin{aligned} p(Y|\Theta,\Omega,X,Z) &\propto p(X|Y,\Omega)p(Y|\Theta) \\ p(\Theta|Y,\Omega,X,Z) &\propto p(Y|\Theta)p(\Theta|Z) \\ p(\Omega|Y,\Theta,X,Z) &\propto p(X|Y,\Omega)p(\Omega) \end{aligned} \tag{2.36}$$

Simplifications from the first to the second line of equation (2.35) are possible because of assumptions made in the system and data models. Recalling that parameter sets $\Theta$ and $\Omega$, corresponding respectively to $Y$ and $X$, do not share any parameter, the following decomposition is possible for their joint distribution:

$$p(\Theta,\Omega|Z) = p(\Theta|Z)p(\Omega) \tag{2.37}$$

From assumption in equation (2.37), conditional independence for the joint distribution of $Y$ and $X$ follows:

$$p(Y,X|\Theta,\Omega,Z) = p(Y|\Theta,Z)p(X|Y,\Omega) \tag{2.38}$$

Moreover, recalling equation (2.10), in the system model it is possible to “drop” the covariate term $Z$ from $p(Y|\Theta,Z)$ as it is already considered in the estimation of the parameter set $\Theta$ whose distribution includes covariates $Z$, (third term in the second line of equation (2.35)).

As mentioned in section 2.3.2, the data model is conditioned on the value of population of the demographic account $Y$ so the model has form

$$p(X|Y,\Omega) = \prod_{m=1}^{M} p(X_m|Y^{[m]}, \Omega_m) \tag{2.39}$$

where $X = \{X_1, \cdots, X_M\}$ (as in equation (2.11)), and $Y^{[m]}$ is the demographic array referred to dataset $X_m$ and accordingly collapsed in order to have the same dimension as $X_m$. This part involving datasets only appears in the conditioning part of the joint posterior (equation (2.35)).

As already pointed out in section 2.3.1, an important difference in the updating process is the one between *point* and *interval* demographic arrays. In the first case no exposure term is involved whereas in the second exposure is normally used. Furthermore, if for point values, typically population arrays, only the initial values are updated, for interval values each cell has the same probability to be updated and population is updated only as a consequence to balance the demographic equation. When updating a term using exposure, then also the new population implied by the potential change has to be taken into account during the updating process. Recalling the demographic account division introduced in section 2.3.1, $Y = \{N, C_1, \cdots, C_L\}$, with all the demographic series assumed as independent, the distribution of the whole demographic account $Y$ is

$$\begin{aligned} p(Y|\Theta_Y, Z) &= p(N|\Theta_N, Z) \prod_{l=1}^{L} p(C_l|N, \Theta_l, Z) \\ &= p(N|\Theta_N) \prod_{l=1}^{L} p(C_l|N, \Theta_l) \end{aligned} \tag{2.40}$$

## 2.6 Updating process

The updating process of the whole model is quite complex and long, this section gives an overview of the main passages. Appendix contains more details about the whole process.

The DAM is a Bayesian model and it calculates the posterior distribution through standard Markov Chain Monte Carlo methods. Most of the time a Metropolis-Hastings algorithm is implemented and when possible Gibbs sampling is used. Despite methods being standard, the complexity of the model requires care during the updating process, and for the demographic account updating the process for generating proposals is customised. As shown in equation (2.36), there are three main full conditional distributions: the first for the demographic account $Y$, the second for the parameter set $\Theta$ of the system model, and the third for $\Omega$, the parameter set of the data model.

- $\mathbf{p}(\mathbf{Y}|\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{X}, \mathbf{Z})$: the updating process of the demographic account $Y$ is a challenging part of the model because $Y$ can be very large and because the balance equations must always be satisfied. This last constraint slows down the whole process as the equation needs an exact result. The respect of the equation is guaranteed by the proposal distribution. Once candidate terms are drawn from the proposal a Metropolis-Hastings acceptance step is performed. It turns out that candidate values both respecting the constraint and accepted by the Metropolis-Hastings algorithm are often values very close to current ones and this causes chains slow mixing and high auto-correlation.

  Despite it, the model performs quite well and it is reasonably fast. Following the description in (Bryant and Graham, 2013), there are five steps to generate a candidate value for the Metropolis-Hastings algorithm:

  1. A demographic series is randomly selected from the demographic account. Whereas any of the interval arrays $C_l$ can be chosen, only the initial population $N_0$ can be directly changed. Intermediate and last population values are updated during the following steps for balancing the account equations. Note that for the constraint to be satisfied it is impossible to change one cell at the time, at each iteration at least two cells need to changed.
  2. A cell among the chosen array is randomly selected. All the cells have the same probability to be selected. The "probability does not depend on the starting point, so it cancels out when the Metropolis-Hastings ratio is taken" (Bryant and Graham, 2013).
  3. The proposed value is drawn from one of the models of section 2.3. If the model includes an exposure term then the corresponding expected exposure is considered and not the current one.
  4. As a change in an array affects the consistency of the demographic account, then all population counts affected by the proposed value have to change in order to re-balance the account.
  5. Check that the subsequent population contains no negative values. If they do, return to step 3.

Once the proposal is ready, the Metropolis-Hastings ratio is calculated and accepted with probability

$$a(Y) = \min\left(1, \frac{p(Y^*|\Theta, \Omega, X, Z)}{p(Y^{(z)}|\Theta, \Omega, X, Z)} \frac{Q(Y^{(z)}|Y^*)}{Q(Y^*|Y^{(z)})}\right) \tag{2.41}$$

where $Q(\cdot)$ is the proposal density, $Y^*$ is the demographic account proposed with the candidate value and $Y^{(z)}$ is the current one with $z$ being the number of the iteration. Details and decomposition of ratio in equation (2.41) is in the Appendix.

- $p(\Theta|Y, \Omega, X, Z)$ and $p(\Omega|Y, \Theta, X, Z)$ contain different kind of parameters with different distributions and belonging to different level of the hierarchy. Following the division in section 2.4, it is possible to mainly divide them in five groups:

  1. Parameters $\gamma$: according to the model chosen for the data and the corresponding link function, the updating process algorithm is either a Metropolis-Hastings in the Poisson and Poisson-Binomial case or a Gibbs sampler, when the model is Normal and the link function is the identity function so that conjugacy property hold.
  2. Standard deviation parameters ($\phi$, $\sigma$, $\tau$): they all have a Half-t distribution and, because of the form and the property of the density, the best way to update them is through a slice sampler (Radford, 2003). Slice sampler is rather straightforward and performs well in the model. The Appendix gives further details.
  3. Parameters $\beta$: because of the model structure for both exchangeable and DLM prior assumptions, they are conjugate with the parameters $g(\gamma)$ therefore a Gibbs sampling is always possible for their update. This has changed since the first version of the model (Bryant and Graham, 2013). The 2013 model had a Poisson-Gamma conjugate model, i.e. $\gamma$s were assumed Gamma distributed instead of Lognormal as now. Because *beta*s can have more complex structures and, unlike $\gamma$s, they have identifiability problems, it was better to ease the update of the *beta*s rather than the $\gamma$s shifting conjugacy from the first to the second level of the hierarchy.
  4. DLM parameters $a_h^{(k)}$, $\delta_h^{(k)}$, $\zeta_k$ and $s_h^{(k)}$: the damping term $\zeta_k$ has a simple prior structure and can be updated through a simple Metropolis-Hastings algorithm. The others all have an autoregressive structure and their updating process relies on a Forward-Filtering Backward-Sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). This method has been

widely tested and it is a standard and efficient choice in Bayesian inference for linear state space models.

## 2.7 Model strengths and limitations

The field of Bayesian demography is developing fast, the investigation area is wide, demographic phenomena can be difficult to understand, and models develop and change quickly, especially since the last two centuries. Among the models proposed until now the DAM is complex and still need testing before being ready for a wide use but the aim is it to make it be available for National statistical institutes in order to perform population size estimation instead of traditional censuses as soon as possible. In this sense authors also developed a package for the open-source statistical programme `R` called `demest`. The package is available on the Statistics New Zealand GitHub repository at the link `https://github.com/StatisticsNZ/demest` but still needs further testing before being finally released.

Model complexity carries both advantages and disadvantages. The flexible hierarchical structure, along with the range of options available provides high adaptability to different needs and it is able to include a wide range of *a priori* information. The model can estimate both a whole demographic account and a single demographic array, and it is meant to work with different qualities of data. Bryant and Zhang (2018) provide examples with both reliable and unreliable data and show how it is possible to perform inference in both cases. In this sense the importance and the novelty of Bayesian demography is to provide results that reflect the quality of data, not aiming to give a precise answer about the population size but rather a sensible credible interval. The main idea is that it is more useful to have an interval that might be wide but right, rather than a wrong but precise estimate. Results in Bryant and Zhang (2018) provide very good and precise estimate for reliable data with rather narrow credible intervals and sensible but wider intervals for unreliable sources.

Drawbacks include the difficulty to estimates $\beta$ parameters. A first problem with the $\beta$s regarding their updating process has been solved by shifting from the Poisson-Gamma conjugate model in Bryant and Graham (2013) to the new model Bryant and Graham (2015) where conjugacy hold between parameters $\beta$ and $g(\gamma)$. The original model in Bryant and Graham (2013) was a Poisson-Gamma model so that $y_i$ and $\log(\gamma_i)$

distributions were

$$\begin{aligned} y_i|\gamma_i &\sim Pois(\gamma_i) \\ \gamma_i|\xi, \mu_i &\sim \Gamma(\xi, \xi/\mu_i) \\ \log \mu_i &= \sum_{k=1}^{K} \beta^{(k)}_{h_i^k} \text{where} \\ \beta^{(k)}_{h_i^k} &\sim N(\eta^{(k)}, \tau_k^2) \end{aligned} \tag{2.42}$$

The new version in Bryant and Graham (2015) considers a Poisson-Lognormal structure, as presented in section 2.3. With respect to the Poisson-Gamma model, it loses conjugacy in the first level of the model, but it simplifies the updating process in the second and third level involving $g(\gamma)$ and $\beta$. The Poisson-Lognormal model is a common structure in hierarchical modelling and, from a computational point of view, it has proven to be a better choice than the previous one (Bryant and Graham, 2013). Another problem that has been reduced but not completely solved is the weak identifiability of many parameter. In section 2.4.3 $\beta^{(k)}$s identifiability problems have been mentioned, but also other parts of the model suffer of it. All the model aims to produce a coherent result but sometimes there is more than one value satisfying this requirement. With $\beta$s and migration count/rate cells this is particularly clear. For instance, when estimating immigration and emigration, if they both increase of the same amount this leaves the balance constraint satisfied but values assumed by both migration values could be very far from its actual value. A very simple example, let population at time 0 be $P_0 = 100$ and population at time 1 be $P_1 = 110$, let births be $B = 5$, deaths $D = 3$ and the true value of immigration and emigration be respectively $I = 14$ and $E = 6$ the balance equation (2.2) becomes $P_1 = P_0 + B - D + I - E = 100 + 5 - 3 + 14 - 6 = 110$ but if instead of $I = 14$ and $E = 6$ values estimated are $I = 24$ and $E = 16$ the equation is still satisfied, but with values much larger than the true ones ($P_1 = P_0 + B - D + I - E = 100 + 5 - 3 + 24 - 16 = 110$). Something similar happens with the $\beta$s. For each $g(\gamma_i) \sim N(\mu_i, \sigma^2)$ the parameter $\mu_i$ is the sum of $k$ $\beta$ parameters (equation (2.8)) but, as in the migration example there are many combinations of $\beta$s that can provide a sensible result for $\mu_i$. For reducing this problem, a centring process has been introduced so that, at each iteration, the level of the $\beta$s is embedded in the intercept value. Another way, more practical, is to carefully act on prior assumptions, e.g. reducing variance, assuming a range for the intercept.

Other aspects of the model that can slow down the estimation process or limit its efficiency are the high number of parameters with respect to the number of data, and the balance equations. The problem of high number of parameters with respect to data is quite common in Bayesian inference, and when data information is low prior distributions are important. In a field like demography, a Bayesian framework allowing for *a priori* information inclusion is ideal and it somehow compensates the lack of data. Nevertheless, *a priori* choices can cause distortion and harm the estimation. Furthermore, if the model is well defined, MCMC models should reach convergence even with few data although the estimation might take longer than with more accurate data. The demographic balance equation constraint requires that at each iteration all the cells of the demographic account must be consistent making sometimes a candidate value hard to find. To overcome this limitation, and in order to have a consistent candidate value and a good acceptance probability, the algorithm tends to propose candidate values close to the current ones. This increases the acceptance rate but it often causes high auto-correlation within the chains requiring sometimes high thinning. Overall, this solution performs well but it is possible that other updating techniques could improve the algorithm. An alternative could be the Adaptive multiple importance sampling (AMIS) proposed by Cornuet *et al.* (2012) where the proposal density can change and adapt at each iteration. The idea is to initially set a loose constraint (e.g. constraint must be satisfied with an error of ±10%) so that it only shrinks the parameter space and then gradually narrowing it until the estimate perfectly satisfies the balance equation. This method would heavily transform the method and at the moment it is just an idea, but the optimisation of the candidate value generation process is an area to further investigate.

Another difficult task is to identify the right model structure both for the system and the data model. The system model has to identify the general features of the population, e.g. what effects and interactions to include, presence of trends, seasonality or important covariates. If these choices could depend on previous knowledge, and are common to any generalised linear model, the choice of prior distribution on the variance terms is less evident because, depending on it, convergence can slow down or speed up. Despite being a delicate point convergence is usually reached and Bryant and Graham (2013) performed sensitivity tests with encouraging results.

Unlike the system model, the task of the data model is to incorporate prior knowledge of each dataset. Each dataset has its specific structure and data model resulting in a highly customizable structure. For example, if there are reasons to believe that coverage or accuracy change across a dimension of the dataset, then this information can be included in the prior model. Data model allows for much more flexibility than the system model (see section 2.3.2), but including too much flexibility and letting parameters vary for too many dimensions can eventually affect estimation more than it helps. Because the estimation of the data model has a conditioning on the true value of the population $Y$, if the data model is too complicated, it becomes difficult to control what features reflect actual characteristics of populations and what only arises because of the choice of the model. For example, given a dataset on births, if the mean value function includes a dependence on time and age of the mother, it is like saying that the dataset accuracy depends on these factors. If the reliability of data depends on many factors, then the estimation of all the parameters becomes very complicated. In this sense for data model "less is more", unless there is a true bias linked to a factor (as time or age), then keeping data model as simple as possible usually gives better results.

At the moment model checking is performed through replicate data and held-back data. Despite held-back data are typically used for forecasting, the technique works also for estimation. Calculation of criteria like Watanabe-Akaike information criterion (WAIC) , Widely applicable Bayesian Information Criterion (WBIC) Watanabe (2010), Akaike Information Criterion (AIC) or Deviance Information Criterion (DIC) with such a complicate posterior distribution is challenging and not possible to calculate with the current `demest R package`. When the whole model estimation will be more stable, `demest` package extensions including such functions will be implemented.

When choosing a Poisson distribution as prior for the data model, an implicit assumption of equi-dispersion is made on the data, since mean and variance have the same value in this distribution. The equi-dispersion assumption can affect the measurement process as the level of dispersion, and hence of variance, can be higher or lower than the mean value of the cell. Dispersion can vary according to the structure and characteristics of a dataset (e.g. number and length of dimensions, periodicity, accuracy, data collection strategy, updating and revisions), and the population considered. Usually a higher level of dispersion is expected for heterogeneous datasets. Variance corresponding to a population specified in a cell of the demographic account can be higher than the population mean if people have different demographically relevant characteristics as life style, revenue, propensity to declare life events to administrative offices, fertility, mortality and migration rates. The opposite happens if the population is homogeneous, in this case variance should be lower than the mean, as people share many characteristics. This make it easier to make assumptions and a low dispersion can reasonably expected. Unfortunately the Poisson distribution always implies equal mean and variance. The solution suggested for this last point is the use of the Conway-Maxwell Poisson (CMP) distribution (Conway and Maxwell, 1962), in chapter 4. Originally introduced in the 1960s and then proposed again in mid 2000s in Shmueli *et al.* (2005), this distribution is a discrete distribution, similar to a Poisson distribution but it has an additional parameter that allows for modelling under-, equi- and over-dispersion.

# Chapter 3

# Application to the Italian case

## 3.1 Introduction to Italian data

The Italian national statistical institute (Istat) is continuously improving its services and harmonising procedures according to European and international standards. The harmonisation, integration and digitalisation of data sources are priorities for all the aspects Istat takes care of. For resident population and other related data (births, deaths and migrations), Istat cooperates with regional and local offices and has started systematic procedures to reduce formal and substantial errors, i.e. respectively errors or omitted information in the documents and information inconsistency with respect to other data available. One of the greatest effort concerns census data. In October 2018, the “permanent census for population and housing” is starting and it represents a revolution for Italian official statistics. Also for the last traditional census in 2011, important innovations were introduced. An important effort was made for the Post Enumeration Survey (PES, introduced in Commission Regulation n. 1151/2010), a three-years process for checking census information and correcting over- or under-coverages due either to errors committed during the census or by administrative delays or errors. During the PES, municipalities compare Census data with their lists (*lista anagrafica comnunale* LAC) and correct errors using on-line softwares like SIREA (*Sistema di revisione delle anagrafi*). For census data, municipalities cooperate with Istat and make a major effort during all the process and, eventually, data are available for further studies, trustworthy and provide information not only on population size but also on its characteristic (Istat, 2016). For other data, Istat follows Eurostat standards and checks data from a qualitative and quantitative point of view, cooperating with other administrative offices and taking advantage of technology, for a direct link and comparison between sources, and also of new methods like record linkage (see for example Tancredi and Liseo (2011)).

Despite efforts and improvements, it is difficult and long to change and modernise procedures according to the latest standards, especially in big institutions and delays and errors have always to be accounted for. If, on the one hand, it is normal to find inconsistencies in the datasets even after revisions and checks, on the other hand this has to be avoided as much as possible as it can be misleading for the NSIs estimations and cause problems. With respect to Italian data, births and deaths errors are quite low and most of the time corrected within one year. More difficult is to have a clear idea of migration and resident population. Migration problems are addressed in section 3.4. Resident population data are kept in municipality offices and suffer from all other series inconsistencies as it depends on the update of other series. Municipalities calculate the total population year after year through the demographic balance equation so that counts should always be consistent, at least in theory. Despite this procedure might appear straightforward, the regularly updated data hardly ever match with census data. For the last census, for example, at January, $1^{st}$, 2011 population was estimated to be 60.626.442 according to the municipalities registers, but the census, referring to population at the $9^{th}$ of October 2011 reported only 59.433.744 people resident in Italy (results after PES). Population between the last two censuses (2001 and 2011) has then been completely reconstructed according to census results and was estimated at 59.364.690 people for the beginning of 2011, a difference of almost 1.300.000 people, i.e. $\sim 2\%$ less than the former value. The following sections describe results of the model proposed by (Bryant and Graham, 2013) applied to Italian data. The demographic system considered is the population resident in Italy, with classification varying according to the model. With the datasets considered dimensions available include region, five years age group, sex and time. Even if not all the dimensions are available in the same dataset, the missing dimensions can be found in another dataset referring to the same series. A person enters the demographic system either by birth or immigration (international or interregional) and exits by death or emigration.

The next two sections contains examples for deaths and births modelling with preliminary analysis and model selections steps. Death counts are addressed before births as their modelling is less problematic, then the more complex case of births is considered. Italian migration data are presented in section 3.4 with problems related to its estimation. Eventually section 3.5 provides examples of complete demographic account estimation according to different dimensions. In each section datasets used are introduced with a preliminary analysis using direct estimates, then the actual estimation process is performed, checked and results presented.

## 3.2 Death counts model

After describing the data and performing preliminary analysis, the section presents the structure of the chosen model with its *a priori* assumptions. Then results of parameter estimation follow with graphs and interpretation. Model checking and selection are performed via held-back data technique. Results provide not only estimated mortality rates and counts but also life expectancies from different models. Life expectancy is calculated following steps in Preston *et al.* (2001) and refers to periods and not to cohorts.

Dimensions considered in these analysis are age, time and sex. Region dimension was also available, but the effect checked in preliminary analysis and also in few attempts does not show clear or regular patterns and did not provide better results than those displayed here. In order to not over-complicate the model and to decrease computational costs for the analysis region dimension is not considered in this death counts model.

### 3.2.1 Data and preliminary analyses

In order to estimate death rates, two different datasets have been compared, both coming from the Italian national statistical institute (Istat). The most detailed dataset is the table by death causes, denoted by $X_2$. In this dataset age, sex, region of residence and cause of death are available. The data collection process Istat has implemented is described in figure 3.1. There are three main steps: (i) the distribution of official forms from Istat to Istat regional offices or directly to municipalities and then from municipalities to hospitals, *Aziende sanitarie locali* (ASL, literally local health companies) and general practitioners; (ii) the information collection. Doctors are responsible for filling the documents sent by Istat with information about death circumstances and personal details of the dead person. The documents are then sent to the municipalities, to the ASL and from the municipalities to the prefectures; (iii) the forms are eventually returned to Istat offices. When documents go from the health system to the administrative system they have local qualitative and quantitative control, plus an additional last control at central level (last Istat box in figure).

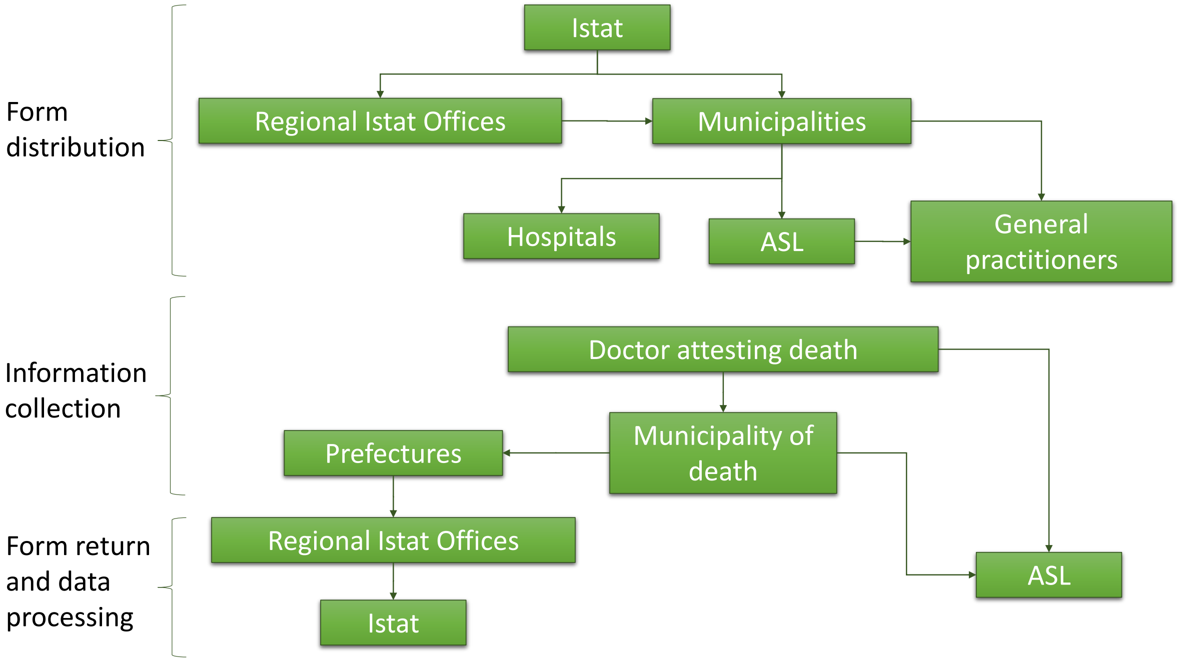
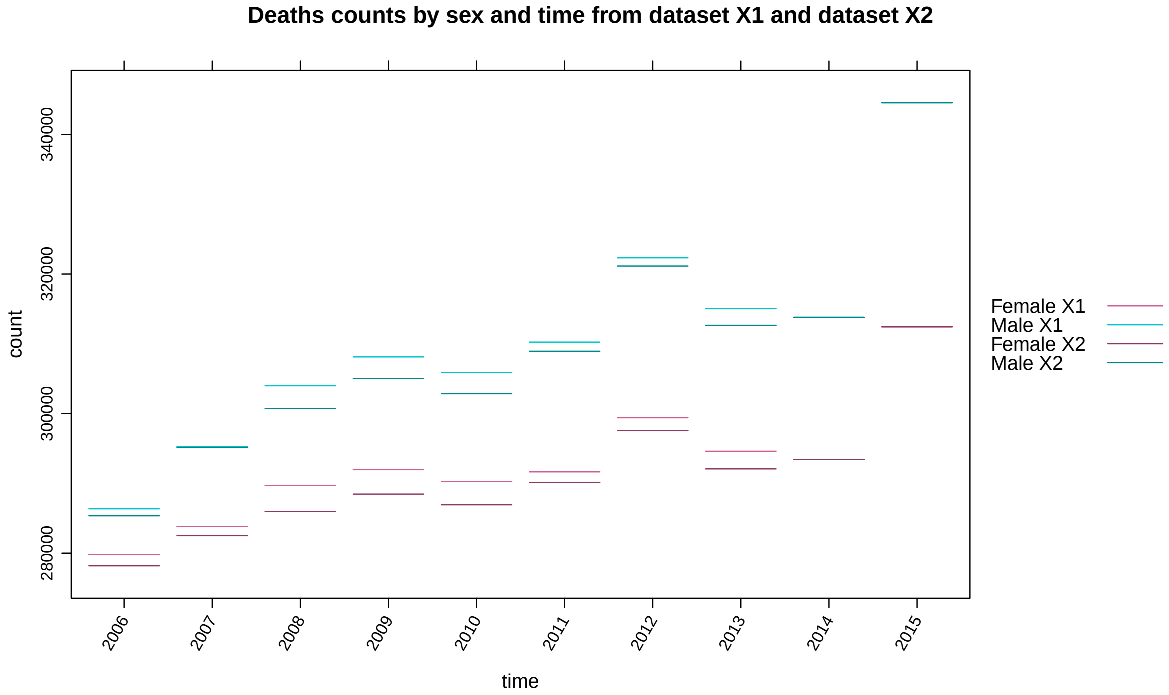
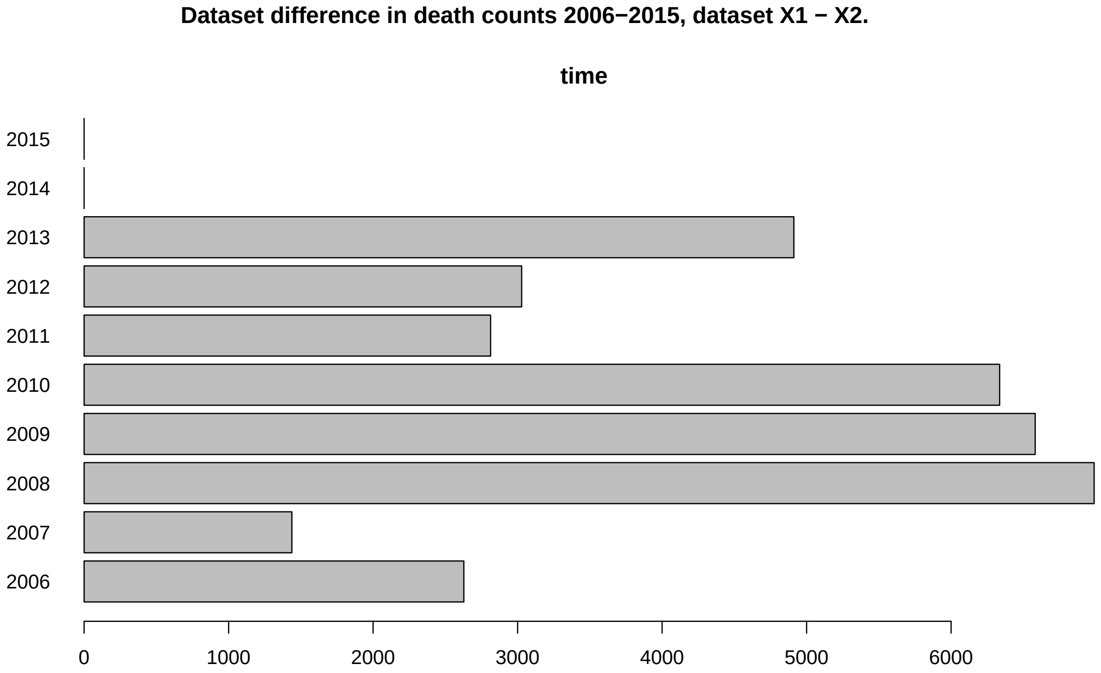


FIGURE 3.1: Information flow for death causes survey. Three phases of (i) form distribution, (ii) information collection and (iii) form returning and data processing. Source: `www.istat.it`.

Preliminary results on death count data are published after one year and final data after two years. These data only account for people dying in Italy, people resident in Italy but who have died abroad are not included. A more complete dataset is the one provided by municipalities to Istat and inserted in the national and regional demographic balance before being published on the Istat official annual documents (*Annuario statistico italiano*). Unfortunately this dataset does not provide age dimension. This dataset from *Annuario statistico italiano* is denoted by $X_1$. Figures 3.2 show a comparison between these two datasets by sex and time (upper figure), and then the differences between the dataset at national level during the years $(X_1 - X_2)$. Only in years 2014 and 2015 they coincide because procedures have been harmonised. As they provide different dimensions, both datasets are needed for a complete analysis.



(a) Death counts for years 2006-2015 from dataset $X_1$ (*Annuario statistico italiano*), males in light turquoise and females in light pink, and from dataset $X_2$ (Death causes survey), males in dark turquoise and females in dark pink.



(b) Differences by time between dataset $X_1$ (*Annuario statistico italiano* and dataset $X_2$ (Death causes survey), i.e. $X_1 - X_2$, from 2006 until 2015.

FIGURE 3.2: Comparisons between datasets $X_1$ (*Annuario statistico italiano*) and dataset $X_2$ (Death causes survey) from 2006 until 2015 at national level.

The most reliable dataset is the one containing values published on Istat official annual documents (*Annuario statistico italiano*). They are the data used for calculation of the national demographic balance. The dataset provides region, sex and time dimensions and is denoted by $X_1$. The second dataset has overall lower counts but it has also age dimension by five years groups, the first group is "$0 - 4$" and the last "90+". The dataset is denoted by $X_2$. Analysis considers 10 years, from 2006 until 2015. The last age group considered (90+) is quite a low age considering the current longevity and the increase in life expectancy in general. The population reaching the age of 90 and surviving until much older ages is increasing as figure 3.3 shows.

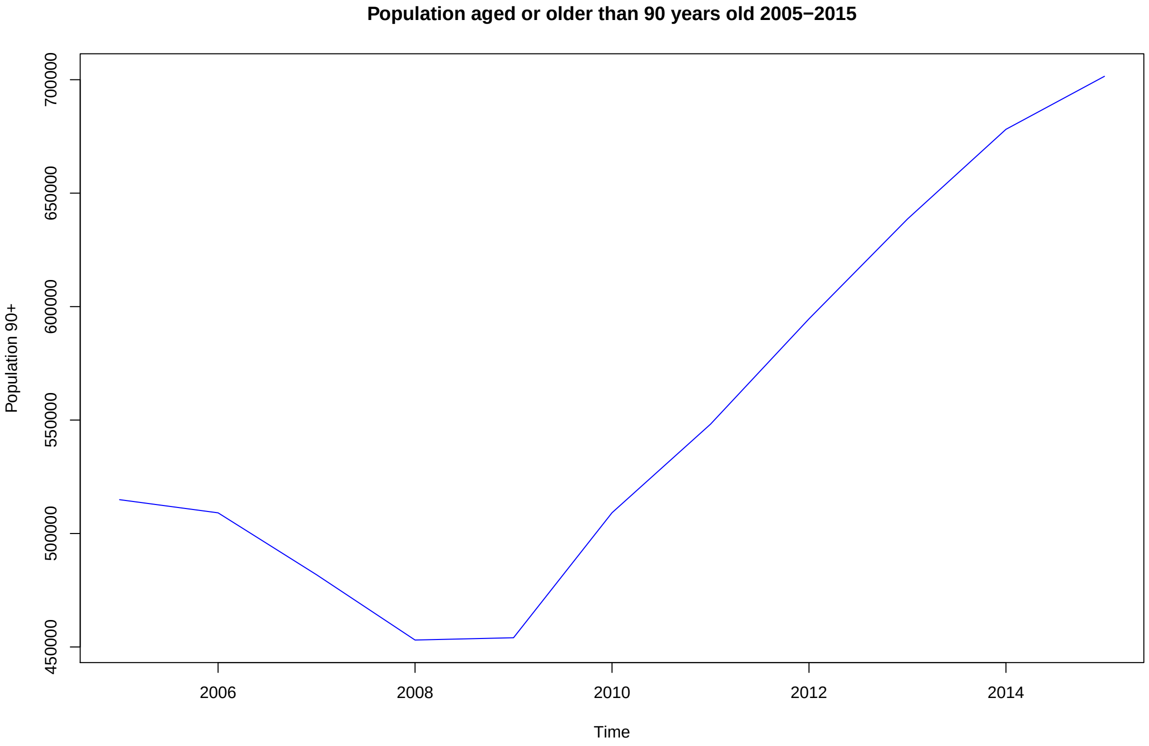


FIGURE 3.3: Population belonging to age group 90+ from 2005 to 2015. From the end of 2009 the size of this groups has constantly increased.

Unfortunately datasets provided by Istat have different last age groups depending on the year. From 2006 to 2009 the last age group is "100+", for years 2010-2013 is "90+" [1], and from 2014 on is "95+". Given this initial heterogeneity, the choice of considering the lowest class ("90+") for all the datasets has been made to simplify results and computations. At a larger scale, the debate of projected mortality trends is

[1] On the new Istat data website `http://dati.istat.it` now class "95+" is available also for year 2011-2012. At the time data have been provided by Istat these table were not available and, in any case, data for year 2010 still considers only until class "90+". The choice of considering "90+" as the last class stays therefore unchanged.

worth mentioning. If life expectancy will keep increasing or if it will reach a “mortality plateau” Vaupel (2014) is still debated among demographers, the topic is increasingly being analysed at international level, and different measure of longevity are investigated, see Lee (2006); Oeppen (2006); Oeppen and Vaupel (2002, 2006) for further discussion. For life table calculations, the last age group causes a problem as it is an open-ended age interval. Closing the life table, i.e. inserting the final estimate of the person-years of life expected in the last open-ended age group (usually denoted by ${}_nL_x$), is a problem, as everybody dies in the last interval, but no information is provided about the number of years lived after the last age group is reached. The usual approaches are (i) the imputation of a value taken from another suitable life table, (ii) using empirical death rates, (iii) make assumptions about the oldest reachable age and apply a trend to estimate ${}_nL_x$. Once ${}_nL_x$ is calculated, it is possible to obtain life expectancies. In the following applications direct death rates are used to close life tables.

The analysis and estimation of mortality rates considers only the two mentioned datasets. The choice to work only with these two datasets has mainly three reasons: (i) differences are relatively low ($\sim 1\%$ difference between $X_1$ and $X_2$) so keeping only two datasets allows not to overcomplicate the estimation; (ii) model is meant to work with sources and datasets of different reliability and level of completeness so it is a good way to test it; (iii) other institutes supplying data on Italian death rates (Eurostat, Human mortality database, OCSE...) still rely +on Istat data therefore a better quality cannot be expected or justified using them. Exposure terms have all the dimensions required by deaths data (age, sex, region and time), and they are calculated from population counts based on municipalities data on resident population according to equation (2.6). Figure 3.4 shows mortality rates by age, sex and time on a logarithmic scale. They are direct estimates, calculated dividing deaths counts by exposures. Working with rates on a log-scale is useful when dealing with small values and with different scales as it emphasizes relative differences rather than absolute differences and it is an usual procedure in demography as in other fields.

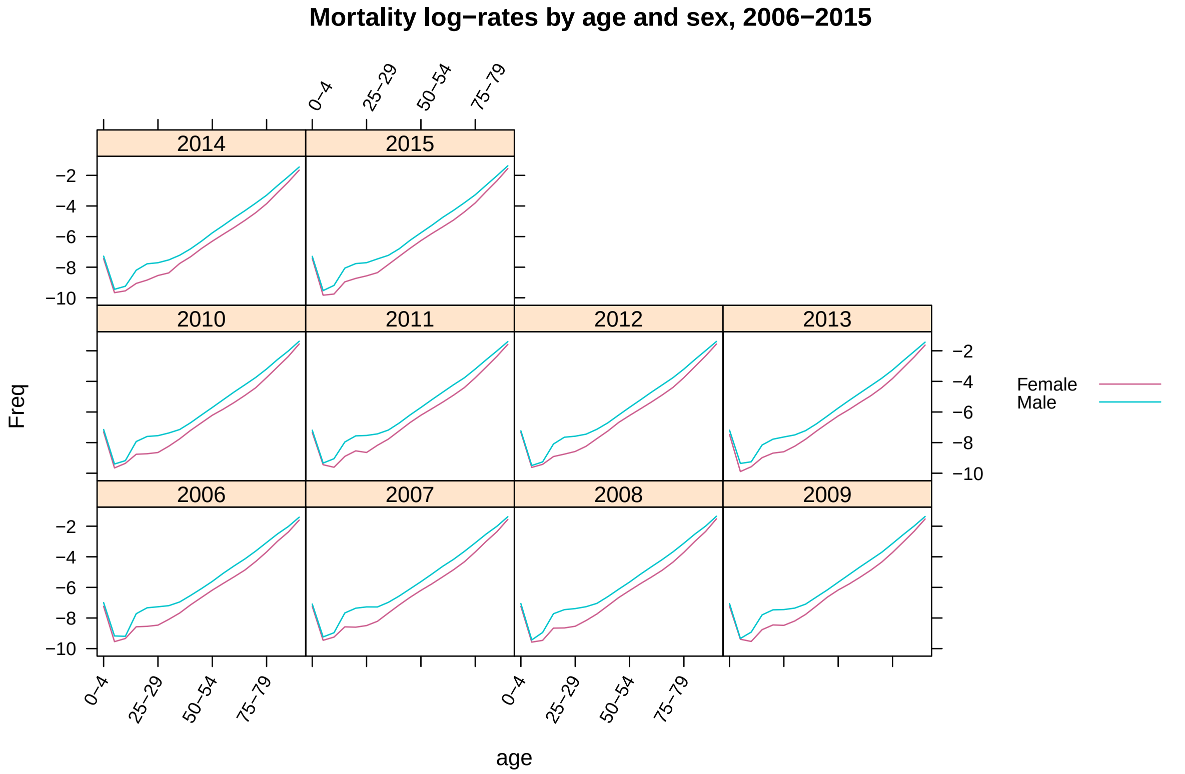


FIGURE 3.4: Mortality log-rates by 5 years age group (from $0-4$ until 90+) , sex (female in pink and male in light blue) and year (2006-2015). Males have higher mortality rates in every age group but in the extremes ($0-4$ and 90+) when the level is almost the same for both sexes.

Before the estimation, it is useful to conduct some preliminary analyses in order to identify what effects could be included in the model. First of all, a graphical analysis can be helpful. From figure 3.4, for example, an age and sex effect are quite clear. Besides this intuitive procedure, it is also possible to decompose log-rates in order to isolate and quantify effects in a more formal way. Bryant and Zhang (2018) report the following formulas:

“Let $m_{ast}$ denote the direct estimate of log mortality rate for age group $a$, sex $s$, and year $t$, where $a = 1, ..., A$, $s = 1, 2$, and $t = 1, ..., T$. The overall average log mortality rate is then”

$$\lambda_0 = \frac{1}{2AT} \sum_{a=1}^{A} \sum_{s=1}^{2} \sum_{t=1}^{T} m_{ast} \tag{3.1}$$

the age, sex and time effects are

$$\lambda_a^{age} = \frac{1}{2T} \sum_{s=1}^{2} \sum_{t=1}^{T} m_{ast} - \lambda_0 \tag{3.2}$$

$$\lambda_s^{sex} = \frac{1}{AT}\sum_{a=1}^{A}\sum_{t=1}^{T} m_{ast} - \lambda_0 \tag{3.3}$$

$$\lambda_t^{time} = \frac{1}{2A}\sum_{a=1}^{A}\sum_{s=1}^{2} m_{ast} - \lambda_0 \tag{3.4}$$

and interactions

$$\lambda_{as}^{age:sex} = \frac{1}{T}\sum_{t=1}^{T} m_{ast} - \lambda_0 - \lambda_a^{age} - \lambda_s^{sex} \tag{3.5}$$

$$\lambda_{at}^{age:time} = \frac{1}{2}\sum_{s=1}^{2} m_{ast} - \lambda_0 - \lambda_a^{age} - \lambda_t^{time} \tag{3.6}$$

$$\lambda_{st}^{sex:time} = \frac{1}{A}\sum_{a=1}^{A} m_{ast} - \lambda_0 - \lambda_s^{sex} - \lambda_t^{time} \tag{3.7}$$

Figures 3.5 and 3.6 show main effects and interactions decomposition. For main effects it is clear that time, age and sex effects are all significant which will be confirmed in the model estimation. Except from the first age group "$0 - 4$" rates increase with age, women values are lower than for men and rates generally decrease with time even if over a quite small range $(-0.1 - 0.1)$. The small time effect is mainly due to the little number of years considered. Also interactions considering time are not very significant.

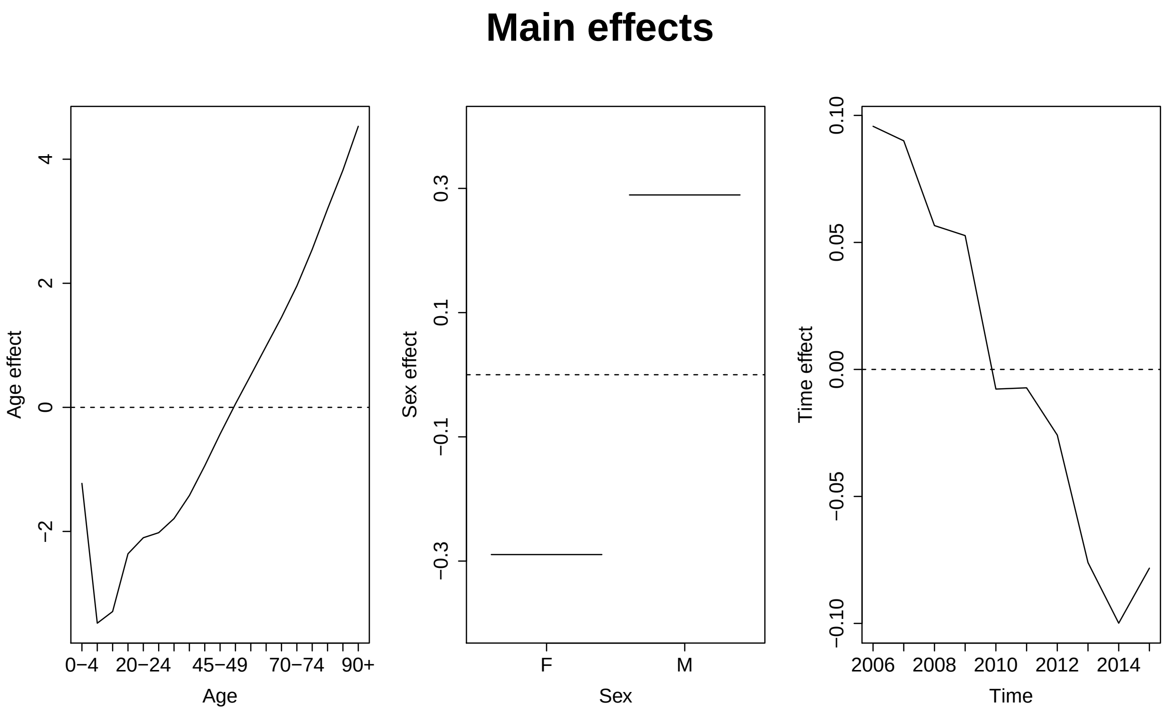


FIGURE 3.5: Computation of main effects from overall mortality log-rates. *Left*: Age effect, dropping from the first age group $(0 - 4)$ to the second $(5 - 9)$ and then constantly rising with a bump at age groups $20 - 24$ and $25 - 29$. *Centre*: Sex effect, with lower mortality for females. Note, the values are symmetrical for males and females. *Right*: Time effect, lowering from 2006 until 2014 and rising again in 2015.

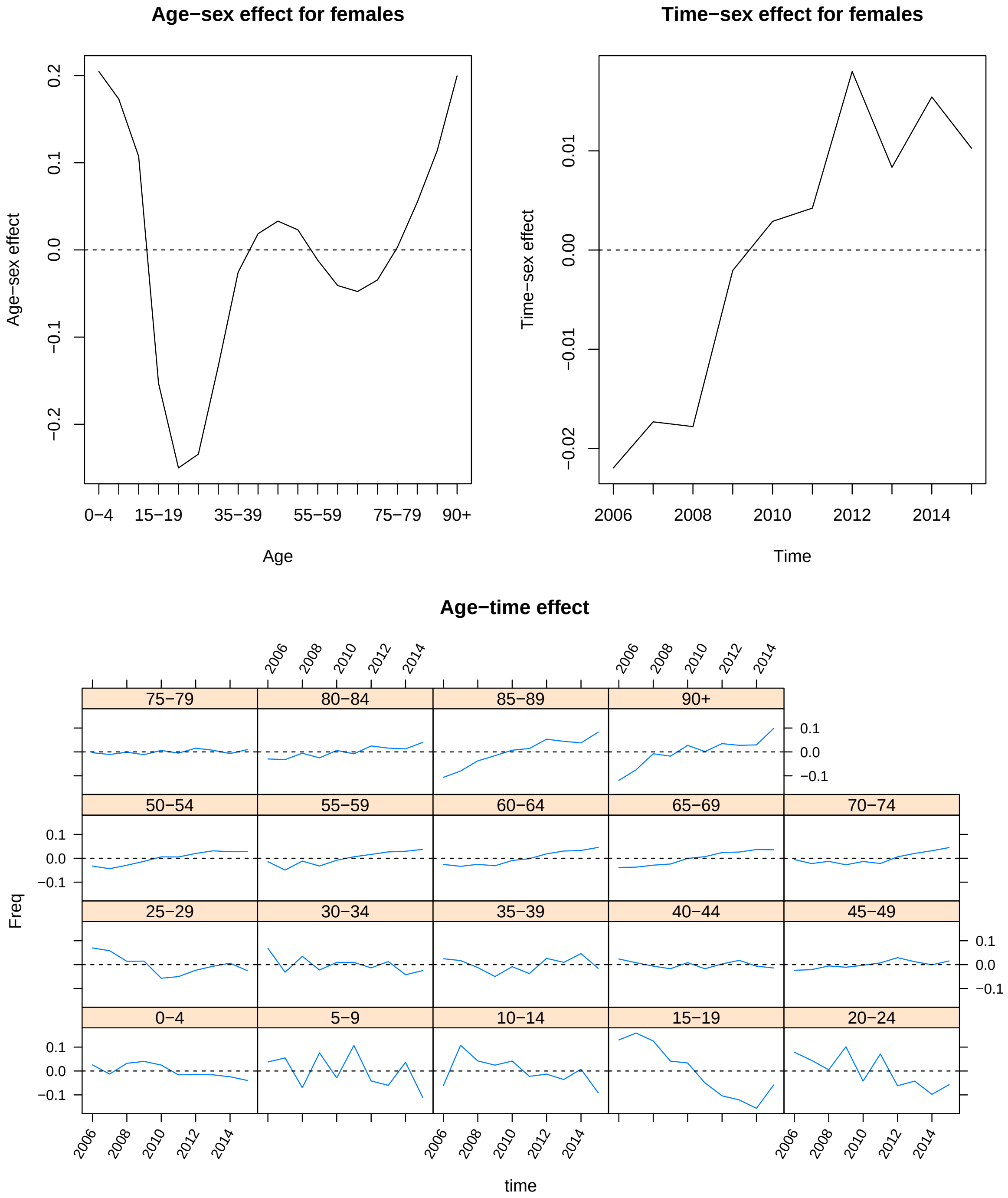


Figure 3.6: Computation of interactions from overall mortality log-rates. *Up-left*: Age-sex interaction for females has the highest magnitude among interactions. 5 years age group (0 – 4 - 90+). *Up-right*: Time-sex interaction for females with increasing pattern but very low magnitude. *Bottom*: Age-time interaction with irregularities in the young ages mainly due to the small counts in these cells.

Age-sex interaction is quite significant, depending on the age females mortality rates are higher or lower than males' one and the range goes from $-0.2$ until $0.2$. For time-sex interaction an effect exists but its range $(-0.02, 0.02)$ is close to zero and therefore it appears to have little impact on rates. Almost the same happens with the age-time effect. For the two oldest age groups the effect is a bit higher but elsewhere it is either very irregular, until 24-29 years old, or very small. The fact that time effect and interactions including time are small is probably linked to the few number of years considered. There is no reason to expect death trend and level to significantly change in Italy during the period $2006-2015$.

### 3.2.2 System and data models

The system model chosen for modelling deaths in Italy from 2006 until 2015 has the form displayed in equation (3.8). Deaths counts, classified by age, sex and time, $y_{ast}$, are i.i.d. [2] random variables from a Poisson distribution with exposures $\omega_{ast}$ and death rates $\gamma_{ast}$.

$$\begin{aligned}
&y_{ast} \sim Poisson(\gamma_{ast}\omega_{ast}),\ a=1,...,7,\ s=1,2,\ t=1,...,10\\
&\log(\gamma_{ast}) \sim N(\mu_{ast},\sigma^2)\\
&\mu_{ast} = \beta^0 + \beta_a^{age} + \beta_s^{sex} + \beta_t^{time} + \beta_{as}^{age:sex}\\
&\sigma \sim t_7^*(1)
\end{aligned} \tag{3.8}$$

Mean $\mu_{ast}$ includes intercept, main effects of age, sex and time and age-sex interaction. Intercept and sex effect have a fixed exchangeable prior whereas age, time and age-sex effects have a DLM prior:

$$\begin{aligned}
&\beta^0 \sim N(0,10^2)\\
&\beta_s^{sex} \sim N(0,1)\\
&\beta_a^{age} \sim N(\alpha_a,\tau_{age}^2)
\end{aligned}$$

with level and trend terms: $\alpha_a \sim N(\alpha_{a-1}+\delta_a,\tau_\alpha^2)$, $\delta_a \sim N(\delta_{a-1},\tau_\delta^2)$

$$\beta_t^{time} \sim N(\alpha_t,\tau_{time}^2)$$

with level and trend terms: $\alpha_t \sim N(\alpha_{t-1}+\delta_t,\tau_\alpha^2)$, $\delta_t \sim N(\delta_{t-1},\tau_\delta^2)$

$$\beta_{as}^{age:sex} \sim N(\alpha_{as},\tau_{age:sex}^2)$$

[2] independent identically distributed

with level and trend terms: $\alpha_{as} \sim N(\alpha_{as-1} + \delta_{as}, \tau_\alpha^2)$, $\delta_{as} \sim N(\delta_{as-1}, \tau_\delta^2)$ damping term and standard deviation have the same form in all priors:

$$\begin{aligned} &\phi \sim Unif(0.8, 1) \\ &\tau_{age}, \tau_{time}, \tau_{age:sex}, \tau_\alpha, \tau_\delta \sim t_7^+(1) \end{aligned}$$

The datasets have two different data models. The more reliable one, $X_1$, has prior:

$$x_{1st} \sim N(\gamma_{1st}, \phi_{1st}^2) \tag{3.9}$$

Parameters $\gamma_{1st}$ and $\phi_{1st}$ are respectively $\gamma_{1st} = y_{[1]st}$ and $\phi_1 = 0.0025 \times x_{1st} \approx 0.0025 \times y_{[1]st}$. According to properties of Normal distribution, the chosen variance term for $X_1$ implies that data are at $\sim$ 95% within 0.5% of the true population count $Y$ and $x_{1st}$ is considered as a proxy for $y_{[1]st}$. In this case standard deviation $\phi_{1st}$ is fixed and different for every cell because it depends on population. Also, the choice of the mean makes the model heavily driven by corresponding values $y_{[1]st}$s and does not assume any effect. In fact there is no reason to think the mean has any bias for any dimension. Coverage and accuracy should *a priori* be the same for all age, sex and time. The same happens for $X_2$ where coverage is assumed to be approximately the same for all dimensions but, as trust on $X_2$ is lower than on $X_1$, a Poisson data model is chosen

$$\begin{aligned} x_{2ast} &\sim Poisson(\gamma_{2ast}\omega_{2ast}) \qquad (3.10) \\ \log(\gamma_{2ast}) &\sim N(\mu_{2ast}, \sigma_2^2) \\ \mu_{2ast} &\sim N(0, 1) \\ \sigma_2 &\sim t_7^*(1) \end{aligned}$$

Prior distributions for $\mu_{2ast}$ and $\sigma_2$ are weakly informative and provide good and quick convergence for both parameters.

Before selecting these models several other models have been tested, both for the system model and for the data model. Overall results seem to be robust to different choices of data model. Eventually choices reflect datasets accuracy, and the Normal assumption on $X_1$ speeds up convergence.

In addition to the chosen model (equation (3.8)), also other system models have been tested with $\mu_{ast}$ modelled as follows

$$\begin{aligned} \mu_{ast} &= \beta^0 + \beta_a^{age} + \beta_s^{sex} \\ \mu_{ast} &= \beta^0 + \beta_a^{age} + \beta_s^{sex} + \beta_t^{time} \end{aligned}$$

$$\mu_{ast} = \beta^0 + \beta_a^{age} + \beta_s^{sex} + \beta_t^{time} + \beta_{st}^{sex:time}$$

$$\mu_{ast} = \beta^0 + \beta_a^{age} + \beta_s^{sex} + \beta_t^{time} + \beta_{as}^{age:sex} + \beta_{st}^{sex:time} + \beta_{at}^{age:time}$$

Results are all somehow comparable to the chosen model. On the one hand this is good because if results completely change according to the model then it means that there is a too high sensitivity to *a priori* assumptions, on the other hand convergence, credible intervals and held-back data checking were better with the chosen model (equation (3.8)).

#### 3.2.2.1 Results and model checking

For all the models three parallel chains are run with a burn-in of 50% of iterations. The number of iterations varies but, on average, 10.000 is enough to reach convergence. Convergence is checked analysing trace plots and calculating $\hat{R}$, the Gelman and Rubin convergence diagnostic between chains (Gelman and Rubin, 1992). If after 10.000 iterations chains do not reach convergence for all the parameters other attempts are made with more iterations. Sometimes an increase in the number of iteration helps but sometimes chains only diverge more and more often implying a problem with the model specification. This happens, for example, with the model considering three interactions (age-sex, sex-time and age-time). Probably this model is over-fitting data, i.e. it considers also random variations as part of the model making convergence harder to reach.

For the other system models the results are all very similar to the chosen one in terms of parameter values but convergence was worse for some parameters especially for standard deviation of data model. An important difference among models is the inclusion of time effect in the system model (3.8). Despite being small, as also preliminary analysis shows (figure 3.5), the effect is not so small to be ignored. The importance to include time effect is also confirmed by figure 3.7.

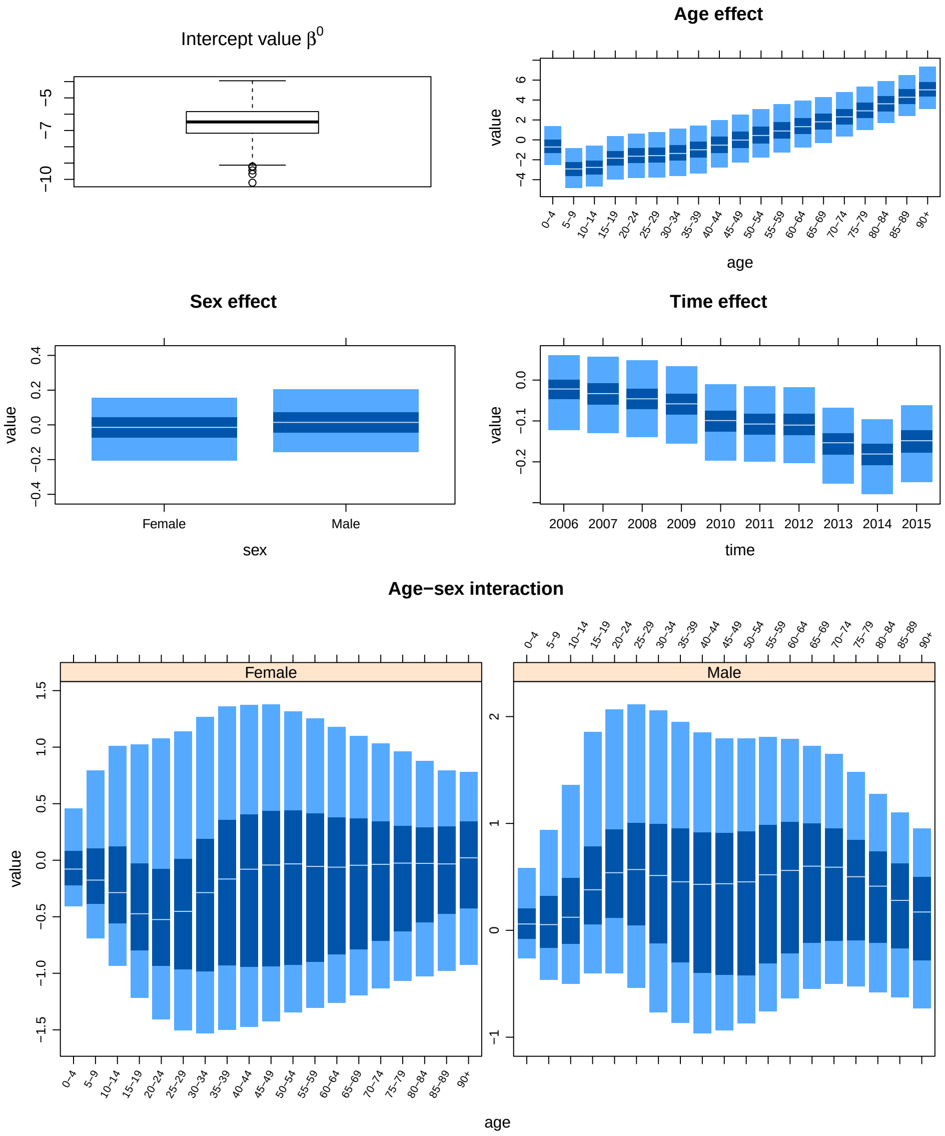


FIGURE 3.7: *Up-left*: Box-plot of model intercept estimation, $\beta^0$. *Up-right*: Age effect estimation, $\beta_a^{age}$. *Centre-left*: Sex effect estimation, $\beta_s^{sex}$. *Centre-right*: Time effect estimation,$\beta_t^{time}$. *Bottom*: Age-sex interaction estimations for females (left) and males (right), $\beta_{as}^{age:sex}$. Except for the box-plot in , all other plots show the medians in white, the 50% C.I.s (credible intervals) in blue and the 95% C.I.s in light blue.

The largest effect, as expected is the age effect, it has interval between $-4$ and 6. Sex effect of preliminary analyses (figure 3.5) seems now almost completely embedded in the age-sex effect (figure 3.7), it might be that the interaction is more important than sex effect itself or it could be an identification problem. This second hypothesis is less plausible as the interaction is essential for the model, as the model checking shows. For time effect, a clearly decreasing trend is confirmed even if low in magnitude. Age-sex interaction has a clear pattern when it comes to median values but it has very large credible intervals. Despite this, the importance of age-sex interaction is strongly confirmed by the held-back data analysis. Hyper-parameter models for time and age effect reflect what expected and estimated parameters are in figures 3.8 and 3.9. The level term for time is decreasing but with a very low decreasing trend, also mitigated by the damping term $\phi^{time}$. As values are all very small and close to each other, the standard deviation term is very low but with a heavy right tail. This is the same for all effects and reflects the difficulty of convergence of these parameters.

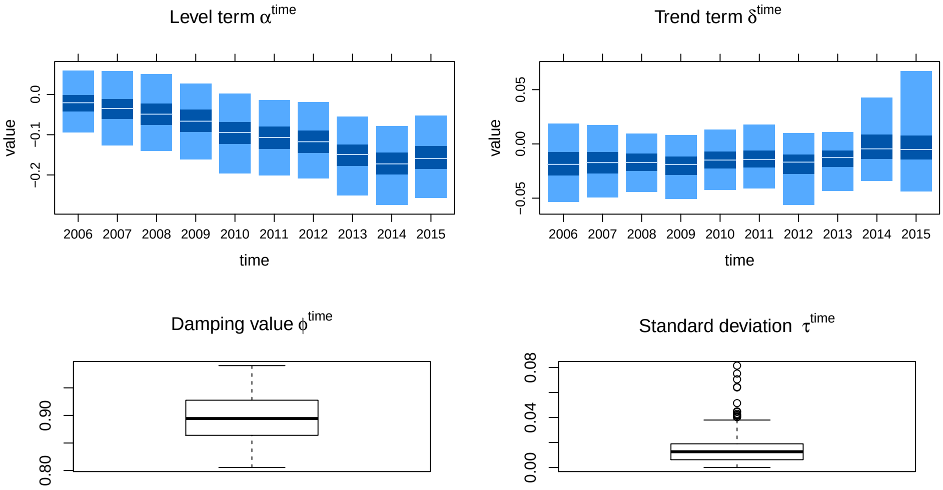


Figure 3.8: Time hyper-parameters. *Up*: Level, $\alpha^{time}$, and trend, $\delta^{time}$, terms, respectively on the left and on the right, with medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue). *Bottom*: Box-plots of the damping term, $\phi^{time}$, on the left, and of the standard deviation, $\tau^{time}$, on the right.

Age hyper-parameters confirm the prior local level model with positive trend except for the first age group because of infant mortality. For infant mortality an encouraging trend is the 25% decrease between 2006 and 2015 (Fig. 3.11).

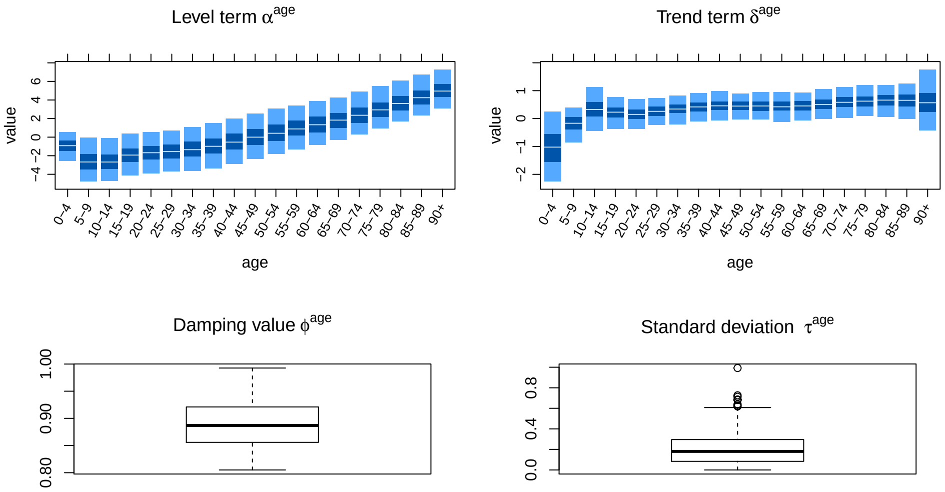


FIGURE 3.9: Age hyper-parameters. *Up*: Level, $\alpha^{age}$, and trend, $\delta^{age}$, terms, respectively on the left and on the right, with medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue). *Bottom*: Box-plots of the damping term, $\phi^{age}$, on the left, and of the standard deviation, $\tau^{age}$, on the right.

Age-sex interaction hyper-parameters in figure 3.10 reflect uncertainty in level term, even if it shows a constant higher level for men. Differences between men and women level terms tends to be higher during the 20s age groups and to disappear in the oldest age groups. The trend changes according to the age considered, an increasing trend for men in young ages and from 35 until 60 years old, whereas almost the opposite happens for women.

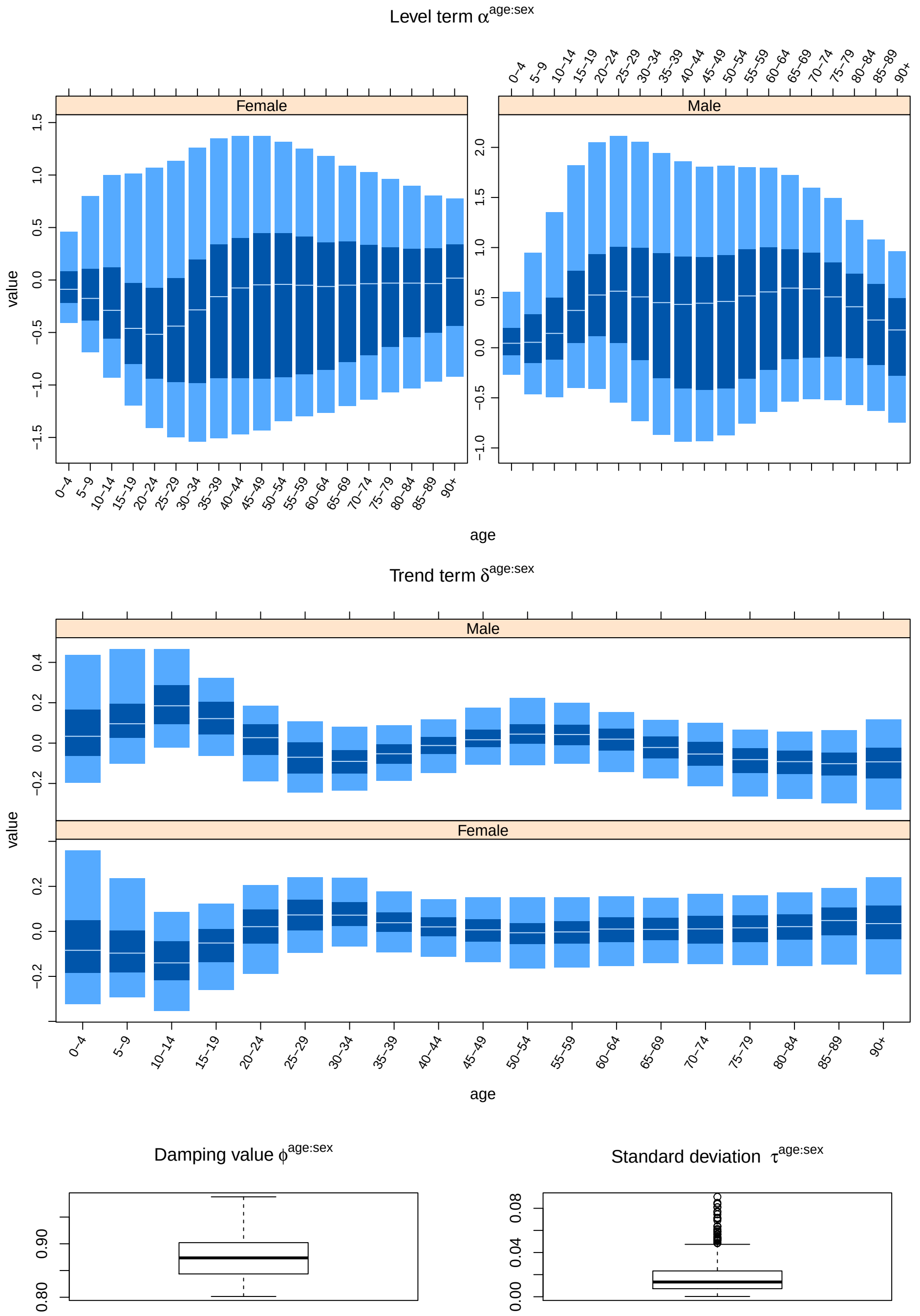


FIGURE 3.10: Age-sex hyper-parameters, *Up*: Level term, $\alpha^{age:sex}$ for females (left) and males (right). *Centre*: Trend term $\delta^{age:sex}$ for males (upper plot) and females (lower plot). Plots show medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue). *Bottom*: Box-plots of the damping term, $\phi^{age:sex}$, on the left, and of the standard deviation, $\tau^{age:sex}$, on the right.

Death counts estimated by the model are very close to the data (red and black lines in the graph) as it was expected by the data model specification (3.9). Figure 3.11 show the median (white line), 50% and 95% credible intervals respectively in dark and light blue. Width of credible intervals seem to mainly depend on the number of registered deaths. For young ages $(5 - 14)$, intervals are wide as counts are quite low (less than 200) then intervals get narrower. An exception stands for age group $80 - 84$ where the number of deaths for men and women is almost of the same magnitude and it is more difficult to well estimates the counts by sex. After this age groups, the number of deaths for women exceed by far the number of men as many more women reach this age. This is a problem only for counts as rate estimation, figure 3.11, does not show this irregularity, rates for men are always higher than for women. The rates estimation is smoother than data as it does not take into account the larger differences due to random death counts variations. Rates are super-population quantities, hence they reflect the theoretical model rather than the random data. Anyway, direct estimates (in red) are most of the time included in the intervals. The estimated mean coverage rate for the dataset $X_2$ is 98.5% with very low variability, confirming the homogeneity of the dataset.

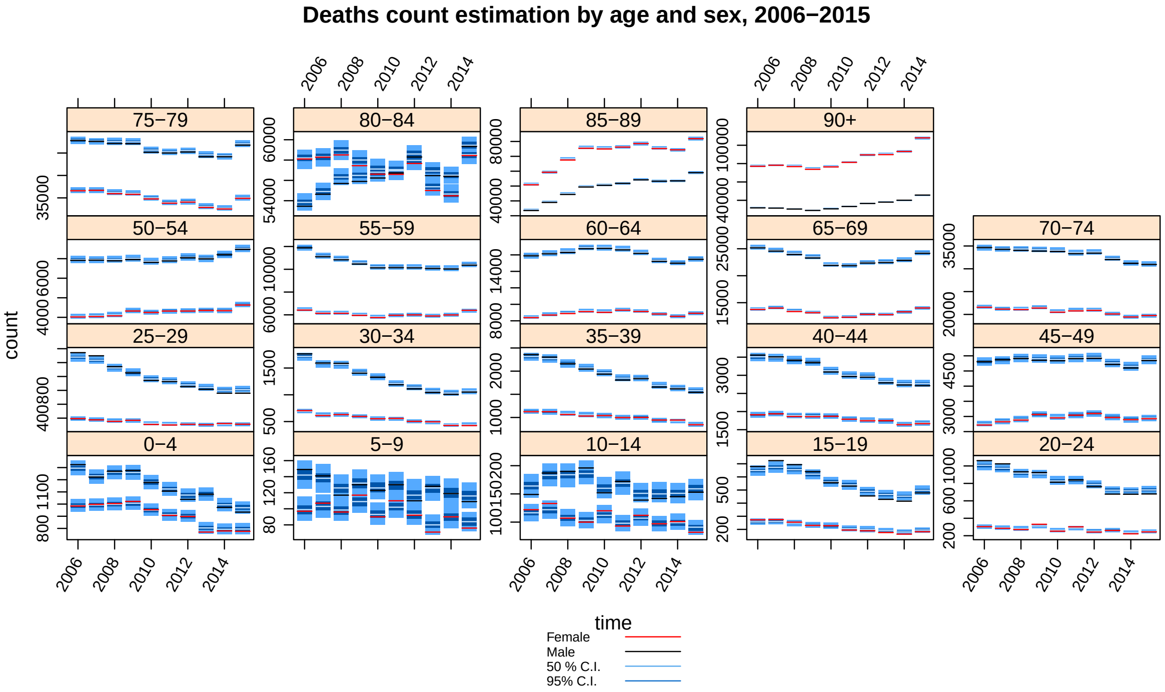


Figure 3.11: Death counts estimation by sex, time and age. Every age group block shows the medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue) for male and female for years 2006-2015. In addition values from dataset $X_1$ (*Annuario statistico italiano*) for males (black) and females (red).

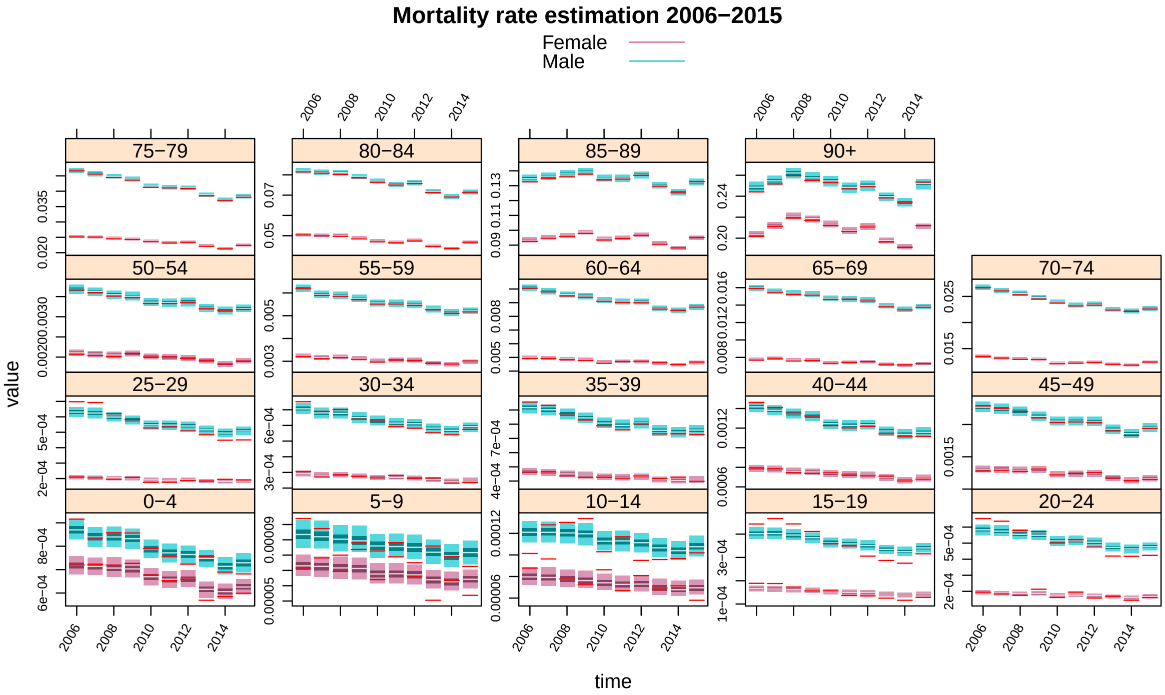


FIGURE 3.12: Mortality rates estimation by sex, time and age. Every age group block shows the medians (white line), 50% C.I.s for males (dark turquoise) and females (dark pink), and 95% C.I.s for male (light turquoise) and female (light pink) for years 2006-2015. Direct estimates of mortality rates from dataset $X_1$ (*Annuario statistico italiano*) for males and females are added in red.

In addition to convergence results, the model choice is also confirmed by the calculation of life expectancy with held-back data. Figure 3.14 shows the estimation of life expectancy calculated from held-back data. The held-back data technique consists of estimating the model only on a subset of data ("training" dataset) and to test the estimated model on the remaining part ("test" dataset). In this case the training part are data from 2006 until 2011 and the rest (2012-2015) is the test dataset.

In figure 3.14 there are results for life expectancy estimation according to model (3.8) chosen as baseline. Lines in light blue are the estimates calculated starting from median values of the death count estimates. In black and red the life expectancies respectively for men and women, computed from actual data. Direct estimates of life expectancy for year 2015 are lower than the estimates. This can be explained because more deaths than expected occurred in 2015, probably related to the unusually high and low temperatures registered respectively summer and winter time. So far, 2015 has been considered a record year for deaths and, therefore, the direct estimation of life expectancy dropped for this year. As 2015 can be considered an exceptional year, this drop in direct estimates of life expectancy does not appear in life expectancy estimation as the drop is just a random variation that is not actually affecting life expectancy. Nevertheless, the exceptional nature of 2015 has to be carefully considered and it is still to be confirmed. Looking at the data, until 2014 they are quite regular (598.364 deaths occurred in 2014), then in 2015 the number rose to 647.571 to decrease again in 2016 (615.261 deaths) but the number rose again in 2017 up to 649.061 deaths. Therefore, in year 2017 deaths exceeded 2015 ones, it can be argued that 2017 was also an exceptional year (summer was exceptionally hot, more than in 2015) but time will confirm this hypothesis. Italian population is getting older and the natural balance has been negative for more than 10 years now (figure 3.13) and, if on the one hand life expectancy is increasing as people live longer, on the other hand population is getting older and deaths naturally exceed births in absolute values. It is likely that in the next years the mortality rate analysis considering as last age group the class "90+" will not longer be acceptable and there will be more elements for the discussion on how life expectancy is evolving (as mentioned earlier, the discussion is about the constant increase or the reach of a plateau of life expectancy).

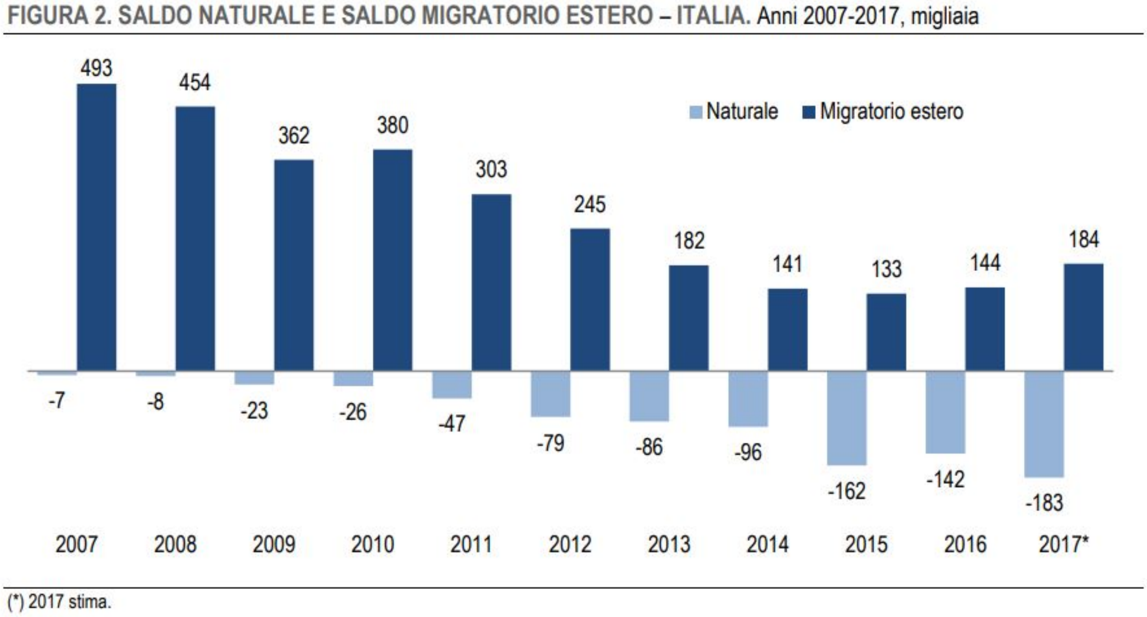


Figure 3.13: Natural balance (light blue) and external migration balance (blue) in thousands, from 2007 to 2017. Source: Demographic indicators, 2017 estimations, www.istat.it.

Except for year 2015, estimated life expectancy values in figure 3.14 are close to the direct estimates. Results from the baseline model including intercept, age, sex and time main effect and age-sex interaction (recalling equation 3.8) are compared with two other models in figures 3.15 and 3.16. Figure 3.15 show the estimation of life expectancy of a model including in $\mu_{ast}$ also age-time and sex-time interactions in addition to the baseline (equation 3.8). This model provides much higher life expectancies than the direct estimates which are quite unrealistic. In this case, how it is often suggested in literature, a more complicated model does not provide a better result. Figure 3.16 instead shows results for life expectancies from a model which only includes intercept and main effects without the age-sex interaction present in the baseline model. Comparing the results in figure 3.16 with the baseline model in figure 3.14, it can be noticed that without the age-sex interaction the model systematically overestimates women life expectancy and the difference between the model and the direct estimates increases for older ages. The result confirms the importance of including the age-sex interaction in the model.

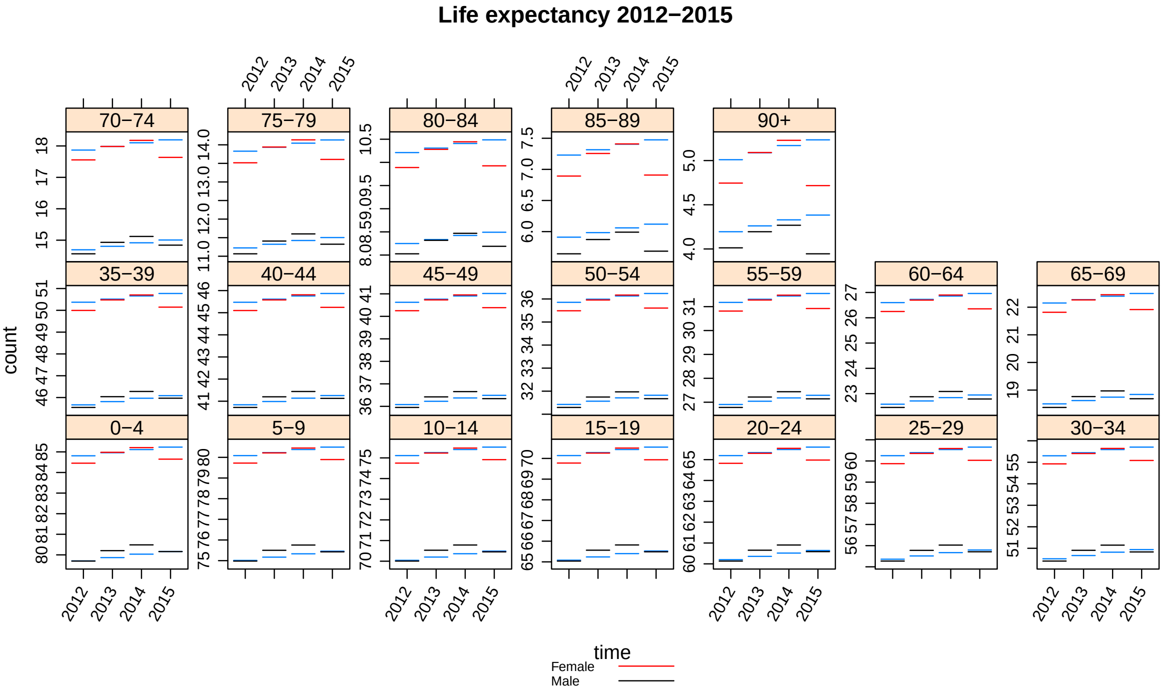


FIGURE 3.14: Median life expectancy estimation (light blue) by time, age and sex calculated using held-back data from the baseline model. Direct estimates for male (black) and female (red) are added for comparison.

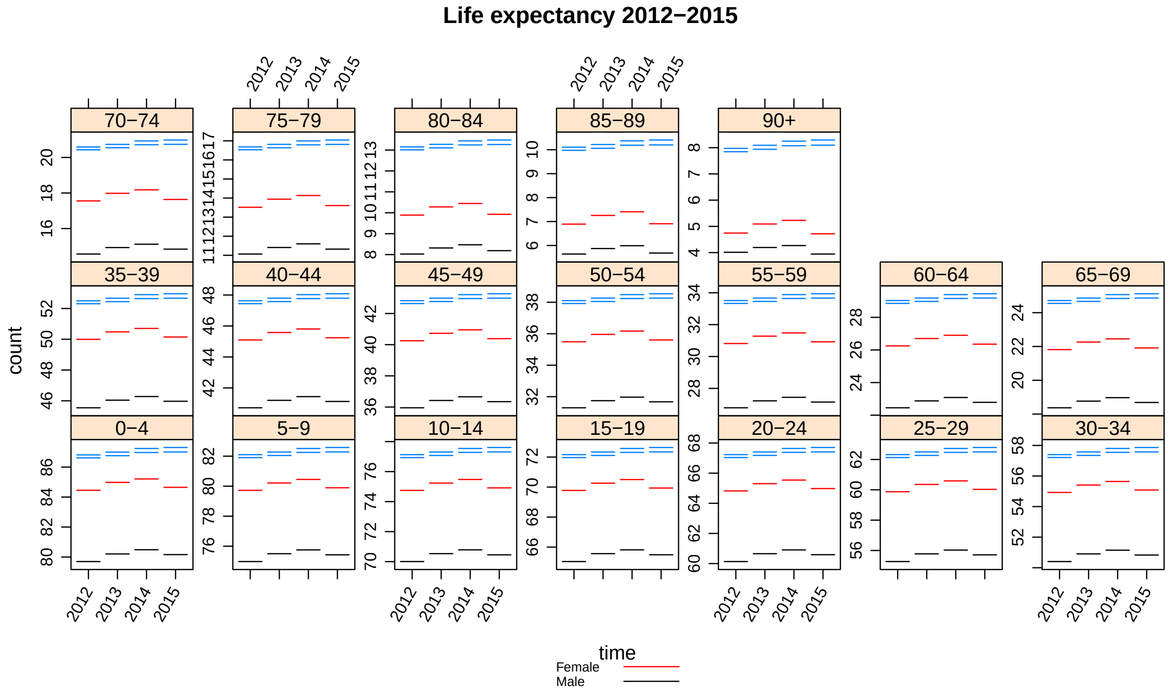


FIGURE 3.15: Median life expectancy estimation (light blue) by time, age and sex calculated using held-back data from the model considering three interactions. Direct estimates for male (black) and female (red) are added for comparison.

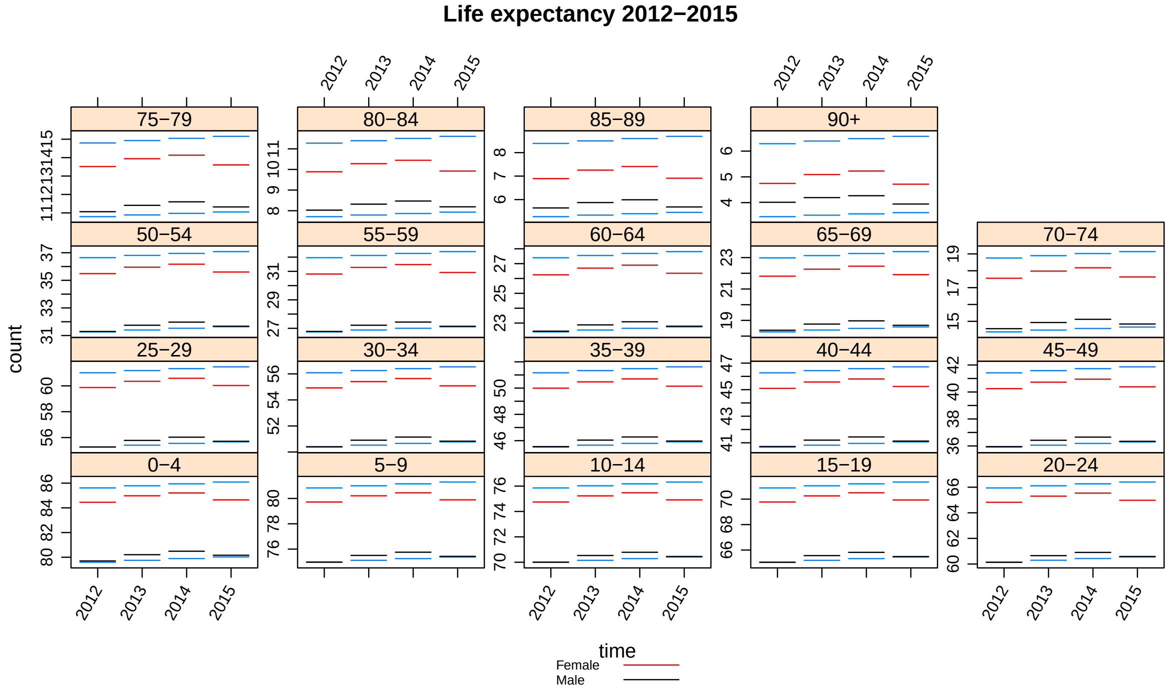


FIGURE 3.16: Median life expectancy estimation (light blue) by time, age and sex calculated using held-back data from the model only including main effects. Direct estimates for male (black) and female (red) are added for comparison.

## 3.3 Birth counts model

For births estimation the model considers age, region and time dimensions. According to the difficulties encountered, there must be variables not included in the model that might have an impact on age-specific fertility rates. Attempts including “marital status”, “parents citizenships” and “regional consumer households’ disposable income per inhabitant” have been made but no significant result or improvement was found. Marital status has almost no impact, parents citizenship is significant in the sense that, overall, immigrants tend to have more children than Italians so region with higher immigration could have higher fertility rates but this has not been proven to be significant in the model estimation. For disposable income the relationship between income and fertility is not clear and this additional information does not help to get better results.

Direct estimates of fertility rates [3] are the ratios of births on women population, by five years age groups from 15 to 49 years old (in the last age group also births from 50+ women are considered). The model used is the female dominant model (Preston *et al.*, 2001), therefore only the number of women influence the number of births, regardless the number of men. The female population is then used as exposure in the model. There are two models outperforming the others, both include intercept, age, time, region and age-time effects and one also consider age-region interaction. Eventually the simpler one without age-region interaction has been preferred.

### 3.3.1 Data and preliminary analyses

Data on births released by Istat come from Italian municipalities administrative registers. As for deaths, there are different values for different tables because of revisions and comparison between datasets. An explanation about how Istat is planning to integrate sources for newborns can be found in Tuoto *et al.* (2015). Data quality is a major issue for statistical institutes and plans to improve are a priority. In order to have all the needed dimensions (region, time and mothers’ age), at least two datasets have to be used: the datasets from *Annuario statistico italiano*, providing Istat ultimate data after corrections and controls, and the datasets from administrative registers where the newborn has to be registered within ten days from the birth. Between the two chosen datasets, differences at national level go from around 5.000 units in 2015 to more than 15.000 in 2011 and 2013. At a percentage level, these differences represent between 1% and 3% of the total births. Figures 3.18 and 3.17 show the differences at national and regional level. Dataset 1 ($X_1$) refers to data from *Annuario statistico italiano* which

[3] we always refer to *age-specific* fertility rates

has only time and region dimensions; dataset 2 ($X_2$) refers to the datasets from the municipalities where the newborn is registered and provide, in addition to time and region dimensions, also mothers' age. Figure 3.17 shows that non-negligible differences exist among regions. Quite remarkable are differences in regions Lazio, Abruzzo and Molise, they are wider and more irregular than in other regions.

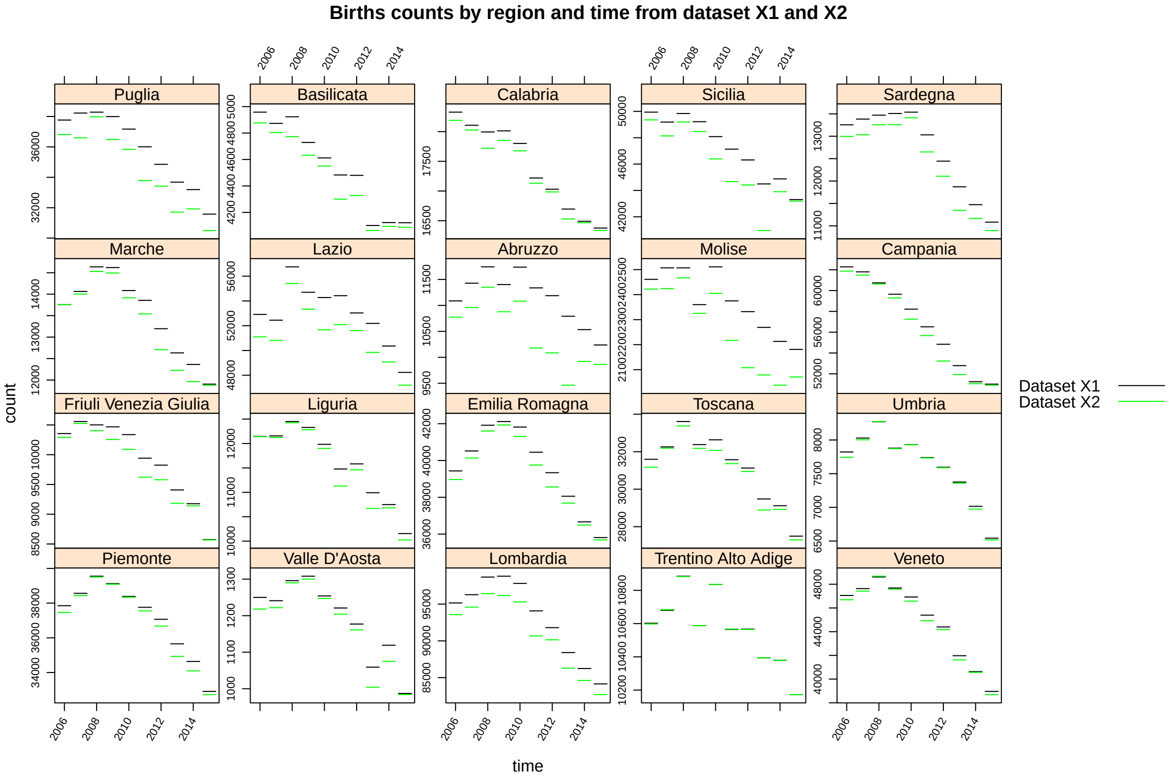


FIGURE 3.17: Birth counts at regional level for years 2006-2015 from dataset $X_1$ (*Annuario statistico italiano*) in black, and from dataset $X_2$ (regional tables by parents' age) in green.

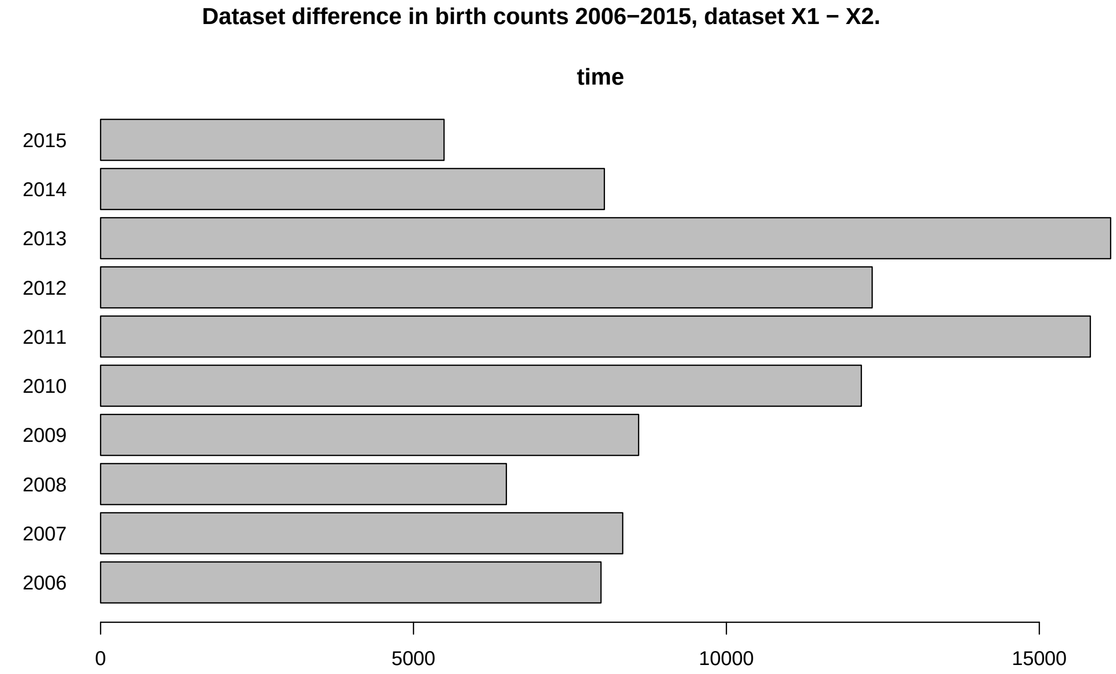


FIGURE 3.18: Differences by time between dataset $X_1$ (*Annuario statistico italiano* and dataset $X_2$ (regional tables by parents' age), i.e. $X_1 - X_2$, from 2006 until 2015 at national level.

In figure 3.19 the central panel shows the computation of region effect from preliminary data analyses. Preliminary analyses and calculation of main effects and interactions follow the same process as for deaths rates in paragraph 3.2.1. Besides the age (first panel on the right in figure 3.19), which is well-known for being a very important variable in fertility studies, the region effect has a span of 0.3 (from $-0.2$ until more than 0.1) wider than time effect in the third panel of figure 3.19, with a span of almost 0.15 (from $-0.1$ to 0.05). The time effect is also noticeable in figure 3.20 where rates increase for age groups over 30 years and decrease for groups before 30. Despite the general trend, from the graph it is also clear that the increase is stronger in older ages, stays almost the same for the age group $30 - 34$ and gets wider again in younger ages but still with lower differences than for older age groups. This suggests that there is an age-time interaction and that, while in young ages the rates are little decreasing, a major change is happening for women in their 40s. Figures 3.22, 3.21 and 3.23 show all the interaction effects computed from overall fertility log-rates. Age-time and age-region effects seem to be more important than region-time interaction and this is confirmed by the results of the model.

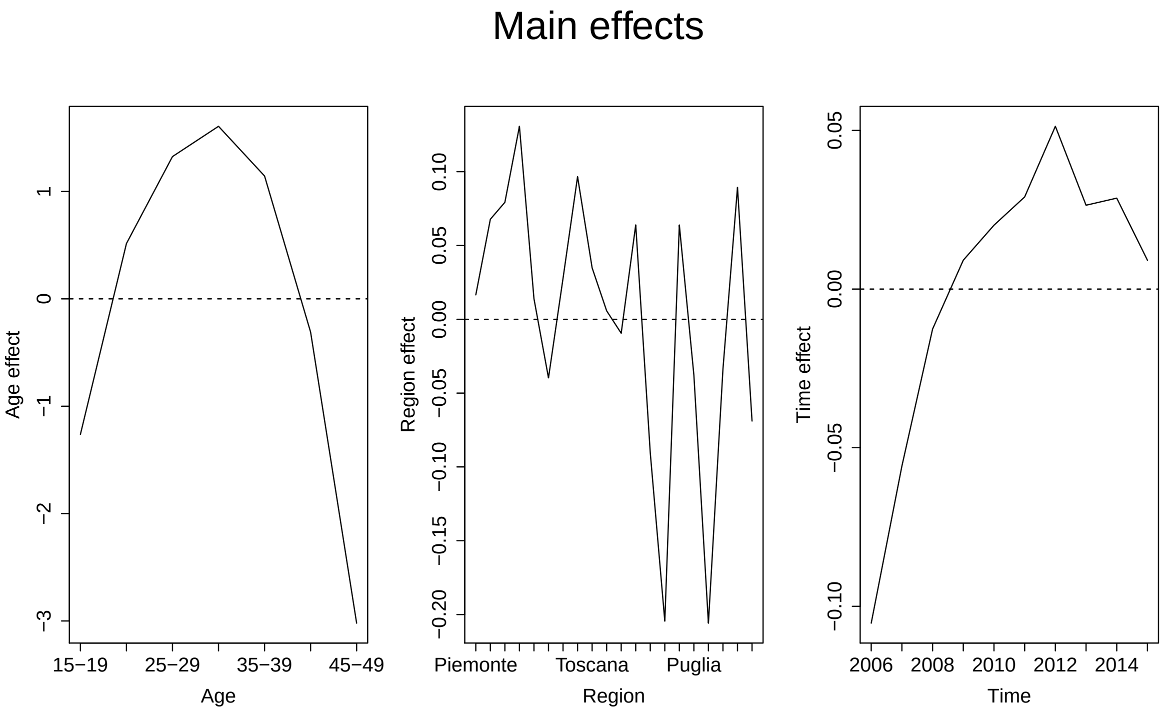


FIGURE 3.19: Computation of main effects from overall age fertility log-rates. *Left*: Age effect, with peak at age 30 − 34. *Centre*: Region effect, with irregular pattern. *Right*: Time effect, growing from 2006 until 2012 and dropping after.

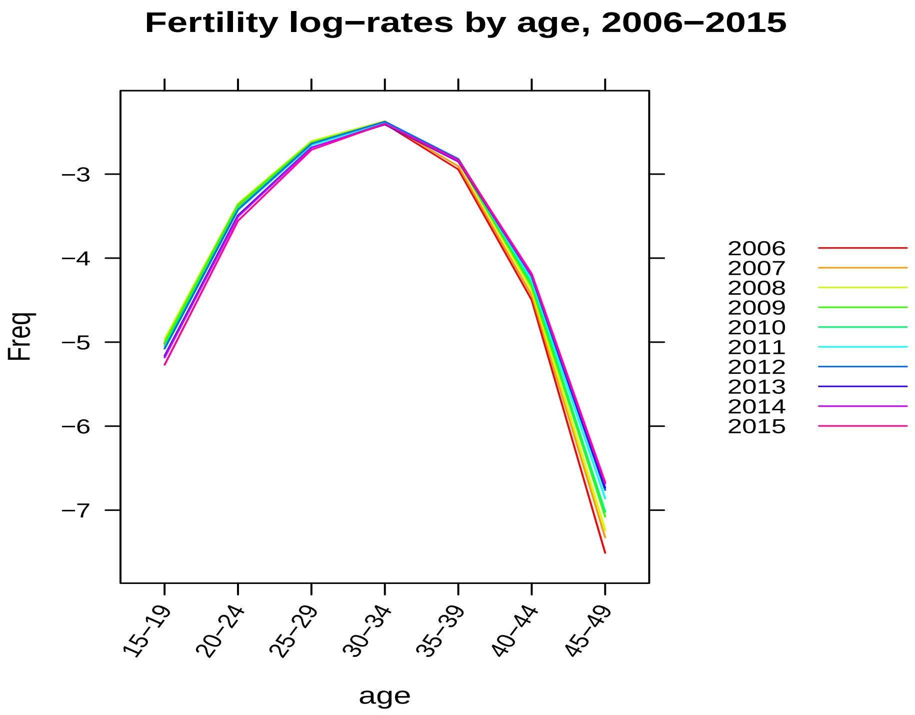


FIGURE 3.20: Age-specific fertility log-rates from 2006 to 2015, mothers' age groups from 15 − 19 until 45 − 49. Each year has a different color.

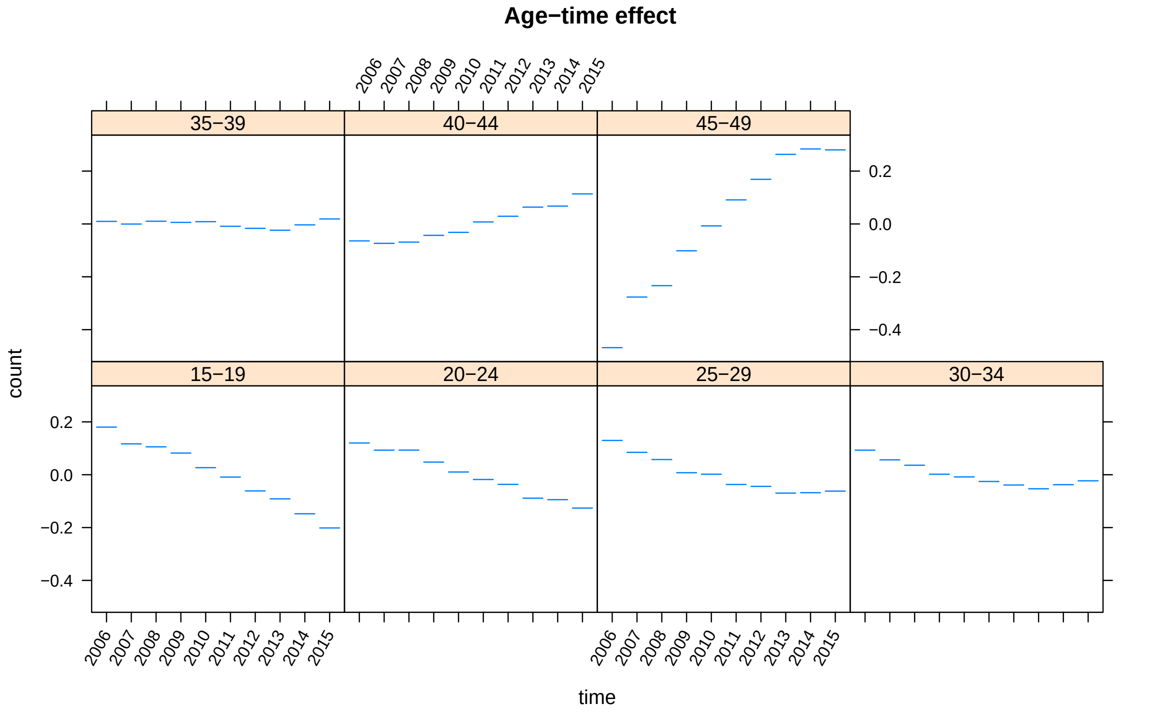


FIGURE 3.21: Computation of age-time interaction from overall fertility log-rates. Each block corresponds to a mothers' age group and each light blue line to a year, from 2006 until 2015.

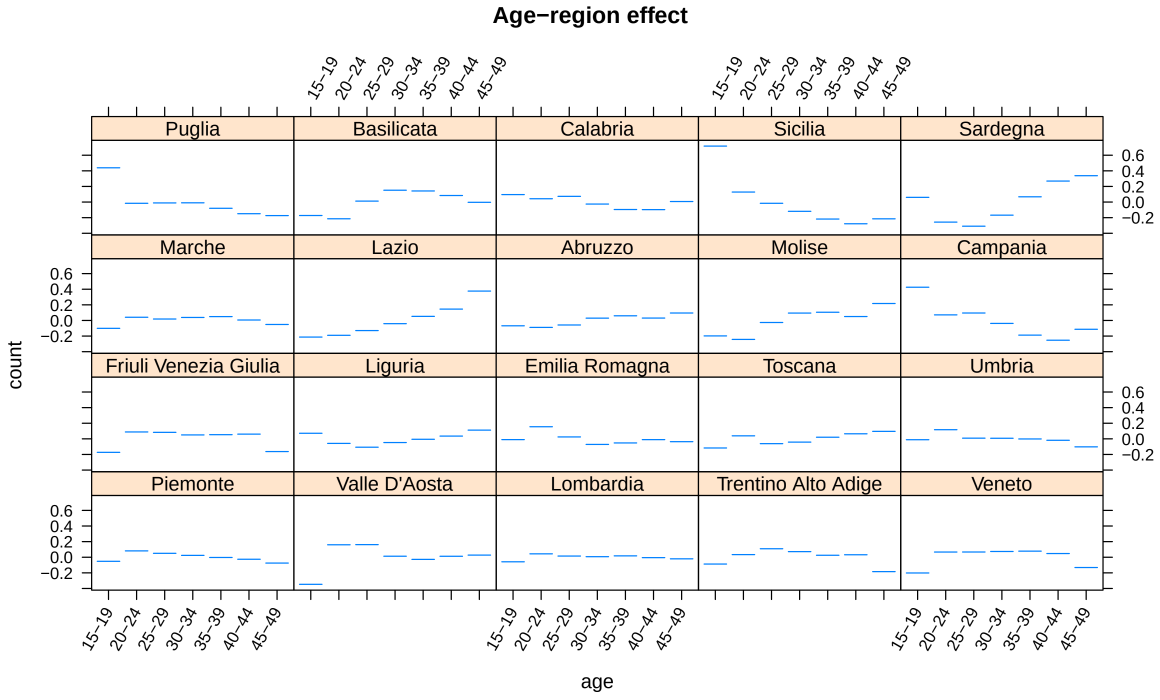


FIGURE 3.22: Computation of age-region interaction from overall fertility log-rates. Each block corresponds to a region and each light blue line to an age group, from $15-19$ until $45-49$.

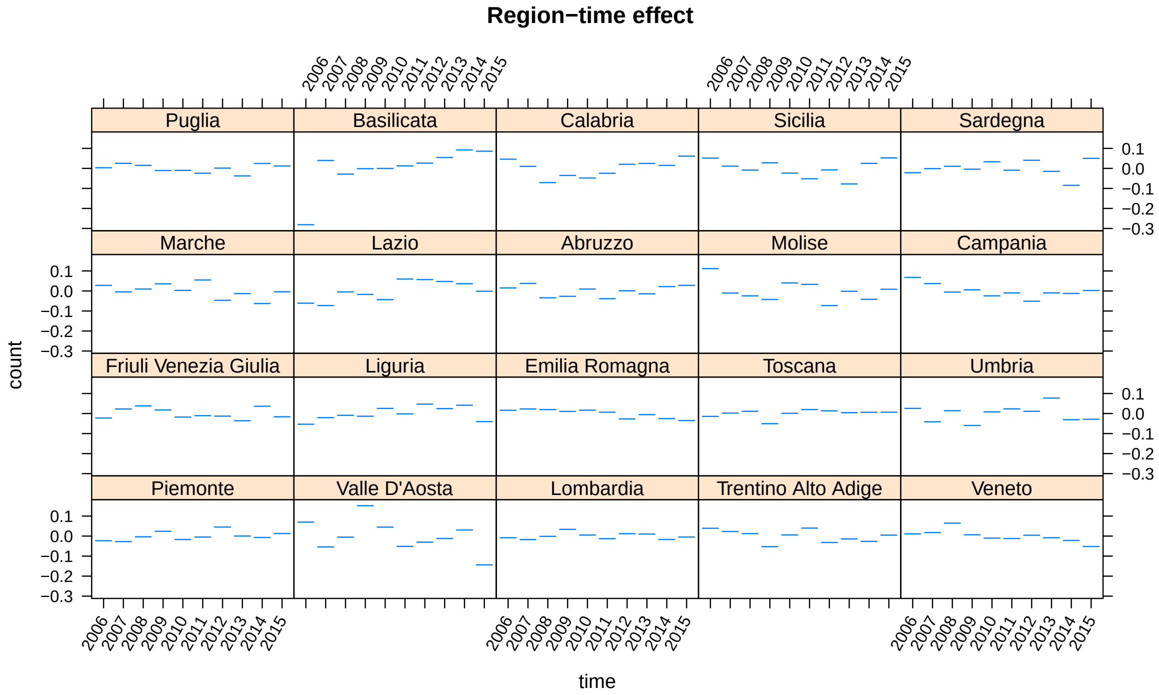


FIGURE 3.23: Computation of region-time interaction from overall fertility log-rates. Each block corresponds to a region and each light blue line to a year, from 2006 until 2015.

Dimensions considered and appearing in the graphs are:

- Age of the mother $(a)$ with seven age groups: $15-19$, $20-24$, $25-29$, $30-34$, $35-39$, $40-44$ and $45-49$. The first and the last groups include respectively women giving birth before 15 or after 49, but this does not significantly affect estimation for the years considered although, according to the trend, in few years age group $50-54$ might be needed.

- Region $(r)$, the twenty Italian regions: Piemonte, Valle D'Aosta, Lombardia, Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Liguria, Emilia Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna. Technically region Trentino Alto Adige is divided in two autonomous provinces, Provincia autonoma di Trento and Provincia autonoma di Bolzano, but as they have very similar characteristics it has been chosen to consider them together in the traditional unique region [4].

- Time $(t)$, ten years: from 2006 to 2015.

[4] The opposite choice is made in section 3.5

### 3.3.2 System and data models

Dimensions considered for the model on births are: “age”, “region” and “time”. Sex dimension for newborns has not been considered. There are two main motivations for this choice: (i) there is no sex effect on the fertility rates in Italy, proportions between male and female does not show different pattern from what has already been investigated in literature; (ii) in order to consider sex another database was needed, with different numbers from the two other datasets adding further complication to an already difficult model.

The model choice for births is delicate both for the system and data model. For the system model different mean specification have been tested and eventually the best one is the one in equation (3.11) with parameter $\mu_{art}$ including intercept and main effects for all the dimensions plus only age-time interaction.

$$\begin{aligned}
&y_{art} \sim Poisson(\gamma_{art}\omega_{art}),\, a = 1, ..., 7,\, r = 1, ..., 20,\, t = 1, ..., 10\\
&\log(\gamma_{art}) \sim N(\mu_{art}, \sigma^2)\\
&\mu_{art} = \beta^0 + \beta_a^{age} + \beta_r^{region} + \beta_t^{time} + \beta_{at}^{age:time}\\
&\sigma \sim t_7^*(1)
\end{aligned} \tag{3.11}$$

Prior distributions on $\beta$ parameters are exchangeable priors on region effect and for intercept whereas a DLM is assumed for age, time and age:time interaction. DLMs do not include any trend and damping terms, there are only local-level models.

$$\begin{aligned}
&\beta^0 \sim N(0, 10^2)\\
&\beta_r^{region} \sim N(0, 1)\\
&\beta_a^{age} \sim N(\alpha_a, \tau_a^2)
\end{aligned}$$

with level term: $\alpha_a \sim N(\alpha_{a-1}, \tau_\alpha^2)$.

$$\beta_t^{time} \sim N(\alpha_t, \tau^2)$$

with level term: $\alpha_t \sim N(\alpha_{t-1}, \tau_\alpha^2)$.

Age-time interaction:

$$\beta_{at}^{age:time} \sim N(\alpha_{at}, \tau_{at}^2)$$

with level term: $\alpha_{as} \sim N(\alpha_{as-1}, \tau_\alpha^2)$.

Standard deviation has the same form in all priors:

$$\tau_a, \tau_t, \tau_{at}, \tau_\alpha \sim t_7^+(1)$$

Data model reflects the one used for modelling death rates. On the most trusted dataset a Normal prior is assumed (equation (3.12)) with mean $\gamma_{art} = y_{[1]rt}$ and standard deviation $\phi_{1rt} = 0.0025 \times x_{1rt} \approx 0.0025 \times y_{[1]rt}$ chosen so that data are at $\sim 95\%$ within 0.5% of the true population count $Y$ whose proxy is $X_1$ as the data are considered accurate.

$$x_{1st} \sim N(\gamma_{1st}, \phi_{1st}^2) \tag{3.12}$$

A Poisson model is set as prior for the second dataset $X_2$, with a weak prior on the mean with no systematic change of coverage according to the dimensions.

$$\begin{aligned} & x_{2rt} \sim Poisson(\gamma_{2art}\omega_{2art}) \\ & \log(\gamma_{2art}) \sim N(\mu_{2art}, \sigma_2^2) \\ & \mu_{2art} \sim N(0,1) \\ & \sigma_2 \sim t_7^*(1) \end{aligned} \tag{3.13}$$

Distribution for $\mu_{2art}$ and $\sigma_2$ are weakly informative, more informative priors on the mean or inclusion of upper and lower limits did not improve results. A prior including a region effect, as suggested by the preliminary analysis, despite performing well in the model with all the data, it performs worse than model in (3.13) when applied to held-back data.

Other system models with more informative priors, trend and damping terms have been tested. Models changing mean $\mu_{art}$ specification are:

$$\begin{aligned} & \mu_{art} = \beta^0 + \beta_a^{age} + \beta_t^{time} \\ & \mu_{art} = \beta^0 + \beta_a^{age} + \beta_r^{region} + \beta_t^{time} \\ & \mu_{art} = \beta^0 + \beta_a^{age} + \beta_r^{region} + \beta_t^{time} + \beta_t^{time} : \beta_a ge^{age} + \beta_r^{region} : \beta_a ge^{age} \\ & \mu_{art} = \beta^0 + \beta_a^{age} + \beta_r^{region} + \beta_t^{time} + \beta_a^{age} : \beta_r^{region} + \beta_t^{time} : \beta_r^{region} + \beta_a^{age} : \beta_t^{time} \end{aligned}$$

They all provided worse results for convergence and with held-back data tests but still comparable ones.

### 3.3.3 Results and model checking

The estimation of the model is quite quick, within 10.000 iterations all chains reach convergence. For each parameter three chains are run in parallel and convergence is checked via trace plot and Gelman and Rubin diagnostic (Gelman and Rubin, 1992). In figure 3.24 there is an example of trace plots for fertility rates estimation $\gamma_{art}$ and the corresponding kernel densities. Chains seem to be robust to starting points. When starting point is far it can take longer to converge because of small steps required to obtain a sufficient acceptance rate, but chains move in the right direction in all the attempts made.

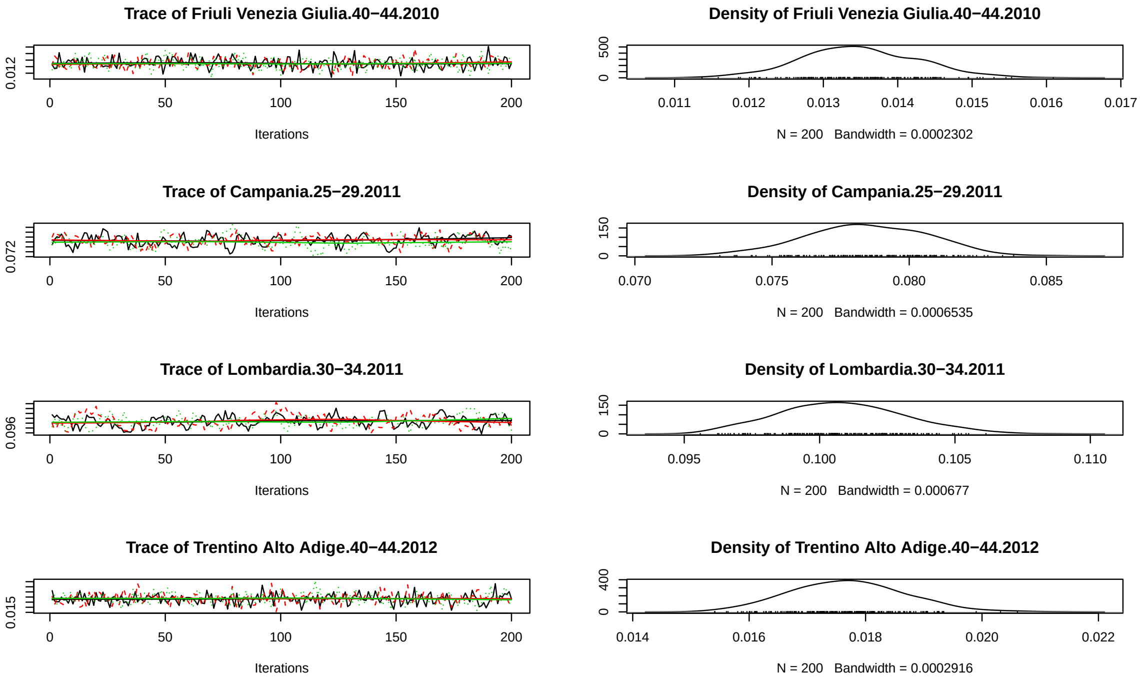


FIGURE 3.24: *Right side*: Trace plots for fertility rates of different age, time and region. Each coloured line (black, green and red) corresponds to a chain resulting from the MCMC algorithm. *Left side*: Kernel density plots of posterior draws with all chains merged.

Figure 3.25 shows plots for estimation of effects the model includes. For level terms results are different from those in figures 3.19, but the width is comparable except for time effect whose width is smaller than in preliminary analysis, and for age-time interaction where width is usually bigger than expected, meaning that the interaction incorporates much of the main time effect.

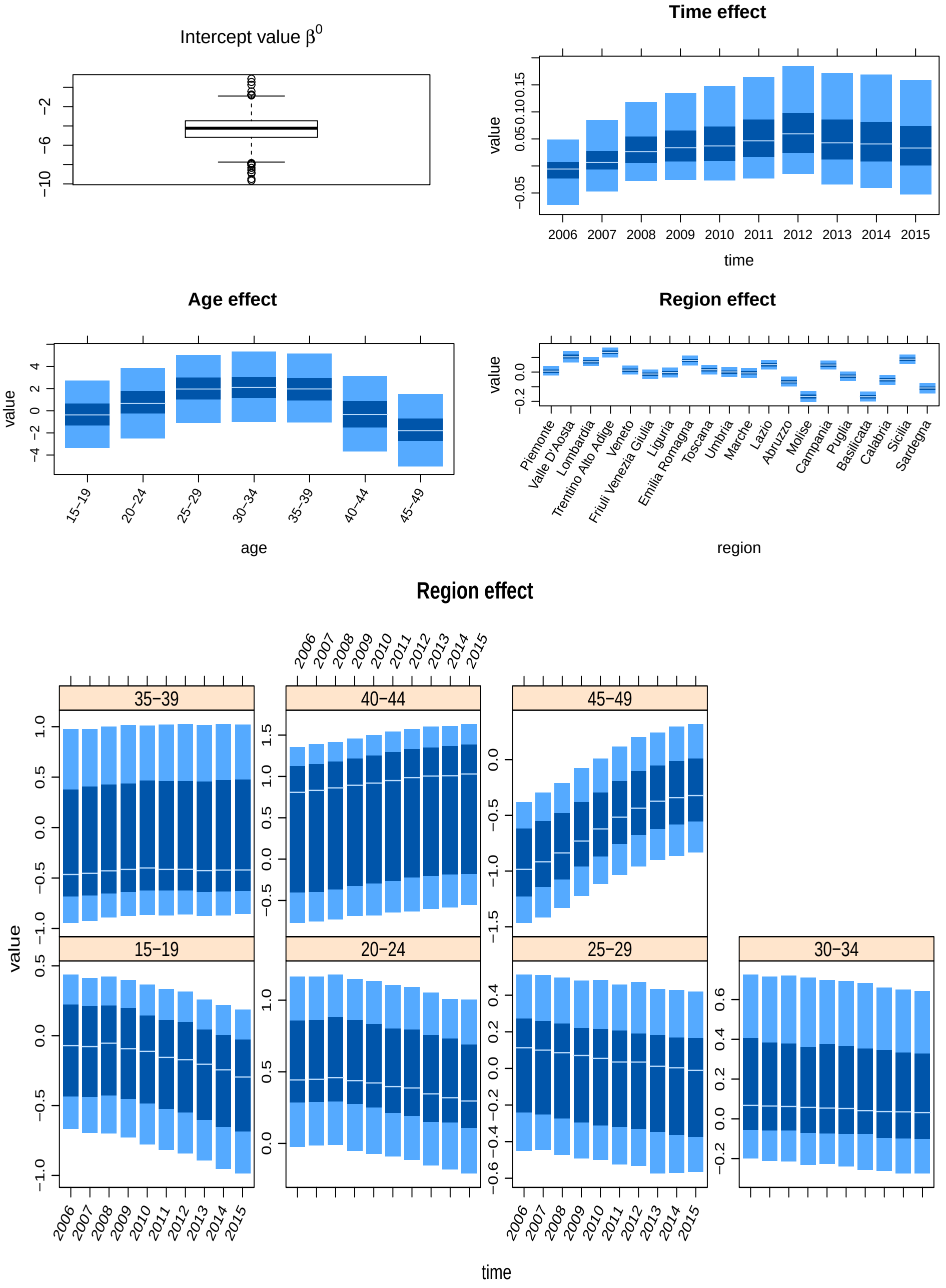


FIGURE 3.25: *Up-left*: Box-plot of model intercept estimation, $\beta^0$. *Up-right*: Time effect estimation, $\beta_t^{time}$. *Centre-left*: Age effect estimation, $\beta_a^{age}$. *Centre-right*: Region effect estimation $\beta_r^{region}$. *Bottom*: Age-time interaction estimation, $\beta_{at}^{age:time}$, each block is an age group. Except for the box-plot in , all other plots show the medians in white, the 50% C.I.s (credible intervals) in blue and the 95% C.I.s in light blue.

Fertility rate estimates for the whole period $2006 - 2015$ are in figure 3.26. The 50% credible intervals are in blue, the 95% credible intervals are in light blue, and the red lines are the direct estimates. A slightly higher level, especially for age group $30 - 34$, is recurrently estimated, neighbouring groups have similar pattern but less evident, whereas for other age groups model estimations reflect direct estimates. The difference can be seen also in the count estimates in figure 3.27 and reflect what happens in the central age groups for rates. This higher uncertainty for age groups with higher rates is found also in others works (e.g. Alkema *et al.* (2008); Bryant and Zhang (2018)) and depends on the higher number of births occurring for women at these ages. Coverage estimated for the second dataset has mean $\sim 98.7\%$.

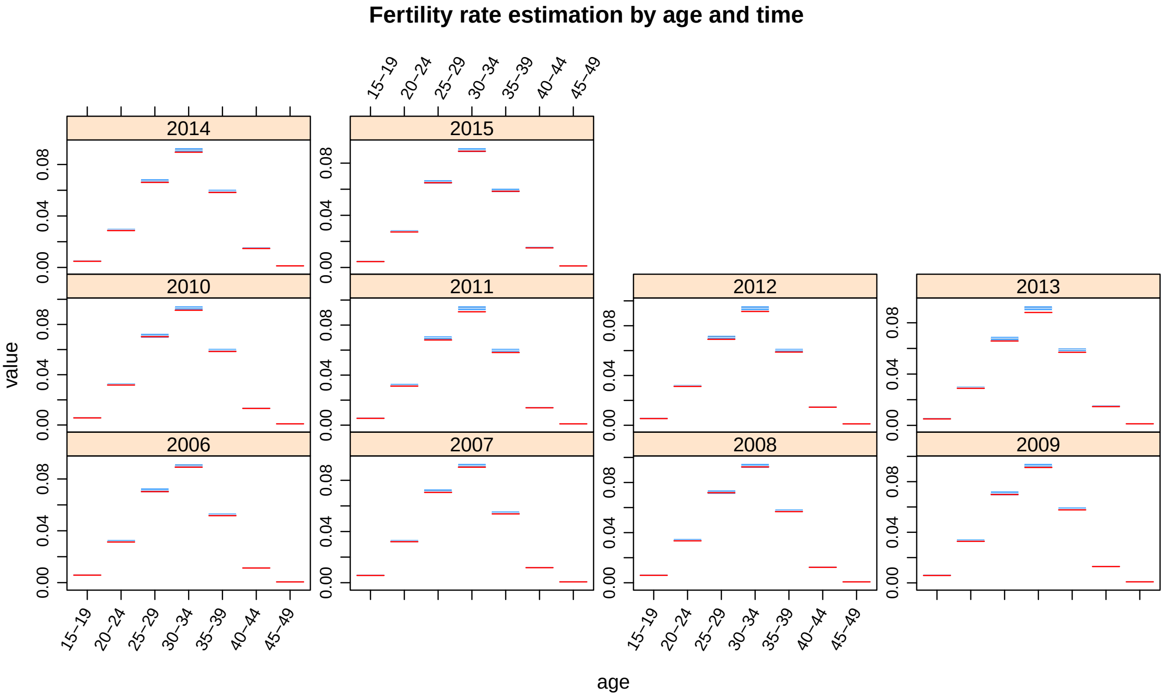


FIGURE 3.26: Age-specific fertility rates estimation by mothers' age and time at national level. Every time block shows the medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue) for mothers' age groups. Direct estimates of fertility rates from dataset $X_1$ (*Annuario statistico italiano*) are added in red.

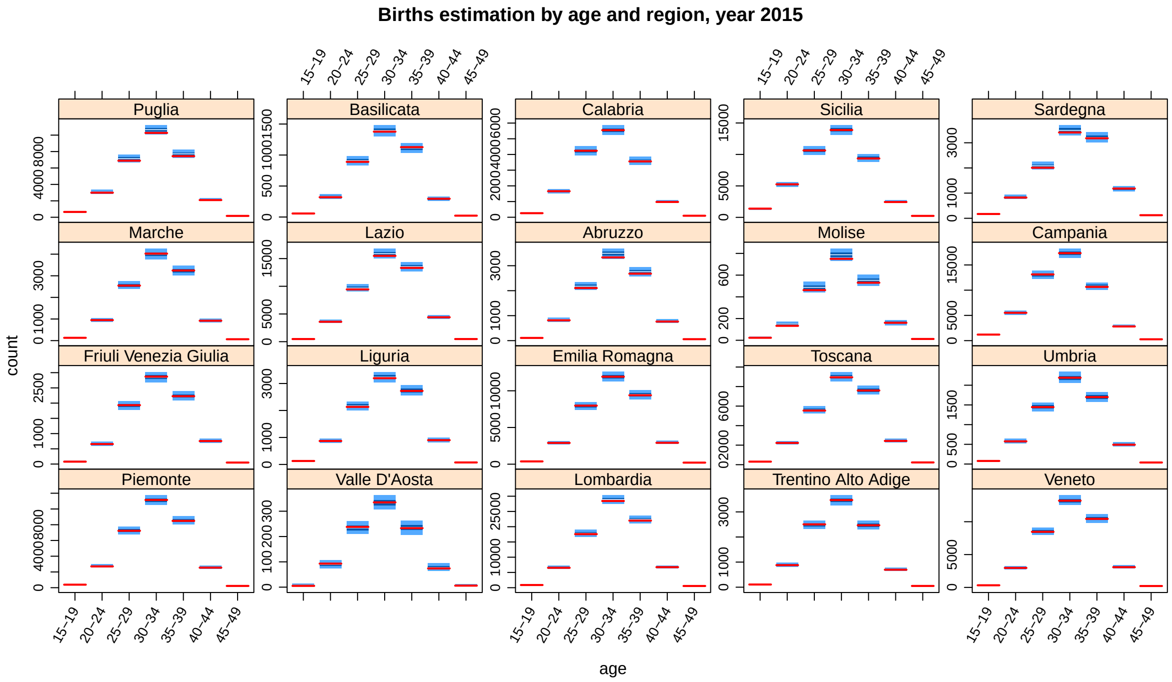


Figure 3.27: Birth counts estimation by mothers' age and region for year 2015. Every region block shows the medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue) for mothers' age groups. In addition values from dataset $X_1$ (*Annuario statistico italiano*) are added in red.

Model selection is made via held-back data. The training dataset includes births from 2006 until 2012 and the test part from 2013 until 2015. Fertility rates results are shown in figures 3.28 with usual 50% and 95% credible intervals. Results reflect quite accurately the direct estimates. The second best model whose results are in figures 3.29 includes age-region interaction in addition to age-time one. In this case credible intervals are much wider than with the baseline model, and convergence is worse. Therefore, the simpler model is preferred.

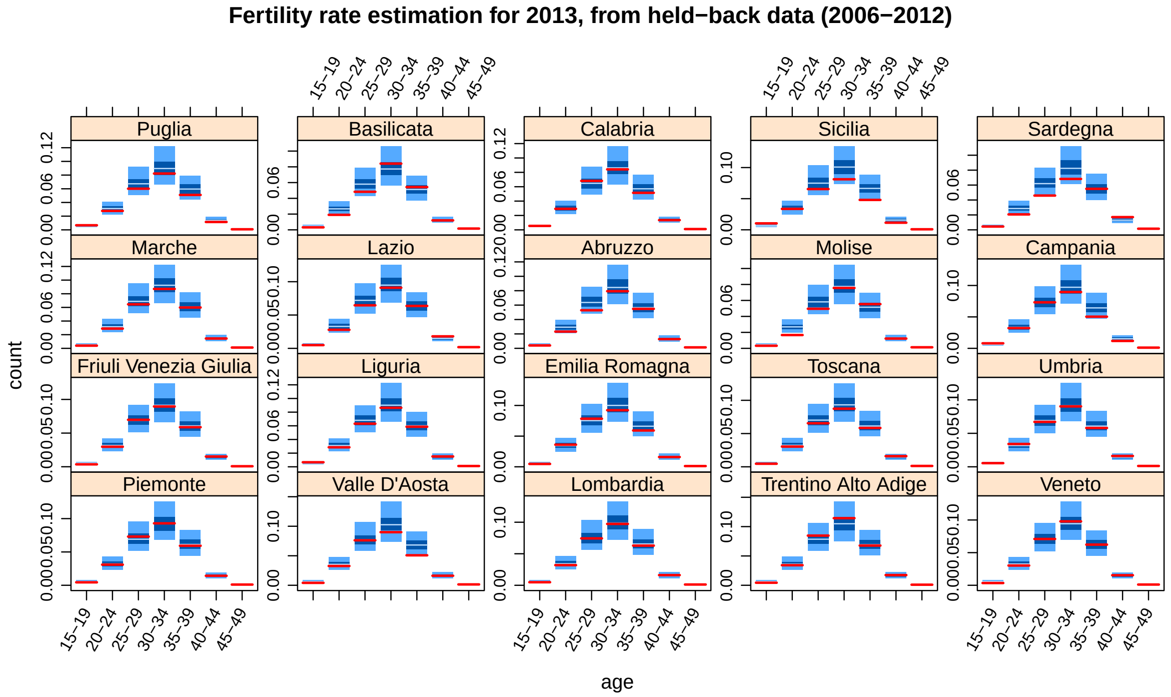


FIGURE 3.28: Estimation of age-specific fertility rates by region using held-back data for 2013 from the baseline model. Every region block shows the medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue) for mothers' age groups. In addition direct estimates from dataset $X_1$ (*Annuario statistico italiano*) are added in red.

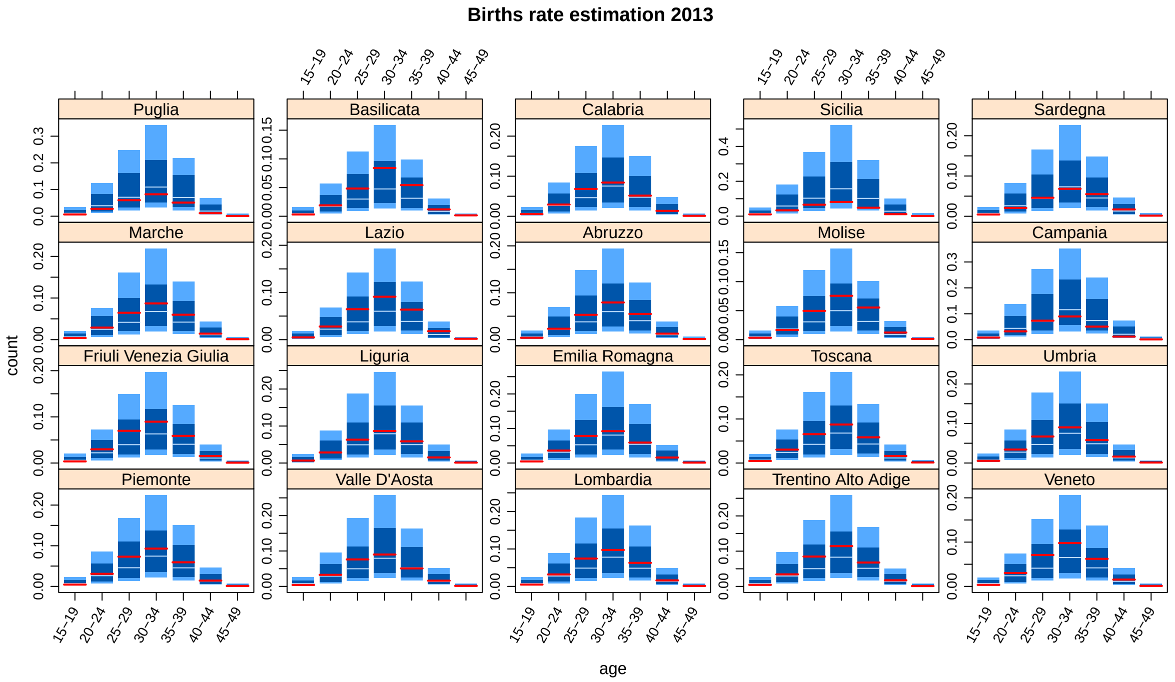


FIGURE 3.29: Estimation of age-specific fertility rates by region using held-back data for 2013 from the alternative model. Every region block shows the medians (white line), 50% C.I.s (blue) and 95% C.I.s (light blue) for mothers' age groups. In addition direct estimates from dataset $X_1$ (*Annuario statistico italiano*) are added in red.

## 3.4 Migrations

Migration is a critical point for population estimation even in countries with reliable registers. Migration related problems are important to address as population mobility is constantly increasing, and registration of such movements is not always and not evenly reported. The differences in registering in and out migration is not the only one, there are also problems for specific sub-populations, like students or migrants within the European Community. Also, the phenomenon of illegal migration has relevant political implications and it is the centre of a significant part of the public and political debates in the recent years. Despite there are studies and estimations of the number of illegal migrant coming to Italy and to other European countries, this issue is out of the scope of this work. These applications only consider official data on resident population of Italy. The complexity of illegal migration is high, data quality is difficult to evaluate and results from application of the model would be difficult to interpret especially because the model investigation has not been completed yet and still needs improvements.

In addition to the difficult data quality assessment and accuracy variability depending on time and countries, migration is a complex phenomenon also from other points of view. Especially for countries experimenting a zero or negative natural population growth, migration is an important issue from demographic but also sociological, political and economic perspectives. Therefore, despite migration is "the most complex and most difficult to predict component of population change, bearing high levels of forecast errors" (Kupiszewski, 2002) it is essential, especially for so-called "developed countries", to find methods to estimate and predict migration flows.

Attempts to estimate or forecast migrations flows can be found, in several works adopting different approaches, and Bijak (2010) dedicated a book to migration in Europe. Raymer *et al.* (2013) address the problem of incoherence in migration flows registrations between countries. Their aim is to harmonise and estimate migration flows among 31 countries in the European Union and European Free Trade Association from 2002 until 2008. They integrate a theory-based migration model and a measurement models from both sending and receiving countries. Using Eurostat data and a set of covariates, they model measurement errors considering imbalance between in- and out-migration and estimate under-counting country levels. Expert opinions are used for building prior distributions. Tests on sensitivity to prior information and to partial removal of the data are also accomplished along with a comparison with other approaches. In Congdon (2008) the Author compares the estimation of migration flows through a fully Bayesian and an estimation approach. He applies the method to the migration flows from Scotland to England during the 1990s. He uses the software “WinBugs” for the analysis and comments on benefits of the Bayesian approach and, specifically, on the random effect approach. The comparison between parametric and non-parametric approaches reveals how the first one performs well and gives good results for preliminary smoothing analysis, whereas the second one, having fewer constraints, is able to reveal details that are not so strongly empathised in the first one. An example of projections on net migration with very few data is Azose and Raftery (2016). They perform a Bayesian estimation of correlation matrices with informative priors and show how it outperforms Pearson correlation matrix and simple shrinkage estimators especially when the correlation matrix to estimate is sparse. Putting interpretable and simple priors on correlations is the main innovation of the method. An extension they suggest is to consider a matrix of bilateral migration instead of net migration.

As for deaths and births examples, data on migration for Italy come from Istat and different datasets show differences in counts and dimensions. A first distinction is between migration from or to other countries, international migration, and migration within the country, internal migration. For internal migration, data have different level of detail from migration between municipalities to migration between regions. Municipalities collect data on registrations and cancellations from their registers and communicate them to Istat which publishes customised tables. Data on Istat website are complete in the sense that there is international and internal migration at different levels. There are data about origin and destination, but they only come with large age class groups (”0-17”, ”18-39”, ”40-64”, ”65+”). In order to have more details about migrants age, Eurostat data provide yearly age classes but only for international migration at national level. All datasets have data on migrant sex, and time span is 2006-2015. Note that registrations and cancellations refer to permanent residences, it is then very likely that actual data on migration are much higher than what data report.

### 3.4.1 International migration

Figure 3.30 shows the differences between the two Istat datasets and trend in international immigration. Dataset 1 ($X_1$) comes from Istat demographic balance data whereas dataset 2 ($X_2$) comes from single series provided by municipalities. From 2007-2008 there is a decreasing trend in almost all regions, whereas for emigration the trend is increasing (figure 3.31). These opposite behaviours are probably consequences of the economic crisis.

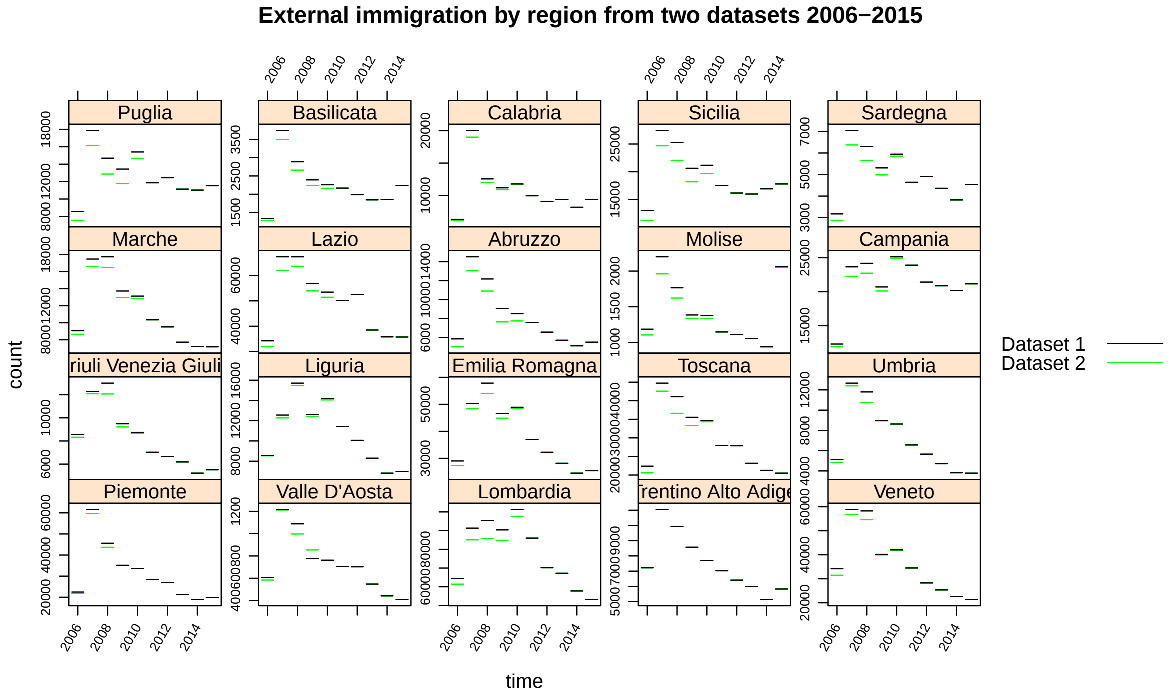


FIGURE 3.30: International immigration data at regional level for years 2006-2015 from dataset $X_1$ (Istat demographic balance) in black, and from dataset $X_2$ (municipalities data) in green.

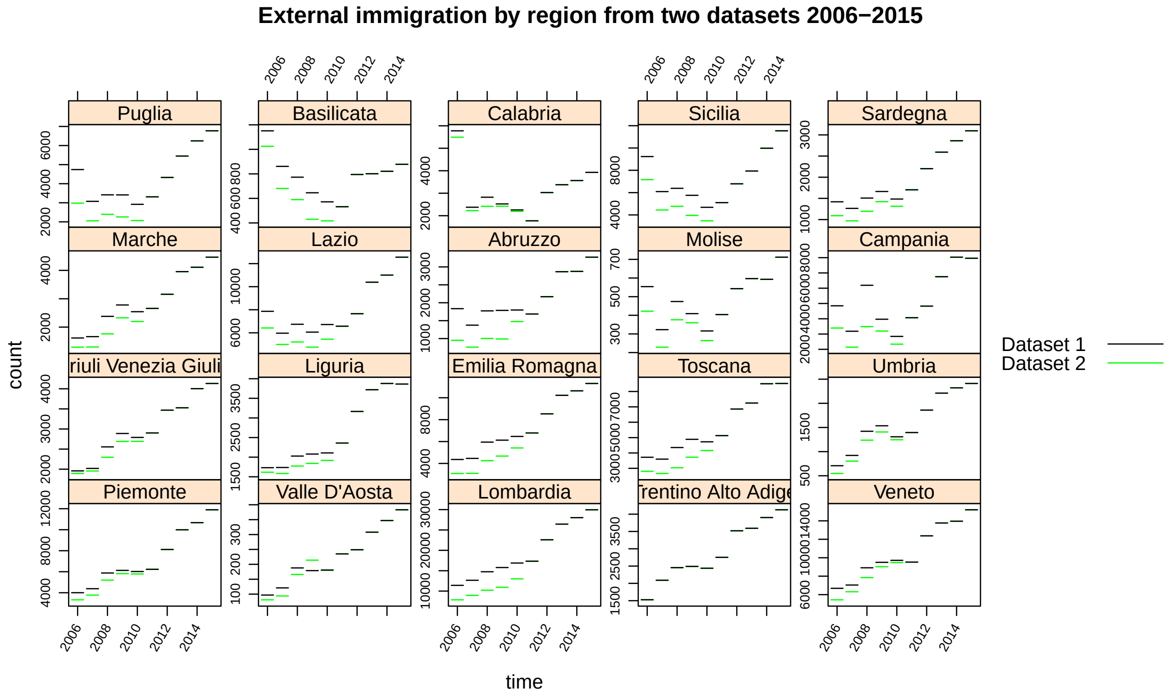


FIGURE 3.31: International emigration data at regional level for years 2006-2015 from dataset $X_1$ (Istat demographic balance) in black, and from dataset $X_2$ (municipalities data) in green.

In figures 3.32 and 3.33 there are respectively numbers on international immigration and emigration by large age groups, sex and year. Not surprisingly the largest group for both is the one for people aged 18-39, followed by 40-64 one. For gender, it seems there is an inversion for immigration, from 18+ years women coming to Italy are always more than men, but from 2011 on, men aged 18-40 become more than women. For emigration instead men are always more than women.

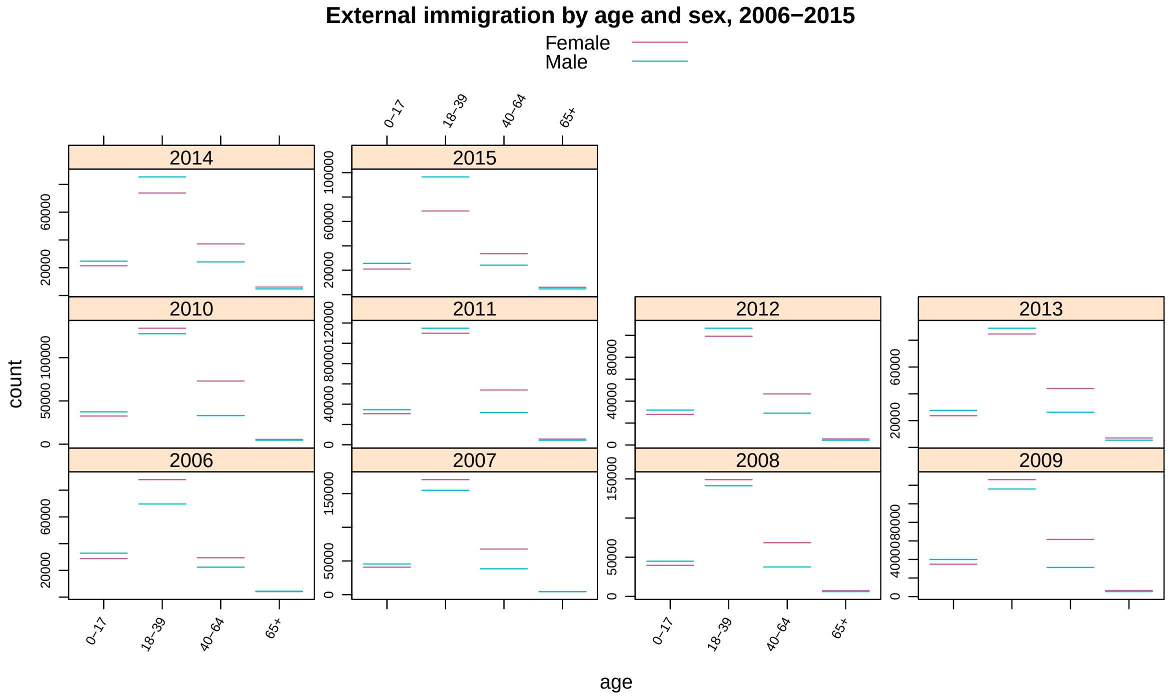


FIGURE 3.32: International immigration data by age, sex and time at national level. Every block represents data for one year (from 2006 until 2015) by large age groups (0 – 17, 18 – 39, 40 – 64, 65+) and sex (pink for females and light blue for males).

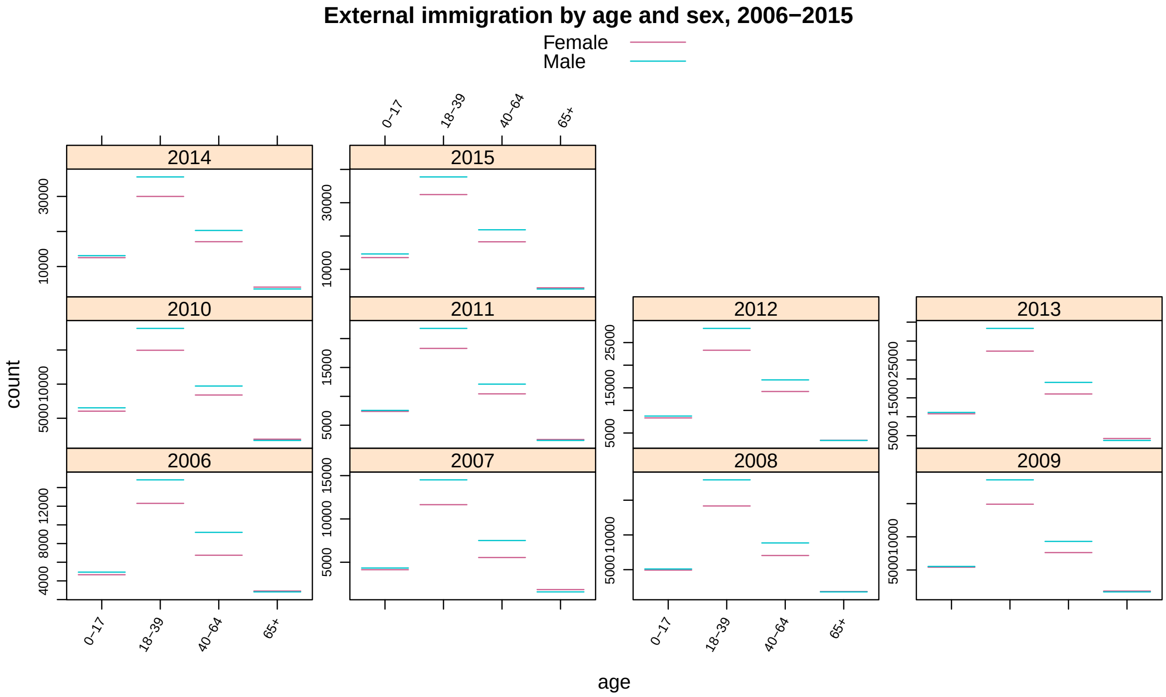


Figure 3.33: International emigration data by age, sex and time at national level. Every block represents data for one year (from 2006 until 2015) by large age groups ($0 - 17$, $18 - 39$, $40 - 64$, $65+$) and sex (pink for females and light blue for males).

Some things noticed in figure 3.32 and 3.33 can be found also analysing migration direct log-rates. Age, time and region effects appear to be strong in both immigration and emigration phenomena. For immigration (figure 3.34) age and sex effects are higher than for emigration (figure 3.35), whereas emigration have higher region and time effects. In both cases log-rates are higher than the overall log-rates for regions in the North and Centre of Italy and lower for the South and Islands. The only exception seems to be the region of Calabria, it is the only Southern region with smaller difference with North and Centre. Time trend confirms to be decreasing for immigration and increasing for emigration.

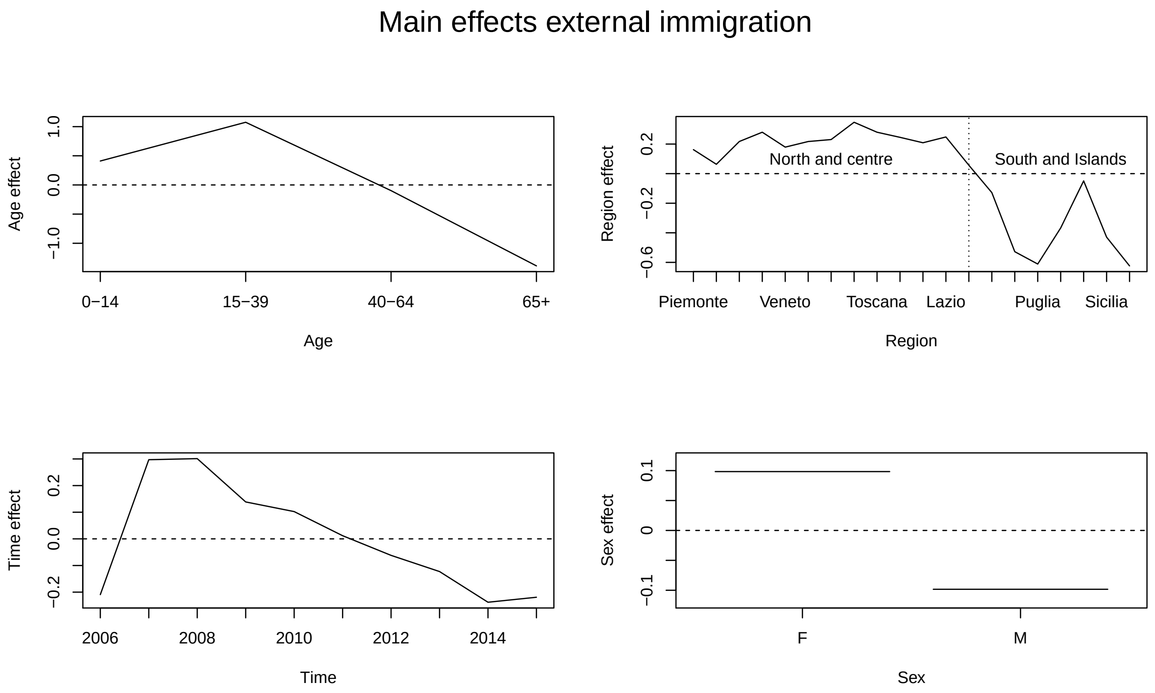


FIGURE 3.34: Computation of main effects from overall immigration direct estimates rates. *Up-Left*: Age effect, by large age groups. *Up-right*: Region effect, with vertical line dividing North and Centre regions from South and Islands regions. *Bottom-Left*: Time effect, lowering since 2008. *Bottom-Right*: Sex effect, higher for females.

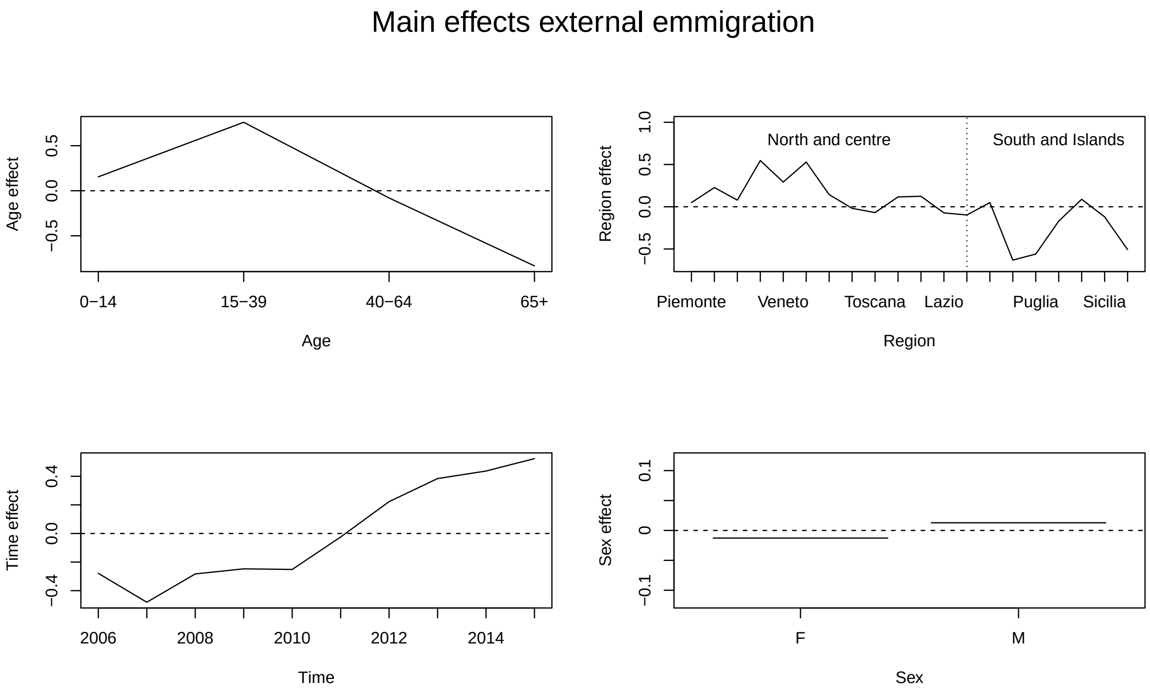


FIGURE 3.35: Computation of main effects from overall emigration direct estimates rates. *Up-Left*: Age effect, by large age groups. *Up-right*: Region effect, with vertical line dividing North and Centre regions from South and Islands regions. *Bottom-Left*: Time effect, rising since 2010. *Bottom-Right*: Sex effect, slightly lower for females.

With respect to interactions, immigration (figures 3.36 and 3.37) do not show strong or clear patterns. Usually women rates are higher in older age groups, and Southern regions have higher rates for older ages, whereas in the North and the Centre, where unemployment rates are lower, rates for older age groups are lower. Age-time interaction in this preliminary analysis does not seem to be significant. The same holds for region-time interaction, few regions show a general trend but it does not seem to be enough to identify an actual super-population effect.

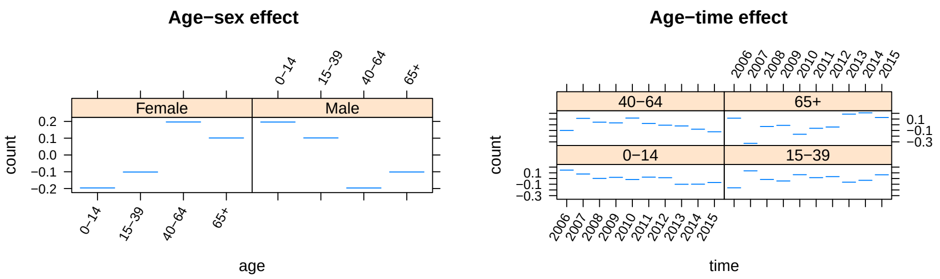


FIGURE 3.36: Computation of interactions obtained decomposing immigration direct estimate. *Left*: Age-sex interaction for females (left) and males (right) by large age groups. *Right*: Age-time interaction, each block shows the effect over time (2006-205) by large age groups.

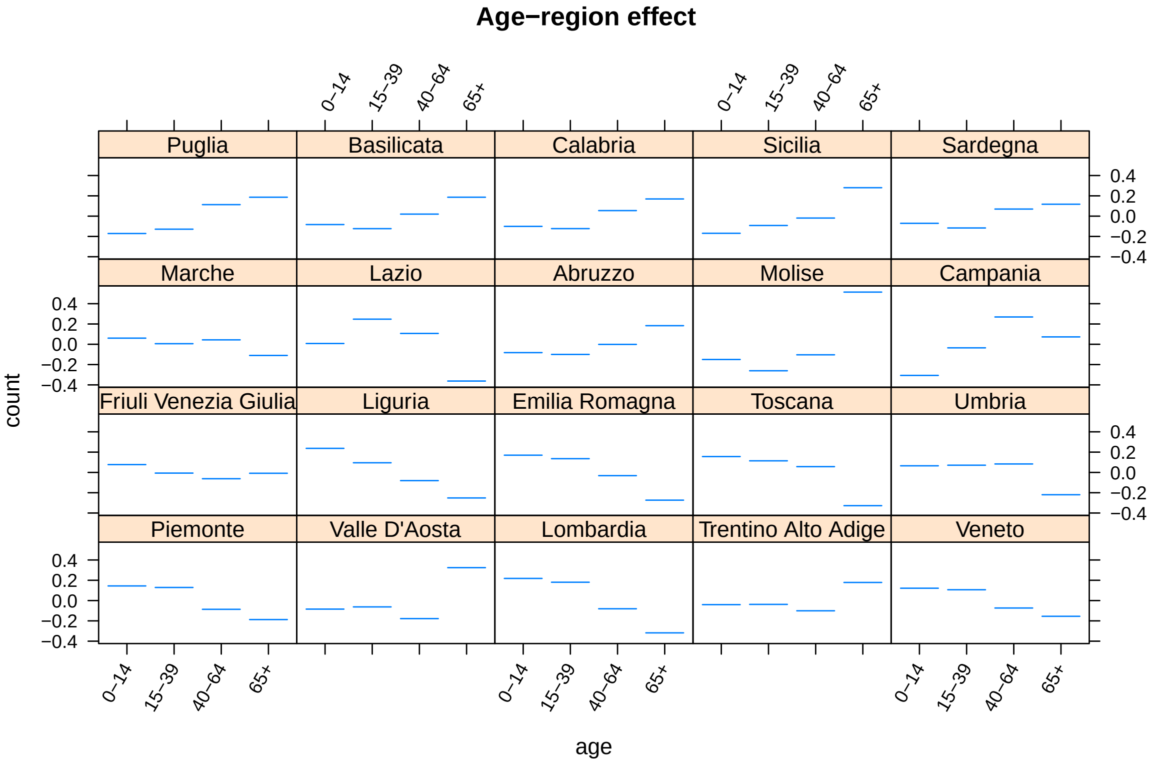


FIGURE 3.37: Computation of age-region interaction obtained decomposing immigration direct estimate. Each block represents a region and each light blue line corresponds to a large age group.

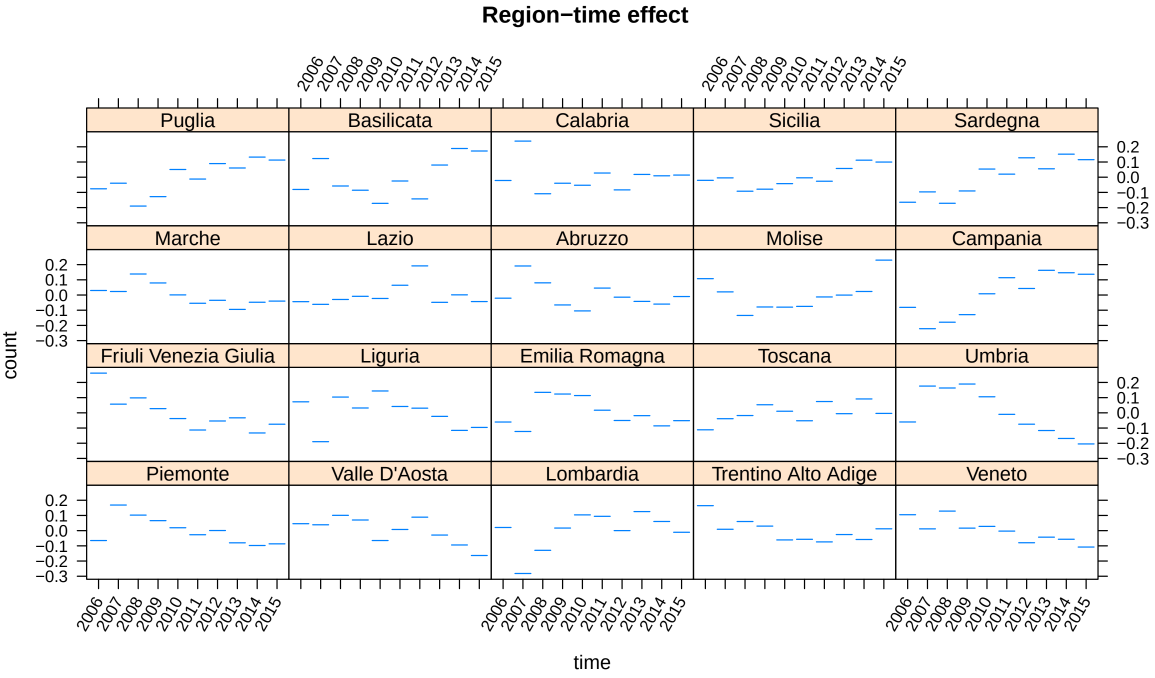


FIGURE 3.38: Computation of region-time interaction obtained decomposing immigration direct estimate. Each block represents a region and each light blue line corresponds to a year.

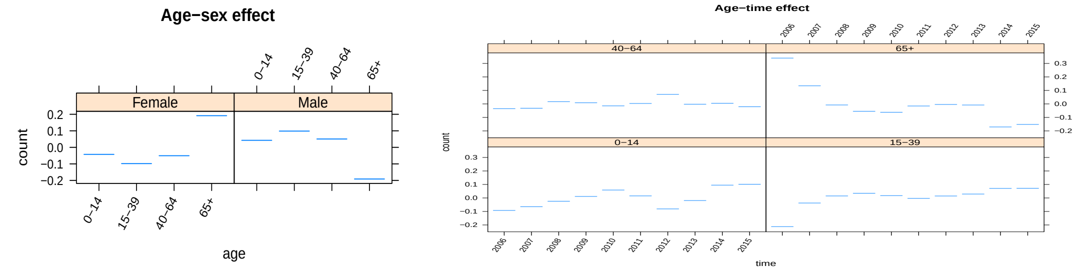


FIGURE 3.39: Computation of interactions obtained decomposing emigration direct estimate. *Left*: Age-sex interaction for females (left) and males (right) by large age groups. *Right*: Age-time interaction, each block shows the effect over time (2006-205) by large age groups.

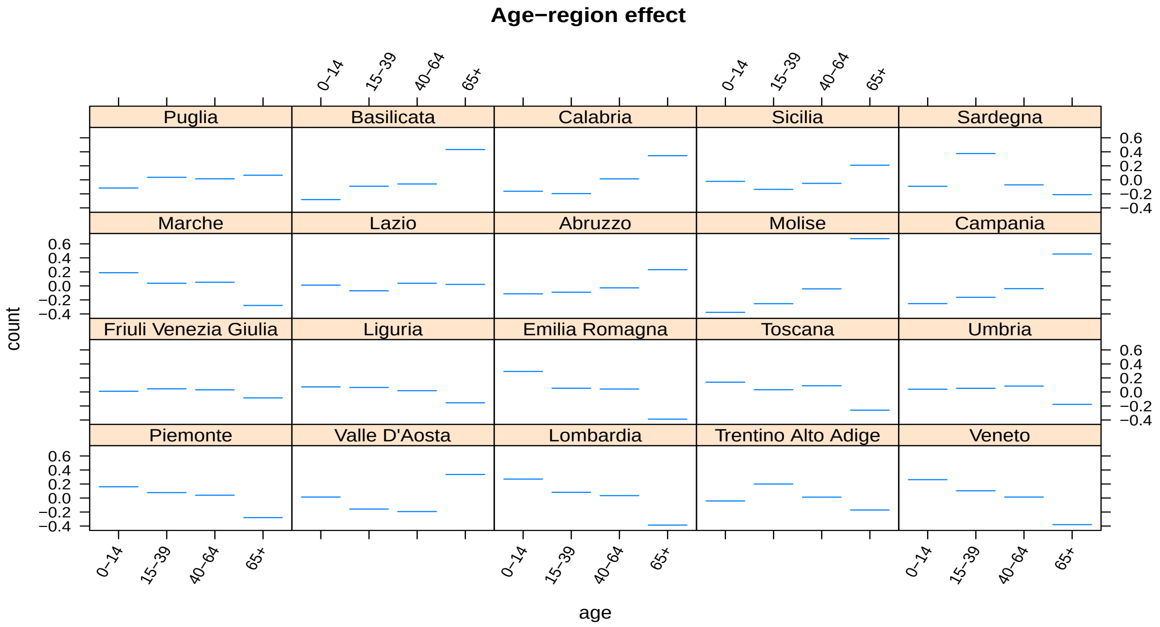


FIGURE 3.40: Computation of age-region interaction obtained decomposing emigration direct estimate. Each block represents a region and each light blue line corresponds to a large age group.

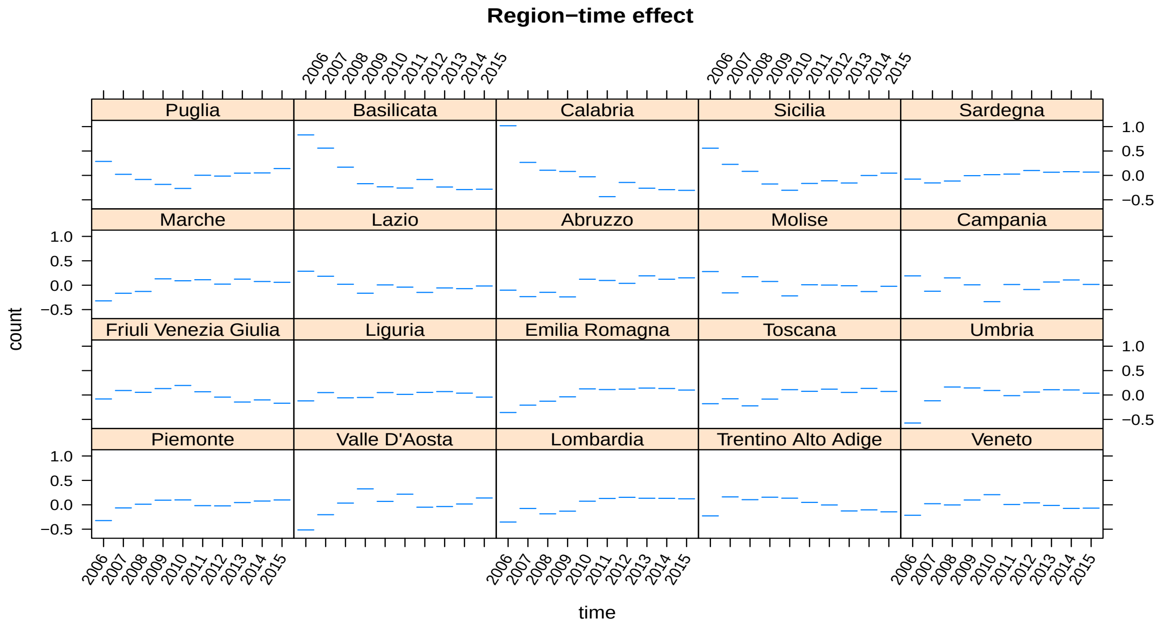


FIGURE 3.41: Computation of region-time interaction obtained decomposing emigration direct estimate. Each block represents a region and each light blue line corresponds to a year.

### 3.4.2 Modelling migration

As mentioned introducing migration topic, difficulties in estimating international flows are a common problem in many countries. Even in a country like New Zealand which is an island with better immigration records than in most of the countries, accuracy on migration data is considered "moderate" (Bryant and Zhang, 2018) whereas births and deaths registration have excellent accuracy. Italian data on migration refer to registrations and cancellations from municipality registers. International migration, especially at European level, is difficult to estimate. European laws and increasing mobility especially for students and workers make available data only partially trustworthy whereas illegal migration topic is not even addressed despite it is a major topic in Italy and Europe nowadays.

Internal migration analysis is similar to the international one but, apart from age effect, all the other effects and interactions do not show any clear pattern and, if they do, magnitude is quite low. Unlike for international migration which has clearer characteristics, a preliminary analysis easing the choice of system model for internal migration is difficult to provide. Clearly preliminary analyses give a glimpse of what could be the driving effects of a phenomenon, but they do not replace the proper estimation procedures.

Another aspect of migration is the format to describe it. There are four formats explained in Bryant and Zhang (2018) each one providing a different level of information. The most complete is the origin-destination format, all the movements are recorded in a square matrix with all the regions of origin and destination. This model provides information about both sending and receiving regions but it is a computationally demanding format. For example, considering the twenty Italian regions a matrix of four hundreds cells would be needed. Another way is the pool structure where only "total outward movements and total inward movements are shown for each status" (Bryant and Zhang, 2018). In this way the number of cells for Italian internal migration would be 40. A third format is the net format, it is efficient for population size estimation and only requires as many cells as status, i.e. twenty for Italy. Net migration only gives the balance between immigration and emigration but it does not provide information about the size of the flows and, as net flows are usually much smaller than inward and outward flows, even small percentage changes in separate flows could produce large percentage changes in net flows.

Data collected from the Istat website have pool format, they do not link origin and destination but only provide the number of registrations and cancellations. An origin-destination format can be obtained but the pool format, more parsimonious, has been chosen. Unlike net migration, pool format allows for separate immigration and emigration estimation but is not as computationally demanding as an origin-destination model. Migration is a complex phenomenon and to only estimate these series based only on the datasets available provide very partial results, especially because they are known not to be very accurate. For this reason migration estimation is only estimated within the demographic account, where demographic balance consistence is always checked and hence provide results that ensure internal consistency. When only estimating migration series results tend stay closer to the data, but they might not reflect the actual situation as balance equation is not considered.

## 3.5 Demographic account estimation

Unlike in sections 3.2 and 3.3, demographic account estimation involves all the demographic series. The whole model is much more complex, and combinations are almost endless. Here, three types of demographic account estimation are presented. The first only involves time dimension. This is the less flexible model, but also the best for comparing results from models with different assumptions. Then models considering time and region follow and, eventually, models with time and age dimensions. Each of them has its own peculiarities and difficulties to tackle. Comparisons are needed but, as it is impossible to give a complete report of all the attempts and changes that can be done, only the most representative results are shown. Unlike for births and deaths counts estimation, where data were trusted more than those on migration and the aim was to find the best model, now the aim is to see how the model performs and how robust it is to different assumptions. For demographic account estimation to choose the best model is not always easy as performance are different and data are sometimes far from the results. Depending on the dimensions needed, one or more datasets are included in the data model referring to the same demographic account series. Usually, there are two datasets for each series, the same compared in the previous sections: (i) one from the municipalities, and (ii) one from the demographic balance published by Istat in the annual report. Only for population there are three datasets: (i) the resident population dataset corresponding to the data published by Istat on the demographic balance and coming from municipalities, i.e. the POSAS form [5]; (ii) census data (only for 2011); (ii) the population reconstruction computed after the 2011 census (only for data from 2006 to 2011). Therefore, there are two datasets for the period 2006-2010, three for census year 2011, and only one for period 2012-2015. Census data, even after the Post

[5] POSAS = *Popolazione residente comunale per Sesso, Anno di nascita e Stato civile*; Resident population for municipalities by sex, year of birth and marital status.

Enumeration Survey (PES), are much lower than the population registered. Therefore there are two choices: either ignore census data and trust residence registers, or trust census data somehow forcing population estimation to closely follow them. To ignore census data is a controversial choice. Even if it is true that census has limitations, its results are presented to be highly reliable by Istat. Istat conducted all census process according to the best international standards, checking census results for three years through the Post enumeration survey (PES), combining census data with administrative registers to correct errors (see for example Istat (2016)). Therefore, following Istat census documents, it has been chosen to consider census data as the most accurate dataset. An accuracy of 98% is assumed, on line with Istat estimated results (Istat, 2015a). Population considered is only the resident population, so legally enrolled in an Italian municipality, and not the present population which is also measured by the census but it is more difficult to estimate both for scarcity of data and much lower accuracy.

Results are presented, as in sections 3.2 and 3.3, through graphs realised using the function `dplot` of `demest R package` and, as a general rule, they show the 95% credible interval in light blue, the 50% credible interval in blue, the median in white and the original data in red.

### 3.5.1 Only time

Estimating the whole demographic account considering only time dimension has pros and cons. The main advantage is that choices for the system model are quite simple. The only dimension that can be included is time for which a DLM prior is the best and almost automatic choice in every model. Therefore, the only thing to tune is the strength of *a priori* assumptions on standard deviation parameters, and the choice of DLM models (local level model, local trend model, with or without dumping term). Setting *a priori* for DLM parameters can be difficult but, in general, results are quite robust to different choices. When the model includes more dimensions, the choice of the system model becomes more complex. As the time dimension is available for all the series and it is the only one considered, time effect has been included in all system models. Equation (3.14) shows the system model assumed for all demographic series with the generic cell denoted by $y_t$. Priors on time coefficient $\beta_t^{time}$ are all DLM with local trend model with no dumping term and informative prior on $\sigma$ with a scale of 0.05.

$$
\begin{aligned}
& y_t \sim Pois(\gamma_t \omega_t),\ t = 1, ..., 10 \\
& \log(\gamma_t) \sim N(\mu_t, \sigma^2) \\
& \mu_t = \beta^0 + \beta_t^{time} \\
& \sigma_2 \sim t_7^*(1) \\
\text{where} & \\
& \beta^0 \sim N(0, 10^2) \\
& \beta_t^{time} \sim N(\alpha_t, \tau_t^2) \\
& \alpha_t \sim N(\alpha_{t-1} + \delta_t, \tau_\alpha^2) \\
& \delta_t \sim N(\delta_{t-1}, \tau_\delta^2) \\
& \tau_t, \tau_\alpha, \tau_\delta \sim t_7^+(1)
\end{aligned}
\tag{3.14}
$$

Exposure term $\omega_t$ appears in all the models but the one for population as no exposure term can be used for it.

More options are available for the data models choice. According to the assumptions on each dataset, results can be different and convergence can take much longer. Before choosing the data models, the first problem is what dataset to include since all of them have time dimension. Attempts have been made both considering all the different datasets available for each series, and only with the most trusted ones. Most of the time it is sensible to differentiate data models according to prior beliefs on datasets. Making the same assumption on all datasets, ignoring that some datasets are more trustworthy than others, prevents the whole estimation process to work properly and often makes the estimation much more difficult. Instead, when datasets are differentiated using prior knowledge on their reliability, the less trusted datasets have lower influence on the estimation and results usually reflect prior assumptions.

Eventually datasets used for the estimation are

- for population series: (i) after census population reconstructions for years $2006 - 2011$, (ii) census data for year 2011 and (iii) population from Istat official annual documents and used from demographic balances for years $2012 - 2015$ coming from POSAS forms.

- for births, deaths and external migration data come all from Istat official annual documents (*Annuario statistico italiano*).

Different combinations of data models have been tested. Only model on census data stays the same. As census data are the best data available for the decade considered, a model allowing for very little variation with respect to the dataset is assumed: a Poisson-Binomial distribution with a probability of 0.98, equation (3.15).

$$x_t^{census} \sim \text{PoissonBinomial}(p = 0.98) \tag{3.15}$$

Census data refer to the resident population in Italy on October the 9th, 2011 and official results appeared only in mid 2014. After collection and counting process, data went through a Post Enumeration Survey (PES), i.e. a control of data collected during the census with lists and registers of all Italian municipalities in order to correct errors, over-coverage or under-coverage that might have occurred during the census. The PES process started in March 2012 and ended on June the 30th 2014. Births and deaths datasets have a normal prior assuming data are at 95% within 2% of the true births and deaths counts. In equation (3.16) only contains the births model but model for deaths series ($x_t^b$) follow the same structure. The standard deviation assumed is higher than the one used in sections 3.2 and 3.3 in order for the model to be more flexible and adapt to balance equations if needed. The difference is not much and posterior estimation closely follow the data anyway.

$$\begin{aligned} & x_t^b \sim N\big(\gamma_t^b, {\phi_t^b}^2\big) \\ & \gamma_t^b = y_t^b \\ & \phi_t^b = 0.01 \times x_t^b \approx 0.01 \times y_t^b \end{aligned} \tag{3.16}$$

Prior distributions for population, external immigration and external emigration are, in all cases, Poisson models with informative priors on standard deviation parameters and no systematic bias on the mean is expected. If in simpler models, like with deaths and births series estimation, standard deviations prior distributions were weak, in more complex model a more informative prior can be helpful. Equation (3.17) presents data model for external (international) immigration ($ei$), but the same holds for external emigration ($ee$) and population.

$$\begin{aligned} & x_t^{ei} \sim Pois(\gamma_t^{ei}\omega_t^{ei}) \\ & \log(\gamma_t^{ei}) \sim N(\mu_t^{ei}, \sigma_{ei}^2) \\ & \mu_t^{ei} \sim N(0, 0.025^2) \\ & \sigma_{ei} \sim t_7^*(0.1) \end{aligned} \tag{3.17}$$

Convergence was reached between 100.000 and 200.000 iterations in all attempts. In all the models, population estimation is lower than the data, even for the period $2006 - 2011$ where data come from the post-census population reconstruction. Figure 3.42

shows results on population estimation for four different combinations of data models, including the one described above. System models are the same as in equation (3.14) but data models change. The upper right graph shows results for population estimation when all the data models are Normal with a prior standard deviation assumption of 10% (i.e. much larger than expected, at least for births, deaths and population). On the upper left plot, population, births and deaths data have a Normal model with standard deviation assumptions of 2%. Bottom left plot shows results for the baseline model (equations 3.16 and 3.17) and the bottom right has all Poisson data models. Results for population do not differ very much from one another, in all cases 95% credible interval (light blue in the figure) hardly reach data. The same also happens for system models with weaker prior, e.g. larger prior standard deviation values $\tau s = 1$.

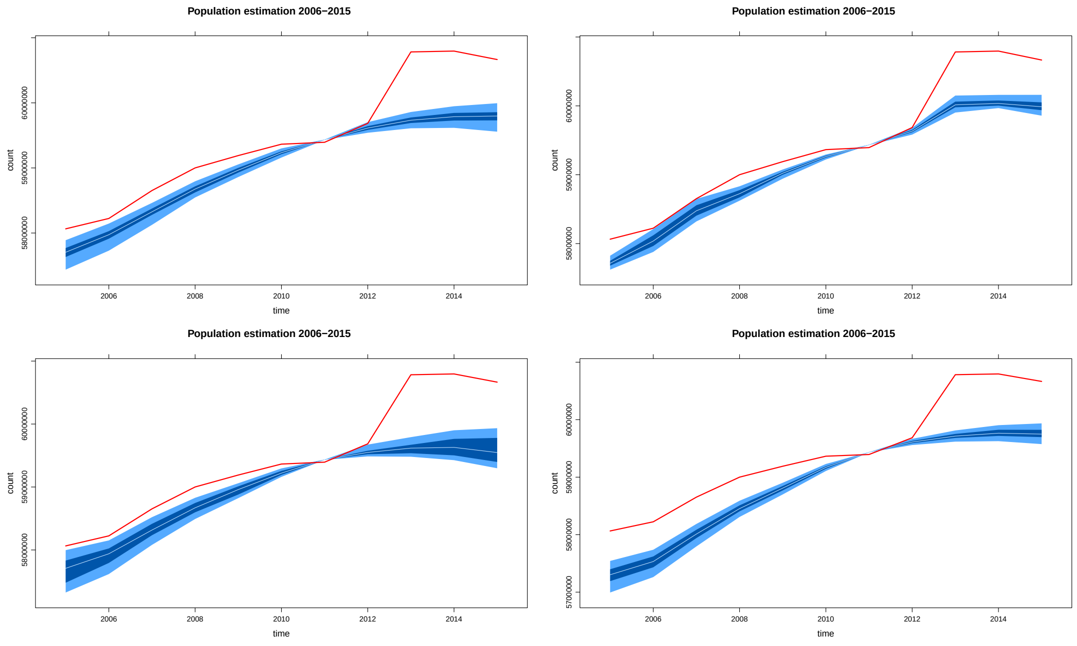


FIGURE 3.42: Population size estimation at national level for four different data models from 2006 until 2015. All four blocks show the medians as a white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue and in red Istat population reconstruction. *Up-left:* Data models consider only Normal distributions. *Up-right:* Data models consider Poisson prior distributions for migrations and all Normal for other series. *Bottom-left:* All distributions are assumed Poisson except for births and deaths where Normal distributions are assumed. *Bottom-right:* Priors for all series are assumed to be Poisson.

Other demographic series results in figures 3.43 and 3.44 show respectively results for births and deaths, and for international immigration and emigration. On the left side there are baseline model results (equations 3.16 and 3.17), on the right side there are results from the model with all Poisson data models. In figure 3.43 the effect of the data model change is only on the credible interval size which is much larger for the Poisson case, but medians are not substantially different.

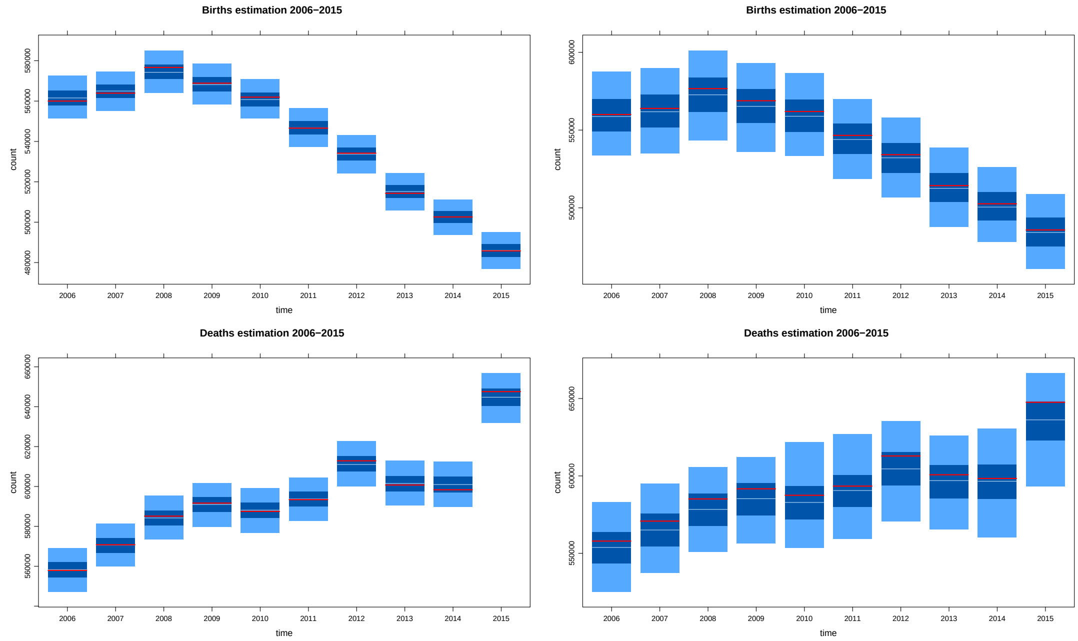


FIGURE 3.43: Comparison between birth and deaths count estimations by time for baseline data models and data models with all Poisson priors. All four blocks show the medians as a white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue and in red data from *Annuario statistico italiano*. *Up-left:* Births, baseline model. *Up-right:* Births, all Poisson. *Bottom-left:* Deaths, baseline model. *Bottom-right:* Deaths, all Poisson.

For migration instead the model is the same, both on the baseline and on the alternative model, and the change on births and deaths models do not affect migration series estimation. For the other two models shown in figure 3.42 results are comparable to results shown, only on the credible intervals size vary a bit. The model appears to be robust to prior choices. Estimates do not substantially change when different priors are assumed, only variations on credible intervals width occur depending on the strength of the prior but no big change on values of the estimates.

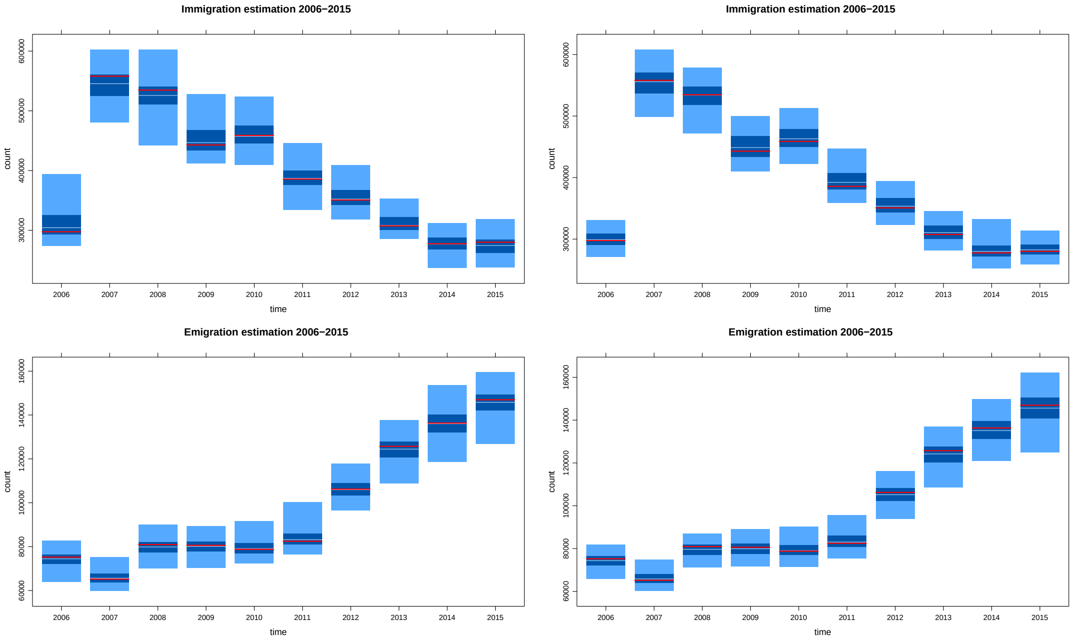


FIGURE 3.44: Comparison between immigration and emigration count estimations by time for baseline data models and data models with all Poisson priors. All four blocks show the medians as a white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue and in red data from Istat demographic balance. *Up-left:* Immigration, baseline model. *Up-right:* Immigration, all Poisson. *Bottom-left:* Emigration, baseline model. *Bottom-right:* Emigration, all Poisson.

An attempt not considering census data has been performed aiming to leave estimation less bounded to the strong prior assumption implied by the use of census data. A Normal prior for data models on both population datasets allowing for a 2% errors is assumed along with baseline model for the other series (equations 3.16 and 3.17). The resulting population size estimation in figure 3.45 is higher than in figures 3.42 but still remains very close to the population reconstruction. Other data series estimations are very close to the data and look very similar to baseline model results in figures 3.43 and 3.44 (right side results of both figures).

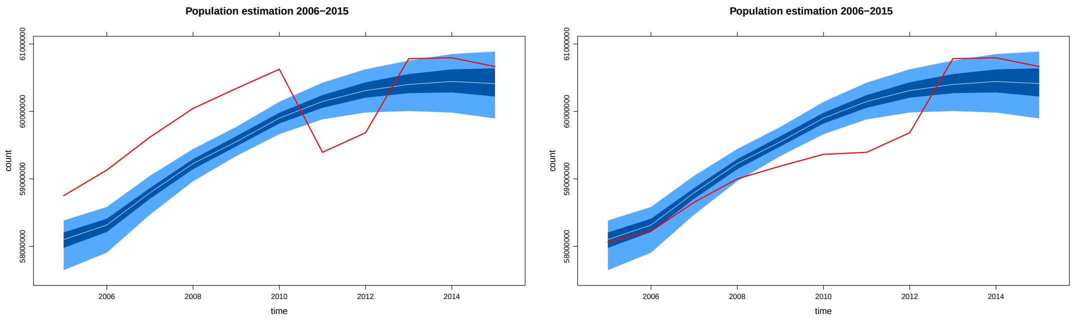
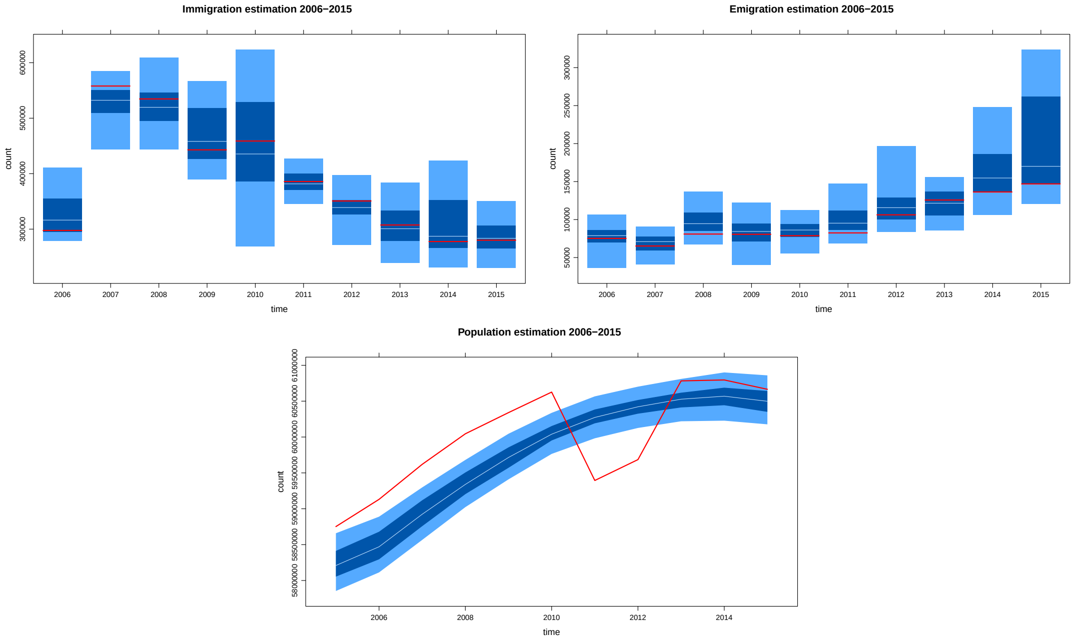


FIGURE 3.45: Both blocks show population size estimation at national level from 2006 until 2015 for baseline model not considering census data, i.e. only with POSAS data and reconstructed population data. The medians are the white lines, the 50% C.I.s in blue and the 95% C.I.s in light blue. Red lines show data from *right:* Administrative (POSAS) data, *left:* Reconstructed data.

Results do not change much when only official data on resident population (POSAS) are included, i.e. ignoring census and reconstructed population data. Keeping the baseline data models for all the remaining series (equations 3.16 and 3.17), results for migration and population are shown in figure 3.46. Births and deaths estimations are not presented as they reflect results for the baseline model in figure 3.43. Population estimation (bottom block of figure 3.46) is higher than in the other cases as census data and reconstructed data are not included, but it still remains lower than POSAS data that are still not included in the 95% credible interval. Only migration appears to be more irregular than in the other cases (compare two upper blocks of figure 3.46 with results in figure 3.44 and figure 3.47). This can be explained by the fact that migration has the weakest assumptions on its series (Poisson), whereas all the other series have Normal models that allow for less variability. The model uses the flexibility allowed by Poisson priors to modify migration series in order to obtain a consistent demographic account and leaves the series on births and deaths substantially unchanged. Results obtained are median lines still very close to the data but with irregular credible intervals.


Immigration estimation 2006–2015

Emigration estimation 2006–2015

Population estimation 2006–2015

Figure 3.46: Estimated results at national level from 2006 until 2015 for the baseline data models only considering POSAS data, i.e. ignoring census and reconstructed population data. The medians are the white lines, the 50% C.I.s are in blue, and the 95% C.I.s in light blue. *Up-Right*: Immigration count estimations with data from Istat demographic balance in red. *Up-Left:* Emigration count estimations with data from Istat demographic balance in red. *Bottom:*Population size estimation with POSAS data in red.

Results for an *ad hoc* model are shown in figure 3.47. Here only resident population dataset is included, deaths and births have Normal prior with 2% prior error and migration is assumed *a priori* under-reported of 10%. Prior distributions for this model are very informative and strongly influence estimation. Results reflect expectations and give a likely scenario for the Italian demographic account, with good birth and death data estimation, immigration data under-reported and asymmetric credible intervals in favour of larger values and a rising and much higher emigration than registered.

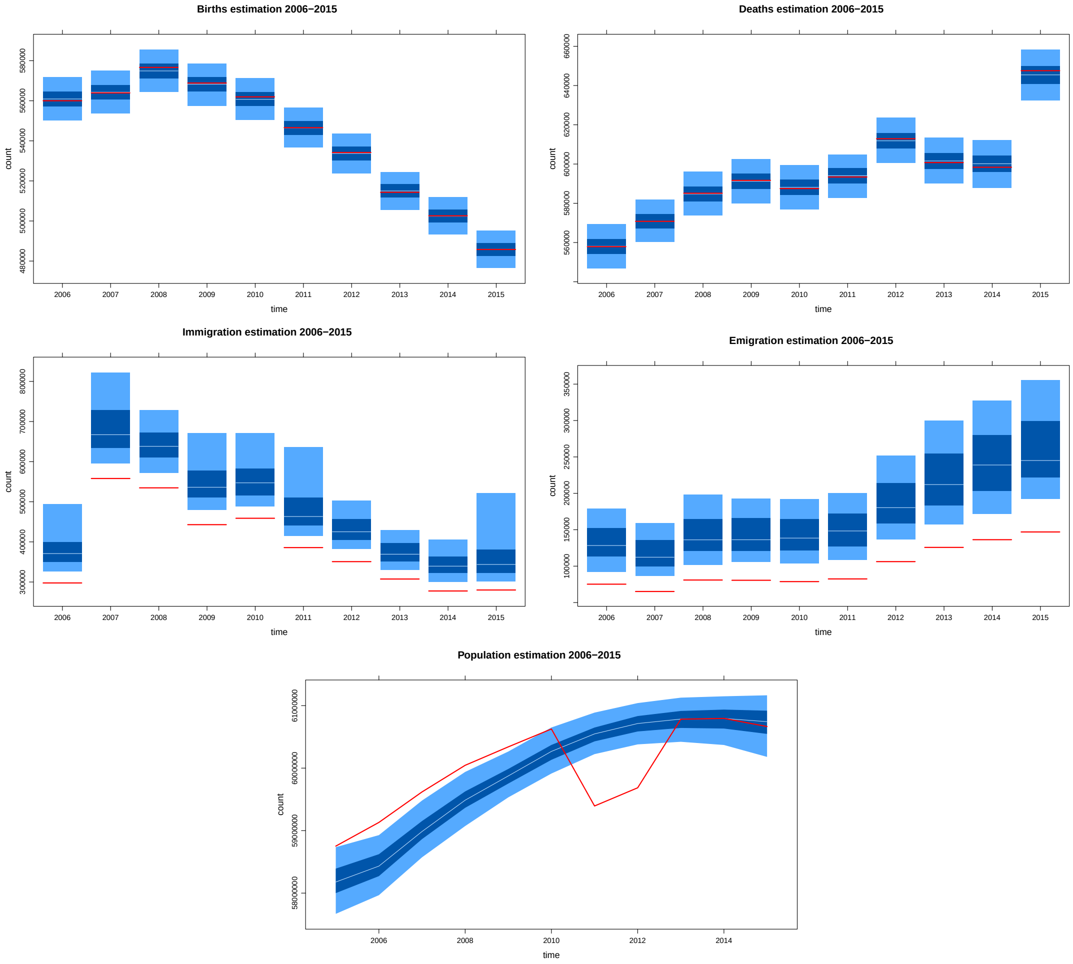


FIGURE 3.47: Demographic account estimation at national level for *ad hoc* data models from 2006 until 2015. Only resident population dataset is included and it is assumed Normal with 2% prior error, as well as births and deaths; both immigration and emigration are considered *a priori* affected by 10% under-reporting. All blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat respectively from: *Up-left*: Birth data from *Annuario statistico italiano*. *Up-right*: Death data from *Annuario statistico italiano*. *Centre-left*: Immigration from Istat demographic balance. *Centre-right*: Emigration from Istat demographic balance. *Bottom*: Resident population from POSAS data.

In general, when only official data on resident population are considered the estimation of demographic series varies much more, results are less robust, with high variation and hard to reach convergence. But, still, results do not show opposite results than in models with more information. Estimations reflect data but also remain consistent and sensible despite changes in prior assumptions and datasets considered. Especially for population, where the higher differences are observed, estimation is always generally lower than data.

### 3.5.2 Time and region

The demographic account considered in this section includes time and region dimensions. As the time dimension, the regional dimension is present in all the considered datasets, therefore only the more accurate ones are included (i.e. Istat regional demographic balance datasets). Italy has twenty regions but instead of region “Trentino Alto Adige” the two autonomous provinces of “Trento” and “Bolzano” are considered, in line with Istat and administrative recommendations. There are therefore twenty-one regions included in the model. Data for “Trentino Alto Adige” simply correspond to the sum of the two provinces. When estimating population size at regional level, a complication arising is the need to include data on internal migration or, more precisely, inter-regional migration.

For historical reasons internal migration has always been hard to trace in Italy. Since the 1990’s the situation has substantially improved and even more after 2013 when paper forms have been replaced by the software ISI-Istatel. ISI-Istatel allows for electronic management of all the forms, it produces tracks of internal migrations and and automatically transmits data. This software has improved all the steps of data management, reducing errors and delays (Istat, 2015b). Nevertheless, internal migration is still difficult to track as many people do not notify their *de facto* change of residency. Reasons for not changing residency when moving somewhere else in the country are diverse. They go from carelessness or ignorance to financial advantages, sometimes the process is considered to be too long therefore people do not notify temporary moving. This phenomenon is particularly true for example for students attending an university outside their region of residence. They study and live for years in a different city, sometimes also starting working there, without notifying it to the authorities. The problem of internal migration has been addressed and studied also in other countries and it has sometimes been considered even more serious than international migration (Bryant and Graham, 2013; Egidi and Ferruzza, 2009; Spencer *et al.*, 2017; Yildiz and Smith, 2015). Hence, as international migration, also internal migration it is very likely to suffer from under-reporting, and, despite the data quality has improved, reliability is still not high enough. Furthermore, when estimating internal migration there is an additional constraint the model has to account for. Internal migration needs to be consistent at national level, i.e. the sum of all inter-regional immigration and emigration must be zero at national level. This further complicates the whole demographic account model (DAM) as in addition to the population balance constraint also the migration constraint (equation (3.18)) needs to be satisfied at every updating of the demographic account.

$$\sum_{r=1}^{R} y_{rt}^{ii} = \sum_{r=1}^{R} y_{rt}^{ie} \tag{3.18}$$

where $r = 1, ..., R$ is the number of regions, $t = 1, ..., T$ is the number of years considered and $ii$ and $ie$ are respectively internal immigration and emigration.

The analysis presented in this section compares three different DAMs denoted by M1, M2 and M3. The three models share the same system model but different data models. In the system model, common to the three models, each series has its system model with different structure of parameter $\mu_{rt}$. Despite standard deviations are small in all cases (0.05 for DLM and 0.25 for Exchangeable priors), the effects included in $\mu_{rt}$ are different. The first part is the same for all series (equation (3.19)), but the definition of $\mu_{rt}$ depends on the series it relates to. Equation (3.20) shows the specification for population, births, deaths, external immigration, external emigration, internal immigration and internal emigration, respectively referred to by $p$, $b$, $d$, $ei$, $ee$, $ii$ and $ie$.

$$\begin{aligned} &y_{rt} \sim Pois(\gamma_{rt}),\, r = 1, ..., 21,\, t = 1, ..., 10 \\ &\log(\gamma_{rt}) \sim N(\mu_{rt}, \sigma^2) \\ &\sigma_2 \sim t_7^*(1) \end{aligned} \tag{3.19}$$

For deaths only time effect is included as no big regional effect has been identified (see section 3.2), births include time and region effects but no interaction to avoid overfitting, whereas for population and all migration series also a region-time interaction is considered. The reason is the evidence during preliminary analyses for a considerable region-time effect, see figures 3.38 and 3.41. Moreover, this is also consistent with quick changes in migration flows and regional differences. In each model intercept $\beta_0$ has the same weak prior $N(0, 10^2)$, whereas $\beta^{time}$ and $\beta_{rt}^{region:time}$ (if included) have both a local trend model without damping term, and $\beta_r^{region}$ has an exchangeable prior with 0.25 standard deviation scale.

$$\mu_{rt}^{b} = \beta^0 + \beta_t^{time} + \beta_r^{region} \tag{3.20}$$

$$\mu_{rt}^{d} = \beta^0 + \beta_t^{time} \tag{3.21}$$

$$\mu_{rt}^{p} = \beta^0 + \beta_t^{time} + \beta_r^{region} + \beta_{rt}^{region:time} \tag{3.22}$$

$$\mu_{rt}^{ei} = \mu_{rt}^{ee} = \mu_{rt}^{ii} = \mu_{rt}^{ie} = \mu_{rt}^{p} \tag{3.23}$$

Models M1 and M2 include reconstructed population datasets and census data for population series, and for all the other series datasets from the demographic balance are used. M3 includes also the dataset on resident population (POSAS) from the demographic balance for population series. Data models for M1 are similar to the one presented in section 3.5.1 but there are more cells due to regional additional dimension. Census data have a Poisson-Binomial prior as in equation (3.15), birth and death counts are Normal as in equation (3.16), and the rest (population and migrations) are all Poisson with a informative priors on the intercept to avoid large identification problems, see equation(3.17). M2 assumes Normal prior on all the series with standard deviation of 1% for population, births and deaths, and 5% for migrations. In M3, census data, births and deaths have the same data models as in M1 and M2, Poisson models are assumed on both population datasets as in equation (3.17), and migration models are again Poisson but with larger standard deviations on hyper-parameters ($\sigma = 1$ and $\tau = 1$). Models M1, M2 and M3 are reported in table 3.1.

| **Model** | **Series** | **Data model** |
|---|---|---|
| M1 | Census | $x_{rt}^{census} \sim \text{PoissonBinomial}(p = 0.98)$ |
| | Births<br>Deaths | $x_{rt}^{b} \sim N(\gamma_{rt}^{b}, \phi_{rt}^{b^2}), \quad \gamma_{rt}^{b} = y_{rt}^{b}, \quad \phi_{rt}^{b} = 0.01 \times x_{rt}^{b}$ |
| | Rec. pop.<br>Ext. imm.<br>Ext. emi.<br>Int. imm.<br>Int. emi. | $x_{rt}^{ei} \sim Pois(\gamma_{rt}^{ei}\omega_{rt}^{ei})$<br>$\log(\gamma_{rt}^{ei}) \sim N(\mu_{rt}^{ei}, \sigma_{ei}^{2}),$<br>$\mu_{rt}^{ei} \sim N(0, 0.025^2),$<br>$\sigma_{ei} \sim t_7^*(0.1)$ |
| M2 | Census | $x_{rt}^{census} \sim \text{PoissonBinomial}(p = 0.98)$ |
| | Births<br>Deaths<br>Rec. pop. | $x_{rt}^{b} \sim N(\gamma_{rt}^{b}, \phi_{rt}^{b^2}), \quad \gamma_{rt}^{b} = y_{rt}^{b}, \quad \phi_{rt}^{b} = 0.01 \times x_{rt}^{b}$ |
| | Ext. imm.<br>Ext. emi.<br>Int. imm.<br>Int. emi. | $x_{rt}^{ei} \sim N(\gamma_{rt}^{ei}, \phi_{rt}^{ei^2})$<br>$\gamma_{rt}^{ei} = y_{rt}^{ei}$<br>$\phi_{rt}^{ei} = 0.05 \times x_{rt}^{ei}$ |
| M3 | Census | $x_{rt}^{census} \sim \text{PoissonBinomial}(p = 0.98)$ |
| | Births<br>Deaths | $x_{rt}^{b} \sim N(\gamma_{rt}^{b}, \phi_{rt}^{b^2}), \quad \gamma_{rt}^{b} = y_{rt}^{b}, \quad \phi_{rt}^{b} = 0.01 \times x_{rt}^{b}$ |
| | Rec. pop.<br>POSAS | $x_{rt}^{p} \sim N(\gamma_{rt}^{p}\phi_{rt}^{p^2}), \quad \gamma_{rt}^{p} = y_{rt}^{p}, \quad \phi_{rt}^{p} = 0.05 \times x_{rt}^{p}$ |
| | Ext. imm.<br>Ext. emi.<br>Int. imm.<br>Int. emi. | $x_{rt}^{ei} \sim Pois(\gamma_{rt}^{ei}\omega_{rt}^{ei})$<br>$\log(\gamma_{rt}^{ei}) \sim N(\mu_{rt}^{ei}, \sigma_{ei}^{2}),$<br>$\mu_{rt}^{ei} \sim N(0, 1), \quad \sigma_{ei} \sim t_7^*(1)$ |

Table 3.1: Data models for the three models tested (M1, M2, M3). For each model the data model assumed on each series is presented. If data model are the same for more than one series, then only one is shown with the name of the series the example refers to. Series are census, births, deaths, reconstructed population and POSAS population, external immigration, external emigration, internal immigration and internal emigration, respectively referred to by *census*, *b*, *d*, *p* (for both reconstructed and POSAS), *ei*, *ee*, *ii* and *ie*.

In figure 3.48 there are plots for population estimation of the three different models. For M1 and M2 estimation is quite similar, but in M1 intervals follow much closely the data than in M2. Especially for the last years M1 seems to fit better to the slower population growth whereas M2 sticks more to the general trend (both upward or downward, depending on the region). In M3 credible intervals are much narrower than in the other cases. Probably the presence of more than one dataset and the weaker assumptions had a positive impact on the results.

As the data model is the same for all three cases, according to expectations, births and deaths counts have almost identical results in all cases. It means that changes in one or more series assumptions do not significantly affect results for the other series. Being so similar only results of M1 are reported in figure 3.49.

For migration results and interpretation is more challenging, as for the three models there are three different results (figure 3.50). M1 has very large and irregular credible intervals (both 50% in blue and 95% in light blue) that include the data. M2 has smaller credible intervals, but they seem not to respect the data trend at all, whereas in M3 both credible intervals and data trend are respected.

For international emigration credible intervals are in all cases small and regular but models M1 and M2 have higher estimations than data (which is sensible according to *a priori* information Istat has on international emigration). In M3 instead, international emigration closely follow data. From the results on international migration a general phenomenon of under-reporting of emigrations can be deduced, whereas more complicated is to understand international immigration.

For internal migration, both immigration and emigration results reflect the expected low accuracy of the datasets according to Istat warnings (Egidi and Ferruzza, 2009) and the further complication the constraint (3.18) introduces. Credible intervals are wide and often do not even include data, immigration estimation is usually lower than data whereas emigration estimation is almost always higher except for the province of Bolzano where is lower. Estimation of internal migration for the three cases are in figures 3.52 and 3.53.

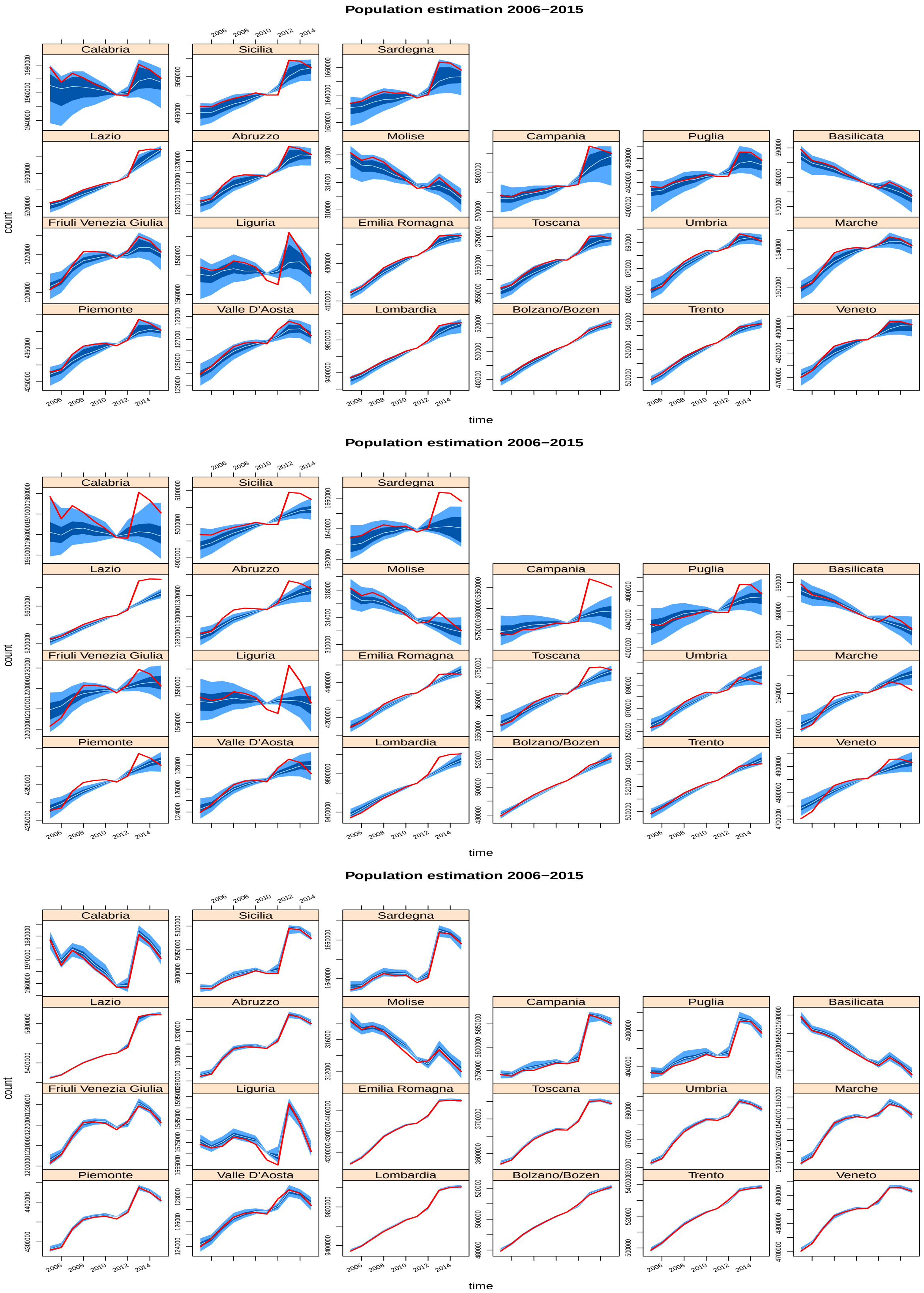


FIGURE 3.48: Population estimation at regional level for years from 2006 until 2015. All regional blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat reconstructed population. *Upper:* Population estimation, M1. *Centre:* Population estimation, M2. *Bottom:* Population estimation, M3.

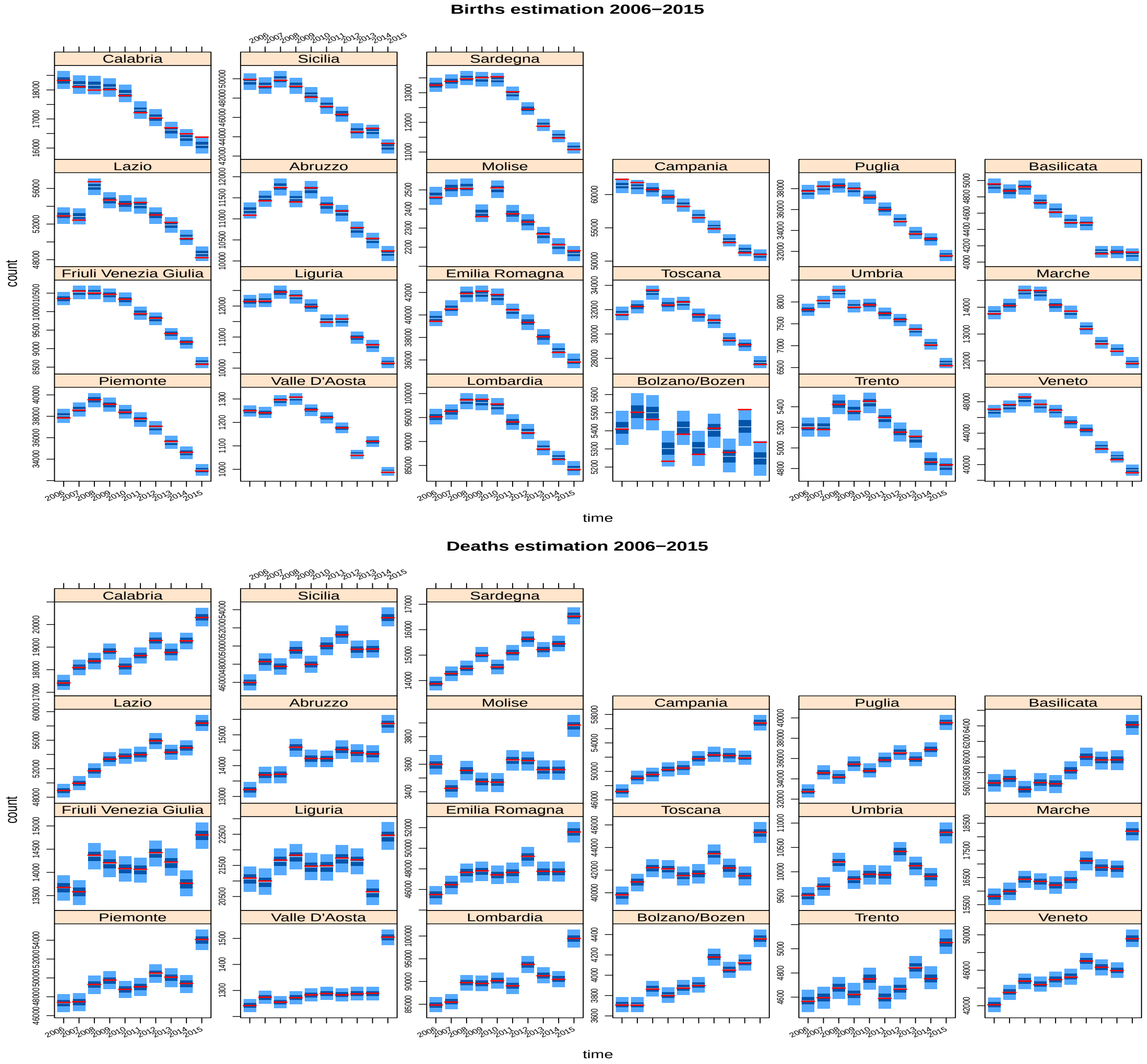


FIGURE 3.49: Birth and death count estimations at regional level for years from 2006 until 2015. All regional blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balances. *Upper:* Birth counts estimation. *Lower:* Death counts estimation.

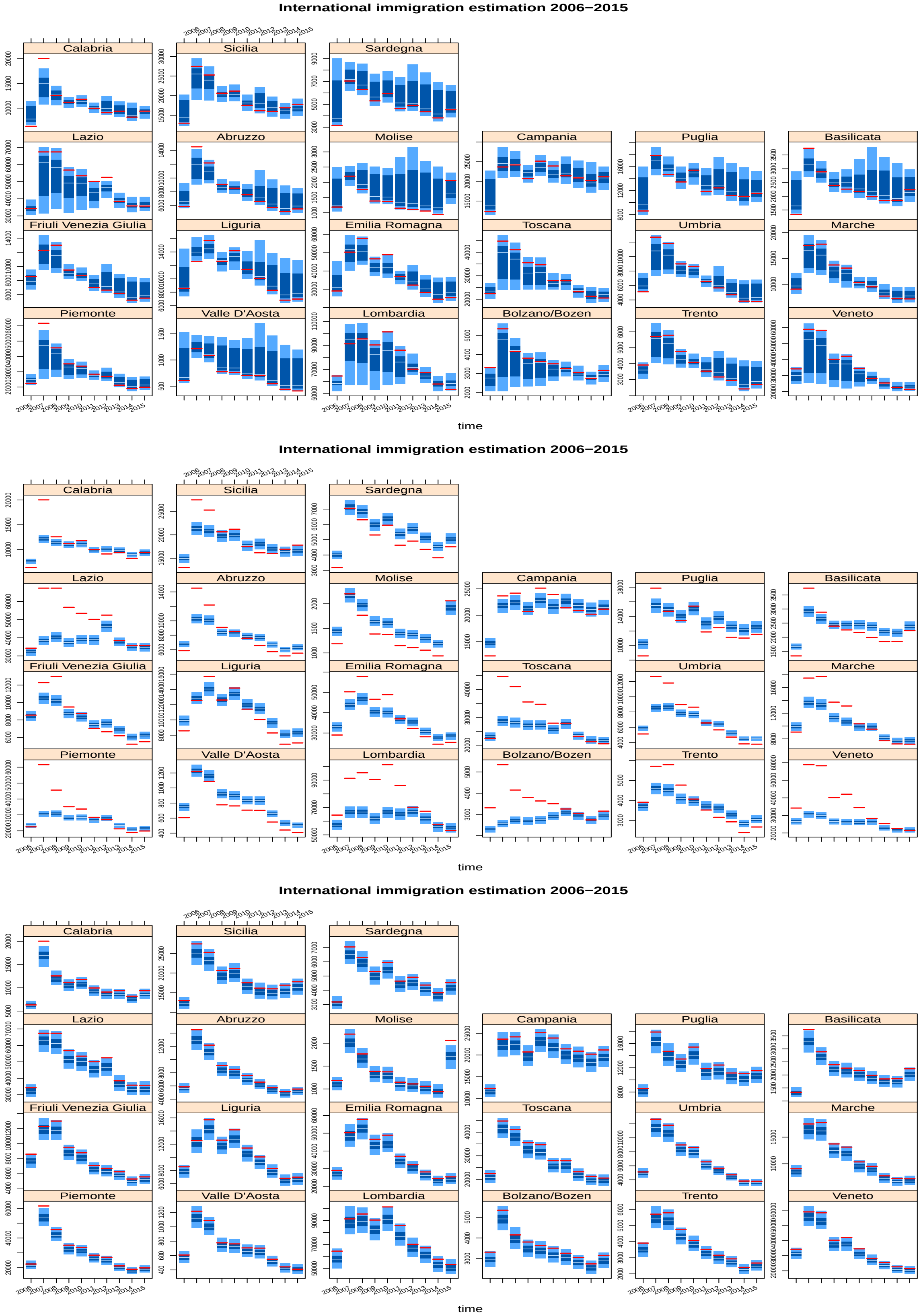


FIGURE 3.50: International immigration count estimations at regional level for years from 2006 until 2015. All regional blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balances. *Upper* International immigration estimation, M1. *Centre:* International immigration estimation, M2. *Bottom:* International immigration estimation, M3.

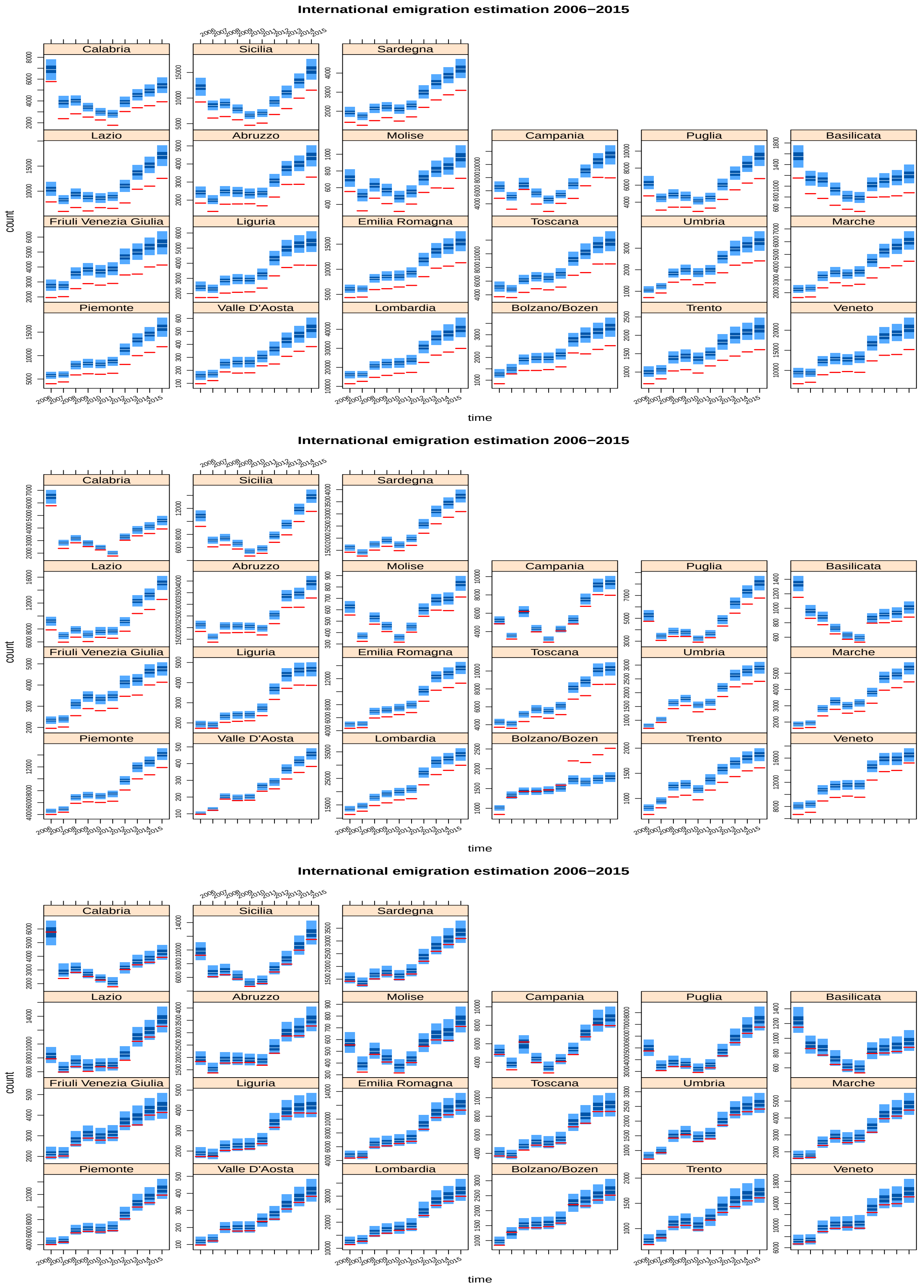


Figure 3.51: International emigration count estimations at regional level for years from 2006 until 2015. All regional blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balances. *Upper* International emigration estimation, M1. *Centre:* International emigration estimation, M2. *Bottom:* International emigration estimation, M3.

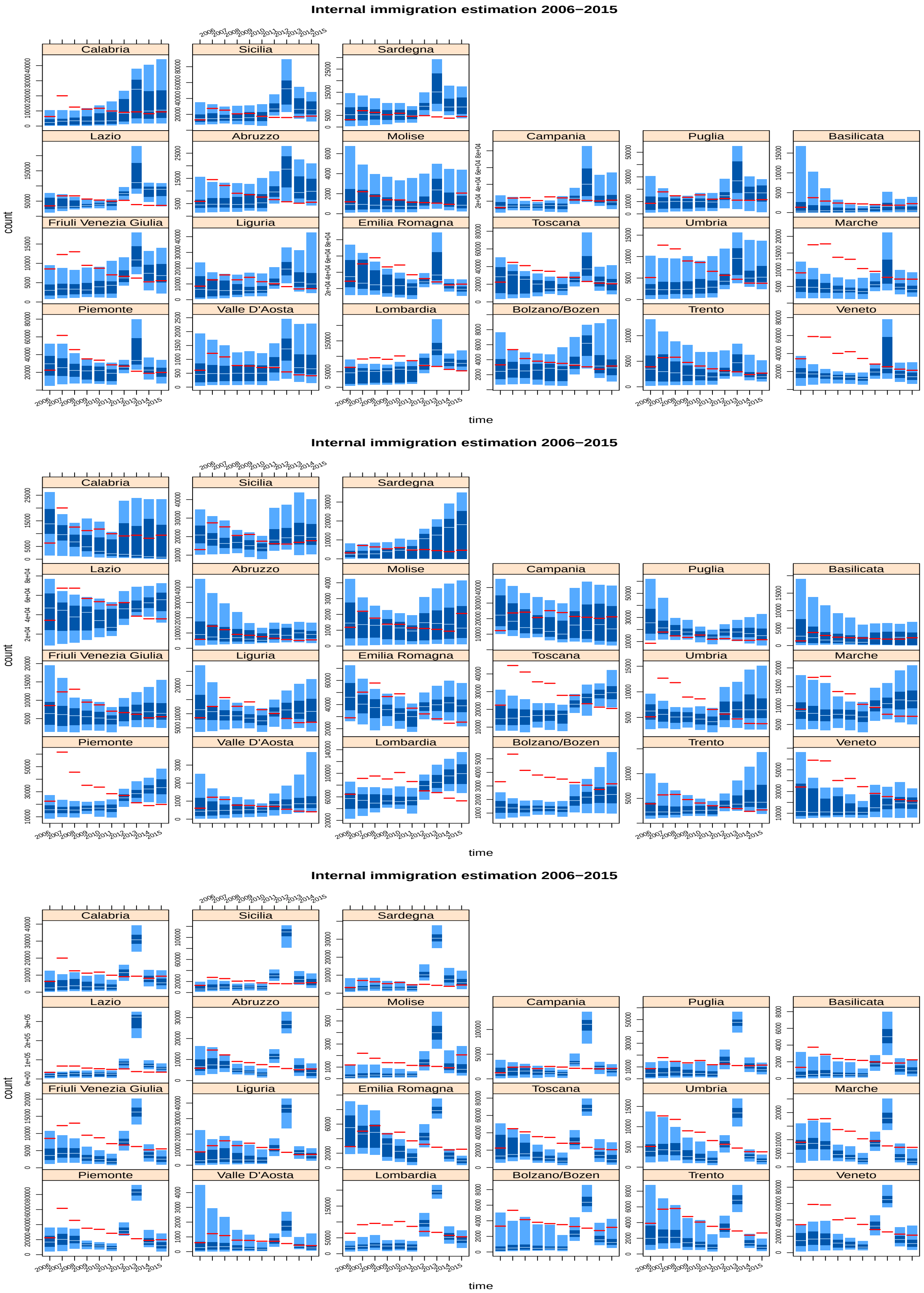


FIGURE 3.52: Internal immigration count estimations at regional level for years from 2006 until 2015. All regional blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balances. *Upper:* Internal immigration estimation, M1. *Centre:* Internal immigration estimation, M2. *Bottom:* Internal immigration estimation, M3.

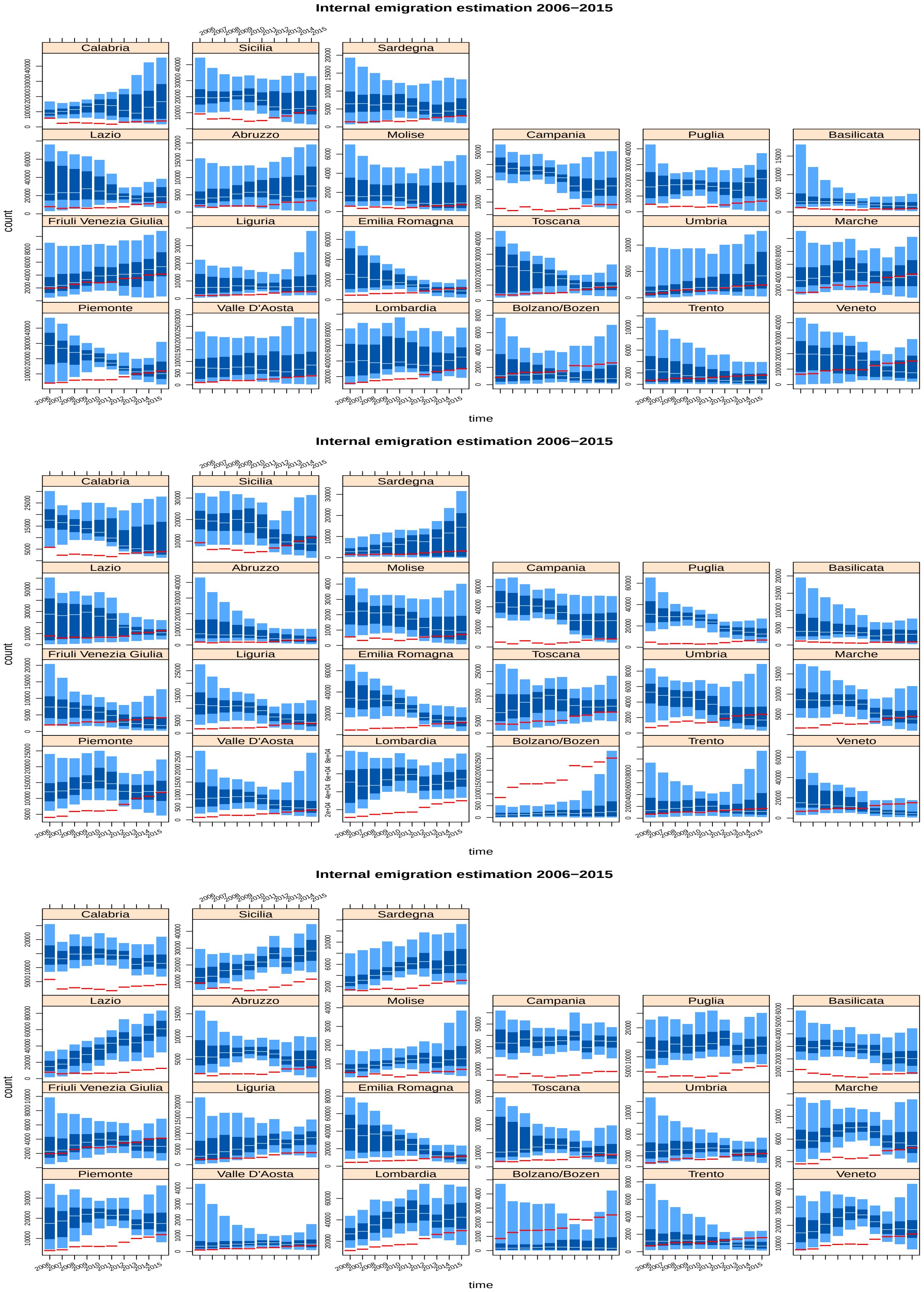


FIGURE 3.53: Internal emigration count estimations at regional level for years from 2006 until 2015. All regional blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balances. *Upper:* Internal emigration estimation, M1.*Centre:* Internal emigration estimation, M2. *Bottom:* Internal emigration estimation, M3.

It is difficult to motivate and compare these results on migration, but estimates clearly show inconsistencies in the data. Istat regularly corrects errors in residence registrations or cancellations, but population movements are not always easy to reconstruct. Even with a Normal prior (in model M2), results do not improve, showing that, even with a prior assuming higher accuracy on data, estimates still deviate from data.

Nowadays migration is a challenging aspect of population studies, here models only include time and region dimensions, without any other covariate. Here only three attempts sharing the same system model are presented, allowing for data model robustness checking, comparisons are harder when too many things change. Changes to the system models do not seem to affect much the estimation, for example with a system model not considering interactions at all or with weaker priors on standard deviations results are almost the same. Other data models and other datasets can also be considered, but attempts with the most accurate datasets were more worth presenting than the others where results were worse and convergence harder to reach.

The territorial level considered in this section is the regional one. There are no theoretical restriction to the level of territorial specification. Application can be performed at provincial or even municipality levels. For Italy only the regional level has been considered for three main reasons: (i) differences between regions are usually more evident that differences between provinces or municipalities; (ii) the number of provinces has been varying multiple times during the time span considered $2006 - 2015$ making difficult datasets comparisons across years; (iii) the analysis of almost a hundred territorial units implies a high computational cost not affordable in the context of the present investigation.

### 3.5.3 Time and age

Estimation by time and age is a particularly challenging task. First of all because not all the datasets come with age, and sometimes age groups are different. For instance, all datasets from Istat demographic balance only come with time, region and sex dimensions, and these are the data having being trusted the most so far; population usually comes in single year groups; mothers' age in births counts and deaths counts depend on the table used. There can be either single or five years groups; migration usually comes in large age groups ("$0-17''$", "$18-39''$", "$40-64''$", "$65+''$") but can also be found by single year. Estimation with single year age groups is prohibitive in terms of computational time, and five-years age groups are commonly used in demography as characteristics and rates are similar in such a small range. More difficult is to deal with large age groups, but considering only age and not the region it is possible to use migration datasets at national level with single year age groups provided by Eurostat. Because some datasets only provide data with five years age groups, the model considers age dimension in five years age groups, as it has already been in the previous models. Considering the whole demographic account all the series must be consistent, and, as the age groups are five years wide, for population (point arrays) only the data at five years interval are considered (i.e. 31/12/2005, 31/12/2010 and 31/12/2015). For other demographic series (interval arrays) data are collapsed for the two period 2006-2010 and 2011-2015.

The population system model (equation (3.24)) includes intercept, age and time effects plus age-time interaction. Intercept is modelled as usual ($\beta^0 \sim N(0, 10^2)$), main effects and the interaction all have DLM priors. Time has a local trend model, whereas age and age-time interaction only assume a local level model. No damping term is included. All other series have the same structure plus exposure term $\omega_{at}$, priors on standard deviations are higher for migration (0.2 instead of 0.1).

$$\begin{aligned}
&y_{at} \sim Pois(\gamma_{at}),\, a = 1, ..., 19,\, t = 1, ..., 10 \qquad (3.24)\\
&\log(\gamma_{at}) \sim N(\mu_{at}, \sigma^2)\\
&\mu_{at} = \beta^0 + \beta_t^{time} + \beta_a^{age} + \beta_{at}^{age:time}\\
&\sigma_2 \sim t_7^*(1)
\end{aligned}$$

where

$$\begin{aligned}
&\beta^0 \sim N(0, 10^2)\\
&\beta_t^{time} \sim N(\alpha_t, \tau_t^2)
\end{aligned}$$

$$
\begin{aligned}
&\alpha_t \sim N(\alpha_{t-1} + \delta_t, \tau_\alpha^2)\\
&\delta_t \sim N(\delta_{t-1}, \tau_\delta^2)\\
&\tau_t, \tau_\alpha, \tau_\delta \sim t_7^+(0.1)\\
&\beta_a^{age} \sim N(\alpha_a, \tau_a^2)\\
&\alpha_a \sim N(\alpha_{a-1}, \tau_\alpha^2)\\
&\tau_a, \tau_\alpha, \tau_\delta \sim t_7^+(0.1)\\
&\beta_{at}^{age:time} \sim N(\alpha_{at}, \tau_{at}^2)\\
&\alpha_{at} \sim N(\alpha_{at-1}, \tau_\alpha^2)\\
&\tau_{at}, \tau_\alpha, \tau_\delta \sim t_7^+(0.1)
\end{aligned}
$$

Taking advantage of the fact that datasets can have different dimensions, both datasets including age (but less accurate) and without age (but more accurate) have been included. Data model for census population in this case cannot be used as census correspond to year 2011 which does not correspond to the year needed. This choice also aims to show results when census data are not available, and there are only two different datasets on population treated in the same way. The two population datasets both have the two needed dimensions and a Normal prior with a standard deviation of 2.5% is assumed on both. For births and deaths, a Normal prior is assumed on demographic balance data, with a standard deviation of 1% (same as (3.16) but with age dimension). Less detailed but more accurate datasets get a more informative prior reflecting information on accuracy despite the lack of age dimension. For dimensionally complete datasets but less accurate, data models are a Poisson distributions with informative prior on the intercept $\beta^0$ and on $\sigma$. Equation (3.25) shows the model only for births, the dataset on death counts has the same structure. No systematic bias is assumed, therefore no effect is included in the mean model. For international immigration and emigration only one dataset for each series is used as time and age dimensions are available and there is no difference with other less detailed datasets. Data models for migration have the same structure of equation (3.25), but higher standard deviation on both $\mu_{at}$ and $\sigma$ respectively 10% and 1% for both immigration and emigration.

$$
\begin{aligned}
&x_{at}^b \sim Pois(\gamma_{at}^b \omega_{at}^b)\\
&\log(\gamma_{at}^b) \sim N(\mu_{at}^b, \sigma_b^2)\\
&\mu_{at}^b \sim N(0, 0.025^2)\\
&\sigma_b \sim t_7^*(0.1)
\end{aligned}
\tag{3.25}
$$

Results on the demographic account are reported from figure 3.54 to figure 3.60. Population has two different graphs, one showing pictures of the population estimates at the 31$^{th}$ of December for years 2005, 2010 and 2015 by age groups (figure 3.54), whereas figure 3.55 shows estimation of the evolution by age groups during the ten years. Interpolation between years is made by a simple straight line in figure 3.54. Differences in credible intervals width, especially for age groups $10 - 14$, $15 - 19$ and $20 - 24$, mainly depend on irregularities within the age groups. As shown in figure 3.56, some age groups have regular evolution over time but others, especially the three groups mentioned, have quite irregular patterns, making estimation harder, and affecting credible intervals widths.

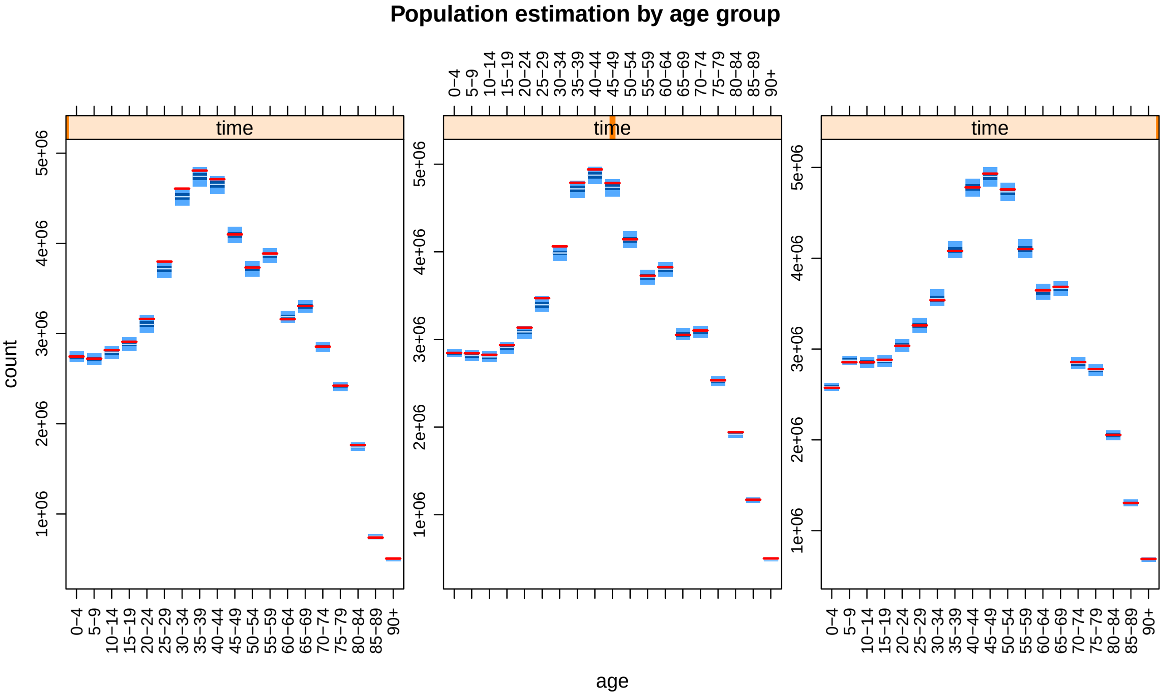


FIGURE 3.54: Population counts estimation by five years age groups in 2005 (left), 2010 (centre) and 2015 (right). The three blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from reconstructed population data.

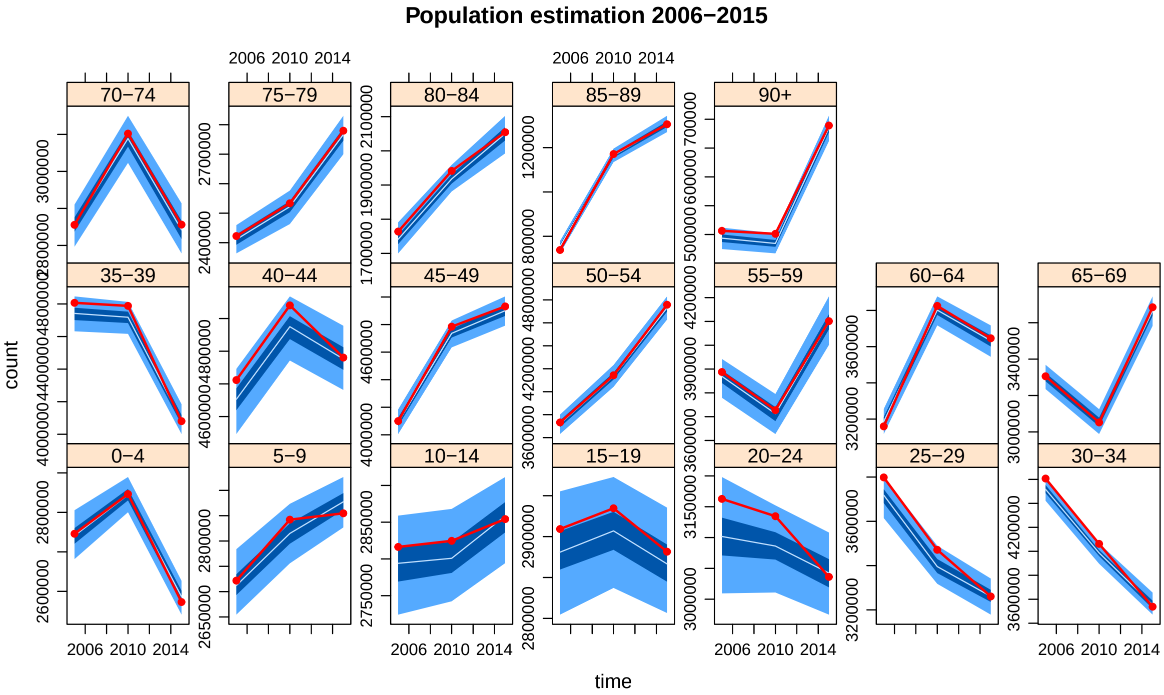


Figure 3.55: Population counts estimation evolution from 31/12/2005 until 31/12/2015 by five years age groups. The age blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from reconstructed population data. Interpolation between the estimated years (2005,2010 and 2015) is linear.

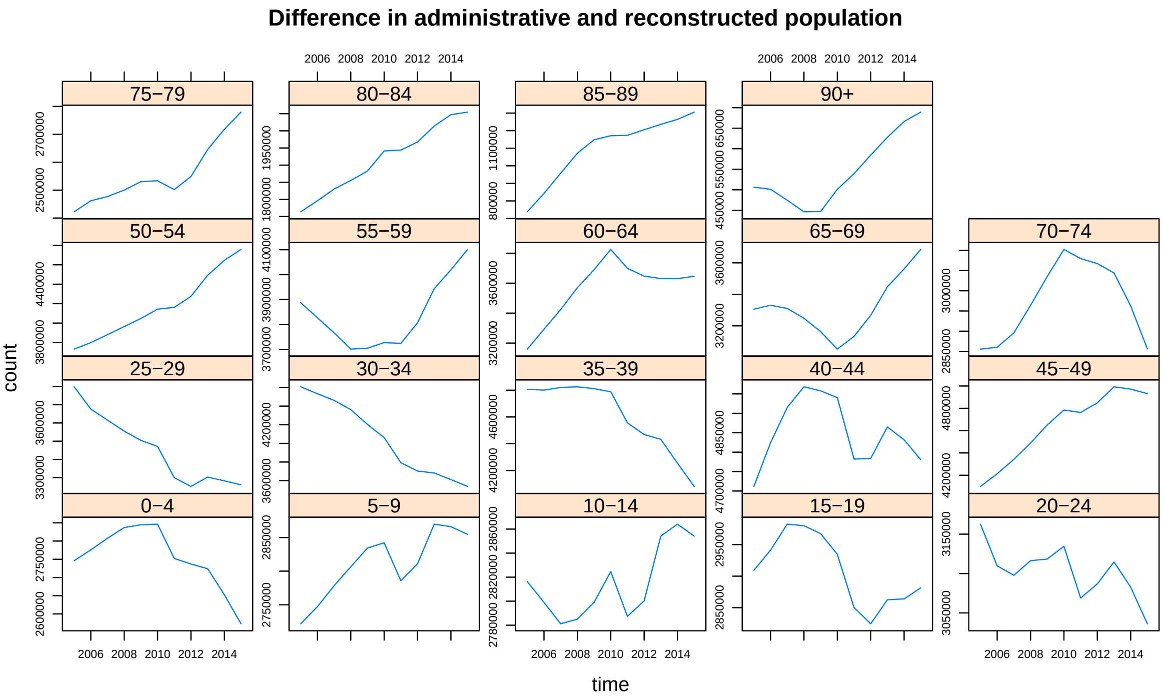


Figure 3.56: Resident population by age group for all years from 31/12/2005 until 31/12/2015. Yearly variations are shown here that are ignored in the model that only considers population at time points 31/12/2005, 31/12/2010 and 31/12/2015.

For births, deaths and migrations estimation, credible intervals are larger for more numerous age groups. Despite this could appear counter-intuitive, larger age groups are also those experiencing larger variations over time; a larger credible interval width can therefore be justified as estimation considers a five years period.

Birth and death counts are quite close to data with rather small credible intervals (figures 3.57 and 3.58), whereas migration have wider ones as it is also expected from the Poisson priors assumed.

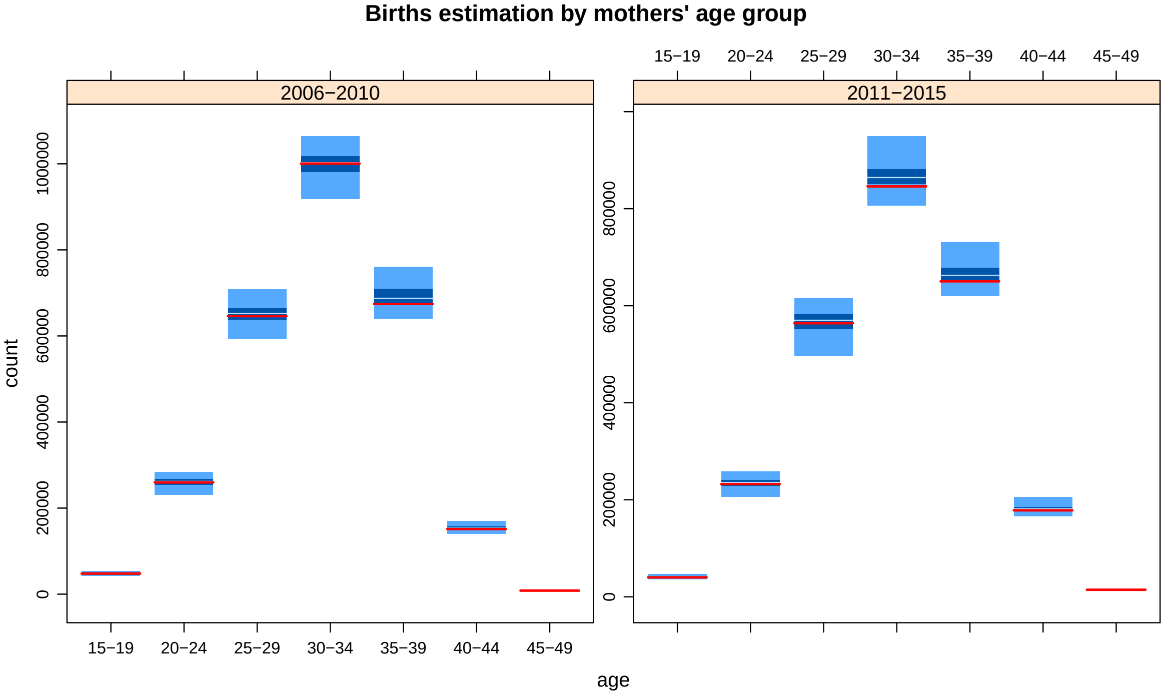


FIGURE 3.57: Births estimation by mothers' age group and time for periods 2005-2010 (left) and 2011-2015 (right). The period blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balance.

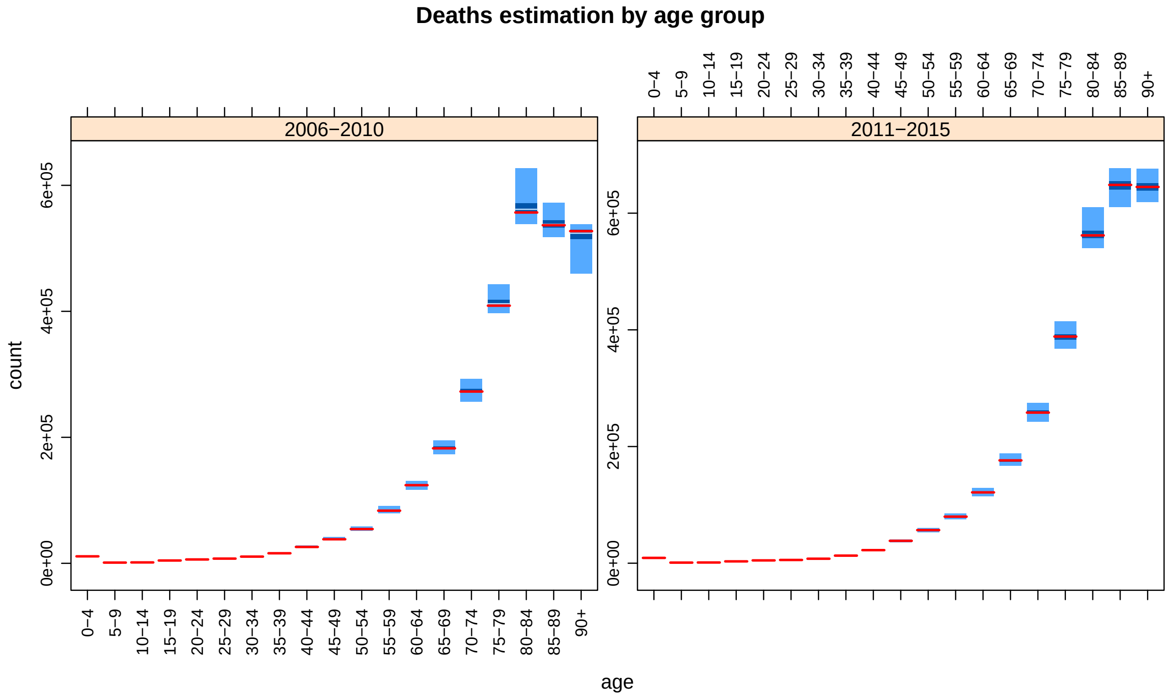


FIGURE 3.58: Deaths estimation by age group and time for periods 2005-2010 (left) and 2011-2015 (right). The period blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balance.

Estimation for international emigration has the widest credible intervals (figure 3.60. Overall emigration tends to be overestimated, whereas immigration (figure 3.59) is closer to data. This is what is expected as emigrants have usually very low interest in being cancelled from their former residency, whereas immigrants are usually interested in registering in their new country or municipality mainly for civil rights and practical reasons (Egidi and Ferruzza, 2009). Results by age and time for migration are much more regular, an more on line with expectations than in the other models (only with time dimension, section 3.5.1, and with region and time dimensions, section 3.5.2). This can be mostly explained by the importance of age dimension in the modelling of migration.

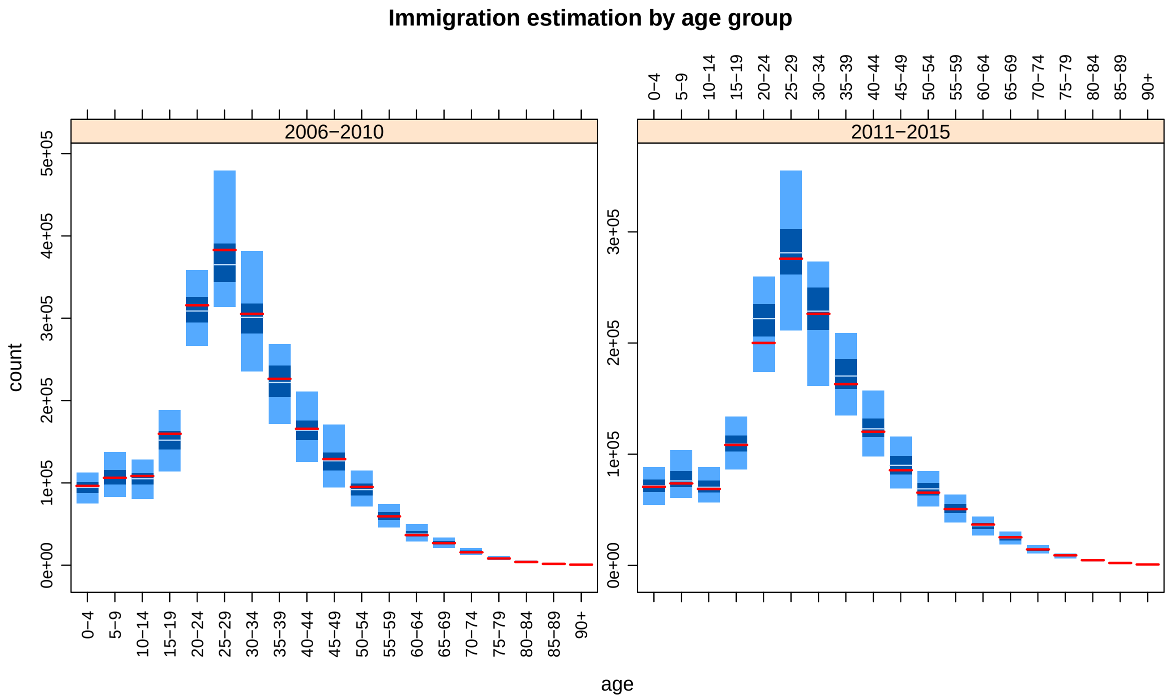


FIGURE 3.59: Immigration estimation by age group and time for periods 2005-2010 (left) and 2011-2015 (right). The period blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balance.

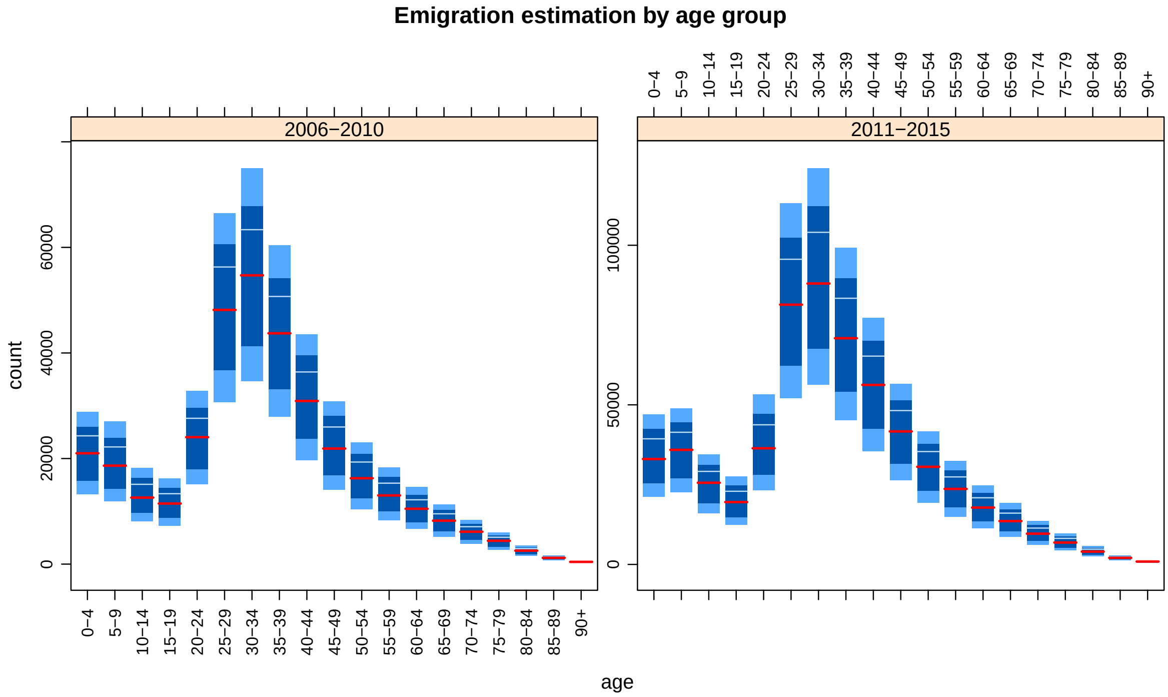


FIGURE 3.60: Emigration estimation by age group and time for periods 2005-2010 (left) and 2011-2015 (right). The period blocks show the medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from Istat demographic balance.

Comparing the demographic account estimation for age and time dimensions with the one considering region and time it is clear that region dimension is important in Italy and that strong regional differences exist. For some regions data and estimates are smooth, whereas there are regions for which data have high variability and credible intervals are much wider. Examples of regions with smooth results and data are for example Tuscany, Emilia Romagna, Umbria, Lombardia, Veneto, Trento and Bolzano. Regions with high variability are Calabria, Campania, Sardegna, Liguria and Friuli Venezia Giulia. The estimation process for regions could benefit from priors embedding expert opinions. To a certain extent this is true also for the model considering age and time and only time but problems are less obvious than with region dimension.

From all the applications the need of keeping uncertainty into account for migration processes arises. The estimation of migration processes has been problematic in all scenarios both considering different dimensions and considering different prior models. The effect is lower for the age and time model as age is an important dimension, internal migration is not considered and data are grouped in five years time spans creating some smoothing both for data and estimation. Even if lower in some cases, the uncertainty related to migration series estimation is shown in all models and proves the importance of providing uncertainty measure when dealing with demographic quantities (especially if not reliable) and the need of further information for a better estimation. Information needed can be characteristics of the migrants, but also the efficiency of the registration and cancellation process or the differences between countries or regions for definitions and administrative processes. The demographic account model allows for much flexibility and inclusion of prior information. In these paragraphs only a small part of what is possible has been shown, accordingly to the status of the R package `demest` and of the data available. The R package `demest` is available on the GitHub repository of Statistics New Zealand (`https://github.com/StatisticsNZ/demest`). The package still needs to be completed in some parts and experimented but many functions are already mature. These applications to the Italian data at the moment are the most advanced applications for the demographic account estimation and needs more work to be fully ready for external use. Functions for single series estimation as in sections 3.2 and 3.3 have been developed and tested for a longer time than those for the whole demographic account estimation. For this reason a deeper analysis has been possible for deaths and births series.

For better demographic account estimation there are many possible improvements as described in section 2.7. The next chapter presents an extension for available data model options introducing the Conway-Maxwell Poisson distribution (Conway and Maxwell, 1962).

# Chapter 4

# Model extension: Conway-Maxwell Poisson distribution

Options for data model distribution proposed by Bryant and Graham (2013) and reported in section 2.3.2 include the Poisson, the Normal and the Poisson-Binomial mixture distributions. The last two options are distributions typically concentrated around the mean, hence they represent a suitable choice for good quality data where expected variance is rather small. Even when high prior variance or low probability are assumed respectively for the Normal and the Poisson-Binomial mixture, the results are still closer to the initial data than when choosing the Poisson distribution. The Poisson prior for data model is the main choice for all the other data whose quality is not very high or is unknown. Unfortunately the Poisson distribution has limitations that do not always suit the population it refers to. As with $\beta$ parameters prior distributions, also the number of available data model has to be increased and improved in order to provide more suitable and customisable options for the data. One of the main limitations of the Poisson distribution is that mean and variance have the same value, implying data equi-dispersion. This is not always the case as population characteristics could require variance higher than the mean or lower, depending if population is more heterogeneous or homogeneous than in the equi-dispersion case, i.e. if it is over-dispersed or under-dispersed. Causes of homogeneity or heterogeneity depend on various aspects. It can be the propensity of an age group to be less traceable in its movements, e.g. students migration; or of a sex to register less in some registers or having different habits, e.g. men health care consumption is usually lower than for women (Friberg *et al.*, 2016; Wang *et al.*, 2013); for regions to typically be sending or receiving regions and therefore populations typically migrating or receiving migrants, e.g. for Italian internal migration the South of Italy is typically a sending region and the North is typically a receiving region. When differences are not considered in the model and hence population is considered as a whole, heterogeneity is high and for this reason variance could be higher than the mean. Conversely when population characteristics are defined in the model and information included through prior distribution, the population considered can be homogeneous and have lower variance than the mean. In both cases a Poisson distribution is not suitable for the model.

The alternative proposed in this chapter is the use of the Conway-Maxwell Poisson distribution (Conway and Maxwell, 1962) which, unlike the Poisson, allows for over- and under-dispersion of data with respect to the mean. The distribution is similar to the Poisson but with an additional parameter for modelling dispersion. With this distribution heterogeneous populations are expected to have dispersion parameters suggesting over-dispersion and, conversely, when more homogeneous populations are considered to suggest under-dispersion, i.e. a lower variance with respect to the mean.

## 4.1 Theory, properties and application examples

Poisson distribution is a common and suitable choice for modelling count data, it has many good properties and it is the basis for Poisson regression. A limitation, which is also one of its property, is the assumption of equi-dispersion of data with respect to the mean, as distribution mean and variance are equal: if $Y \sim Pois(\lambda)$, then $E[Y] = \lambda$ and $\text{Var}[Y] = E[Y] = \lambda$. This assumption works in many cases but, when the assumption of equi-dispersion does not hold, there might be problems with the estimation of Poisson parameters. A solution when data are over-dispersed is to use a Negative Binomial distribution, but it does not work if data are under-dispersed. Under-dispersion is less frequent than over-dispersion, but if data are homogeneous enough, under-dispersion is a possibility and it allowed to explain some phenomena, among others, in ecology, insurance, marketing, and spatio-temporal count data in general.

An extension of the Poisson distribution, accounting for both over- and under-dispersion, is the Conway-Maxwell Poisson (CMP) distribution. It is not the only one extension, among other solutions, the main ones are the weighted Poisson distributions of Del Castillo and Pérez-Casany (2005) and the generalised Poisson (GP) distribution of Consul (1989). Until its "revival" in 2005 (Shmueli *et al.*, 2005), the CMP distribution remained substantially unused after its first introduction in the 60s in Conway and Maxwell (1962), mainly because of computational problems due to its structure. The CMP probability mass function includes a normalisation constant creating problems for parameter estimation, but with new methods and technology it has been possible to reduce the impact of this issue. Equation (4.1) shows the original CMP parametrisation (Conway and Maxwell, 1962; Shmueli *et al.*, 2005). A more intuitive equation has been proposed in Guikema and Goffelt (2008). In order to show distribution properties the original parametrisation is used, but for Bryant and Graham (2013) model extension the latest one is considered.

Let $Y \sim CMP(\lambda, \nu)$, then the probability mass function is

$$P(Y = y) = \frac{1}{Z(\lambda, \nu)} \frac{\lambda^y}{(y!)^\nu} \tag{4.1}$$

where:

- $Y$ is a count variable, $y \in \mathbb{Z}^+ = \{0, 1, 2, ...\}$.
- $\lambda$, as in the Poisson case, can be any positive real number, $0 < \lambda \leq \infty$
- $\nu$, dispersion (or shape) parameter, can take any positive real number, but, according to its value, the properties of the distribution change:

$$\begin{cases} \nu = 0, \quad \lambda \geq 1 & \text{distribution is undefined} \\ \nu = 0, \quad \lambda < 1 & \text{geometric distribution} \\ \nu < 1 & \text{over-dispersion} \\ \nu = 1 & \text{reduction to a Poisson distribution} \\ \nu > 1 & \text{under-dispersion} \\ \nu = \infty & \text{reduction to a Bernoulli distribution} \end{cases}$$

- $Z(\lambda, \nu) = \sum_{i=0}^{\infty} \frac{\lambda^i}{(i!)^\nu}$, normalising constant, for $\lambda > 0$ and $\nu \geq 0$

The term $Z(\lambda, \nu)$ is a normalising constant depending on both parameters and, being an infinite sum, it needs to be either approximated or simplified. Approximation methods are used in the Maximum Likelihood Estimation (MLE) approach, whereas in the Bayesian approach simplification has been possible with the exchange algorithm (Moller *et al.*, 2006; Murray *et al.*, 2006). Approximations, upper bound of $Z(\lambda, \nu)$, and related quantities are provided in Minka *et al.* (2003). Due to the normalisation constant, moments, cumulants and quantiles can only be approximately calculated. Mean and variance are respectively

$$E[Y] \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \tag{4.2}$$

$$\text{Var}[Y] \approx \frac{1}{\nu} \lambda^{1/\nu} \tag{4.3}$$

where approximations are good for $\nu \leq 1$ and $\lambda > 10^{\nu}$.

The moment generating function (MGF) is

$$E[Y^{r+1}] = \begin{cases} \lambda E[Y+1]^{1-\nu}, & \text{If } r = 0, \\ \lambda \frac{d}{d\lambda} E[Y^r] + E[Y]E[Y^r], & \text{If } r > 0, \end{cases} \tag{4.4}$$

and the approximation for the $n$-th cumulant $\kappa_n$ for any $n \geq 1$ is

$$\kappa_n \approx \frac{1}{\nu^{n-1}} \lambda^{1/\nu} + O(1) \tag{4.5}$$

as $\lambda \to \infty$.

An asymptotic approximation of $Z(\cdot, \cdot)$ was finally found, after an initial hint from Shmueli *et al.* (2005), by Gillispie and Green (2015):

$$Z(\lambda, \nu) \sim \frac{\exp(\nu \lambda^{1/\nu})}{\lambda^{(\nu-1)/2\nu} (2\pi)^{(\nu-1)/2} \sqrt{\nu}} \big(1 + O\big(\lambda^{-1/\nu}\big)\big) \tag{4.6}$$

for $\nu$ fixed and for $\lambda \to \infty$. Combining this asymptotic result with the MGF in equation (4.4), it is possible to obtain another approximation for CMP expected value and variance, i.e. when, respectively, in equation (4.4) $r = 0$ and $r = 1$.

$$\begin{aligned} E[Y] &= \lambda \frac{d[\log(Z(\lambda, \nu))}{d\lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \\ \text{Var}[Y] &= \lambda \frac{dE[Y]}{d\lambda} \approx \frac{1}{\nu} \lambda^{1/\nu} + O(1) \end{aligned} \tag{4.7}$$

whereas the median approximation is $m \approx \lambda^{1/\nu} + O(\lambda^{1/2\nu})$, as $\lambda \to \infty$. These approximations too are good for $\nu \leq 1$ and $\lambda > 10^{\nu}$ as in equations (4.2).

Another property of the CMP distribution is that its probability function allows for a non-linear decrease in ratios of successive probabilities in the form

$$\frac{P(Y = y - 1)}{P(Y = y)} = \frac{x^{\nu}}{\lambda} \tag{4.8}$$

As well as the Poisson distribution, the CMP is also suitable for regression models with discrete variables and it belongs to the exponential and the two parameter power series families, inheriting all their properties. The likelihood in equation (4.9) is presented according the typical likelihood formulation for the exponential family: let $Y$ be a non-degenerate random variable with density $p_0(y)$, $s(Y)$ statistic, $K_S(\theta)$ cumulant generating function for $\theta \in \tilde{\Theta}_S$. Then $p(y; \theta) = \exp\{\theta s(y) - K_S(\theta)\} p_0(y)$, $y \in \mathcal{Y}$, $\theta \in \tilde{\Theta}_S$ is a density for every $\theta \in \tilde{\Theta}_S$. Or more generally, if parameter $\theta$ is expressed as a function of a parameter $\phi \in \Phi$, with $\theta(\Phi) \subset \tilde{\Theta}_S$, then $p(y;\phi) = \exp\{\theta(\phi)s(y) - G(\phi)\}h_0$.

Let $y_1, \cdots, y_n$ be a set of $n$ independent identically CMP distributed observations, then the likelihood is

$$\begin{aligned} L(y_1, \cdots, y_n|\lambda, \nu) &= \frac{\prod_{i=1}^{n} \lambda^{y_i}}{\left(\prod_{i=1}^{n} y_i!\right)^{\nu}} Z^{-n}(\lambda, \nu) \\ &= \underbrace{\lambda^{\sum_i = 1^n y_i}}_{= h_0 \text{ (first part)}} \exp\Big(\underbrace{-\nu \sum_{i=1}^{n} \log(y_i!)\Big)}_{= G(\cdot)} \underbrace{Z^{-n}(\lambda, \nu)}_{= h_0 \text{ (second part)}} \end{aligned} \tag{4.9}$$

this likelihood form shows that the CMP belongs to the exponential family and has sufficient statistics $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} \log(y_i!)$.

There are three main methods to perform CMP parameter estimation, explained in Sellers *et al.* (2011). These methods are

- *Weighted least square method.* It is an easy method, it fits the regression of the logarithm of equation (4.8) on $\log(y)$ considering model heteroscedasticity and first-order dependence. The method performs well when there are not many zero counts.

- *Maximum likelihood estimation.* At the point of maximum likelihood, parameters $\lambda$ and $\nu$ satisfy $E[Y] = \lambda\bar{Y}$ and $E[\log(Y!)] = \overline{\log(Y!)}$ where each of these equations is an infinite sum of form $E[f(Y)] = \sum_{j=0}^{\infty} f(j)\lambda_j/(j!)^{\nu} Z(\lambda, \nu)$. As it is not possible to compute them analytically, a Newton-Raphson method is usually used. This method is accurate but computationally intensive.

- *Bayesian estimation.* Shmueli *et al.* (2005) define this method as “immediate and simple” because, belonging to the exponential family, the CMP has a conjugate prior distribution. Kadane *et al.* (2006) provide an analysis of the conjugate prior. Let the joint conjugate prior density for $\lambda$ and $\nu$ have form:

  $$h(\lambda, \nu) = \lambda^{a-1} \exp(-\nu b) Z^{-c}(\lambda, \nu) \kappa(a, b, c) \tag{4.10}$$

  for $\lambda > 0$ and $\nu \geq 0$, where $\kappa(a, b, c)$ is the integration constant, then the posterior on $\lambda$ and $\nu$ is of the same form of (4.10) with parameters $a' = a + \sum_{i=1}^{n} Y_i$, $b' = b + \sum_{i=1}^{n} \log(Y_i!)$ and $c' = c + n$. Prior in (4.10) “can be thought of as an extended bivariate Gamma distribution” (Kadane *et al.*, 2006). Its density is

> proper only if values of $a$, $b$ and $c$ lead to a finite $\kappa^{-1}(a,b,c)$ where
>
> $$\kappa^{-1}(a,b,c) = \int_0^\infty \int_0^\infty \lambda^{a-1} e^{-b\nu} Z^{-c}(\lambda,\nu) d\lambda d\nu \tag{4.11}$$
>
> Despite the conjugate model simplifies the estimation, practitioner are usually not familiar with this distribution, parameters $a$, $b$ and $c$ are difficult to tune, and their meaning is not very clear. Kadane *et al.* (2006) made efforts to elicit them, but the conjugate prior for the CMP distribution has not been much used until now. In practice, Bayesian applications of CMP regression model do not use the conjugate prior, but prefer assuming a Lognormal prior distribution on both $\lambda$ and $\nu$. The Bayesian framework is also the only one allowing for estimation without approximating the normalising constant $Z(\lambda,\nu)$.

If with Poisson random variables $Y \sim Pois(\lambda)$ the parameter *lambda* is a location parameter immediately available, with the CMP, $Y \sim CMP(\lambda,\nu)$, parameter $\lambda$ does not provide the same understanding of the distribution. A more intuitive re-parametrisation is found in (Guikema and Goffelt, 2008). They substitute parameter $\lambda$ in the original CMP probability mass function (equation (4.1)) with parameter $\mu = \lambda^{1/\nu}$, $0 < \mu \leq \infty$. In this version parameter $\mu$ has a more intuitive meaning than $\lambda$ in the first parametrisation because, as in the Poisson distribution, $\mu$ is a location parameter. Let $Y \sim CMP(\mu,\nu)$ then the probability mass function is

$$\begin{aligned} P(Y=y) &= \frac{1}{S(\mu,\nu)}\left(\frac{\mu^y}{y!}\right)^\nu, \qquad && y \in \mathbb{Z}^+ = \{0,1,2,...\} \\ S(\mu,\nu) &= \sum_{i=0}^{\infty}\left(\frac{\mu^i}{i!}\right)^\nu, \qquad && \text{for } \mu > 0 \text{ and } \nu \geq 0 \\ E[Y] &\approx \mu + \frac{1}{2\nu} - \frac{1}{2}; \qquad && \text{Var}[Y] \approx \frac{\mu}{\nu} \end{aligned} \tag{4.12}$$

Figure 4.1 compares the histograms of three samples of 1000 i.i.d. CMP distributed random variables with $\mu = 10$ and dispersion parameter $\nu$ respectively equal to 0.5 (over-dispersion), 1 (equi-dispersion) and 5 (under-dispersion). A Poisson density with mean equal $\mu$ is overlapped in red.

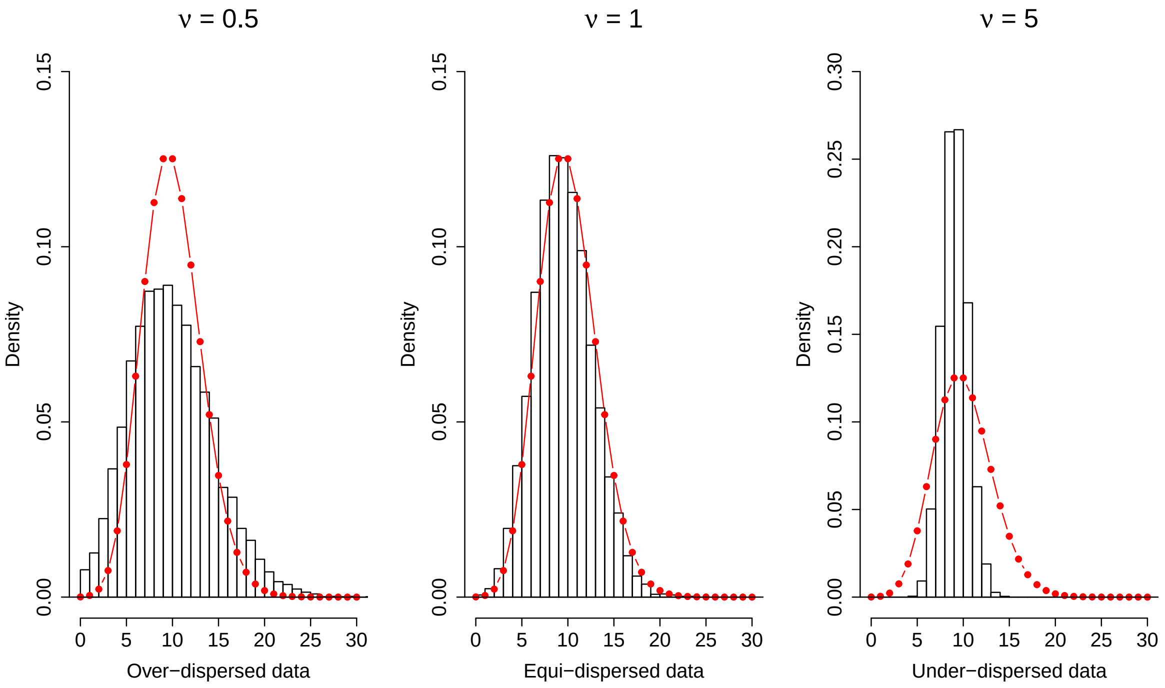


Figure 4.1: Histograms of a thousand CMP random variable samples compared with corresponding Poisson densities. *Left*: Over-dispersed CMP data. $Y \sim CMP(10, 0.5)$. *Centre*: Equi-dispersed CMP data. $Y \sim CMP(10, 1)$. *Right*: Under-dispersed CMP data. $Y \sim CMP(10, 5)$.

Parameter $\nu$ controls the shape of the distribution, there is over-dispersion if $\nu < 1$, i.e, variance is higher than mean, on the contrary, under-dispersion implies $\nu > 1$, and variance is lower than mean. Parameter $\lambda$ in the original version is immediately related to the parameter of the Poisson distribution, but, in the CMP case, it is less interpretable and does not provide a clear centring parameter. Parameter $\mu$ in the second version is a better option, especially when the aim is to perform a CMP GLM on the data. The ceiling of $\mu$ ($\lceil \mu \rceil$) is the mode of the distribution and it provides, along with the dispersion parameter, a good approximation for the mean (equation (4.12)). Approximations of both mean and variance, as shown in the third line of equation (4.12), are good especially when $\mu \geq 10$. In equation (4.1), values are more accurate when $\nu \leq 1$ and $\lambda^{1/\nu} \geq 10$. From now on the text refers to the second parametrisation (Guikema and Goffelt, 2008).

As with the Poisson distribution, also with the CMP it is possible to implement a regression model (CMP regression). In most of the applications (Chanialidis *et al.*, 2014; Huang, 2017; Sellers and Shmueli, 2010), the chosen link function is the logarithmic one for both $\mu$ and $\nu$. As discussed in Sellers *et al.* (2011), the dispersion parameter can be unique for all the data (*constant dispersion*), differ at group-level (*group-level* *dispersion*), or change at each observation (*observation-level dispersion*). The same is true also for parameter $\mu$. Sellers *et al.* (2011) also mention the implementation of a CMP cure-rate model, and a CMP model with censoring.

CMP regression has been used in several works. Applications to marketing, linguistics, biology (Sellers *et al.*, 2011), transportation (Lord *et al.*, 2008), healthcare (Cancho *et al.*, 2012) and also population studies (both in demography (Handock *et al.*, 2014) and ecology (Wu *et al.*, 2013; Lynch and Brown, 2010)) are few examples. Particularly innovating Bayesian implementations of CMP regression are Chanialidis *et al.* (2017) and Benson and Friel (2017). In addition to CMP regressions, in both cases authors introduce a new method to sample from a CMP distribution with exact algorithms. The complication due to the normalising constant also affects the distribution sampling algorithm. Chanialidis *et al.* (2017) construct a rejection sampler using a piecewise geometric distribution which has a closed form normalisation constant, whereas Benson and Friel (2017) propose a "more efficient and less computationally intensive method" (Benson and Friel, 2017), using a single envelope distribution depending on the dispersion parameter. A description of the Benson and Friel (2017) algorithm is presented in section 4.1.1 where modifications introduced to add the CMP distribution to the DAM are also explained. Once it becomes possible to sample from a CMP, a MCMC algorithm avoiding the computation of the normalising constant can also be implemented. This algorithm is the *exchange algorithm* and it was first proposed in Moller *et al.* (2006) and Murray *et al.* (2006). Unlike standard MCMC algorithms, the exchange algorithm allows generating from *doubly-intractable* posterior distributions, i.e. where there are intractable normalising constants to their dependence on the parameter(s) of interest.

The idea of the exchange algorithm is to simplify the normalisation constant in the Metropolis-Hastings ratio by enlarging the state of the Markov chain with auxiliary variables drawn from a suitable proposal distribution. Formally, let a random variable $Y$ be distributed as a double intractable distribution of density $f(\cdot)$ and parameter $\gamma$, where $\gamma$ can be either a single parameter or a vector. Let the posterior distribution be denoted by $pi(\cdot)$ so that $\pi(\gamma|y) = \mathcal{L}(y|\gamma)\pi(\gamma)$ where $\mathcal{L}(y|\gamma)$ is the likelihood and $\pi(\gamma)$ the prior on $\gamma$. The novelty of the exchange algorithm is in the posterior construction. Instead of a standard posterior, an *augmented* posterior distribution is calculated. The augmented posterior includes auxiliary draws ($y^*$) in addition to the data. Let $\gamma^*$ be the candidate parameter generated by the proposal distribution $h(\gamma^*|\gamma^i)$, and let $y^*$ be draws from the sampling model $\pi(y^*|\gamma^*)$, then the augmented posterior for $\gamma^*$ has the form

$$\pi(\gamma^*|y, y^*, \gamma^i) \propto \mathcal{L}(y|\gamma^*)\mathcal{L}(y^*|\gamma^i)\pi(\gamma^*) \tag{4.13}$$

where $\gamma^i$ is the current value of the parameter at iteration $i = 1, ..., N$, $y = y_1, ..., y_n$ is the data vector, and $y^* = y_1^*, ..., y_n^*$ is the vector of auxiliary draws of same length as the data vector $y$. Therefore, the augmented posterior considers two likelihoods: one with the original data and one with auxiliary draws. The augmented posterior for current parameter $\gamma^i$ is

$$\pi(\gamma^i|y, y^*, \gamma^*) \propto \mathcal{L}(y|\gamma^i)\mathcal{L}(y^*|\gamma^*)\pi(\gamma^i)$$

. The exchange algorithm also assumes that the probability density function $f(\cdot)$ of $Y$ can be factorised as

$$f(y|\gamma) = q(y|\gamma)(S(\gamma))^{-1} \tag{4.14}$$

with $S(\gamma)$ intractable normalising constant, which is the case of the CMP distribution. Starting from a Metropolis-Hastings acceptance ratio with augmented posterior distribution, equation (4.15) shows how it becomes possible to simplify the normalising constant.

$$\begin{aligned} a &= \frac{\pi(\gamma^*|y, y^*, \gamma)h(\gamma|\gamma^*)}{\pi(\gamma|y, y^*, \gamma^*)h(\gamma^*|\gamma)} \qquad \text{expliciting } \pi() \text{ by eq. (4.13)} \\ &= \frac{\mathcal{L}(y|\gamma^*)\mathcal{L}(y^*|\gamma)\pi(\gamma^*)h(\gamma|\gamma^*)}{\mathcal{L}(y|\gamma)\mathcal{L}(y^*|\gamma^*)\pi(\gamma)h(\gamma^*|\gamma)} \qquad \text{expliciting likelihoods} \\ &= \frac{\prod_{i=1}^n f(y_i|\gamma^*)f(y_i^*|\gamma)\pi(\gamma^*)h(\gamma|\gamma^*)}{\prod_{i=1}^n f(y_i|\gamma)f(y_i^*|\gamma^*)\pi(\gamma)h(\gamma^*|\gamma)} \end{aligned} \tag{4.15}$$

replacing $f(\cdot)$ by equation (4.14)

$$\begin{aligned} &= \frac{\prod_{i=1}^n \frac{q(y_i|\gamma^*)}{\cancel{(S(\gamma^*))}} \frac{q(y_i^*|\gamma)}{\cancel{(S(\gamma))}} \pi(\gamma^*)h(\gamma|\gamma^*)}{\prod_{i=1}^n \frac{q(y_i|\gamma)}{\cancel{(S(\gamma))}} \frac{q(y_i^*|\gamma^*)}{\cancel{(S(\gamma^*))}} \pi(\gamma)h(\gamma^*|\gamma)} \\ &= \frac{\prod_{i=1}^n q(y_i|\gamma^*)q(y_i^*|\gamma)\pi(\gamma^*)\cancel{h(\gamma|\gamma^*)}^{\nearrow \text{If symmetric}}}{\prod_{i=1}^n q(y_i|\gamma)q(y_i^*|\gamma^*)\pi(\gamma)\cancel{h(\gamma^*|\gamma)}} \\ &= \frac{\prod_{i=1}^n q(y_i|\gamma^*)q(y_i^*|\gamma)\pi(\gamma^*)}{\prod_{i=1}^n q(y_i|\gamma)q(y_i^*|\gamma^*)\pi(\gamma)} \end{aligned} \tag{4.16}$$

Using the augmented posterior the normalising constants cancel out, and therefore, a usual acceptance step can be performed. This solution is among those that Robert (2015) call pseudo-marginals solutions. They “may prove difficult to calibrate” and performances“ depend on the quality of the estimators and are always poorer than when using the exact target” but, in both Chanialidis *et al.* (2017) and Benson and Friel (2017), the Authors use this algorithm with apparently encouraging results.

The CMP distribution has advantages and disadvantages; among the advantages there are flexibility and parsimony, it belongs to the exponential family, it generalises some important discrete probability distribution, and it can be used to perform regression models as in the Poisson and logistic cases, but with dispersion parameter adaptable to the required level of specificity. Disadvantages are the absence of an easy closed form, difficulties to simulate from the distribution, difficult regression coefficients interpretation, and computational issues.

Due to the similarities with the Poisson distribution and the high flexibility, the CMP distribution seemed an attractive model for population data that, depending on the aggregation level, could be homogeneous or heterogeneous, i.e. under- or over-dispersed, and, therefore, not suitable for a Poisson distribution.

### 4.1.1 Rejection sampling implementation

To include the CMP distribution in the DAM within a Bayesian framework the approach of Chanialidis *et al.* (2017) using the exchange algorithm has been implemented, but replacing their rejection sampling method with the one in Benson and Friel (2017). Existing `R` functions for sampling from a CMP have been also considered but, when dealing with a national population, it is often necessary to sample with location parameter of the order of thousands or millions and, in many cases, existing functions were not able or too slow to deal with the Italian population. Therefore, a rejection sampling algorithm has been implemented starting from quantities explained in Benson and Friel (2017), but considering their logarithmic versions. This modification speeds up computations and allows dealing with large numbers.

For their rejection sampler, Benson and Friel (2017) use two envelope distributions and, depending on the value of the dispersion parameter, either one of them is used. Let $f(\cdot|\theta)$ be the target distribution (CMP$(\mu, \nu)$) with $\theta = (\mu, \nu)$, and $g(\cdot|\gamma)$ be the envelope distribution where

$$g(\cdot|\gamma) = \begin{cases} g_1(y|\gamma = p) = p(1-p)^y, & \text{if } \nu < 1 \text{ (, Geometric envelope)} \\ g_2(y|\gamma = \mu) = \frac{\mu^y}{e^\mu y!}, & \text{if } \nu \geq 1 \text{ (, Poisson envelope)} \end{cases} \tag{4.17}$$

An essential quantity in the algorithm, is the tractable bounding constant $B^{\nu}_{f/g}$ which can be interpreted as the "upper bound on the ratio of the unnormalised densities $q_f(\cdot|\theta)$, and $q_f(\cdot|\gamma)$" (Benson and Friel, 2017), where the unnormalised densities are densities showed in equation (4.14). As for the envelope distributions, there are two tractable bounding constants, one in case of over-dispersion ($\nu < 1$) and one for equi- or under-dispersion ($\nu \geq 1$):

- Logarithmic enveloping bounds for over-dispersion, $B_{f/g}^{\nu<1}$

$$
\begin{aligned}
B_{f/g}^{\nu<1} &= \frac{1}{p}\frac{\mu^{\nu\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}}{(1-p)^{\nu\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}}\left(\frac{\mu^{\lfloor\mu\rfloor}}{\lfloor\mu\rfloor!}\right)^{\nu-1} \qquad (4.18)\\
\log(B_{f/g}^{\nu\geq 1}) = b_{f/g}^{\nu\geq 1} &= -\log(p) + \nu\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\log(\mu)\\
&\quad - \left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\log(1-p) - \nu\sum_{n=1}^{\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}\log(n)
\end{aligned}
$$

- Logarithmic enveloping bounds for equi- and under-dispersion, $B_{f/g}^{\nu\geq 1}$

$$
\begin{aligned}
B_{f/g}^{\nu\geq 1} &= \left(\frac{\mu^{\lfloor\mu\rfloor}}{\lfloor\mu\rfloor!}\right)^{\nu-1} \qquad (4.19)\\
\log(B_{f/g}^{\nu\geq 1}) = b_{f/g}^{\nu\geq 1} &= (\nu-1)\Big(\log(\mu^{\lfloor\mu\rfloor}) - \log(\lfloor\mu\rfloor!)\Big)\\
&= (\nu-1)\Big(\lfloor\mu\rfloor\log(\mu) - \sum_{n=1}^{\lfloor\mu\rfloor}\log(n)\Big)
\end{aligned}
$$

It is now possible to consider the acceptance ratio in its logarithmic form. The initial formulas for $\alpha^{\nu<1}$ and $\alpha^{\nu\geq 1}$ directly come from Benson and Friel (2017), and passages are reported in Appendix.

- Logarithmic acceptance ratios, $\nu < \mathbf{1}$:

$$
\begin{aligned}
\alpha^{\nu<1} &= \frac{\left(\mu^{y'}/y'!\right)^{\nu}}{B_{f/g}^{\nu<1}(1-p)^{y'}p} \qquad (4.20)\\
\log(\alpha^{\nu<1}) &= \nu\log\left(\mu^{y'}/y'!\right) - \Big(\log(B_{f/g}^{\nu<1}) + y'\log(1-p) + \log(p)\Big)\\
&= \left(y' - \left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\right)\Big(\nu\log(\mu) - \log(1-p)\Big)\\
&\quad + \nu\left(\sum_{n=1}^{\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}\log(n) - \sum_{m=1}^{y'}\log(m)\right)
\end{aligned}
$$

where $y'$ is drawn from a Geometric distribution ($y' \sim Geo(p)$), and the optimal value for $p$ is $p = \frac{2\nu}{2\mu\nu+1+\nu}$ (Benson and Friel, 2017).

- Logarithmic acceptance ratios, $\nu \geq \mathbf{1}$:

$$\begin{aligned} \alpha^{\nu\geq 1} &= \frac{\big(\mu^{y'}/y'!\big)^{\nu}}{B^{\nu\geq 1}_{f/g}\big(\mu^{y'}/y'!\big)} \\ \log(\alpha^{\nu\geq 1}) &= \nu \log\big(\mu^{y'}/y'!\big) - \Big( \log(B^{\nu\geq 1}_{f/g}) + \log\big(\mu^{y'}/y'!\big)\Big) \\ &= (\nu - 1)\Big( \log(\mu)\big(y' - \lfloor\mu\rfloor\big) - \sum_{m=1}^{y'} \log(m) + \sum_{n=1}^{\lfloor\mu\rfloor} \log(n)\Big) \end{aligned} \tag{4.21}$$

  where $y'$ is drawn from a Poisson distribution ($y' \sim Po(\mu)$).

Once calculated the acceptance ratio, the algorithm works as other Metropolis-Hastings algorithms: a random variable is generated $u \sim U(0, 1)$ and, if $u \leq \alpha$, the sampled random variable $y'$ is accepted, otherwise the procedure is repeated with a new $y'$ until the value is accepted. In order to avoid infinite loops a maximum value of attempts is set.

The function for sampling CMP random variables has been compared with other functions available. It gives good results for small values of $\mu$ and, unlike the other `R` functions, it is also efficient with higher values. A comparison between the modified Benson and Friel (2017) sampler (implemented in a GitHub package called `demCMP` in addition to `demest` package), `CRAN R Packages: COMPoissonReg` and `compoisson`, and Chanialidis *et al.* (2017) package available on GitHub `combayes` is presented in figures 4.2 and 4.3. All of them show histograms of 1000 CMP random variables samples with different packages. When possible, all four packages results are shown. The bottom row of 4.2 only shows results for `combayes` and `demCMP` as the two others could not simulate with $\lambda = 1,000$ and $\lambda = 10,000$. In these last two cases `demCMP` was slower than `combayes`, but when it comes to under-dispersion, `demCMP` out-performs `combayes`, see figures 4.3. For values up to $\lambda = 1,000$ all packages work, but for higher values `demCMP` definitely provides the best samples, despite being much slower than its counterpart in `combayes`. As demographic values in the demographic account are large, the rejection sampling algorithm in the `demCMP` package has been preferred. It is worth noting that parameter $\mu$ depends on both $\lambda$ and $\nu$, therefore, for different $\lambda$ and $\nu$ combinations, the maximum values of $\lambda$ the algorithms can tolerate changes accordingly but the plots presented provide an idea of the advantages of the chosen algorithm.

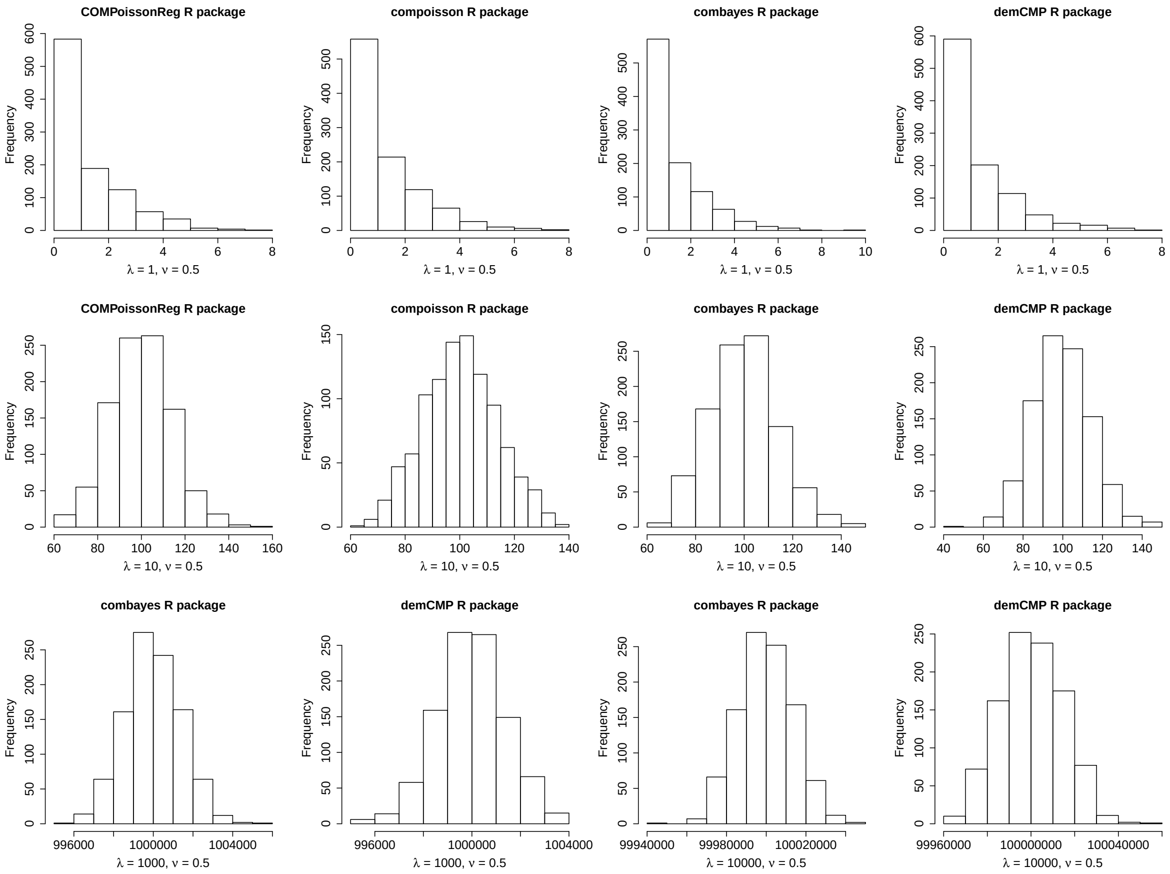


FIGURE 4.2: Histograms of over-dispersed ($\nu = 0.5$) CMP draws from four different R package (column from left to right: COMPoissonReg, compoisson, combayes and demCMP, except the last row where only combayes and demCMP worked) with $\lambda = 1$ (first row), $\lambda = 10$ (second row), $\lambda = 1,000$ and $\lambda = 10,000$ (third row).

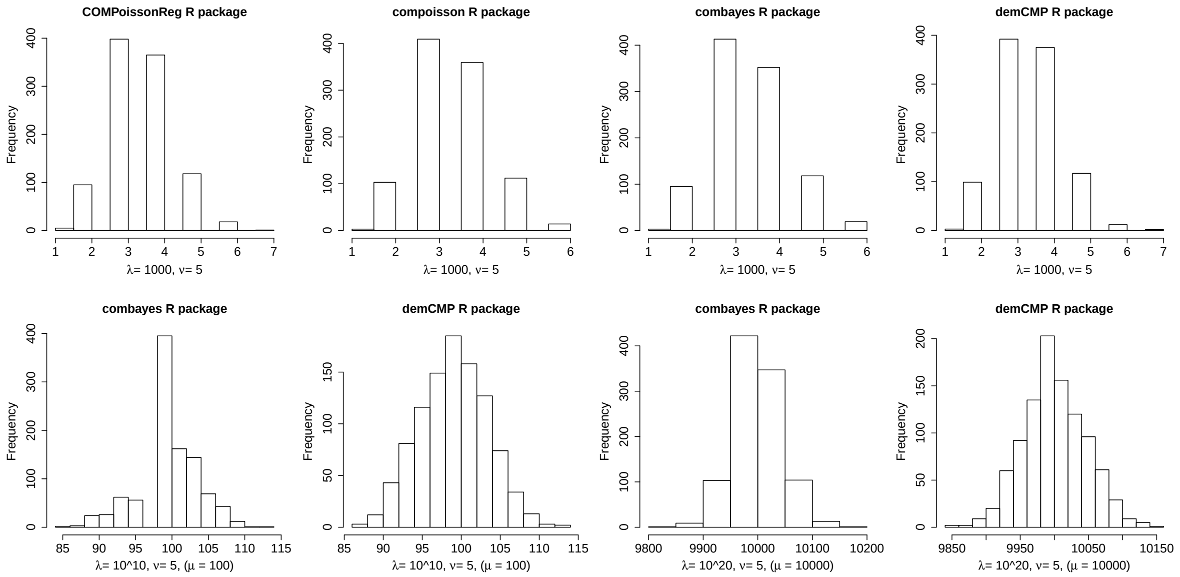


FIGURE 4.3: Histograms of under-dispersed ($\nu = 5$) CMP draws from four different R package (column from left to right: COMPoissonReg, compoisson, combayes and demCMP, second row only combayes and demCMP worked) with $\lambda = 1000$ (first row), $\lambda = 10^{10}$ and $\lambda = 10^{20}$ (second row).

## 4.2 Integration in the Demographic Account model

Once solved the normalising constant computational issues, including the CMP distribution among the data model choices is quite straightforward as the model is very similar to the Poisson case except for dispersion parameter $\nu$. Data model with the CMP distribution is displayed in equation (4.23) without exposure term, which can always be included. Notation now gets back to the one used in other chapters. For CMP introduction the notation usually adopted in other articles is used to ease comparability.

$$\begin{aligned} x_{jm} &\sim CMP(\gamma_{jm}, \nu_{jm}) \\ \log(\gamma_{jm}) &\sim N(\mu_{jm}, \sigma^2_\gamma) \\ \mu_{jm} &= \textstyle\sum_{k=0}^{K} \beta^{(k)} \\ \sigma_\gamma &\sim t^+_{df_{\sigma_\gamma}}(A^2) \\ \log(\nu_{jm}) &\sim N(\eta, \sigma^2_\nu) \end{aligned} \tag{4.22}$$

where $\gamma$ is the location parameter (as $\mu$ in the Guikema and Goffelt (2008) parameterisation) and $\nu$ dispersion parameter. As in the Poisson case the parameter $\gamma$ is assumed to be Lognormal distributed, the same holds for both $\gamma$ and $\nu$ with the CMP.

For parameter $\gamma$ the model follows the structure of chapter 3 with the usual regression structure. For $\nu$, a simpler approach is adopted, with just a simple Normal prior with hyper-parameters $\eta$ and $\sigma_\nu$ for all cells. So, there is a $\nu_{jm}$ for each cell of the demographic series, but they all have same mean and variance. This choice is possible as there is no reason to expect much difference in the dispersion parameters of each series, and a simpler structure allows not to further increase the number of parameters, which is already high, and hence to further complicate the model. The posterior calculation for $\nu$ is straightforward as there is only the CMP-Lognormal structure and $\eta$ and $\sigma_\nu$ are set *a priori*. Usually the prior on the mean is $\eta = 0$, i.e. $\nu = 1$, meaning that the initial assumption is a Poisson model (equi-dispersion), then, during the estimation, over- or under-dispersion can be identified. For $\sigma_\nu$ a small value is recommendable, at least within this framework, to help convergence and for simple numeric reasons. For examples, in the following applications a prior $\sigma_\nu = 0.5$ is chosen so that values of $\log(\nu)$ hardly go beyond the interval $[-2, 2]$, i.e. $0.13 \leq \nu < 7.4$ which is already a high level of over- and under-dispersion. Especially for under-dispersion, as shown in figure 4.1, a value of $\nu = 5$ is already high, despite $\nu$ being theoretically any positive real number ($0 \leq \nu < \infty$). Therefore, in practice, there is no need to let $\nu$ assume very large values.

In an earlier version of the model, each $\nu_{jm}$ was assumed to have its prior mean and variance, but as results were not satisfying, and all values were very similar, a simpler model with unique mean and variance has been implemented. Dispersion parameter estimation still needs further developments and testing. Developments could be either towards a further simplification with a unique value for all cells, or, possibly, towards a more complicated model assuming a regression as with parameter $\gamma$. The first option is likely to be implemented first as it is simpler and less complicated data models have been performing better so far. In any case, CMP regression applied to demographic series estimation needs to be further investigated.

## 4.3 Applications

Two applications to Italian death and birth counts follow with encouraging results but, at the moment, they are only applications to single series, as applications to the whole demographic account provides unlikely results suggesting the CMP model is not stable enough yet. Examples can then be considered as an extension of sections 3.2 and 3.3.

### 4.3.1 CMP data model for death counts

For death count series a comparison between a model including only time dimension and another including time, age and sex is performed. The model with time has a simple system model, only with a DLM prior on time effect with scale parameters of 0.1 for standard deviations, an informative prior on the intercept ($\beta^0 \sim N(0, 0.025^2)$), and a low scale parameter on variance $\sigma \sim t^+_{df_{\sigma_\gamma}}(0, 0.1^2)$. Two datasets are included as in section 3.2, a Poisson-Binomial with $p = 0.98$ is assumed on the data coming from the demographic balance dataset, and a CMP on the death causes dataset. No effect is included in the CMP data model and dispersion parameter prior is $\log(\nu) \sim N(0, 0.5)$, as previously suggested. The more complicated model includes effects on time, age, sex, and an age-sex interaction. The system model is similar to the previous one plus the prior on the additional effects. Time and age have the same DLM prior as in the model with only time, whereas age-sex interaction has a DLM prior with weaker standard deviation priors (scale parameter is 1), and sex has a fixed Normal prior $\beta^{sex} \sim N(0, 1)$. The data models are the same as in the previous model to allow for comparability. Results on count estimation are shown respectively in figures 4.4 and 4.5. For the model with only time, estimated counts are much higher than data, whereas when sex and age are included estimates fit data much better. The most important point is the estimation of dispersion parameters shown in tables 4.1 and 4.2. Comparing $\nu$s values for both tables a weak under-dispersion is present in both models but in the second values are usually higher with a mean for time model of $\bar{\nu}^t = 1.07$ and $\bar{\nu}^{ast} = 1.10$ for the second one. When more dimensions are included in the demographic series, cells have lower values and the population is likely more homogeneous. The difference is not very high but it still corresponds to expectations. Probably there are other dimensions not included in the model that could increase even more the homogeneity than age and sex.

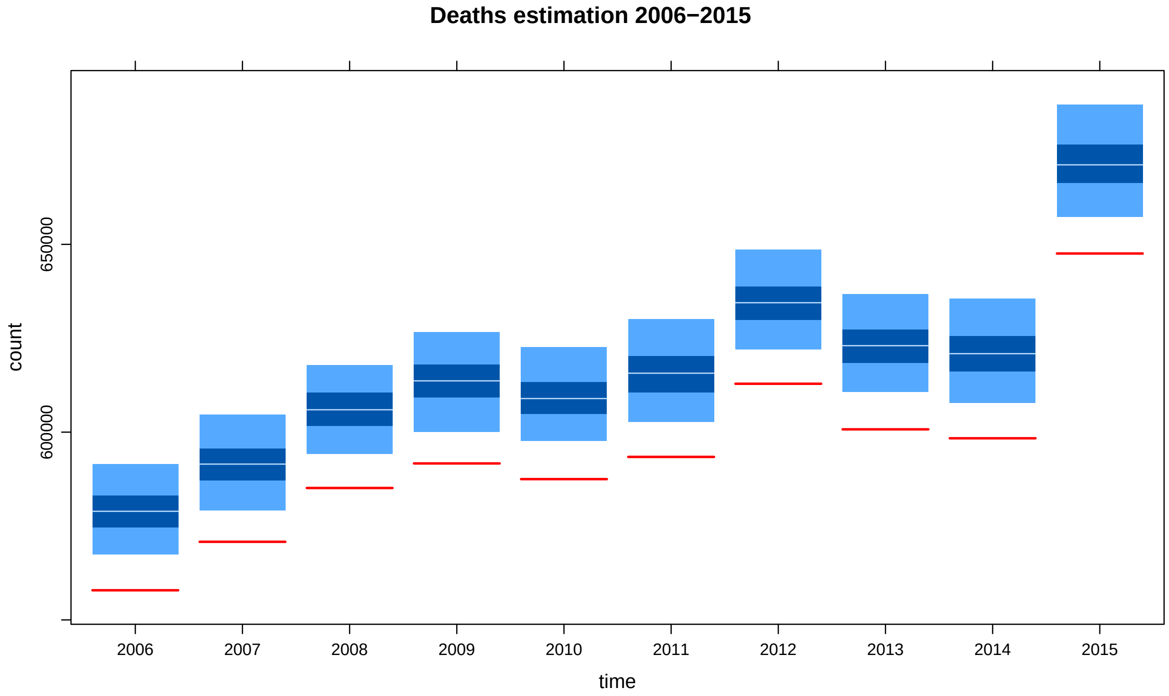


FIGURE 4.4: Death counts estimation for CMP model with only time dimension. The medians are the white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from *Annuario statistico italiano*.

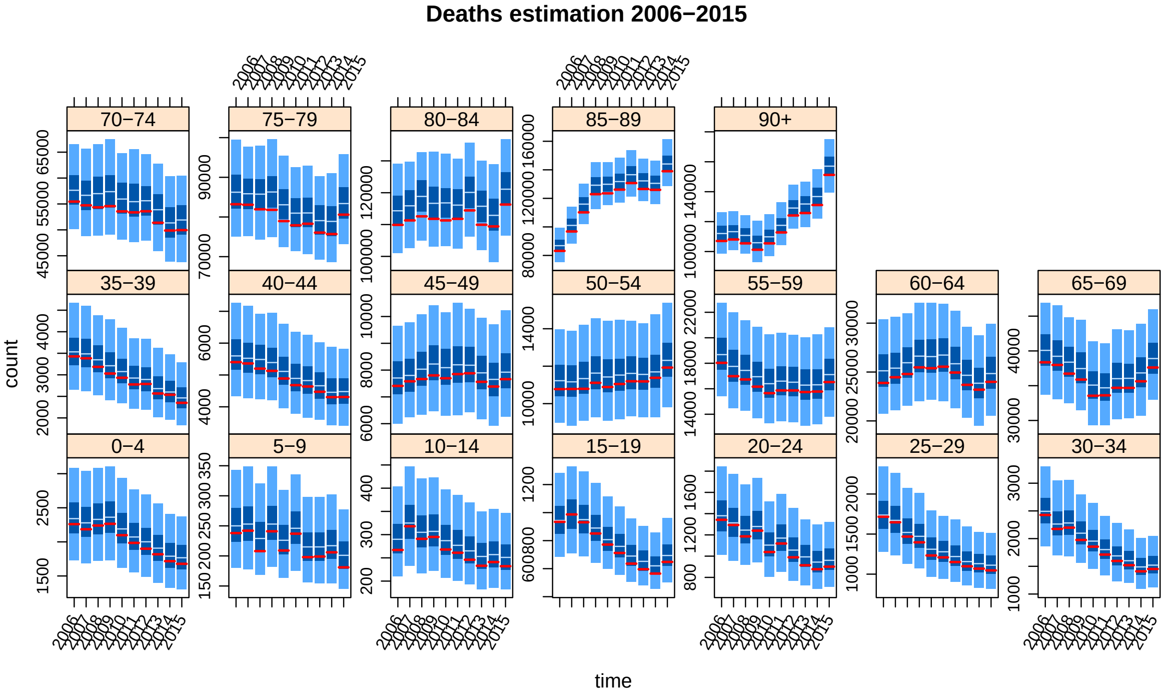


FIGURE 4.5: Death counts estimation for CMP model by age group and time dimensions, from 2006 until 2015. Each age block shows medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from *Annuario statistico italiano*.

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | 1.04 | 1.04 | 1.06 | 1.10 | 1.06 | 1.05 | 1.06 | 1.10 | 1.08 | 1.09 |

TABLE 4.1: Mean value of $\nu$ for death counts model with only time dimension, from 2006 until 2015.

| | 2006 | | 2007 | | 2008 | | 2009 | | 2010 | | 2011 | | 2012 | | 2013 | | 2014 | | 2015 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male |
| 0-4 | 1.05 | 1.07 | 1.13 | 1.12 | 1.09 | 1.15 | 1.09 | 1.07 | 1.09 | 1.06 | 1.10 | 1.11 | 1.07 | 1.16 | 1.08 | 1.10 | 1.08 | 1.06 | 1.09 | 1.04 |
| 5-9 | 1.14 | 1.09 | 1.15 | 1.14 | 1.08 | 1.14 | 1.07 | 1.09 | 1.14 | 1.07 | 1.09 | 1.10 | 1.12 | 1.12 | 1.14 | 1.12 | 1.11 | 1.10 | 1.10 | 1.10 |
| 10-14 | 1.10 | 1.14 | 1.05 | 1.08 | 1.13 | 1.08 | 1.13 | 1.06 | 1.13 | 1.09 | 1.15 | 1.07 | 1.10 | 1.14 | 1.12 | 1.07 | 1.11 | 1.10 | 1.10 | 1.12 |
| 15-19 | 1.09 | 1.05 | 1.09 | 1.12 | 1.10 | 1.06 | 1.14 | 1.11 | 1.09 | 1.09 | 1.10 | 1.08 | 1.08 | 1.14 | 1.13 | 1.14 | 1.14 | 1.10 | 1.08 | 1.09 |
| 20-24 | 1.11 | 1.09 | 1.12 | 1.08 | 1.10 | 1.09 | 1.09 | 1.09 | 1.10 | 1.10 | 1.10 | 1.08 | 1.11 | 1.10 | 1.11 | 1.08 | 1.09 | 1.09 | 1.12 | 1.08 |
| 25-29 | 1.10 | 1.08 | 1.09 | 1.10 | 1.11 | 1.13 | 1.13 | 1.14 | 1.12 | 1.07 | 1.07 | 1.08 | 1.08 | 1.15 | 1.11 | 1.07 | 1.10 | 1.09 | 1.12 | 1.09 |
| 30-34 | 1.08 | 1.08 | 1.09 | 1.14 | 1.10 | 1.10 | 1.10 | 1.09 | 1.11 | 1.10 | 1.14 | 1.07 | 1.15 | 1.12 | 1.08 | 1.15 | 1.11 | 1.10 | 1.08 | 1.14 |
| 35-39 | 1.06 | 1.07 | 1.09 | 1.10 | 1.11 | 1.10 | 1.07 | 1.11 | 1.10 | 1.10 | 1.11 | 1.12 | 1.03 | 1.11 | 1.11 | 1.14 | 1.12 | 1.08 | 1.14 | 1.09 |
| 40-44 | 1.15 | 1.09 | 1.15 | 1.09 | 1.08 | 1.05 | 1.13 | 1.08 | 1.06 | 1.07 | 1.03 | 1.20 | 1.09 | 1.05 | 1.08 | 1.09 | 1.13 | 1.03 | 1.09 | 1.08 |
| 45-49 | 1.12 | 1.09 | 1.08 | 1.15 | 1.05 | 1.09 | 1.33 | 1.12 | 1.11 | 1.14 | 1.12 | 1.18 | 1.17 | 1.11 | 1.13 | 1.09 | 1.10 | 1.05 | 1.06 | 1.08 |
| 50-54 | 1.06 | 1.15 | 1.17 | 1.09 | 1.10 | 1.07 | 1.10 | 1.10 | 1.09 | 1.11 | 1.06 | 1.07 | 1.07 | 1.11 | 1.11 | 1.08 | 1.11 | 1.05 | 1.12 | 1.09 |
| 55-59 | 1.17 | 1.14 | 1.08 | 1.13 | 1.07 | 1.11 | 1.10 | 1.07 | 1.14 | 1.09 | 1.05 | 1.09 | 1.09 | 1.11 | 1.23 | 1.12 | 1.08 | 1.09 | 1.03 | 1.09 |
| 60-64 | 1.09 | 1.05 | 1.17 | 1.04 | 1.05 | 1.11 | 1.07 | 1.09 | 1.10 | 1.11 | 1.09 | 1.06 | 1.12 | 1.10 | 1.03 | 1.07 | 1.15 | 1.08 | 1.05 | 1.17 |
| 65-69 | 1.09 | 1.13 | 1.07 | 1.81 | 1.09 | 1.10 | 1.12 | 1.04 | 1.05 | 1.12 | 1.07 | 1.06 | 1.10 | 1.62 | 1.07 | 1.08 | 1.10 | 1.06 | 1.08 | 1.08 |
| 70-74 | 1.07 | 1.13 | 1.14 | 1.10 | 1.08 | 1.11 | 1.12 | 1.03 | 1.13 | 1.07 | 1.10 | 1.14 | 1.06 | 1.04 | 1.10 | 1.07 | 1.10 | 1.18 | 1.12 | 1.08 |
| 75-79 | 1.06 | 1.10 | 1.09 | 1.10 | 1.13 | 1.12 | 1.08 | 1.15 | 1.06 | 1.09 | 1.09 | 1.07 | 1.12 | 1.08 | 1.09 | 1.06 | 1.07 | 1.12 | 1.13 | 1.15 |
| 80-84 | 1.15 | 1.14 | 1.09 | 1.11 | 1.04 | 1.13 | 1.09 | 1.12 | 1.09 | 1.13 | 1.11 | 1.13 | 1.11 | 1.11 | 1.10 | 1.05 | 1.12 | 1.05 | 1.05 | 1.08 |
| 85-89 | 1.15 | 1.05 | 1.14 | 1.13 | 1.06 | 1.14 | 1.16 | 1.11 | 1.09 | 1.13 | 1.09 | 1.37 | 1.11 | 1.06 | 1.15 | 1.14 | 1.08 | 1.21 | 1.05 | 1.11 |
| 90+ | 1.07 | 1.04 | 1.07 | 1.09 | 1.09 | 1.17 | 1.09 | 1.02 | 1.10 | 1.08 | 1.07 | 1.08 | 1.11 | 1.06 | 1.15 | 1.07 | 1.08 | 1.09 | 1.14 | 1.13 |

TABLE 4.2: Mean value of $\nu$ for death counts model with age, sex and time dimensions, from 2006 until 2015.

### 4.3.2 CMP data model for birth counts

Almost the same applies to birth counts examples. Three system model are compared: (i) one only considering time, (ii) one with time and region, (iii) and the last one with age and time. As in the death count model, all system models assume a DLM prior on time with standard deviation scale of 0.1 for all parameters, an informative prior on the intercept, and a prior on the variance with scale 0.1. Age effect, region-time and age-time interactions have the same DLM prior as time, whereas on region effect an exchangeable prior is assumed. Also, as with death count model, two datasets are considered, the one from the demographic balance and the one including mothers' age. The first has a Poisson-Binomial prior with $p = 0.98$, and the second a CMP prior, as in the death counts example, with informative variance on the intercept for location parameter and $\nu \sim N(0, 0.5)$. Results are comparable with death counts models. In figure 4.6, birth counts estimation is much higher than data and parameters $\nu$s for each year are showed in table 4.3. The best fit is for the model including age and time (figure 4.7). Parameters $\nu$s for model with age and time, and region and time are shown respectively in tables 4.4 and 4.5. If regional dispersion parameters by year are not much higher than the one considering only time, the difference is more evident when comparing $\nu$s computed by age and time in table 4.4 with table 4.3. The mean values for the three models are: $\bar{\nu}^t = 1.035$, $\bar{\nu}^{rt} = 1.09$ and $\bar{\nu}^{at} = 1.19$. Even if differences amount to few decimals only, it is hard to quantify what is a high or low difference between two dispersion parameters, but this difference exists and reflects what was reasonably expected.

Despite the number of region (21) is three times higher than the number of age groups (7), $\nu^{at}$ is still higher than $\nu rt$ suggesting that age is a more important dimension in modelling births than region. This result that might seem counter-intuitive, as for the same amount of people groups are smaller in the model considering regions, it is actually not surprising from a demographical perspective as, in Italy, birth regional differences are as not as important as mothers' age ones, which has also been confirmed by the preliminary analyses and results for age and region effects in section 3.3.

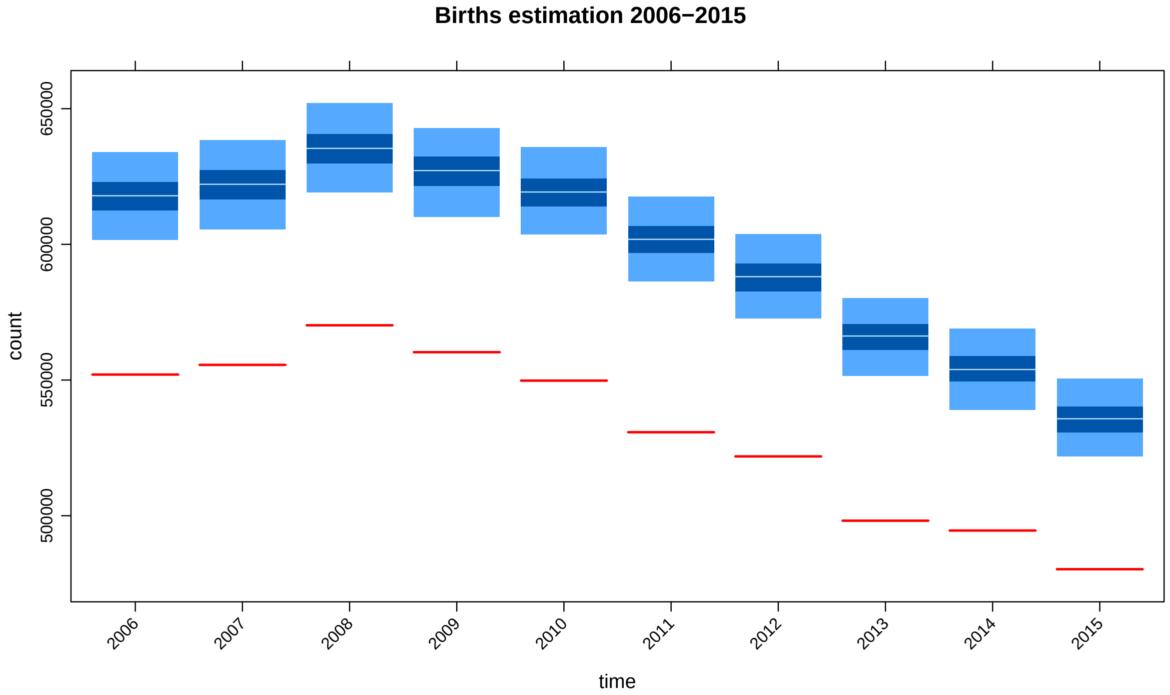


FIGURE 4.6: Birth counts estimation for CMP model with only time dimension, from 2006 until 2015. The medians are the white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from *Annuario statistico italiano*.

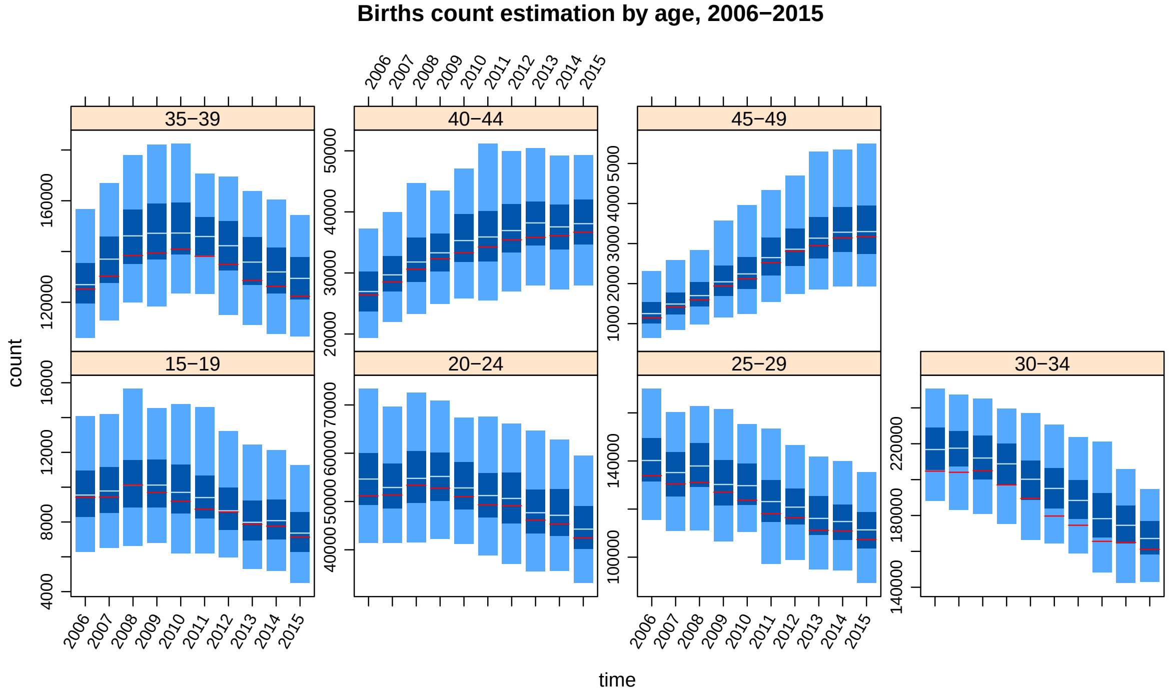


FIGURE 4.7: Birth counts estimation for CMP model by age group and time dimensions, from 2006 until 2015. Each age block shows medians as white lines, the 50% C.I.s in blue, the 95% C.I.s in light blue, and in red data from *Annuario statistico italiano*.

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | 1.04 | 1.04 | 1.03 | 1.03 | 1.04 | 1.04 | 1.04 | 1.02 | 1.04 | 1.03 |

Table 4.3: Mean value of $\nu$ for birth counts model with only time dimension, from 2006 until 2015.

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|
| 15-19 | 1.11 | 1.08 | 1.10 | 1.18 | 1.29 | 1.21 | 1.15 | 1.06 | 1.26 | 1.13 |
| 20-24 | 1.23 | 1.14 | 1.06 | 1.30 | 1.22 | 1.21 | 1.09 | 1.11 | 1.17 | 1.23 |
| 25-29 | 1.35 | 1.14 | 1.17 | 1.15 | 1.13 | 1.17 | 1.12 | 1.13 | 1.09 | 1.21 |
| 30-34 | 1.29 | 1.29 | 1.10 | 1.16 | 1.15 | 1.22 | 1.17 | 3.19 | 1.18 | 1.20 |
| 35-39 | 1.08 | 1.13 | 1.11 | 1.19 | 1.32 | 1.20 | 1.13 | 1.15 | 1.15 | 1.17 |
| 40-44 | 1.11 | 1.26 | 1.22 | 1.14 | 1.21 | 1.31 | 1.12 | 1.17 | 1.09 | 1.09 |
| 45-49 | 1.12 | 1.18 | 1.06 | 1.11 | 1.12 | 1.11 | 1.11 | 1.17 | 1.10 | 1.02 |

Table 4.4: Mean value of $\nu$ for birth counts model with mothers' age and time dimension, from 2006 until 2015.

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|
| Piemonte | 1.06 | 1.06 | 1.09 | 1.07 | 1.07 | 1.04 | 1.06 | 1.08 | 1.09 | 1.08 |
| Valle D'Aosta | 1.15 | 1.20 | 1.19 | 1.19 | 1.18 | 1.20 | 1.20 | 1.08 | 1.16 | 1.15 |
| Lombardia | 1.03 | 1.03 | 1.07 | 1.07 | 1.05 | 1.04 | 1.06 | 1.05 | 1.04 | 1.07 |
| Bolzano/Bozen | 1.16 | 1.11 | 1.17 | 1.13 | 1.14 | 1.12 | 1.14 | 1.15 | 1.11 | 1.11 |
| Trento | 1.13 | 1.13 | 1.15 | 1.12 | 1.10 | 1.13 | 1.13 | 1.14 | 1.14 | 1.13 |
| Veneto | 1.09 | 1.07 | 1.07 | 1.06 | 1.11 | 1.06 | 1.09 | 1.10 | 1.04 | 1.06 |
| Friuli Venezia Giulia | 1.09 | 1.10 | 1.13 | 1.14 | 1.13 | 1.09 | 1.11 | 1.12 | 1.10 | 1.13 |
| Liguria | 1.10 | 1.10 | 1.09 | 1.10 | 1.11 | 1.09 | 1.14 | 1.10 | 1.11 | 1.12 |
| Emilia Romagna | 1.06 | 1.10 | 1.07 | 1.04 | 1.07 | 1.08 | 1.10 | 1.09 | 1.06 | 1.12 |
| Toscana | 1.07 | 1.08 | 1.06 | 1.07 | 1.10 | 1.07 | 1.07 | 1.10 | 1.08 | 1.09 |
| Umbria | 1.14 | 1.12 | 1.11 | 1.14 | 1.11 | 1.10 | 1.10 | 1.12 | 1.12 | 1.13 |
| Marche | 1.10 | 1.09 | 1.08 | 1.09 | 1.13 | 1.11 | 1.08 | 1.11 | 1.10 | 1.10 |
| Lazio | 1.06 | 1.07 | 1.08 | 1.08 | 1.06 | 1.06 | 1.05 | 1.07 | 1.04 | 1.07 |
| Abruzzo | 1.10 | 1.11 | 1.12 | 1.09 | 1.10 | 0.81 | 0.87 | 0.67 | 1.07 | 1.11 |
| Molise | 1.14 | 1.15 | 1.16 | 1.13 | 1.15 | 0.99 | 0.83 | 0.92 | 0.94 | 1.10 |
| Campania | 1.05 | 1.08 | 1.07 | 1.07 | 1.08 | 1.09 | 1.09 | 1.08 | 1.09 | 1.06 |
| Puglia | 1.09 | 1.07 | 1.09 | 1.09 | 1.06 | 1.05 | 1.07 | 1.06 | 1.08 | 1.08 |
| Basilicata | 1.14 | 1.12 | 1.14 | 1.15 | 1.13 | 1.08 | 1.13 | 1.11 | 1.16 | 1.15 |
| Calabria | 1.10 | 1.12 | 1.11 | 1.09 | 1.10 | 1.08 | 1.10 | 1.11 | 1.08 | 1.08 |
| Sicilia | 1.03 | 1.08 | 1.07 | 1.05 | 1.05 | 1.06 | 1.07 | 1.02 | 1.06 | 1.05 |
| Sardegna | 1.11 | 1.09 | 1.08 | 1.11 | 1.12 | 1.11 | 1.11 | 1.11 | 1.11 | 1.11 |

Table 4.5: Mean value of $\nu$ for birth counts model with region and time dimension, from 2006 until 2015.

The application to the demographic account has not given satisfying results yet and further developments are still needed in order to better understand what is the best suitable model for the estimation of the CMP parameters. Dependence between $\gamma$ and $\nu$ have only started being studied and, as mentioned earlier, the form of $nu$ hyper-parametrisation can also be changed. Nevertheless, results provided so far are encouraging and reflect expectations despite the magnitude of the difference among the $\nu$s is not very high. Literature on CMP distribution and CMP regression models is not wide but this extension of the Poisson distribution needs to be further investigated and understood especially for demographic application where Poisson distribution is a common and natural choice but it implies limitations that do not always suit data.

# Conclusions

## Discussion

Results provided in the previous chapters show advantages and limitations of the Bayesian demographic account approach for population size estimation. Literature proposing models to correct or limit data problems and getting closer to the true population size is wide and includes different approaches. Among the reasons to implement and develop models to provide estimates of population size there is, first of all, the uncertainty statements included in the estimation that are not provided by data. Statistical models provide results in terms of confidence or credible intervals, implying that their answers are probabilistic and not deterministic as data are. When dealing only with data, their accuracy can be checked by comparison with other sources (whose accuracy have to be verified too), or by other information not coming directly from data; instead, the model proposed and, in general, Bayesian or Frequentist approaches to population size estimation provide confidence and credible intervals giving an idea of the accuracy and reliability of their results also with respect to available data. As a general rule, when intervals are wide it means uncertainty is high and, conversely, narrower intervals mean results are more reliable so the notion of uncertainty is clearly included and provided in the results. Bayesian approaches, including *a priori* knowledge in the model, help the estimation process and to obtain likely results. The model initially proposed in Bryant and Graham (2013) embed these characteristics. The empirical analysis on Italian case in chapter 3 shows results according to different assumptions and highlights both the importance of these assumptions along with the robustness of the model. Assumptions strength must always be motivated and reasonable as the model adjusts to them, despite being in general quite robust. Constraints limiting parameters to some values or assuming very low standard deviations have an impact on the results but, for instance, the choice of distribution for the data models does not usually have a high impact on the results if assumptions on parameters are not too strong. The same holds for the system model: some models are better than others but, if assumptions are reasonable, usually results are too. A third point is the possibility to correct for systematic biases in the data. With Italian data there has been no need to include any correction for systematic bias, but with other data there can be biases depending on regions, age groups or other dimensions. For instance, in Bryant and Zhang (2018) a correction for mothers'age has been included in the data model to estimate age-specific fertility rates in Cambodia, because there were coverage problems with older age groups. A novelty of the demographic account model is also to estimate all the series at the same time, each with its own models and data, whereas other models often extrapolate one series from the demographic balance equations. Nonetheless, the demographic balance equations consistency is still satisfied due to constraints included in the MCMC algorithm. Being developed in the Bayesian framework, Bryant and Graham (2013) model is one of the most flexible models proposed for population size estimation because it allows estimating models with far more parameters than data, giving demographers much space for decision and inclusion of *a priori* information. Flexibility makes the model useful for a large range of data of different quality, and allows for inclusion of more than one dataset for each series. Eventually, the CMP model implementation also seems promising and worth to be developed.

From the results, model limitations can be deduced. Firstly, the difficult convergence, especially when data quality is low, and the high autocorrelation of the Markov chains. High autocorrelation is mainly due to the constraints of the demographic balance equations that force candidate values to be very close to the current ones in order to be accepted. Secondly, the identifiability problems that sometimes lead to unlikely results especially with the CMP model and with migration series estimates.

Dealing with data problems is also not always straightforward, because the model naturally provides smooth results and privileges more accurate datasets, but it is sometimes hard to combine these two features. For instance, it was difficult to combine 2011 Italian census data with administrative data on the resident population because census data, without any information about their nature, would be considered as outlier with respect to the other data. On the one hand, according to Istat (Egidi and Ferruzza, 2009; Istat, 2016, 2015a), final census data after PES are very accurate, they are checked and compared with administrative registers, therefore, if people are absent in the census count but present in the registers or, vice versa, if someone is in the census count but not in the registers, then the error is corrected. In the other hand census data were not matching the administrative registers trend, hence obtaining the actual evolution of population size was not easy. With the current model it is also difficult to compare results as no standard criterion has been implemented yet in the `R package demest`. Therefore, results coming from different models are compared, but it is difficult to assess what is the best one. Checking with held-back data have been made when estimating single series of birth and death counts but a more general approach is needed, also for comparison with other methods. The high complexity and flexibility have also important drawbacks. Tuning the model can be sometimes difficult due to the high number of parameters and flexibility might over-fit rather than improve the model.

Because of the problem complexity, the aim of the application to the Italian data, especially for the demographic account estimation, does not propose a “right” model. It is difficult to choose one model as, the true population is unknown and results, even if somehow robust, need to be interpreted and, in any case, they also reflect what the model structure suggests. The level of uncertainty provided often reflects the accuracy assumed on data and the strength of *a priori* choices. What datasets are considered more reliable, how much variability is allowed on parameters, everything influences the estimation process and trying to compare or implement all the possible models is impossible. Therefore, examples included are those considered representative to explain model characteristics, and useful to give an idea of the different possibilities the model provides. More than on results themselves the focus is on their comparison.

The model needs further investigations before being ready for being used by NSIs, especially for the demographic account estimation part, but some parts are more mature and are already used by Statistics New Zealand. The extension with the CMP model is ambitious, it is still in its early stages, and it needs to be further developed and understood.

# Future directions of research

The area of Bayesian demography is still relatively young and it is developing fast. Organisations as United Nations and more and more NSIs are starting implementing some of these methods and keep on improving and working on them. Research is also growing in this area, therefore it is reasonable to expect fast progress in the whole field. The model initially proposed by Bryant and Graham (2013) and further developed and discussed in this thesis will be no exception. What aspects of the model need improvement have been pointed out discussing advantages and limitations of the model, and steps to address issues and improve the model have also been planned. Especially for the whole demographic account estimation, the model certainly needs more applications for a better understanding of parameters interaction, to identify what modifications are mostly needed, and how changes impact the whole structure and estimation process.

Especially data model structures need to be simplified in order to avoid over-fitting and speed up computations. Bryant and Zhang (2018) suggest the implementation of models allowing for specifying if "some dimensions are measured more accurately than others", or to base data models on indirect models (Preston *et al.*, 2001).

Prior distributions model set can be increased with both computationally convenient and purpose-specific distributions; for instance, there were attempts using an adapted version of the model proposed in Kunihama and Dunson (2013) for modelling interactions, *ad hoc* distributions like the one used for migration in Wiśniowski *et al.* (2013), or simple objective priors to use with high quality data.

From a computational point of view the demographic balance equation constraints sometimes makes the MCMC algorithm inefficient, causing convergence and autocorrelation issues. Working with a more flexible framework could help to tackle this issues and increase robustness to chains starting points. Now starting points need to be initially consistent, i.e. the initial demographic account must satisfy demographic balance equations. The same consistency is required at each iteration. An idea for improving this aspect is to use an adaptive proposal distribution that, starting from looser constraints, gets closer to the target posterior iteration after iteration, until it perfectly satisfies the constraints. Methods considered are linked to importance sampling and particle filtering. The best, so far, seems to be the Adaptive Multiple Importance Sampling (Cornuet *et al.*, 2012).

A lot of work is also planned to improve the `R package demest`. A first example of function worth implementing is the disaggregation function. To disaggregate a dataset by given dimensions, a model must account for random variation, information, trends or counter-trends and interactions. Sometimes this requires high complex models, but it is important to keep the model as simple as possible and try to manage complexity efficiently. Bryant and Zhang (2018) suggest, "decompositions and graphs to guide the construction of each piece". Also, a function to compute standard model estimation criteria as the Watanabe-Akaike information criterion (WAIC) and/or other information criteria Watanabe (2010) to compare different models is planned. Another issue is the high computation power required for the estimation of very large models with hundreds of thousands or millions of parameters. The examples presented in the thesis only require a desktop computer, but for the estimation of larger models with a reasonable number of iterations a much faster and powerful machine is needed.

The package also allows for forecast and for a wide range of applications related to population studies. Bryant and Zhang (2018) suggest topics such as "modelling disease prevalence, forecasting future labour supply, and studying promotion within organisations". The model experimentation is only starting but it is promising and hopefully it is providing a productive contribution to statistical methods and demography.

# Appendix

## Account updating

Let $N$ be the population array, which is usually the only one which refers to a point on time (e.g. population on December $31^{st}$, 2017) whereas all the other are values referring to an interval in time (e.g. births in 2017, or in December). The other array are denoted with $C_l$, $l = 1, \cdots, L$, and with the population array they compose the demographic account $Y = \{N, C_1, \cdots, C_L\}$. Typically, there are two population arrays, one for population at the beginning of the period considered and one for the population at the end of the period and all the account cells are linked by the account identity. The demographic account $Y$ is an unobserved quantity and its distribution is

$$\begin{aligned} p(Y|\Theta_Y, Z) &= p(Y|\Theta_Y) \\ &= p(N|\Theta_N, Z) \prod_{l=1}^{L} p(C_l|N, \Theta_l, Z) \\ &= p(N|\Theta_N) \prod_{l=1}^{L} p(C_l|N, \Theta_l) \end{aligned} \tag{A.1}$$

when conditioned on the parameter set $\Theta_Y$ then the $Y$ is independent of covariates $Z$ as they are already included in the estimation of $\Theta_Y$.

The data model has a conditioning on the demographic account $Y$ so that the model has the form

$$p(X|Y, \Omega) = \prod_{m=1}^{M} p(X_m|Y^{[m]}, \Omega_m)$$

where $X = \{X_1, \cdots, X_M\}$ and $Y^{[m]}$ is the demographic array the dataset $X_m$ refers to collapsed in order to have the same dimensions as $X_m$. Then each sub-model is $p(X_m|Y^{[m]}, \Omega_m) = \prod_j p(x_{jm}|y_{jm}, \Omega_{jm})$.

For the set of parameters associate to $Y$ and $X$, respectively $\Theta$ and $\Omega$, they are assumed not to share any parameter and their distribution can therefore be decomposed as:

$$p(\Theta, \Omega|Z) = p(\Theta|Z)p(\Omega)$$

Therefore

$$p(Y, X|\Theta, \Omega, Z) = p(Y|\Theta, Z)p(X\ Y, \Omega)$$

From all these parts it is now possible to introduce the complete joint posterior the whole model aims to estimate and that can be divided in three marginal models:

1. $p(Y|X, Z)$, for the demographic account itself
2. $p(\Theta|X, Z)$ which estimates parameters of the system model and it gives information about the super-population quantities
3. $p(\Omega|X, Z)$ which relates to the data model and potentially helps understanding characteristics or possible problems of the datasets.

the joint posterior is therefore

$$\begin{aligned} p(Y, \Theta, \Omega|X, Z) &\propto p(X|Y, \Theta, \Omega, Z)p(Y|\Theta, \Omega, Z)p(\Theta, \Omega|Z) \\ &= p(X|Y, \Omega)p(Y|\Theta)p(\Theta|Z)p(\Omega) \end{aligned} \tag{A.2}$$

with full conditionals

$$\begin{aligned} p(Y|\Theta, \Omega, X, Z) &\propto p(X|Y, \Omega)p(Y|\Theta) \\ p(\Theta|Y, \Omega, X, Z) &\propto p(Y|\Theta)p(\Theta|Z) \\ p(\Omega|Y, \Theta, X, Z) &\propto p(X|Y, \Omega)p(\Omega) \end{aligned} \tag{A.3}$$

## Updating $Y$

As the form of (A.1) suggests, models for the arrays of the demographic account are independent, therefore, simplifications are possible in the updating process.

1. When it comes to the Metropolis-Hastings algorithm, when updating starting values for $N$, the proposal distribution $Q$ takes the form

   $$\frac{Q(Y^{(z)}|Y^*)}{Q(Y^*|Y^{(z)})} = \frac{p(n^{(z)})}{p(n^*)} \tag{A.4}$$

   where $Y^{(z)}$ and $n^{(z)}$, $z = 1, \cdots, Z$, are respectively the current value of the demographic account $Y$ and the population cell $n$ to update at iteration $z$ whereas $Y^*$ and $n^*$ are the proposed value.

Let $N = (N_0, N_{1+})$ where $N_0$ is the population array at time 0 and $N_{1+}$ is the set of arrays from time 1 onward, the acceptance probability $a$ is

$$a = \min\left(1, \frac{p(Y^*)}{p(Y^{(z)})} \frac{Q(Y^{(z)}|Y^*)}{Q(Y^*|Y^{(z)})}\right)$$

As the cells for $t = 0$ simplify, the ratio becomes

$$\begin{aligned} \frac{p(Y^*)}{p(Y^{(z)})} \frac{Q(Y^{(z)}|Y^*)}{Q(Y^*|Y^{(z)})} &= \frac{p(N^*)}{p(N^{(z)})} \left(\prod_{l=1}^{L} \frac{p(C_m^{(z)}|N^*)}{p(C_m^{(z)}|N^{(z)})}\right) \frac{p(n^{(z)})}{p(n^*)} \\ &= \frac{p(N_{1+}^*)}{p(N_{1+}^{(z)})} \prod_{l=1}^{L} \frac{p(C_m^{(z)}|N^*)}{p(C_m^{(z)}|N^{(z)})} \end{aligned} \tag{A.5}$$

The first ratio in (A.6) is

$$\frac{p(N_{1+}^*)}{p(N_{1+}^{(z)})} = \prod_{t=1}^{T} \frac{\text{Poisson}(n^*(t)|\lambda(t))}{\text{Poisson}(n^{(z)}(t)|\lambda(t))}$$

where $n^{(z)}(t) = N^{(z)}[i, \min(a+t, A), t]$, $n^*(t) = N^*[i, \min(a+t, A), t] = n^{(z)}(t) + \Delta$, with $\Delta = n^* - n^{(z)}$, $a = 0, \cdots, A$ age classes and $t$ age class width.

2. In order to update a component $C_u$, $u \in \{1, \cdots, L\}$, the proposal density depends on the expected population $\hat{N}$ rather than on the parameter $\gamma^u$. This conditioning is convenient as both $C_u$ and $\hat{N}$ are distributed according to a Poisson distribution and it makes possible to suppress the dependence on $\gamma^u$ which would make calculations more complex.

   In this case (A.4) becomes $\frac{Q(Y^{(z)}|Y^*)}{Q(Y^*|Y^{(z)})} = \frac{p(c^{(z)}|\hat{N})}{p(c^*|\hat{N})}$ and, using (A.1), equation (A.6) takes the form:

$$\begin{aligned}
\frac{p(Y^*)}{p(Y^{(z)})}\frac{Q(Y^{(z)}|Y^*)}{Q(Y^*|Y^{(z)})} &= \frac{p(N^*)\prod_{l=1}^{L} p(C_l^*|N^*)}{p(N^{(z)})\prod_{l=1}^{L} p(C_l^{(z)}|N^{(z)})} \\
&= \frac{p(N^*)}{p(N^{(z)})}\frac{p(C_u^*|N^*)}{p(C_u^{(z)}|N^{(z)})}\left(\prod_{l\neq u}\frac{p(C_l^{(z)}|N^*)}{p(C_l^{(z)}|N^{(z)})}\right)\frac{p(c^{(z)}|\hat{N})}{p(c^*|\hat{N})} \\
&= \frac{p(N^*)}{p(N^{(z)})}\frac{p(C_u^*|N^*)}{p(C_u^{(z)}|N^{(z)})}\frac{p(C_u^{(z)}|N^{(z)})}{p(C_u^{(z)}|N^*)}\frac{p(C_u^{(z)}|N^*)}{p(C_u^{(z)}|N^{(z)})} \qquad (A.6) \\
&\left(\prod_{l\neq u}\frac{p(C_l^{(z)}|N^*)}{p(C_l^{(z)}|N^{(z)})}\right)\frac{p(c^{(z)}|\hat{N})}{p(c^*|\hat{N})} \\
&= \frac{p(N^*)}{p(N^{(z)})}\frac{p(C_u^*|N^*)p(c^{(z)}|\hat{N})}{p(C_u^{(z)}|N^*)p(c^*|\hat{N})}\prod_{l=1}^{L}\frac{p(C_l^{(z)}|N^*)}{p(C_l^{(z)}|N^{(z)})}
\end{aligned}$$

Through the simplification in (A.6), it is possible to calculate separately the three terms on the last line of (A.6).

The first term

$$\frac{p(N^*)}{p(N^{(z)})} = \prod_{t=0}^{T} -t\frac{\text{Poisson}(n^*(t)|\lambda(t))}{\text{Poisson}(n^{(z)}(t)\ \lambda(t))}$$

or, if there is an array containing two flows, in and out, like internal migration with “origin-destination” format, or “pool” format or “net” format then

$$\frac{p(N^*)}{p(N^{(z)})} = \prod_{t=0}^{T} -t\frac{\text{Poisson}(n^*_{out}(t)|\lambda(t))}{\text{Poisson}(n^{(z)}_{out}(t)\ \lambda(t))}\frac{\text{Poisson}(n^*_{in}(t)|\lambda(t))}{\text{Poisson}(n^{(z)}_{in}(t)\ \lambda(t))}$$

The second term depends on the form of the model if it contains or not exposure. In the model without exposure the estimated and proposed populations $\hat{N}$ and $N^*$ coincide, as there is no new proposition for $N$ but only for $C_u$, therefore only the estimation of the population deriving from the new proposal $C_u^*$ is calculated. Consequence of this is the complete simplification of the second term

$$\frac{p(C_u^*|N^*)p(c^{(z)}|\hat{N})}{p(C_u^{(z)}|N^*)p(c^*|\hat{N})} = \frac{\cancel{\text{Poisson}(c^*|n)}\cancel{\text{Poisson}(c^{(z)}|n)}}{\cancel{\text{Poisson}(c^z|n)}\cancel{\text{Poisson}(c^*|n)}}$$

When the exposure terms are included in the model they are first calculated

$$\begin{aligned}
e^z &= E_i^z \qquad (A.7) \\
e^* &= e^{(z)} + 1/2\Delta \\
\gamma &= \gamma_{li}
\end{aligned}$$

and then the ratio becomes

$$\frac{p(C_l^*|N^*)p(c^{(z)}|\hat{N})}{p(C_l^{(z)}|N^*)p(c^*|\hat{N})} = \frac{\text{Poisson}(c^*|\gamma e^*)\text{Poisson}(c^{(z)}|\gamma \hat{e})}{\text{Poisson}(c^{(z)}|\gamma e^*)\text{Poisson}(c^*|\gamma \hat{e})} \\ = (e^*/\hat{e})^{(c^- c^{(z)})} \tag{A.8}$$

The third term of (A.6) does not depend on the term is being updated and the product components have all the same form reflecting the previous terms structure.

The starting point of the whole updating process of the demographic account is to choose the starting population to update. All the cells $N$ at time 0 have the same selection probability. Once the cell has been selected, a subset is created with the cells in need to be updated because of the change of the starting population. Subset includes: other population cells, exposures, expected values and corresponding cells in the datasets $X$. The proposed value $n^*$ is drawn from a left-truncated Poisson distribution, where truncation point is set at value $v = \min_i n_i^{(z)}$ to avoid unlikely or impossible values.

$$p(y) \propto \begin{cases} e^{-\gamma}\gamma^y & \text{If } y \geq n^{(z)} - v \\ 0 & \text{otherwise} \end{cases}$$

Candidate value $n^*$ acceptance or refusal is performed through a Metropolis-Hastings algorithm. Metropolis-Hastings ratios $r$s for population updating are formed by:

1. Likelihoods ratio:
$$r_{X|N} = \prod_{m \in \mathcal{M}_X} \prod_{i \in \mathcal{N}_m} \frac{p(x_{jm}|n_{jm}^*)}{p(x_{jm}|n_{jm}^{(z)})} \tag{A.9}$$
where $n_{jm}^* = n_{jm}^{(z)} + \Delta$, $\Delta = n^* - n^{(z)}$ and $\mathcal{M}_X$ and $\mathcal{N}_m$ are the subset of indexes affected by the potential change of $n^{(z)}$.

2. Prior probabilities ratios:
$$r_{N|\lambda} = \prod_{i \in \mathcal{N}} \frac{p(n_i^*|\lambda_i)}{p(n_i^{(z)}|\lambda_i)} \tag{A.10}$$
and
$$r_{C_l|\lambda,E} = \prod_{i \in \mathcal{C}_l} \frac{p(c_{il}|\lambda_{il}e_{il}^*)}{p(c_{il}|\lambda_{il}e_{il}^{(z)})} \tag{A.11}$$
the first one, (A.10), directly involves population, the second, (A.11), follows from variation in the exposure term due to potential change in population.

3. Proposal probabilities ratio:
$$r_J = \frac{p(n^{(z)}|\lambda)}{p(n^*|\lambda)} \tag{A.12}$$

Hence, a Metropolis-Hastings ratio $r$ is

$$r = r_{X|N} r_{N|\lambda} r_{C_l|\lambda,E} r_J$$

and therefore

$$n^{(z+1)} = \begin{cases} n^* & \text{If } r > U \sim U(0,1) \\ n^{(z)} & \text{Otherwise} \end{cases}$$

If $n^*$ is accepted, then $e^*_{il}$ automatically is too.

When updating cells from $C$ the process is the same as for cells from $N$ but, in addition to ratios $r_{X|N}$ of (A.9) and $r_{N|\lambda}$ of (A.10), there is

$$r_{X|C} = \prod_{m \in \mathcal{M}_u} \prod_{j \in \mathcal{J}^m_u} \frac{p(x_{jm}|c^*_{jm})}{p(x_{jm}|c^{(z)}_{jm})} \tag{A.13}$$

where $\mathcal{M}_u$ and $\mathcal{J}^m_u$ are the set of indexes affected by a potential change in cell $u$ and

$$r_{C|\lambda} = \prod_{i \in \mathcal{N}_u} \frac{p(c^*_i|\lambda_i)}{p(c^{(z)}_i|\lambda_i)} \tag{A.14}$$

ratios (A.9) and (A.10) are still included because, to a potential change from a cell $c^{(z)}$ to $c^*$, corresponds a change in population and exposure from $n^{(z)}$ to $n^*$ and $e^{(z)}$ to $e^*$ respectively, therefore related ratios participate to the Metropolis-Hastings formation. Proposal probabilities ratio changes accordingly from (A.12) to

$$r_J = \frac{p(c^{(z)}_u|\lambda_u)}{p(c^*_u|\lambda_u)} \tag{A.15}$$

The resulting Metropolis-Hastings ratio is then

$$r = r_{X|C} r_{X|N} r_{C|\lambda} r_{N|\lambda} r_J$$

## Updating $\sigma$ and $\phi$

Considering the log-density of $\sigma$, i.e. logarithm of the Half-t probability distribution function (equation (2.34)),

$$f(\sigma) = -n \log \sigma - \frac{V_\sigma}{2\sigma^2} - \frac{\nu_\sigma + 1}{2} \log(\sigma^2 + \nu_\sigma A^2_\sigma)$$

it is possible to maximise $f$ taking the first derivative

$$\frac{df}{d\sigma} = -\frac{n}{\sigma} + \frac{V_\sigma}{\sigma^3} - \frac{(\nu_\sigma + 1)\sigma}{\sigma^2 + \nu_\sigma A_\sigma^2} = -\frac{1}{\sigma^3(\sigma^2 + \nu_\sigma A_\sigma^2)} h(\sigma)$$

where

$$h(\sigma) = (n + \nu_\sigma + 1)(\sigma^2)^2 + (n\nu_\sigma A_\sigma^2 - V_\sigma)\sigma^2 - V_\sigma \nu_\sigma A_\sigma^2 \tag{A.16}$$

The value $\sigma^*$ maximising $f$, considering (A.16) as a function of $\sigma^2$, is

$$\sigma^* = \sqrt{\frac{V_\sigma - n\nu_\sigma A_\sigma^2 + \sqrt{(V_\sigma - n\nu_\sigma A_\sigma^2)^2 + 4(n + \nu_\sigma + 1)V_\sigma \nu_\sigma A_\sigma^2}}{2(n + \nu_\sigma + 1)}}$$

So that:

$$\begin{cases} h(\sigma) < 0 \text{ for } \sigma^2 < (\sigma^*)^2 \\ h(\sigma) = 0 \text{ for } \sigma^2 = (\sigma^*)^2 \\ h(\sigma) > 0 \text{ for } \sigma^2 > (\sigma^*)^2 \end{cases}$$

Therefore $f(\sigma)$ has a maximum at $\sigma^*$, it is increasing for $\sigma < \sigma^*$, decreasing for $\sigma > \sigma^*$, and for any $z < f(\sigma^*)$ there are two roots for $f(\sigma) = z$ denoted by $\sigma_{min}$ and $\sigma_{max}$. These two roots can be found using a Newton-Raphson algorithm, either finding values satisfying $f(\sigma_{min}) = f(\sigma_{max}) = z$, or minimising the objective function $g(\sigma) = \big(f(\sigma) - z\big)^2$.

From this results is then possible to update $\sigma$ using a slice sampler Radford (2003): set $z = f(\sigma) - e$ with $e \sim Exp(1)$, find $\sigma_{min}$ and $\sigma_{max}$ roots of $f(\sigma) = z$ and draw the new value from $U(\sigma_{min}, \sigma_{max})$. The slice sampling is used also to update the variance of the Normal model $\phi$.

## Updating $\beta$ for exchangeable prior distribution

The estimation of the vector $\beta$ is a complex process. Parameters are weakly identified and, despite the estimation of $\mu$ and hence $\gamma$ might be right overall, the estimation of the single $k$-th components of $\mu_i = \sum_{k=0}^{K} \beta^{(k)}$ might be wrong even if the sum is right.

There are quantities useful to reduce this problem. Let

$$r_i^k = g(\gamma_i) - \mu_i + \beta^{(k)}_{h_i^k}$$

then, recalling that $g(\gamma_i) \sim N(\mu_i, \sigma^2)$ and $\mu_i = \sum_{k=0}^{K} \beta^{(k)}_{h_i^k}$, we have $r_i^k \sim N(\beta^{(k)}_{h_i^k}, \sigma^2)$. Then, let $\delta_h^k = \{i : j_i^k = j\}$ and $n_k$ be the number of elements in $\delta_h^k$, then $\sum_{i \in \delta_h^k} r_i^k \sim$

$N(n_k \beta_h^{(k)}, n_k \sigma^2)$ and

$$\bar{r}_h^{(k)} \sim N\left(\beta_h^{(k)}, \frac{\sigma^2}{n_k}\right) \tag{A.17}$$

where $\bar{r}_h^{(k)} = \left(\sum_{i \in \delta_h^k} r_i^k\right)/n_k$. These quantities are useful to update each $\beta_h^{(k)}$. In case covariates are included:

$$\begin{aligned}
\beta_h^{(k)} &\sim N(z_h^{(k)} \eta^{(k)}, \tau_k^2) \\
\tau_k &\sim t_{\nu_\tau}^{+}(0, A_{\tau k}^2) \\
\eta_1^{(k)} &\sim N(0, A_0^2) \\
\eta_p^{(k)} &\sim N(0, U_{\eta p}^{(k)}), \qquad p = 2, \cdots, P_k \\
U_{\eta p}^{(k)} &\sim \text{Inv-}\chi^2(\nu_\eta, A_{\eta k}^2), \qquad p = 2, \cdots, P_k
\end{aligned}$$

So $\beta_h^{(k)}$ can be updated within a Gibbs sampler drawing from

$$\beta_h^{(k)} \sim N\left(\frac{\frac{n}{\sigma^2}\bar{r}_h^{(k)} + \frac{1}{\tau_k^2} z_h^{(k)} \eta^{(k)}}{\frac{n}{\sigma^2} + \frac{1}{\tau_k^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_k^2}}\right) \tag{A.18}$$

Then, it is possible to update the parameter vector $\eta^{(k)}$ treating the hyper-priors as extra data points. Let

$$\begin{aligned}
y_*^{(k)} &= \begin{pmatrix} \beta^{(k)} \\ 0 \end{pmatrix} \qquad \text{(A.19)} \\
X_*^{(k)} &= \begin{pmatrix} Z^{(k)} \\ I \end{pmatrix} \\
\Sigma_k &= \begin{pmatrix} \tau_k^2 I & 0 \\ 0 & D_{U_\eta^{(k)}} \end{pmatrix} \\
&\text{where } D_{U_\eta^{(k)}} = \text{diag}\{A_0^2, U_{\eta 2}^{(k)}, ..., U_{\eta P_k}^{(k)}\}
\end{aligned}$$

Then

$$\begin{aligned}
y_*^{(k)} &\sim N(X_*^{(k)} \eta^{(k)}, \Sigma_k) \qquad \text{(A.20)} \\
\Sigma_k^{-1/2} y_*^{(k)} &\sim N(\Sigma_k^{-1/2} X_*^{(k)} \eta^{(k)}, I) \\
\eta^{(k)} &\sim N(\hat{\eta}_k, Y_{\beta_k})
\end{aligned}$$

where $\hat{\eta}_k = V_{\beta_k}\frac{1}{\tau_k^2}Z^{(k)T}\beta_k$ and $V_{\beta_k} = \frac{1}{\tau_k^2}Z^{(k)T}Z^{(k)} + D^{-1}_{U_\eta^{(k)}}$. Draws for updating $U_{\eta p}^{(k)}$ come from

$$U_{\eta p}^{(k)} \sim \text{Inv-}\chi^2\left(\nu_\eta + 1, \frac{\nu_\eta A_{\eta m}^2 + \left(\eta_p^{(k)}\right)^2}{\nu_\eta + 1}\right), \quad p = 2, ..., P_k$$

## Updating $\beta$ for DLM prior distribution

When coefficient $\beta$ has a DLM prior then it is updated drawing from a normal distribution with parameters similar to equation (A.18)

$$\beta_h^{(k)} \sim N\left(\frac{\frac{n_k}{\sigma^2}\bar{r}_h^{(k)} + \frac{1}{\tau_k^2}\left(\alpha_h^{(k)} + s_h^{(k)} + z_h^{(k)}\eta^{(k)}\right)}{\frac{n_k}{\sigma^2} + \frac{1}{\tau_k^2}}, \frac{1}{\frac{n_k}{\sigma^2} + \frac{1}{\tau_k^2}}\right) \tag{A.21}$$

Besides parameters $a_h^{(k)}$, $\delta_h^{(k)}$ and $s_h^{(k)}$ are updated using the forward-filtering backward-sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). Let

$$\begin{aligned}
\theta_h &= \begin{pmatrix} \alpha_h^{(k)} \\ \delta_h^{(k)} \end{pmatrix} \\
\tilde{\beta}_h &= \beta_h^{(k)} - s_h^{(k)} - z_j^{(k)}\eta^{(k)} \\
F &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
G &= \begin{pmatrix} 1 & 1 \\ 0 & \phi_k \end{pmatrix} \\
v_h &= \tau_k^2 \\
W_h &= \begin{pmatrix} \omega_{\alpha k}^2 & 0 \\ 0 & \omega_{\delta k}^2 \end{pmatrix}
\end{aligned} \tag{A.22}$$

and if the seasons are being updated, let $s_{qj}^{(k)}$, $q = 1, ..., S_k$, be the seasonal component for element $j - q - 1$. Then

$$\begin{aligned}
\theta_h &= \begin{pmatrix} s_{1h}^{(k)} \\ \vdots \\ s_{S_k,h}^{(k)} \end{pmatrix} \\
\tilde{\beta}_h &= \beta_h^{(k)} - \alpha_h^{(k)} - z_h^{k)}\eta^{(k)}
\end{aligned} \tag{A.23}$$

$$
\begin{aligned}
F &= \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
G &= \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \\
v_j &= \tau_k^2 \\
W_h &= \begin{pmatrix} \omega_{sk}^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}
\end{aligned}
$$

where $F$ is $S_k \times 1$, $G$ is $S_k \times S_k$ and $W_h$ is $S_k \times S_k$. With $m_h$ and $C_h$ being respectively the mean and the variance of $\theta_h$, given $\tilde{\beta}_1, \cdots, \tilde{\beta}_H$, the FFSB algorithm can be applied.

The first part of forward filter is:

$$
\begin{aligned}
a_h &= G m_{h-1} \\
R_h &= G C_{h-1} G^T + W_h \\
q_h &= F_h^T R_h F_h + v_h \\
A_h &= R_h F_h / q_h \\
e_h &= \tilde{\beta}_h - F_h^T a_h \\
A_h &= R_h F_h / q_h \\
m_h &= a_h + A_h e_h \\
C_h &= R_h - A_h A_h^T q_h
\end{aligned}
\tag{A.24}
$$

Then values $\theta_h \sim N(m_h, C_h)$, $h = 1, \cdots, H$, can be drawn and then the second part of backward sampling consist of:

$$
\begin{aligned}
B_h &= C_h G^T R_{j+1}^{-1} \\
m_h^* &= m_h + B_h(\theta_{h+1} - a_{h+1}) \\
C_h^* &= C_h - B_h R_{h+1} B_h^T
\end{aligned}
\tag{A.25}
$$

$$\theta_h \sim N(m_h^*, C_h^*)$$

to obtain updated values for $\theta_h = J-1, \cdots, 0$. The coefficient vector $\eta^{(k)}$ is updated as in the exchangeable prior case except that $\tilde{\beta}_h = \beta_h^{(k)} - \alpha_h^{(k)} - s_h^{(k)}$ takes the place of $\beta_j$in (A.19).

# CMP rejection algorithm calculation

In order to calculate the logarithm of the acceptance ratio $\alpha$ for implementing the Metropolis-Hastings algorithm, the logarithmic forms of the enveloping bounds $B_{f/g}^{\nu<1}$ and $B_{f/g}^{\nu\geq 1}$ are needed and will be used in the calculations of $\alpha$s

· **Logarithmic enveloping bounds $\nu < 1$:**

$$B_{f/g}^{\nu<1} = \frac{1}{p} \frac{\mu^{\nu\left\lfloor \frac{\mu}{(1-p)^{1/\nu}} \right\rfloor}}{(1-p)^{\nu\left\lfloor \frac{\mu}{(1-p)^{1/\nu}} \right\rfloor}} \left( \frac{\mu^{\lfloor \mu \rfloor}}{\lfloor \mu \rfloor !} \right)^{\nu-1} \tag{A.26}$$

$$\begin{aligned} \log(B_{f/g}^{\nu\geq 1}) = b_{f/g}^{\nu\geq 1} &= -\log(p) + \nu \left\lfloor \frac{\mu}{(1-p)^{1/\nu}} \right\rfloor \log(\mu) \\ &= -\left\lfloor \frac{\mu}{(1-p)^{1/\nu}} \right\rfloor \log(1-p) - \nu \sum_{n=1}^{\left\lfloor \frac{\mu}{(1-p)^{1/\nu}} \right\rfloor} \log(n) \end{aligned}$$

· **Logarithmic enveloping bounds $\nu \geq 1$:**

$$B_{f/g}^{\nu\geq 1} = \left( \frac{\mu^{\lfloor \mu \rfloor}}{\lfloor \mu \rfloor !} \right)^{\nu-1} \tag{A.27}$$

$$\begin{aligned} \log(B_{f/g}^{\nu\geq 1}) = b_{f/g}^{\nu\geq 1} &= (\nu-1)\Big( \log(\mu^{\lfloor \mu \rfloor}) - \log(\lfloor \mu \rfloor !) \Big) \\ &= (\nu-1)\Big( \lfloor \mu \rfloor \log(\mu) - \sum_{n=1}^{\lfloor \mu \rfloor} \log(n) \Big) \end{aligned}$$

The logarithmic forms of the acceptance ratios $\alpha^{\nu<1}$ and $\alpha^{\nu\geq 1}$ are:

- **Logarithmic acceptance ratios $\nu < 1$:**

$$
\begin{aligned}
\alpha^{\nu<1} &= \frac{\left(\mu^{y'}/y'!\right)^{\nu}}{B_{f/g}^{\nu<1}(1-p)^{y'}p} \qquad \text{(A.28)}\\
\log(\alpha^{\nu<1}) &= \nu\log\left(\mu^{y'}/y'!\right) - \Big(\log(B_{f/g}^{\nu<1}) + y'\log(1-p) + \log(p)\Big)\\
&= \nu\Big(y'\log(\mu) - \sum_{m=1}^{y'}\log(m)\Big) - \log(B_{f/g}^{\nu<1}) - y'\log(1-p) - \log(p)\\
&= \nu\left(y'\log(\mu) - \sum_{m=1}^{y'}\log(m)\right)\\
&\quad - \left( - \cancel{\log(p)} + \nu\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\log(\mu) - \left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\log(1-p)\right.\\
&\quad \left. - \nu\sum_{n=1}^{\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}\log(n)\right) - y'\log(1-p) - \cancel{\log(p)}\\
&= \nu y'\log(\mu) - \nu\sum_{m=1}^{y'}\log(m)\\
&\quad - \nu\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\log(\mu) + \left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\log(1-p)\\
&\quad + \nu\sum_{n=1}^{\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}\log(n) - y'\log(1-p)\\
&= \nu\log(\mu)\left(y' - \left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\right)\\
&\quad + \log(1-p)\left(\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor - y'\right)\\
&\quad + \nu\left(\sum_{n=1}^{\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}\log(n) - \sum_{m=1}^{y'}\log(m)\right)\\
&= \left(y' - \left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor\right)\Big(\nu\log(\mu) - \log(1-p)\Big)\\
&\quad + \nu\left(\sum_{n=1}^{\left\lfloor\frac{\mu}{(1-p)^{1/\nu}}\right\rfloor}\log(n) - \sum_{m=1}^{y'}\log(m)\right)
\end{aligned}
$$

where $y'$ is drawn from a Geometric distribution ($y' \sim Geo(p)$).

· **Logarithmic enveloping bounds $\nu \geq 1$:**

$$\begin{aligned}
\alpha^{\nu\geq 1} &= \frac{\left(\mu^{y'}/y'!\right)^{\nu}}{B^{\nu\geq 1}_{f/g}\left(\mu^{y'}/y'!\right)} \\
\log(\alpha^{\nu\geq 1}) &= \nu \log\left(\mu^{y'}/y'!\right) - \Big(\log(B^{\nu\geq 1}_{f/g}) + \log\left(\mu^{y'}/y'!\right)\Big) \\
&= (\nu-1)\log\left(\mu^{y'}/y'!\right) - \log(B^{\nu\geq 1}_{f/g}) \\
&= (\nu-1)\Big(y'\log(\mu) - \sum_{m=1}^{y'}\log(m)\Big) - (\nu-1)\Big(\lfloor\mu\rfloor\log(\mu) - \sum_{n=1}^{\lfloor\mu\rfloor}\log(n)\Big) \\
&= (\nu-1)\Big(\log(\mu)\big(y'-\lfloor\mu\rfloor\big) - \sum_{m=1}^{y'}\log(m) + \sum_{n=1}^{\lfloor\mu\rfloor}\log(n)\Big)
\end{aligned} \tag{A.29}$$

where $y'$ is drawn from a Poisson distribution ($y' \sim Po(\mu)$).

# Bibliography

Abel, G., Bijak, J., Forster, J., Raymer, J. and Smith, P. (2010) What do Bayesian methods offer population forecasters? *Technical report, ESRC Centre for Population Change* .

Alkema, L., Gerland, P., Raftery, A. and Wilmoth, J. (2015) The United Nations probabilistic population projections: an introduction to demographic forecasting with uncertainty. *Foresight (Colchester, Vt.)* **37**, 19–24.

Alkema, L., Raftery, A., Gerland, P., Clark, S. and Pelletier, F. (2008) Estimating the total fertility rate from multiple imperfect data sources and assessing its uncertainty. *Technical report, Centre for Statistics and the Social Sciences, University of Washington* .

Azose, J. J. and Raftery, A. (2016) Estimating large correlation matrices for international migration. *arXiv.org 1605.08759* .

Ballin, M., Scanu, M. and Vicard, P. (2005) Bayesian networks and complex survey sampling from finite populations. *2005 FCSM Conference Papers, Federal Committee on Statistical Methodology, US Office of Management and Budget* .

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **53**, 370–418.

Benson, A. and Friel, N. (2017) Bayesian inference, model selection and likelihood estimation using fast rejection sampling: the Conway-Maxwell-Poisson distribution. *ArXiv* .

Bernardo, J. M. (1996) The concept of exchangeability and its applications. *Far East Journal of Mathematical Science* **4**, 111–122.

Bertino, S. and Sonnino, E. (2003) The stochastic inverse projection and the population of Velletri (1590–1870). *Mathematical Population Studies* **10**, 41–73.

Bijak, J. (2010) *Forecasting international migration in Europe: a Bayesian view.* 24. Springer.

Bijak, J. and Bryant, J. (2016) Bayesian demography 250 years after Bayes. *Population Studies* **70**(1), 1–19.

Bijak, J., Courgeau, D., Silverman, E. and Franck, R. (2014) Quantifying paradigm change in demography. *Demographic Research* **30**(32), 911–924.

Billari, F. (2015) Integrating macro- and micro-level approaches in the explanation of population change. *Population Studies* **69**(sup1), S11–S20.

Bryant, J. and Graham, P. (2015) A Bayesian approach to population estimation with administrative data. *Journal of Offcial Statistics* **31**(3), 475–487.

Bryant, J. and Zhang, J. (2018) *Bayesian demographic estimation and forecasting.* Chapman and Hall/CRC.

Bryant, J. R. and Graham, P. J. (2013) Bayesian demographic accounts: subnational population estimation using multiple data sources. *Bayesian Analysis* **8**, 591–622.

Cancho, V. G., De Castro, M. and Rodrigues, J. (2012) A Bayesian analysis of the Conway-Maxwell-Poisson cure rate model. *Statistical papers* **53**, 165–176.

Carter, C. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

Chanialidis, C., Evers, L., Neocleous, T. and Nobile, A. (2014) Retrospective sampling in MCMC with an application to COM-poisson regression. *Statistics* **3**(1), 273–290.

Chanialidis, C., Evers, L., Neocleous, T. and Nobile, A. (2017) Efficient Bayesian inference for COM-Poisson regression models. *Statistics and computing* pp. 1–14.

Congdon, P. (2008) Models for migration age schedules: a Bayesian perspective with an application to flows between Scotland and England. In *International migration in Europe: data, models and estimates*, eds J. Raymer and F. Willekens. Wiley and Sons.

Consul, P. C. (1989) *Generalized Poisson distributions: properties and applications.* Marcel Dekker: New York, NY.

Conway, R. W. and Maxwell, W. L. (1962) A queuing model with state dependent service rates. *Journal of Industrial Engineering* **12**, 132–136.

Cornuet, J., Marin, J.-M., Mira, A. and Robert, C. P. (2012) Adaptive multiple importance sampling. *Scandinavian Journal of Statistics* **39**(4), 798–812.

Daponte, B., Kadane, J. and Wolfson, L. (1997) Bayesian demography: projecting the Iraqi Kurdish population, 1977-1990. *Journal of the American Statistical Association* **92**(440), 1256–1267.

Davison, A. C. (2003) *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

De Finetti, B. (1965) Sull'oppurtunità di perfezionamenti e di estensione di funzioni dei servizi anagrafici. In *Problemi di rilevazione e classificazione dei dati demografici*, ed. U. di Roma, pp. 3–26. Facoltà di scienze statistiche ed attuariali.

Del Castillo, J. and Pérez-Casany, M. (2005) Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference* **134**, 486–500.

Dellaportas, P. and J., F. J. (1999) Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633.

Dunson, D. B. and Xing, C. (2009) Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**, 1042–1051.

Egidi, V. and Ferruzza, A. (2009) *Navigando tra le fonti demografiche e sociali*. Istituto nazionale di statistica.

Fienberg, S. E. (2011) Bayesian models and methods in public policy and government settings. *Statistical Science* **26**(2), 212–226.

Freedman, D. A. and Navidi, W. C. (1986) Regression models for adjusting the 1980 census (with discussion). *Statistical Science* **1**, 1–39.

Friberg, I. O., Krantz, G., Määttä, S. and Järbrink, K. (2016) Sex differences in health care consumption in sweden: A register-based cross-sectional study. *Scandinavian Journal of Public Health* **44**(3), 264–273.

Frühwirth-Schnatter, S. (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**(3), 515–533.

Gelman, A. and Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press Cambridge.

Gelman, A. and Rubin, D. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

Gerland, P., Raftery, A. E., Sevcíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., Bay, G., Buettner, T., Heilig, G. K. and Wilmoth, J. (2014) World population stabilization unlikely this century. *Science* **346**, 234–237.

Gillispie, S. B. and Green, C. G. (2015) Approximating the Conway-Maxwell-Poisson distributionnormalizing constant. *Statistics* **49**, 1062–1073.

Guikema, S. D. and Goffelt, J. P. (2008) A flexible count data regression model for risk analysis. *Risk Analysis* **28**(1), 213–223.

Handock, M. S., Gile, K. J. and Mar, C. M. (2014) Estimating hidden population size using Respondent-Driven Sampling data. *Electronic journal od Statistics* **8**(1), 1491–1521.

Huang, A. (2017) Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts. *ArXiv* .

INE Spain (2011) *Demographic census project 2011.* Instituto Nacional de Estadistica.

Istat (2015a) *Indagine di copertura del 15 censimento generale della popolazione e delle abitazioni. Nota informativa.* Istat, Via Cesare Balbo, 16 - Roma.

Istat (2015b) Storia delle fonti popolazione. *http://seriestoriche.istat.it* .

Istat (2016) *La revisione post censuaria delle anagrafi: 2012-2014.* Istat, Via Cesare Balbo, 16 - Roma.

Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S. and Boatwright, P. (2006) Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis* **1**(2), 363–374.

King, R. and P., B. S. (2001) On the Bayesian analysis of population size. *Biometrika* **88**(1), 317–336.

Kunihama, T. and Dunson, D. B. (2013) Bayesian modeling of temporal dependence in large sparse contingency tables . *Journal of the American Statistical Association* **108**(504), 1324–1338.

Kupiszewski, M. (2002) The role of international migration in the modelling of population dynamics. *Warsaw: Institute of Geography and Spatial Organisation, Polish Academy of Sciences* .

Laplace, P.-S. d. (1781) Mémoire sur les probabilités. *Mémoires de l'Académie Royale des Sciences de Paris, Année 1778* pp. 227–332.

Lee, R. (2006) Mortality forecasts and linear life expectancy trends. In *Perspectives on mortality forecasting. III. The linear rise in life expectancy: history and prospects*, pp. 19–39. Social insurance studies, no. 3, edited by t. bengtsson. stockholm: Swedish social insurance agency edition.

Lee, R. D. (2001) *Demography abandons its core.* Berkeley,CA.

Lee, R. D. and Carter, L. R. (1992) Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association* **87**, 659–671.

Lee, R. D. and Tuljapurkar, S. (1994) Stochastic population forecasts for the United States: beyond high, medium, and low. *Journal of the American Statistical Association* **89**, 1175–1189.

Lehman, E. L. and Casella, G. (1998) *Theory of pount estimation.* Second edition edition. Springer.

Leslie, P. H. (1945) On the use of matrices in certain population mathematics. *Biometrika* **33**, 183–212.

Leslie, P. H. (1948) Some further notes on the use of matrices in population mathematics. *Biometrika* **35**, 213–245.

Lewis, E. G. (1942) On the generation and growth of a population. *Sankhya: The Indian Journal of Statistics (1933–1960)* **6**, 93–96.

Lord, D., Guikema, S. and Reddy Geedipally, S. (2008) Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention* **40**, 1123–1134.

Lynch, S. M. and Brown, J. S. (2010) Obtaining multistate life table distributions for highly refined subpopulations from cross-sectional data: A Bayesian extension of Sullivan's method. *Demography* **47**(4), 1053–1077.

Madigan, D. and York, J. C. (1997) Bayesian methods for estimation of the size of a closed population. *Biometrika* **85**, 19–31.

Minka, T. P., Shmueli, G., Kadane, J., Borle, S. and Boatwright, P. (2003) Computing with the com-poisson distribution. In *Technical Report Series.* Carnegie mellon university department of statistics, pennsylvania edition.

Moller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**(2), 451–458.

Murray, I., Ghahramani, Z. and MacKay, D. J. C. (2006) MCMC for doubly-intractable distributions. In *In Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 359–366. AUAI Press edition.

Oeppen, J. (2006) Life expectancy convergence among nations since 1820: Separating the effects of technology and income. In *Perspectives on mortality forecasting. III. The linear rise in life expectancy: history and prospects*, pp. 55–82. Social insurance studies, no. 3, edited by t. bengtsson. stockholm: Swedish social insurance agency edition.

Oeppen, J. and Vaupel, J. W. (2002) Broken limits to life expectancy. *Science* **296**, 1029–1031.

Oeppen, J. and Vaupel, J. W. (2006) The linear rise in the number of our days. In *Perspectives on mortality forecasting. III. The linear rise in life expectancy: history and prospects*, pp. 9–18. Social insurance studies, no. 3, edited by t. bengtsson. stockholm: Swedish social insurance agency edition.

Pflaumer, P. (1988) Confidence Intervals for Population Projections Based on Monte Carlo Methods. *International Journal of Forecasting* **4**, 135–142.

Prado, R. and West, M. (2010) *Time series: modeling, computation, and inference.* CRC Press.

Preston, S., Heuveline, P. and Guillot, M. (2001) *Demography: modelling and measuring population processes.* Oxford: Blackwell.

Radford, N. (2003) Slice sampling. *Annals of Statistics* **31**(3), 705–767.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P. and Heilig, G. K. (2012) Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences* **109**(35), 13915–13921.

Rao, J. N. K. (2003) *Small area estimation.* Wiley, New York. MR1953089.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W. F. and Bijak, J. (2013) Integrated modeling of european migration. *Journal of the American Statistical Association* **108**(503), 801–819.

Rees, P. H. (1979) Regional population projection models and accounting methods. *Journal of the Royal Statistical Society, Series A: General* **142**, 223–255.

Rendall, M. S., Handcock, M. S. and Jonsson, S. H. (2009) Bayesian estimation of Hispanic fertility hazards from survey and population data. *Demography* **46**(1), 65–83.

Rettaroli, R. (2011) The evolution of demographic studies. an overview from the '80s: from a macro descriptive perspective to dynamic analysis of biographies. *Statistica* **71**(1), 9–22.

Robert, C. P. (2015) *The Metropolis–Hastings algorithm*, pp. 1–15. American Cancer Society.

Salgado, D. (2016) *A modern vision of official statistical production*. Instituto Nacional de Estadistica.

Sellers, K. F., Borle, S. and Shmueli, G. (2011) The COM-Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Businees and Industry* **28**, 104–116.

Sellers, K. F. and Shmueli, G. (2010) A flexible regression model for count data. *The Annals of Applied Statistics* **4**(2), 943–961.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005) A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistics Society. Series C (Applied Statistics)* **54**(1), 127–142.

Spencer, B. D., May, J., Kenyon, S. and Seeskin, Z. (2017) Cost-benefit analysis for a quinquennial census: The 2016 population census of South Africa. *Journal of Official Statistics* **33**(1), 249–274.

Stone, R. (1986) Nobel memorial lecture 1984: The accounts of society. *Journal of applied econometrics* **1**, 5–28.

Tancredi, A. and Liseo, B. (2011) A hierarchical bayesian approach to record linkage and size population problems. *The Annals of Applied Statistics* **5**, 1553–1585.

Toti, S., Lipsi, R. M. and Giavante, S. (2017) Regional population estimation with Italian administrative data: Preliminary results of the Bryant and Graham approach. Presented at SIS Bayes 2017 Rome.

Trevisani, M. and Torelli, N. (2004) Small area estimation by hierarchical Bayesian models: Some practical and theoretical issues. In *Atti della XLII Riunione Scientifica, Società Italiana di Statistica*, p. 273–276.

Tuljapurkar, S. and Boe, C. (1999) Validation, probability-weighted priors, and information in stochastic forecasts. *International Journal of Forecasting* **15**, 259–271.

Tuljapurkar, S. and Lee, R. (1997) Demographic uncertainty and the stable equivalent population. *Mathematical and Computer Modelling* **26**(6), 39–56.

Tuoto, T., Attili, M., Burgio, A., Cotroneo, R., Iaccarino, C., Prati, S., Rinesi, F., Rottino, F., Tosco, L. and Valentino, L. (2015) Fecondità e maternità: un sistema integrato per la misurazione di fenomeni sanitari e socio-demografici. *Rivista di Statistica Ufficiale* **3**.

UNECE report (2006) *Conference of European statisticians recommendations for the 2010 censuses of population and housing.* United Nations publication.

Vaupel, J. ad Missov, T. (2014) Unobserved population heterogeneity: A review of formal relationships. *Demographic research* **31**(22), 659–686.

Wang, Y., Hunt, K., Nazareth, I., Freemantle, N. and Petersen, I. (2013) Do men consult less than women? an analysis of routinely collected uk general practice data. *BMJ Open* **3**(8).

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571–3591.

Wheldon, M., Raftery, A., Clark, S. and Gerland, P. (2010) A Bayesian model for estimating population size and demographic parameters with uncertainty. *Annual Meeting Population Association of America, Dallas* .

Wheldon, M., Raftery, A., Clark, S. and Gerland, P. (2012) Reconstructing population dynamics of the recent past, with uncertainty, from fragmentary data. *Annual Meeting Population Association of America, Dallas* .

Wheldon, M., Raftery, A., Clark, S. and Gerland, P. (2013) Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association* **108**(501), 96–110.

Wheldon, M. C., Clark, S. J., Raftery, A. E. and Gerland, P. (2015) Bayesian reconstruction of two-sex populations by age: estimating sex ratios at birth and sex ratios of mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**(4), 977–1007.

Wheldon, M. C., Raftery, A. E., Clark, S. J. and Gerland, P. (2016) Bayesian population reconstruction of female populations for less developed and more developed countries. *Population Studies* **70**(1), 21–37.

Wiśniowski, A., Bijak, J., Christiansen, S., Forster, J. J., Keilman, N., Raymer, J. and Smith, P. W. F. (2013) Utilising expert opinion to improve the measurement of international migration in Europe. *Journal of Official Statistics* **29**(4), 583–607.

Wu, G., Holan, S. H. and Wikle, C. K. (2013) Hierarchical Bayesian Spatio-Temporal Conway-Maxwell Poisson models with dynamic dispersion. *Journal of Agricultural, Biological, ad Environmental Statistics* **18**(3), 335–356.

Yildiz, D. and Smith, P. W. (2015) Models for combining aggregate-level administrative data in the absence of a traditional census. *Journal of Official Statistics* **31**(3), 431–451.

Zhang, L. (2011) A unit-error theory for register-based household statistics. *Journal of Official Statistics* **27**, 415–432.

# Charlotte Taglioni

CURRICULUM VITAE

## Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: charlotte.taglioni@phd.unipd.it

## Current Position

*Since October 2015; (expected completion: February 2019)*
**PhD Student in Statistical Sciences, University of Padova.**
*Thesis title: Bayesian hierarchical modelling for population size estimation: application to Italian data*
Supervisor: Prof. Brunero Liseo

## Research interests

- Bayesian modelling
- Bayesian demography
- Copula modelling
- R programming

## Education

*October 2011 – January 2015*
**Master (***laurea specialistica/magistrale***) degree in Finance**.
University of Rome "La Sapienza", Faculty of Economics
Title of dissertation: "Analisi delle dipendenze nei rendimenti finanziari mediante "vine copula"
"Dependence analysis of financial returns using vine copulas "
Supervisor: Prof. Brunero Liseo
Final mark: 110/110 *cum laude*

*October 2008 – October 2011*
**Bachelor degree (***laurea triennale***) in Business and Economics**.
University of Rome "La Sapienza", Faculty of Economics
Title of dissertation: "Problematiche di finanziamento dell'avvio delle libere professioni"
"Funding problems in starting a new business"
Supervisor: Prof. Brunero Liseo
Final mark: 110/110 *cum laude.*

## Visiting periods

*September 2016 – June 2017*
University of Rome “La Sapienza”
Rome, Italy.
Supervisor: Prof. Brunero Liseo

*June 2017 – March 2018*
University of Canterbury,
Christchurch, New Zealand.
Supervisor: Prof. John Bryant

## Research experience

*May 2015 – August 2015*
Erasmus+ Traineeship
Plymouth University
Working on copula models
Supervisor: Prof. Julian Stander

## Computer skills

- Office suit
- R
- LateX
- GitHub repository
- Matlab and Python (basics)

## Language skills

Language1: Italian (native); Language2: French (native); Language3: English (fluent); Language3: Spanish (good).

## Publications

### Working papers

Stander, J., Dalla Valle, L., (2018). Analysis of paediatric visual acuity using Bayesian copula models with sinh-arcsinh marginal densities. *Statistics in Medicine.*

## Conference presentations

Stander, J., Dalla Valle, L., (2016). Using statistical models to understand child eye development. (poster) *(ISBA 2016 World Meeting)*, Cagliari, Italy, 13-17 June 2016.

## Teaching experience

*January 2017 – February 2017*
Introduction to R programming
European Doctorate School of Demography
Lab and exercises, 30 hours
University of Rome "La Sapienza"

## Other Interests

Volunteering, indoor and outdoor exercise, economics, social and environmental debate.

## References

**Prof. Brunero Liseo**
University of Rome "La Sapienza"
MEMOTEF Department, Faculty of Economics, Via del Castro Laurenziano, 9, 00161 Rome, Italy
Phone: +390649766973
e-mail: brunero.liseo@uniroma1.it

**Prof. John Bryant**
Statistics New Zealand
120 Hereford St, Christchurch Central, Christchurch 8011, New Zealand
Phone: +64 3-964 8700
e-mail: john.bryant@stats.govt.nz