Small Cohort Population Forecasting via Bayesian Learning

SIMON WALLIN

Master's Degree Project
Stockholm, Sweden June 9, 2017

Examiner and Supervisor at KTH: Tatjana Pavlenko
Collaborators: BizOne, Täby, Värmdö and Haninge Kommun

Contents

Abstract

A set of distributional assumptions regarding the demographic processes of birth, death, emigration and immigration have been assembled to form a probabilistic model framework of population dynamics. This framework was summarized as an Bayesian network and Bayesian inference techniques are exploited to infer the posterior distributions of the model parameters from observed data. The birth, death and emigration processes are modelled using a hierarchical beta-binomial model from which the inference of the posterior parameter distribution was analytically tractable. The immigration process was modelled with a Poisson type regression model where posterior distribution of the parameters have to be estimated numerically. This thesis suggests an implementation of the Metropolis-Hasting algorithm for this task. Classification of incomings into subpopulations of age and gender is subsequently made using an Dirichlet-multinomial hierarchic model, for which parameter inference is analytically tractable. This model framework is used to generate forecasts of demographic data, which can be validated using the observed outcomes. A key component of the Bayesian model framework used is that is estimates the full posterior distributions of demographic data, which can take into account the full amount of uncertainty when forecasting population growths.

Befolkningsprognoser för små kohorter genom Bayesiansk inlärning

Sammanfattning

Genom att använda en mängd av distributionella antaganden om de demografiska processerna födsel, dödsfall, utflyttning och inflyttning har vi byggt ett stokastiskt ramverk för att modellera befolkningsförändringar. Ramverket är kan sammanfattas som ett Bayesianskt nätverk och för detta nätverk introduceras tekniker för att skatta parametrar i denna uppsats. Födsel, dödsfall och utflyttning modelleras av en hierarkisk beta-binomialmodell där parametrarnas posteriorifördelning kan skattas analytiskt från data. För inflyttning används en Poissontyp regressionsmodell där parametervärdenas posteriorifördelning måste skattas numeriskt. Vi föreslår en implementation av Metropolis-Hastings algoritmen för detta. Klassificering av subpopulationer hos de inflyttande sker via en hierarkisk Dirichlet-multinomialmodell där parameterskattning sker analytiskt. Ramverket användes för att göra prognoser för tidigare demografisk data, vilka validerades med de faktiska utfallen. En av modellens huvudsakliga styrkor är att kunna skatta en prediktiv fördelning för demografisk data, vilket ger en mer nyanserad pronos än en enkel maximum-likelihood-skattning.

1 Introduction

1.1 Background

Since the 17th century when John Graunt pioneered the field of demography with life tables [1], population forecasting has been one of its core problems. The ability to make accurate predictions of future populations is paramount for societal planning. Several Swedish municipal governments have faced difficulties in making accurate predictions, in large part due to difficulty to model the complex dynamics of population levels. The current procedures used by the municipal governments are mostly point estimations of future population, with occasional scenario testing based on heuristic methods [25]. Advances in computational power has made many statistical learning methods feasible to a larger range of problems in the last few decades, and the purpose of this study is to apply these techniques to create a model that accurately predicts future population and their uncertainties. This thesis is a part of a joint project between business intelligence consultant firm *BizOne*, the three municipalities *Värmdö*, *Täby* and *Haninge kommun* of Stockholm county and we hope that this project will result in a tool that can be used by Swedish municipal governments that will facilitate parts of their planning operations. The goal is to develop a general framework that can be applied to predict future probability distribution of a wide ranges of populations going down to specific age groups in small districts, as long as appropriate demographic data is available.

1.2 Previous work

A common method in population forecasting is the *cohort component model* [21]. A *cohort* is in general defined as a subset of a population that share a set of common traits. Gender, age and place of residence are common traits that is used to partition a population into cohorts. The cohort component model makes cohort-specific point estimates of the rates of the demographic processes, which are birth, death and migration. The main issue is to estimate these rates and to project their future values. Assuming that the rates will remain at the same level as previous few years yields somewhat accurate predictions of fertility rates and mortality rates. It has been proven to be a too blunt tool to accurately model migration on local levels since these processes vary from year to year, as is shown in figure 1. Another limitation is that it does not contain any information about uncertainties of the estimates, which is argued to essential for population forecasts by [12].

Time series analysis have been used to model long term large scale population dynamics. One of the more prominent examples of this is the non-parametric Lee-Carter model for national mortality rates in the U.S., which is an ARIMA-model of historic rates with age specific impact factors [7]. This model has been extended into the Bayesian realm by some authors [15][24]. These authors [20] have used a Bayesian framework for estimating population levels based on demographic data in an approach similar to this thesis. They use the framework for modelling between-census populations, and they discuss the possibilities of extending their framework into predictive estimations as well. These authors have all developed models have been readily applied to large scale population pre-

dictions, i.e. on national level or national subdivisions with large populations. However this study focuses on local prediction with mainly small population. Many assumptions of these models become hard to justify, such as using the Poisson distribution as a limiting case of the Binomial distribution for modelling individual probabilities of giving birth, dying etc. They however provide a strong foundation for demographic modelling and this study will largely build upon modified versions of the work of these authors.

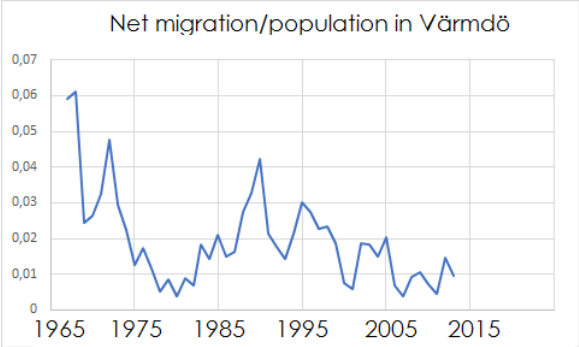

Figure 1: Very unstable migration rate is a typical situation on local level, this example comes from Värmdö municipality outside of Stockholm

1.3 Thesis overview

A general framework for population prediction has been developed based on the cohort component method fortified with Bayesian statistical methods. The populations, number of births, deaths and number of migrations for each cohort are modelled as random variables whereas their values in past time are seen as realizations of these random variables. Bayesian inference is then used to estimate posterior distributions of these random variables given some suitably chosen prior distributions and the expected value of the posterior distribution is taken as the forecast of that cohort-component. Information about uncertainties are captures by confidence intervals which can be generated from the posterior distributions. The population prediction framework developed in this study has been applied on historic data from three Swedish Municipalities in order to make projections for cohort specific population levels. These cohorts span from having having population levels of zero to 2.5 million individual. Hence, the model has been evaluated on a varied range of populations. The projections has been evaluated in discrete time, where one sample point is the last of December of each year.

In the section *population dynamics*, we present the key definitions regarding demography that is used in this thesis. We also present the mathematical relation between these definitions, and define the stochastic processes and distributional assumptions for the demographic entities. In the section *parameter estimation* we present the how the key parameters that govern the demographic entities are estimated using historic data and theoretically justify the methods used for statistical inference. In the section *applications*, we show the model framework

used on real demographic data for population levels of different subpopulations of Stockholm county. We also compare the forecasts made by the model to actual outcomes of population levels to asses the accuracy of the assumptions made. In the section *discussion*, we discuss the results and possible improvements of the framework. Finally in the *appendix* section, we summarize the definitions, notation and introduce and prove concepts that would otherwise disrupt the flow of the thesis.

2 Population dynamics

2.1 Definitions

A *cohort* is defined as a set of individuals that share a set of common traits.[1] A *population*, is defined as a non-negative integer representing the cardinality of the cohort, ie the number of individuals that make up the cohort. Populations vary through time and can be seen as functions from continuous time to the natural numbers $\mathbb{N}$. Let $\mathbf{N}$ denote a set of populations in some future time with unknown values and $\mathbf{n}$ denote a set of populations in past or present times with known values. $\mathbf{N}$ can hence be regarded as a multivariate random variable and $\mathbf{n}$ is a realization of a multivariate random variable. The goal of the project is to model the probability of future populations given past populations, i.e. $p(\mathbf{N}|\mathbf{n})$.

A *demographic process* D is defined as the factors that causes changes in populations and they occur at specific points in continuous time. The demographic processes considered in this study are aging, birth denoted F, death denoted M, immigration denoted P and emigration denoted Q.[2]. The *population at risk* of a demographic process is defined as the set of individuals with a non-zero probability of going through the demographic process. The population at risk are different for various demographic processes. Using the standard female dominant model of birth [10], the population at risk for the birth process is the set of females of fertile age, which in this study is regarded as 14-49 years of age. Births outside of this range do occur but are rare enough to be neglected without significantly affecting the accuracy of the model.

Each demographic process has a different effect on populations. Aging transfers individuals to populations consisting of people one year older. Migration transfers people to populations at new locations. Birth creates new individuals to the population consisting of people of age zero having the same location as the mother (usually) and death will decrease the population. Aging is a deterministic process that occurs to every member of a population. The other processes are however modelled as stochastic processes.

Note that within the scope of this project, immigration and emigration is defined as people moving from one location to another, independently of how the geographic location at interest is defined. This definition will coincide with the

[1] Examples of traits includes being of a set of ages, living in some location, having completed some level of education etc.

[2] This is not an exhaustive list of all possible demographic processes since they depend on the conditions used to define a population. For example, if we consider the population of married people, divorce is demographic process that decreases its population count.

common definition of emigration and immigration where people move in and out of different countries, if the populations at interest are the population of countries, but this is not necessarily the case and will not be the case in the application of the model framework presented in section 4.

Depending on which conditions are used to define a population, different demographic processes will affect the dynamics of the population. To mention a few examples of this, a population covering all ages will not change over time due to aging, a population covering all geographic locations will not change over time due to immigration or emigration and populations that exclusively consist of ages larger than zero will not be affected by birth.

A *demographic variable* is a positive integer that for each cohort indicates the number of occurrences of the corresponding demographic process between two points in time, such that the effect of the demographic process has changed the population of a cohort.[3] The sampling times for demographic variables and population counts is conveniently set to once per year, which will have the advantage that for each step in the process, every individual have aged exactly one year. This is also the sampling rate of the demographic data that the application of this framework is used on. In this study, I will use integers t as an index to denote time, where . The different demographic variables and their interpretations are:

- Birth counts: Denotes the number of new born individuals that between times $t-1$ and t where registered in the location of the cohort at the time of birth. This is regardless of whether they still live there at sampling time t.
- Death counts: Denotes the number of reported deaths of individuals between times $t-1$ and t that would have reached the age of the cohort at time t if they would not have died and who at the time of death were registered in the location of the cohort.
- Immigration counts: Denotes the number of occasions where individuals have registered to live at a new address within the location of the cohort at any point between times $t-1$ and t. This includes individuals who move between locations within the cohort.
- Emigration counts: Denotes the number of occasions where individuals who are registered on an address within the location of the cohort register at any other address at any point between times $t-1$ and t. This includes individuals who move between locations within the cohort.

The reason that migration within cohorts is included in the immigration and emigration counts is to obtain the property that the sum of any demographic variables of a set of mutually disjoint cohorts will equal the value of the demographic variable for another cohort, defined as the union of the original cohorts, which is mathematically expressed as (6) in the next section. This is also how the demographic data used in this study was collected. The consequence of this is that if an individual moves between two locations within the cohort, it will

[3]Note that this definition differs from the definition used by [24]

add one to the emigration count and one to the immigration count.

In analogy with the notation for population counts, let $\mathbf{D}$ denote a set of demographic variables in time periods containing future time with unknown values and $\mathbf{d}$ denote a set of demographic variables in past time periods with known values. Ones again, $\mathbf{D}$ can be regarded as a multivariate random variable and $\mathbf{d}$ can be regarded as a realization of a multivariate random variable. The purpose of introducing demographic variables is to be able to work with the joint probability of future population counts and demographic variables given old realizations, i.e. $p(\mathbf{N}, \mathbf{D}|\mathbf{n}, \mathbf{d})$ which can be marginalized to obtain $p(\mathbf{N}|\mathbf{n})$. As will become clear in the subsequent sections of this paper, the joint probability $p(\mathbf{N}, \mathbf{D}|\mathbf{n}, \mathbf{d})$ is more intuitive to model compared to modelling $p(\mathbf{N}|\mathbf{n})$ directly.

2.2 Model of populations

In this study, the qualifying variables used to define cohorts are age (denoted by $j \in \wp(0, 1, ..., \mathcal{J})$ where $\wp(*)$ denotes the power set of a set $*$, each integer denote the maximum age achieved during the given time period and $\mathcal{J}$ denotes the cap age category that includes people older than or equal to $\mathcal{J}$ years old), gender (denoted by $k \in \{0, 1, \{0, 1\}\}$, where 1 indicate *male*, 0 indicates *female* and $\{0, 1\}$ indicate both genders are included in the cohort) and geographic location (denoted by $l \in \wp(L)$, where L denote some set of non-overlapping locations). Let N_{jklt} denote the population count of a cohort consisting of people of ages j, genders k living in locations l at time t and let $D_{jklt} \in \{F_{klt}, M_{jklt}, P_{jklt}, Q_{jklt}\}$ denote a demographic variable for cohort $jklt$, where where F_{klt} denotes the number of people being born, M_{jklt} denotes the number of people dying, P_{jklt} denotes the number of people moving into an area and Q_{jklt} denotes the number of people moving out between time periods t and $t-1$ in cohort $jklt$. The missing age index from F_{klt} corresponds to every one being born is of age zero. In later parts of the thesis, I will sometimes denote birth counts with an age index as F_{jklt}. In those cases, the age index j denotes the age of the mother and the interpretation is the number of babies of gender k born in location l between times $t-1$ and t by a mother of age k. It hence follows that for all j, j and t, we have $\sum_{j=14}^{49} F_{jklt} = F_{klt}$

If we consider population counts recorded at regular one year intervals between time points, the relation between cohort population counts between two adjacent points in time is, for $j = 1, ..., \mathcal{J} - 1$.

$$N_{jklt} = N_{(j-1)klt-1} - M_{jklt} + P_{jklt} - Q_{jklt} \tag{1}$$

$N_{(j-i)klt-1}$ corresponds to the initial population of the cohort and the $j-1$ accounts for the fact that each member of the cohort was one year younger at the last time point. This decomposition is presented in [24]. For $j = 0$, we have no initial population. Hence:

$$N_{0klt} = F_{klt} - M_{0klt} + P_{0klt} - Q_{0klt} \tag{2}$$

For members of the cap age category, the existing members must also be included since it covers people of age $\mathcal{J}$ and above. Hence, for $j = \mathcal{J}$:

$$N_{\mathcal{J}klt} = N_{(\mathcal{J}-1)klt-1} + N_{\mathcal{J}klt-1} - M_{\mathcal{J}klt} + P_{\mathcal{J}klt} - Q_{\mathcal{J}klt} \tag{3}$$

To generalize these notions, we will introduce the notion of *source cohort* of a cohort γ, which denotes the cohort consisting of the same individuals as γ at time $t-1$ if no demographic process except aging would have occurred. Let γ^* denote the source cohort of γ. If age is not a cohort defining factor, we simply have $\gamma = \gamma^*$, otherwise if $\gamma = \{jkl\}$ then $\gamma^* = \{(j-1)kl\}$ for $j = 1, ..., \mathcal{J}-1$, $\gamma^* = \varnothing$ for $j = 0$ and $\gamma^* = \{jkl\} \cup \{(j-1)kl\}$ for $j = \mathcal{J}$. By using this notation, (1), (2) and (3) can be compactly expressed as:

$$N_{\gamma t} = N_{\gamma^* t-1} + F_{\gamma t} - M_{\gamma t} + P_{\gamma t} - Q_{\gamma t} \tag{4}$$

It follows that $F_{\gamma t} = 0$ for any cohort not containing newborns and $N_{\gamma^* t-1} = 0$ for cohorts that exclusively contains newborns.

Let γ be a cohort and $\gamma_i \quad i = 1, .., I$ be a disjoint partition of γ. Then, for population counts N_γ, $N_{\gamma_i} \quad i = 1, ..., I$ and for any demographic variable D_γ, $D_{\gamma_i} \quad i = 1, ..., I$, we have for all t:

$$N_{\gamma t} = \sum_{i=1}^{I} N_{\gamma_i t} \tag{5}$$

$$D_{\gamma t} = \sum_{i=1}^{I} D_{\gamma_i t} \tag{6}$$

Using (4), (5) (6), we can create networks representing the relations between arbitrary populations in a series of adjacent time steps.

2.3 Model of intrinsic demographic processes

In this thesis, I will use the term *intrinsic* to describe the demographic processes of birth, death and emigration. This is because the populations at risks for them are mainly the source cohorts and they will be modelled similarly. The demographic variables $D_{\gamma t}$ of intrinsic demographic processes can be split into an intrinsic component $D^*_{\gamma t}$ for which the population at risk is the source cohort and an external component $D^{ext}_{\gamma t}$ for which the population at risk is individuals outside of the source cohort:

$$D_{\gamma t} = D^*_{\gamma t} + D^{ext}_{\gamma t} \tag{7}$$

The intrinsic component $D^*_{\gamma t}$ of a demographic variable is modelled by individual probabilities of undergoing a demographic processes. For any cohort γ, each member of its source cohort γ^* is assumed to have some individual probability μ^D_t of going through an intrinsic demographic processes D between time points $t-1$ and t. The individual probability μ^D_t is hereby denoted the *demographic rate*. Hence each individual can be regarded as a Bernoulli trial for each demographic, where the demographic rate is the probability of success.

This Bernoulli trial model of the demographic processes is a somewhat oversimplified representation of reality, since it assumes that for a given individual,

the demographic processes occur independently of each other, which is clearly not the case. The possibility for an individual to undergo multiple demographic processes within the time frame $t-1$ to t depend on the exact times that the processes occured. If an individual dies, no subsequent demographic process can occur and if an individual emigrates, any subsequent demographic process will not enter the demographic count for that cohort. The solution to this problem is to define the demographic rates as the probability for each demographic process to occur before the individual either dies or moves out.

The Bernoulli trial model also implies that an individual cannot undergo a given demographic process more than once per year. This assumption is clearly valid for death and emigration. For births, the assumption valid to a large degree since non-twin births rarely occur more than once per mother within a year and twin births only occur between 9 and 16 times per 1000 births in western Europe according to [17]. The few occurences of multiple births within a year for one woman will create a small bias to the structural interpretation of the individual annual probability of birth μ_t^F, since the model will slightly overestimate the true probability. This will however not affect the predictive power of the model, since the structural interpretation of parameters is irrevant for that purpose.

Let γ denote a cohort that containins individuals that between times t and $t-1$ are modelled to share the demographic rate μ_t^D for some intrinsic demographic variable $D \in \{F, M, Q\}$. Such cohort is hereby denoted a *homogeneous population* or a *homogeneous cohort* with respect to D and their common demographic rate is denoted $\mu_{\gamma t}^D$. Under these assumptions, the intrinsic part of the demographic variable $D_{\gamma t}^*$ is binomially distributed with $N_{\gamma^* t-1}$ number of trials and with $\mu_{\gamma t}$ probability of success, since the sum of any independent Bernoulli trials with a common probability of success is binomially distributed with equivalent parameters [11].

$$D_{\gamma t}^* \in Bin(N_{\gamma^* t-1}, \mu_{\gamma t}^D) \tag{8}$$

This general model is also applicable to cohorts consisting of single individuals, since $Bin(1, \mu^D)$ will simply be Bernoulli distributed with probability μ^D. Trivially, every cohort consisting of single individuals are homogeneous cohort.

2.3.1 Model of non-intrinsic part of intrinsic demographic processes

In this project, the external parts of the death process $M_{\gamma t}^{ext}$ and emigration process $Q_{\gamma t}^{ext}$ are considered neglectable and are hence set to zero for every cohort. This is motivated by exploratory analysis showing that age groups that are likely to die are unlikely to immigrate and by the authors subjective judgement that it is rare for an individual to move more than twice per year. Non-intrinsic death and emigration is probably occurring in some degree, and a slight bias is hence introduced in the structural interpretations of these parameters where the individual probability of death and emigration is slightly underestimated by the parameter. The external component for birth is however not neglected in this project, which will be further developed in the subsequent section.

2.3.2 Model of birth

The non-intrinsic part of birth counts from parents in cohort γ, $F^{ext}_{\gamma t}$ can be modelled using the same individual based approach as for the intrinsic part with the source cohort replaced with the cohort consisting of every individual in γ that does not belong to the source cohort $\gamma*$. The two demographic processes that adds individuals to a cohort except for ageing are birth and immigration. Since new born children are not a population at risk for birth, the immigration count $P_{\gamma_i t}$ will by itself make up the population at risk for non-intrinsic birth counts. If $\gamma_i \quad i = 1, ..., I$ are homogeneous populations that form a partition of γ, then by making use of (5), (6) and (8), we get:

$$F^{ext}_{\gamma t} = \sum_{i=1}^{I} F^{ext}_{\gamma_i t} \tag{9}$$

Where:

$$F^{ext}_{\gamma_i t} \in Bin(P_{\gamma_i t}, \mu^{F^{ext}}_{\gamma_i t}) \tag{10}$$

Where $\mu^{F^{ext}}_{\gamma_i t}$ is the demographic rate of this process, which further on will be assumed to be equal to the birth rate of the source cohort $\mu^{F^*}_{\gamma_i t}$. Now, demographic data seldom comes partitioned into external and internal components. A child is assigned to a cohort regardless of whether or not the parents of the child just moved in. And children that are born to parents that later move into a cohort will not be included in the data. The solution is to include the zero year old people moving into an area into the birth count. By doing that, birth counts of children from parents that are member of a cohort γ, can be modelled as:

$$F_{\gamma t} + P_{\gamma, j=0, t} \in Bin(N_{\gamma^* t-1} + P_{\gamma t}, \mu^F_{\gamma t}) \tag{11}$$

Where $F_{\gamma t}$ denotes the number of registered births in the location of cohort γ between times $t-1$ and t.

2.4 Model of immigration

The population at risk for immigration to a cohort γ consist of every individual in a location within and outside of the source cohort having the same gender and age minus one. Immigration counts can be decomposed into different component based on place origin, a practice which may be usefull if the underlying processes of immigraion from different origins are fundamentally different.[4] The structure of the place of origin will generally determine the appropriate model. If the place of origin is large enough, the immigration rate can be thought of as independent of the number of people living there. This model is used by [1] and [24]. Then the immigration counts $P^{\delta}_{\gamma t}$ for a homogeneous population γ and a place of origin δ can be assumed to be Poisson distributed with rate $\lambda^{\delta}_{\gamma t}$:

$$P^{\delta}_{\gamma t} \in Po(\lambda^{\delta}_{\gamma t}) \tag{12}$$

[4]For example, in the wake of the European 2015 migration crisis, teenagers have been overrepresented in international immigraion compared to national immigration [25].

The parameter is denoted $\lambda^{\delta}_{\gamma\delta t}$ instead of $\mu^{P^{\delta}_{\gamma\delta t}}$ to avoid confusion in structural interpretation. It is argued in [24] that the Poisson distribution is useful for this kind of purpose. If the cohort of origin is too small for the Poisson-limit to be applicable, it is recommended to use (8) where the number of trials is replaces with the source population of the place of origin. Authors such as [20] fixes the Poisson parameters to be proportional to the population of the destination cohort, citing good empirical results using this practice despite lacking a priori motivation for this. This will however severely limit the usefulness of the model framework in cohorts where the majority of influx of people will be driven by new housing development, and author like [24] simply uses an equivalent to (12) to model immigration. An a priori explanation of the correlation between population and immigration counts is that they both correlate with the number of available homes in the area, a factor which clearly can be considered to have an a priori affect on immigration counts. The demographic data used to construct this model framework does not contain information about the number of available homes, and the solution is to use regression models to represent the relationship between the immigration rate and factors that correlate with the number of homes, such as emigration counts, death counts and populations. Emigration counts for each are in the same manner dependent on the number of available homes in the destination cohort. This number can be considered constant for large enough destination cohorts.

Instances of migration cannot be considered independent on one another in the sense that instances of birth and death are. It is assumed that the a posteriori birth rate or death rate of an individual will not change if we include knowledge of outcomes demographic processes for other individuals. They may very well be dependent of demographic rates of other individuals, but the outcome of the demographic processes given the rates contain no information for the birth and death process, as opposed to the migration. For example, the probability of an individual moving into a location will increase for every person actually moving out of that location, since the process creates available housing.

3 Parameter estimation

The model of population dynamics given a set of parameters have been described up till this point. In this section, we are going to explain how the parameters are estimated using historic data. In particular, we will use the Bayesian approach to model the posterior probability of the parameter values given demographic data $p(\boldsymbol{\mu}|\boldsymbol{n}, \boldsymbol{d})$, where $\boldsymbol{\mu}$ denotes a set of parameters. From this, we can derive the probability of future populations and demographic variables given historic data by integrating over the parameters:

$$p(\boldsymbol{N}, \boldsymbol{D}|\boldsymbol{n}, \boldsymbol{d}) = \int p(\boldsymbol{N}, \boldsymbol{D}|\boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{d}) p(\boldsymbol{\mu}|\boldsymbol{n}, \boldsymbol{d}) d\boldsymbol{\mu} \tag{13}$$

Where the fist factor of the integrand is given by equations (4)-(12) and the second factor of the integrand is obtained by the Bayesian inference describes in this section. Neither the inference of the parameters and the forecasting procedure will be analytically tractable and different numerical Monte Carlo simulations will be utilized to perform the calculations. In particular, we will

resort to the *Metropolis-Hastings* algorithm first proposed by [3] and generalized by [6].

Consider the demographic rate μ_t^D of some individual. To infer their probability distribution from historic data, we will resort to two main approaches: Tree based inference and regression. In the tree based inference, a larger population is for each intrinsic demographic variable partitioned into homogeneous cohorts, where each branch having one parameter to fit. The other approach is to use some regression model for the demographic rates $\mu_t^D = g(\gamma, t, \boldsymbol{w}(t))$, where $g(*)$ denotes some function, $\boldsymbol{w}(t)$ denotes the state of environmental factors at time t that is modelled to affect the demographic rates.

3.1 Stationary model

A *stable time period* τ of a parameter is defined as a set of neighbouring time points t where:

- In the tree based model, the demographic rate is considered constant for each point in the time period.
- In the regression model, the regression parameter are considered valid for the entire time period.

For such time period, μ_t^D is written μ_τ^D in the tree based model and we have $g(\gamma, t, \boldsymbol{w}(t)) = g_\tau(\gamma, \boldsymbol{w}(t))$, where the only time dependence within the stable time period comes from fluctuation in the environmental factors. In the stationary model, the forecasting time frame is considered to lay within a stable time period where historic data $\{\boldsymbol{n}, \boldsymbol{d}\}$ exist and probability distributions future parameter values are either the same as when they governed the historic data for the tree based model or deducible through probability distributions of future values of environmental factors for the regression model. The assumption of stable time periods is what makes inference of the future possible and the length of what could be considered a stable time period will vary depending on the model used. Including environmental factors and trend components will generally extend what could be considered a stable time period.

3.2 Tree based hierarchical Beta-Binomial model

For the intrinsic demographic processes, we will consider the tree based hierarchical Beta-Binomial model for inference of the parameters and the specifications of the model will be given in this section. We assume a priori independence between the parameters, meaning that $p(\boldsymbol{\mu}) = \prod_s p(\mu_s)$ where μ_s denotes each element of any parameter vector $\boldsymbol{\mu}$. Using this assumption, we can estimate the posterior distributions of the demographic rates given historic data element-wise for each homogeneous population γ using Bayes theorem [14]:

$$p(\mu_{\gamma\tau}^D | N_{\gamma^*\tau^*}, D_{\gamma\tau}) \propto p(\mu_{\gamma\tau}^D) p(D_{\gamma\tau} | \mu_{\gamma\tau}^D, N_{\gamma^*\tau^*}) \tag{14}$$

In (14), $N_{\gamma^*\tau^*} = \sum_{t\in\tau} N_{\gamma^* t-1}$, representing every person year lived in a homogeneous cohort during a stable time period. We also have $D_{\gamma\tau} = \sum_{t\in\tau} D_{\gamma t}$, corresponding to every occurrence of the demographic period during these person years. As explained in section 2.3.2, we need to include immigration of

new born babies to our demographic variable and the immigration count of potential parents to the source cohort. Hence, when modelling birth we define $N_{\gamma^*\tau^*} = \sum_{t\in\tau} N_{\gamma^* t-1} + P_{\gamma t}$ and $F_{\gamma\tau} = \sum_{t\in\tau} F_{\gamma t} + P_{j=0,\gamma t}$. The expected value of the likelihood function will correspond to the number of person years times the demographic rates. By solving for the cohort specific demographic rate during a stable time period, the interpretation naturally becomes the percentage of a cohort per year that experience a demographic process:

$$\frac{\mathbb{E}[D_{\gamma\tau}|\mu^D_{\gamma\tau}, N_{\gamma^*\tau^*}]}{N_{\gamma^*\tau^*}} = \mu^D_{\gamma\tau} \tag{15}$$

The second factor of the right hand side of 12 represent the likelihood function of the demographic variable given source population and demograpic rate $p(D_{\gamma\tau}|\mu^D_{\gamma\tau}, N_{\gamma^*\tau^*})$. By the model presented in section 2, this is a binomially distributed random variable. When estimating the posterior probabilities of the parameter, we consider the likelihoodfunction for observed historic data that is assumed to lie in the stable time period τ. The first factor of the right hand side of 12 denotes the prior distribution of the parameters. Picking the prior distribution of the demographic rate from the Beta-family will create a closed form expression for the posterior, since it comes from the conjugate family of the likelihood function [11]. We hence pick the following prior:

$$\mu^D_{\gamma_i\tau} \in Beta(\alpha^D_{\gamma_i\tau}, \beta^D_{\gamma_i\tau}) \tag{16}$$

In (16), $\alpha^D_{\gamma\tau}$ and $\beta^D_{\gamma\tau}$ corresponds to the shape parameters of the Beta distribution and are considered hyperparameters. This is what is referred to as a hierarchical model in for example [11]. The subscript and superscript of the shape parameters indicate that they may be unique for each homogeneous cohort, for each demographic variable and for each stable time period. They will henceforth however be dropped for convenience. An illustration of the beta random variable can be found in figure 2. The posterior distribution becomes:

$$\mu^D_{\gamma_i\tau}|N_{\gamma_i^*\tau^*}, D_{\gamma_i\tau} \in Beta(\alpha + D_{\gamma_i\tau}, \beta + N_{\gamma_i^*\tau^*} - D_{\gamma_i\tau}) \tag{17}$$

The proof of this is presented in Appendix A. Using this procedure, closed form posterior distributions of all parameter can be obtained following this pattern. A strength of this approach is that the variance of the posterior distribution will decrease with the number of observation, as the variance of a beta distributed random variable decreases as its parameter values increases, which is clear in figure 2. The beta distribution plotted in the upper right and the two bottom graphs have the same expected value. But as the shape parameter value increases, the variance decreases. This corresponds to the fact that more observations leads to more and more certain assessment of the underlying demographic rate.

3.2.1 Haldane's prior and adaptive stable time periods

Due to the vast number of different parameters in this study that will be estimated on a large set of different populations, I have chosen to work with *noninformative priors* for the tree based hierarchical model. A noninformative prior as described in [14] is a prior distribution that represent no a priori knowledge of the probability distribution of the parameter. An uninformative prior

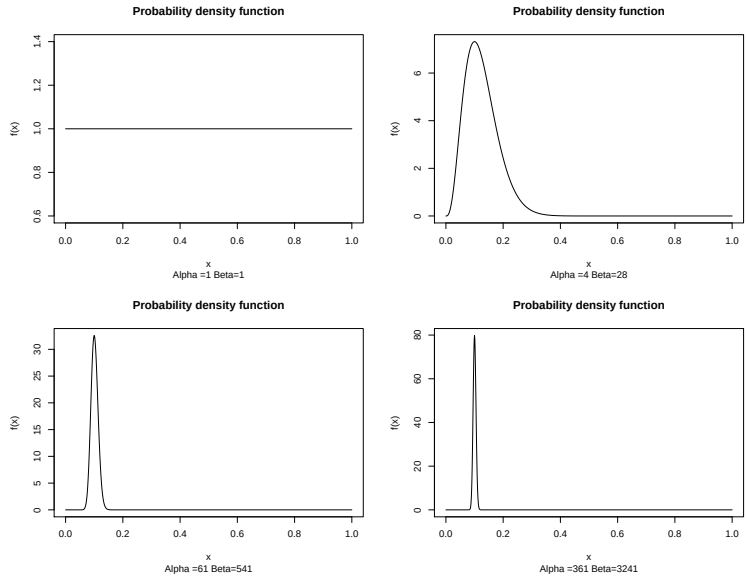

Figure 2: The probability density function of the beta distribtution with four different sets of shape parameters.

proposed by J. B. S. Haldane for the beta-binomial model is to set $\alpha = \beta = 0$ [2]. The beta distribution only allows for strictly positive shape parameters and Haldane's prior can be seen as the limiting case of letting both parameters go to zero. This is an *improper prior* since the prior distribution itself does not constitute a probability distribution whose probability density function integrates to one. This is however no problem as long as the posterior distribution is well defined [19]. For (17) to be a well defined posterior distribution for $\alpha = \beta = 0$ if there is at least one observed occurrence of the demographic processes and at least one member of the source cohort that did not go through the demographic process. In this approach, the posterior probability completely data-driven and it is equivalent to the likelihood function. Further discussion of appropriate prior distribution for beta-binomial model can be found in [19].

Some rare demographic events will have few observed occurrences, such as births by teenagers and deaths of young children. Hence, these events may have no observed occurrence rendering the Haldane's prior infeasible. Even if occurrences is observed, they may be very few, and single occurrences will have a huge impact on the parameter estimated. To avoid this, we will use adaptive stable time period assumptions. If a demographic process have few occurrences for a certain homogeneous cohort, we will include additional years of historic data in the stable time period in hope for finding more observations. In this study, we use a lower limit of four observations of a demographic process for a homogeneous cohort until we stop including more data in the inference process. If no more data is available and there are still fewer than four observations, we use the data we have. If there are no observations after extending the stable time period assumption, we simply assume that the demographic process will

not occur with regard to the stable cohort γ, hence deterministically setting $\mu^D_{\gamma\tau} = 0$. Conversely, if there are only successful trials, we set $\mu^D_{\gamma\tau} = 1$ deterministically. This will only occur occasionally for small homogeneous cohort and is probably a sign that the homogeneous cohort partition is too fine.

3.2.2 The cohort partition problem

A central issue with the tree based approach is which set of individuals that can be considered a homogeneous cohort. In principle, the population of interest for the forecast itself could be considered a homogeneous cohort, and the estimation of the demographic rates for this particular group will accurately reflect the probability of uniformly picking an individual that has gone though a demographic process in a population. This will however completely neglect the differences of demographic rates within the cohort. As the population composition changes with time the demographic rates are likely to shift, making the parameter estimations unstable.

A different approach at the other end of the spectrum is to partition the feature space of individuals such that every individual within the homogeneous cohorts have the same attributes. This is possible for the qualifying variables defined on finite spaces, such as age which is defined as 101 different classes, and gender which are two classes. The location variable and environmental factors are of course difficult to partition in this manner, but one could pick the smallest available unit in the data upon which the parameters are estimated. Demographic data often comes in the form of tables where each cell corresponds to the number of individuals fulfilling conditions specified in the rows and columns, so the feature space is usually already partitioned into some finite set. This lead to a large number of parameter that has to be estimated using few observations, and the model may hence sensitive to overfitting and too strong influence of prior distributions.

The optimal cohort partition clearly is somewhere in between these two extremes. In the population forecasts that is currently conducted by the municipalities of Sweden, each age and gender is considered a homogeneous groups per municipality. Data inspection does in fact reveal that the demographic rates varies significantly between ages, genders and locations, see figure 9.

3.3 Poisson regression model for immigration

The cohort specific immigration counts $\boldsymbol{P}_{\gamma t} = [P^{\delta_1}_{\gamma t}, P^{\delta_2}_{\gamma t}, ...]$ was in the population dynamics sections modelled as a Poison distributed random variable with immigration rate $\boldsymbol{\lambda}_{\gamma t} = [\lambda^{\delta_1}_{\gamma t}, \lambda^{\delta_2}_{\gamma t}, ...]$ for large cohorts of origin $\delta_1, \delta_2, ...$ We will assume that immigration rates are independent of immigration rates of other locations, meaning that $p(\boldsymbol{\lambda}_{l_1}, \boldsymbol{\lambda}_{l_2}...) = \prod_i p(\boldsymbol{\lambda}_{l_i})$, where l_i denotes different non-overlapping cohorts partitioned only on location. This is however not true for the immigration rates for subpopulations within an area. The assumed strong correlation between total migration into an area and the number of available housing there is the reason that each instance immigration can be a priori assumed to decrease the probability of other instances. This is an a priori assumption that may or may not be true for empirical data. But it is enough

to argue that immigration rates within an area should be modelled to have the ability to be dependent on one another. Immigation counts from different source locations to a given cohort can also be considered dependent for the same reason as for subcohorts within a location is dependent, which is motivation for using multivariate models rater than separate univariate models for this.

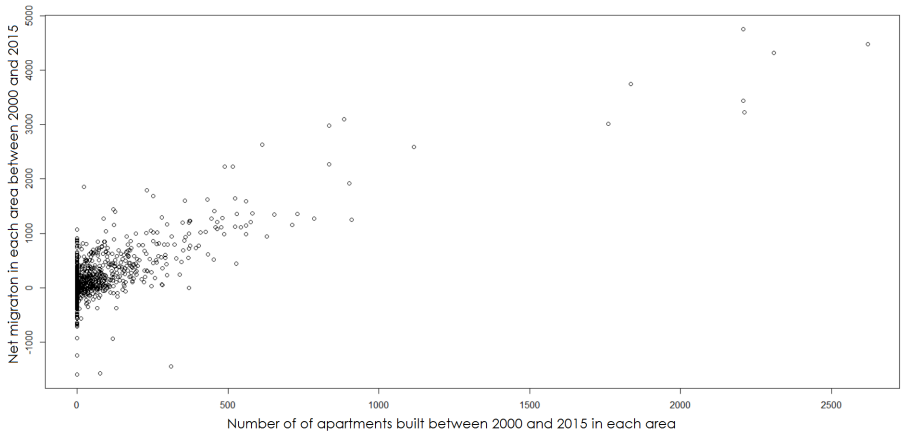

Figure 3: A plot of net migration and number of newly built apartments for 1418 different areas within Stockholm county. The the existence of a correlation is clear

Consider the case where the cohort of interest is simply defined by a location. We will also for simplicity only consider total immigration counts regardless of location of origin. The model could be extended to a multivariate model that capture different behaviours of different location of origins. We wish to estimate the probability distribution of the immigration rate based on the states of environmental factors of the area. Let the vector $\boldsymbol{X}_{lt} = [X_{lt}^1, X_{lt}^2, ..., X_{lt}^K]$ denote the state a set of K environmental factors, or *covariates*, for location l at time t. The standard Poisson linear regression model is usually used for this kind of situation, as stated in [16]:

$$\lambda_{lt} = exp(\beta_{0\tau} + \sum_{k=1}^{K} \beta_{k\tau} X_{lt}^k) \tag{18}$$

In the model above, $\beta_{k\tau}, k = 0, 1, ..., K$ are regression parameters and by using a Bayesian framework, they will be equipped with prior distributions. The subscript τ indicates that the Beta-parameters are assumed to be valid for demographic data from the stable time period τ. The expected value and variance of immigration counts are under this model:

$$\mathbb{E}[P_{lt}] = Var[P_{lt}] = exp(\beta_{0\tau} + \sum_{k=1}^{K} \beta_{k\tau} X_{lt}^k) \tag{19}$$

This model is too strict for our problem due to a number of built in assumptions. The expected immigration counts does not necessarily have a log-linear relationship between the covariates. There is an a priori expected linear relationship

between covariates that contribute to increasing the number of available housing, such as death counts, emigration counts and the number of apartments built.[5] Another limitation is the constraint that the variance is proportional to the expected value. This takes the built in randomness of the Poisson process into account, but neglects randomness of the parameter itself that may arise from unobserved heterogeneity or from intrinsic randomness. The solution is to use a model family called negative binomial models. In particular, the model referred to as the negbin-II model in [16] is particularly handy due to its convenient closed form analytical expression. The negbin-II model assumes that the probability distribution of the immigration rate given covariates is gamma distributed.

$$\lambda_{lt} \in \Gamma(\sigma_\tau, \frac{\sigma_\tau}{g_\tau(\boldsymbol{X}_{lt})}) \tag{20}$$

Where $g_\tau(*)$ is some general positive real-valued regression function that captures the relation between covariates and the immigration rate and σ is a real positive. The subscript indicates that the model parameters are assumed to be valid for the entire stable time period τ. This subscript may be dropped henceforth due to convenience of notation. Central properties is that we have $\mathbb{E}[\lambda_{lt}|\boldsymbol{X}_{lt}, \sigma] = g(\boldsymbol{X}_{lt})$ and $Var[\lambda_{lt}|\boldsymbol{X}_{lt}, \sigma] = g(\boldsymbol{X}_{lt})^2/\sigma$. The structural interpretation of the regression function and the parameter σ becomes clear. Using the law of double expectation [11], we get $\mathbb{E}[P_{lt}|\boldsymbol{X}_{lt}, \sigma] = g(\boldsymbol{X}_{lt})$ and using the law of double variance [11], we get:

$$\begin{aligned} Var(P_{lt}|\boldsymbol{X}_{lt}, \sigma) &= \mathbb{E}[Var(P_{lt}|\lambda_{lt}, \boldsymbol{X}_{lt}, \sigma)] + Var(\mathbb{E}[P_{lt}|\lambda_{lt}, \boldsymbol{X}_{lt}, \sigma]) \\ &= \mathbb{E}[\lambda_{lt}|\boldsymbol{X}_{lt}, \sigma] + Var(\lambda_{lt}|\boldsymbol{X}_{lt}, \sigma) = g(\boldsymbol{X}_{lt}) + \frac{g(\boldsymbol{X}_{lt})^2}{\sigma} \end{aligned} \tag{21}$$

The regression function $g(*)$ can be defined in various ways, as will be presented in the subsequent sections.

3.3.1 Linear regression function

Exploratory data analysis has shown that for subareas of Stockholm county, the number of individuals moving into the area has a strong linear relationship with number of individuals moving out, the number of individuals dying and the number of newly built housing. A crude example of this can be found in figure 3, where the linear trend between new housing development and net migrations is strong. Modelling the relationship between the dependent variable (immigration counts) and these as log-linear would likely create a specification error. Hence, we resort to using a linear regression function. Let $\boldsymbol{H}_{lt} = [H^1_{lt}, H^2_{lt}, ..., H^{K_H}_{lt},]$ denote a vector representing the number of newly built units of housing of types $k_H = 1, 2, ..., K_H$ between times $t-1$ and t. Then, the linear model of environmental variables $\boldsymbol{X}_{lt} = [N_{lt-1}, M_{lt}, Q_{lt}, \boldsymbol{H}_{lt}]$ becomes:

$$g(\boldsymbol{X}_{lt}) = \beta_{0\tau} + \sum_{k=1}^{K} \beta_{k\tau} X^k_{lt} \tag{22}$$

[5] This is as long as the demand for housing somewhat matches the supply, but information regarding this is not present in the value of the variables.

Where $\beta_{k\tau}, \quad k = 1, ...K$ are parameters and $K = 3 + K_H$. The partition of the number of new houses into different categories is used to be able to model different effects of different types of housing. One room studio apartments will probably have a different effect on the population than houses.

The regression function must be positive valued and this fact will impose restrictions on the covariates and the parameters. All the covariates are in the form of positive integers representing count data. Hence, restricting the parameters to non-negative real numbers will be a sufficient condition to assure that the immigration rate λ_{lt} is non-negative. This may sound to restrictive, but there is a priori reasons to allow for this, since all covariates represent processes creating available housing that makes immigration possible. It is hence unreasonable to expect negative parameter values. This restriction can be built in to the prior distributions of $\beta_{k\tau}, \quad k = 1, ..., K$ by setting the prior probability to zero for negative values. Consider a rather common situation in demographic data, where all covariate values are zero. In this situation, we have $\lambda_{lt} = \beta_{0\tau}$ and this will be the structural interpretation of the intercept $\beta_{0\tau}$. That is, the expected number of immigrant for an area without current population and without new housing development.

3.3.2 Priors and posteriors

Let $\boldsymbol{\beta}_\tau = [\beta_{1\tau}, \beta_{2\tau}, ..., \beta_{K\tau}]$ denote the set of regression parameters. The purpose of the parameter estimation is to calculate the posterior parameter distribution function $f(\boldsymbol{\beta}_\tau, \sigma | \boldsymbol{X}_\tau, \boldsymbol{P}_\tau)$, where $\boldsymbol{X}_\tau$ are historic values of the covariates and $\boldsymbol{P}_\tau$ are historic immigration counts for the stable time period τ. This is through Bayes theorem proportional to:

$$f(\boldsymbol{\beta}_\tau, \sigma | \boldsymbol{X}_\tau, \boldsymbol{P}_\tau) \propto f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) f(\boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) \tag{23}$$

The first factor to the right hand side of (23) represents the likelihood function of the data given the parameters. Let $\boldsymbol{\lambda}_\tau$ denote a set of Poisson parameters from the stable time period τ, where the elements of λ matches the elements of $\boldsymbol{P}_\tau$ in terms of the cohorts and times they represent. Now, the law of total probability with regards to the immigration rates can be used to rewrite the likelihood factor:

$$f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) = \int f(\boldsymbol{P}_\tau | \boldsymbol{\lambda}_\tau, \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) f(\boldsymbol{\lambda}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) d\boldsymbol{\lambda}_\tau \tag{24}$$

Immigration counts $\boldsymbol{P}_\tau$ are conditionally independent of the parameters $\boldsymbol{\beta}_\tau$ and σ given the immigration rates $\boldsymbol{\lambda}_\tau$. The probability distribution of immigration counts given immigration rates are independent for each cohort included in the demographic data. The immigration rates given the covariates and parameters are also independent between cohorts. This simplifies the expression to:

$$f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) = \prod_{l,t} \int f(P_{lt} | \lambda_{lt}) f(\lambda_{lt} | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_{lt}) d\lambda_{lt} \tag{25}$$

Each factor of (25) are known from the model and the probability density function of the likelihood function can be calculated with ease. Let $\hat{y}_{lt} =$

$\beta_{0\tau} + \sum_{k=1}^{K} \beta_{k\tau} X_{lt}^{k}$ and let $\Gamma(z)$ denote the gamma function defined by $\Gamma(z) \triangleq \int_0^\infty t^{z-1} e^{-t} dt$. We then have:

$$f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) = \prod_{l,t} \int_0^\infty \frac{\lambda_{lt}^{P_{lt}} e^{-\lambda_{lt}}}{P_{lt}!} \frac{\left(\frac{\sigma}{\hat{y}_{lt}}\right)^\sigma}{\Gamma(\sigma)} \lambda_{lt}^{\sigma-1} exp\Big(- \frac{\sigma \lambda_{lt}}{\hat{y}_{lt}} \Big) d\lambda_{lt} \tag{26}$$

Rearranging and using the identity $P_{lt}! = \Gamma(P_{lt}+1) = P_{lt}\Gamma(P_{lt})$ yields:

$$f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) = \prod_{l,t} \frac{\sigma^\sigma}{P_{lt}\Gamma(P_{lt})\Gamma(\sigma)\hat{y}_{lt}^\sigma} \int_0^\infty \lambda_{lt}^{P_{lt}+\sigma-1} e^{-\lambda_{lt}(1+\sigma/\hat{y}_{lt})} d\lambda_{lt} \tag{27}$$

Now, consider a gamma distributed random variable with shape parameters $P_{lt}+\sigma$ and $1+\sigma/\hat{y}_{lt}$. The probability density function of that random variable will integrate to one, hence:

$$\frac{\left(1+\frac{\sigma}{\hat{y}_{lt}}\right)^{P_{lt}+\sigma}}{\Gamma(P_{lt}+\sigma)} \int_0^\infty \lambda_{lt}^{P_{lt}+\sigma-1} e^{-\lambda_{lt}(1+\sigma/\hat{y}_{lt})} d\lambda_{lt} = 1 \tag{28}$$

From the equation above, we can simplify the integral expression of (27), in that:

$$\int_0^\infty \lambda_{lt}^{P_{lt}+\sigma-1} e^{-\lambda_{lt}(1+\sigma/\hat{y}_{lt})} d\lambda_{lt} = \frac{\Gamma(P_{lt}+\sigma)}{\left(1+\frac{\sigma}{\hat{y}_{lt}}\right)^{P_{lt}+\sigma}} \tag{29}$$

This yields:

$$f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) = \prod_{l,t} \frac{\sigma^\sigma \Gamma(P_{lt}+\sigma)}{P_{lt}\Gamma(P_{lt})\Gamma(\sigma)\hat{y}_{lt}^\sigma (1+\sigma/\hat{y}_{lt})^{P_{lt}+\sigma}} \tag{30}$$

Rearranging and using the relation between the gamma function and the beta-function, where $\Gamma(\sigma+P_{lt})/(\Gamma(\sigma)\Gamma(P_{lt})) = Beta(\sigma, P_{lt})$ we get:

$$f(\boldsymbol{P}_\tau | \boldsymbol{\beta}_\tau, \sigma, \boldsymbol{X}_\tau) = \prod_{l,t} \frac{1}{P_{lt} Beta(\sigma, P_{lt})} \left(\frac{\sigma}{\sigma+\hat{y}_{lt}}\right)^\sigma \left(\frac{\hat{y}_{lt}}{\sigma+\hat{y}_{lt}}\right)^{P_{lt}} \tag{31}$$

We will make use of improper prior distributions of the parameters, but imposing the restriction that only non-negative values are allowed. That is:

$$f(\boldsymbol{\beta}_\tau, \sigma) \propto 1 \quad \forall \boldsymbol{\beta}_\tau, \sigma \geq 0 \tag{32}$$

We can now express the posterior probability distribution of the parameters as, for all non-negative parameter values, we have:

$$f(\boldsymbol{\beta}_\tau, \sigma | \boldsymbol{X}_\tau, \boldsymbol{P}_\tau) \propto \prod_{l,t} \frac{1}{Beta(\sigma, P_{lt})} \left(\frac{\sigma}{\sigma+\hat{y}_{lt}}\right)^\sigma \left(\frac{\hat{y}_{lt}}{\sigma+\hat{y}_{lt}}\right)^{P_{lt}} \tag{33}$$

The omission of $1/P_{lt}$ for the equation above is because this is simply a multiplicative constant when only considering the parameters, which is the only feature of interest of this procedure.

3.4 Inference through Metropolis-Hastings algorithm

As opposed to the Beta-binomial model used for birth, death and emigration, there are no prior distribution family for the parameters $\boldsymbol{\beta}_\tau$ and σ that will yield a closed form posterior distribution of the parameters given the historic data. We must hence resort to approximate the distribution through numerical methods and in particular, we will use the Metropolis-Hastings algorithm for sampling from the joint posterior distribution. The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method which can be used to draw samples from any target distribution for which we know the probability density function up to a constant [6]. The concept of the algorithm is that you start with some initial sample x_0, draw a proposed sample x^* from some proposal distribution that depends on x_0 denoted $r(x^*|x_0)$. Now, we calculate the likelihood ratio of the two samples and if the proposed sample is more likely than the old sample x_0, we set $x_1 = x^*$. If the old sample is more probable, we still set $x_1 = x^*$ with a probability proportional to the likelihood ratio, hence creating a small probability of accepting less likely states. If the proposed sample is not accepted, we set $x_1 = x_0$. We then keep generating new proposed samples from the proposal distribution, but given the last accepted sample [6] for each step. This is a Markov chain method since the probability of the next state is only dependent on the current state. One can show that the Markov chain constructed through this algorithm will have a stationary distribution that coincides with normalized version of the probability distribution we wish to draw from. Each sample is however highly dependent on the previous samples and in order to downplay the role of the arbitrarily picked initial value x_0, one usually discards a number of samples drawn in the beginning. These discarded samples are usually denoted the *burn in*.

Let $\mathcal{X}$ be a random variable, $x \in \Omega$ denote an element of the sample space Ω and $f(x)$ denote the probability density function of the target distribution we wish to draw samples from. The random variable $\mathcal{X}$ may or may not be multivariate. Now define $z(x) = f(x)/C$ for any positive real constant C. The Metropolis Hastings algorithm for drawing samples from is [6]:

Set x_0 to some initial value;
Set n as the number of sample we wish to draw;
Set m as the number of samples to be discarded as burn-in;
for $i \leftarrow 1$ **to** $m + n$ **do**
  Draw $x^* \in r(x^*|x_{i-1})$;
  Set $\alpha = \dfrac{z(x^*)r(x_{i-1}|x^*)}{z(x_{i-1})r(x^*|x_{i-1})}$;
  if $\alpha \geq 1$ **then**
    Set $x_i = x^*$;
  else
    Set $x_i = x^*$ with probability α and $x_i = x_{i-1}$ otherwise;
  end
end
Result: Samples x_i for $i = m + 1, ..., m + n$ can be regarded as n samples from $f(x)$

As the stationary distribution of the Markov-chain coincides with the target

distribution, we know that samples from the Metropolis Hastings algorithm asymptotically approaches samples drawn from the target distribution. This convergence may however be computationally too slow, and the rate of convergence is highly dependent on the structure proposal distribution. Hence, choosing an appropriate proposal distribution is crucial for practical usefulness of the algorithm. A common diagnostics tool of the algorithm is the sample autocorrelation function $\hat{\rho}(h)$ of the samples [8] which is defined as:

$$\hat{\rho}(h) = \frac{1}{(n-h)\hat{\sigma}^2} \sum_{i=1}^{n-h} (x_i - \hat{\mu})(x_{i+h} - \hat{\mu}) \tag{34}$$

Where $\hat{\mu}$ is an estimator of the mean and $\hat{\sigma}^2$ is an estimator of the variance. If the autocorrelation function rapidly goes to zero as h grows, the samples can be considered fairly independent on one another. This situation is called *fast mixing*, as opposed to slow mixing where the autocorrelation function does not decrease as fast. Fast mixing indicates that the Markov chain quickly reaches all feasible states of the feature space Ω which decreases the necessary number of samples needed to accurately represent the target distribution [8]. Another diagnostic tool is the acceptance rate, which is the share of iteration where the algorithm accepts the proposal density. A commonly used heuristics for this is that an acceptance rate of around 23% is good [9]. These factors can be tuned by picking an appropriate proposal distribution.

A common class of proposal distributions are the symmetric proposal distributions [13], where $r(x^*|x_{i-1}) = r(x_{i-1}|x^*)$. Using this reduces the acceptance probability to $\alpha = z(x^*)/z(x_{i-1})$ which is computationally convenient. One symmetric proposal distribution which is used in this study is the random walk proposal, where $x^* = x_{i-1} + \epsilon$ for ϵ drawn from some symmetric zero mean random variable. We will use $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ denotes the multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. The tuning parameter of the algorithm is the covariance matrix $\mathbf{\Sigma}$, which simply reduces to a single variance parameter for single dimensional feature spaces. If the different variables of the multivariate feature space are somewhat independent, there should be no problems picking $\mathbf{\Sigma} = \boldsymbol{I}\boldsymbol{s}$, where $\boldsymbol{I}$ denotes the identity matrix and $\boldsymbol{s}$ denotes some vectors of component-wise variances. If the dependence is so strong that we are unable to obtain fast mixing, we need to consider some covariance matrix $\mathbf{\Sigma}$ having non-zero off-diagonal elements. More on picking proposal distributions can be read in [13].

Consider each element s_k of $\boldsymbol{s}$ that govern how far away the proposed sample typically gets from the last sample. A too small value will make the samples highly dependent of one another. A too large value will often push the proposed sample into improbably values, causing the acceptance rate to drop. This causes many samples to assume identical values, which creates large correlation between sample and slow mixing. There is hence some optimal variance for the proposal distributions of each component, and it is up to the parameter selection procedure to tune them into their optimal value.

A good sanity check for the Metropolis Hastings algorithm is to test different starting samples x_0 and to make sure that the algorithm converges to producing

samples in the same part of the feature space [8]. The algorithms does indeed theoretically converge after infinitely many sample, but if different values of x_0 produce vastly different results, there is reason to believe that the Markov chain has not converged to its stationary distribution yet, meaning that either more samples are needed or that the proposal distribution needs to be tweaked in order to produce better mixing.

3.4.1 Empirical distribution approximation

For every parameter that needs to be inferred through the Metropolis Hastings algorithm, the posterior distribution can only be represented as a large set of samples generated through the algorithm. These samples can be used to define the *empirical distribution*, whose cumulative distribution function $\hat{F}_n(x)$ for n number of samples can be defined as:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i < x\} \tag{35}$$

Where $X_i, i = 1, ..., n$ are the samples drawn from the target distribution and $\mathbb{1}\{X_i < x\}$ denotes the indicator function that assumes value 1 if $X_i < x$ and 0 otherwise. The Glivenko-Cantelli theorem states that the empirical distribution converges almost surely to the target distribution as the number of samples goes to infinity [4]:

$$\sup_{x \in \mathbf{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s}} 0 \tag{36}$$

Future sampling from the posterior distribution of these parameters will be approximated by sampling from its empirical distribution based on the samples of the algorithm, which is justified by the theorem above.

3.5 Dirichlet-multinomial hierachical model

In the previous section, we defined a procedure for the location specific immigration counts, where the rates could be assumed to be a priori independent of the rates of other locations. Given a location l and a location specific immigration count $P_l t$, we wish to make a model of immigration to the populations of any partition of l into I subcohorts, denoted $l_{\gamma_i}, \quad i = 1, ..., I$. In this thesis, the sub-cohorts considered are the genders and ages of the immigrants to a location l. The model used is a multinomial model with P_{lt} number of trials that will fall into any of the categories $l_{\gamma_i}, \quad i = 1, ..., I$, where $P_{l_{\gamma_i} t}, \quad i = 1, ..., I$ denotes the number of trials falling into l_{γ_i}. The parameters of the multinomial model, except for the number of trials P_{lt}, are the individual probabilities of each subcohort $\pi_{l_{\gamma_i}\tau}$, which are assumed to be valid for the stable time period τ. The probability parameters for each subcohort must obey the following constraint:

$$\sum_{i=1}^{I} \pi_{l_{\gamma_i}\tau} = 1 \tag{37}$$

Define $\boldsymbol{\pi}_{l\tau} = [\pi_{l_{\gamma_1}\tau}, \pi_{l_{\gamma_2}\tau}, ..., \pi_{l_{\gamma_I}\tau}]$ and $\boldsymbol{P}_{lt} = [P_{l_{\gamma_1} t}, P_{l_{\gamma_2} t}, ..., P_{l_{\gamma_I} t}]$. The purpose of the inference is to find the posterior probability of the parameters given historic accounts, which according to Bayes' theorem can be expressed as:

$$p(\boldsymbol{\pi}_{l\tau}|P_{lt}, \boldsymbol{P}_{lt}) \propto p(\boldsymbol{P}_{lt}|\boldsymbol{\pi}_{l\tau}, P_{lt})p(\boldsymbol{\pi}_{l\tau}) \tag{38}$$

Where $p(\boldsymbol{\pi}_{l\tau})$ denotes the prior probabilities of subcohort probability parameters, and $p(\boldsymbol{P}_{lt}|\boldsymbol{\pi}_{l\tau}, P_{lt})$ denotes the likelihood of the subcohort specific immigration counts given total location immigration counts and the probability parameters. The multinomial model states that the likelihood function is expressed as [14]:

$$p(\boldsymbol{P}_{lt}|\boldsymbol{\pi}_{l\tau}, P_{lt}) = P_{lt}! \prod_{i=1}^{I} \frac{\pi_{l_{\gamma_i}\tau}^{P_{l_{\gamma_i}t}}}{P_{l_{\gamma_i}t}!} \tag{39}$$

For the prior probability, we will resort to the Dirichlet family. Having parameters $\boldsymbol{\alpha}_{l\tau} = [\alpha_{l\tau}^{(1)}, \alpha_{l\tau}^{(2)}, ..., \alpha_{l\tau}^{(I)}]$, we say that a random variable is Dirichlet distributed, that is $\boldsymbol{\pi}_{l\tau}|\boldsymbol{\alpha}_{l\tau} \in Dir(\boldsymbol{\alpha}_{l\tau})$, if the probability density function is:

$$p(\boldsymbol{\pi}_{l\tau}|\boldsymbol{\alpha}_{l\tau}) = \Gamma\left(\sum_{i=1}^{I} \alpha_{l\tau}^{(i)}\right) \prod_{i=1}^{I} \frac{\pi_{l_{\gamma_i}\tau}^{\alpha_{l\tau}^{(i)}-1}}{\Gamma(\alpha_{l\tau}^{(i)})} \tag{40}$$

It can be shown that given this prior probability, we will have [14]:

$$\begin{aligned} &\boldsymbol{\pi}_{l\tau}|P_{lt}, \boldsymbol{P}_{lt} \in Dir(\hat{\boldsymbol{\alpha}}_{l\tau}) \\ &\hat{\boldsymbol{\alpha}}_{l\tau} = \{\alpha_{l\tau}^{(i)} + P_{l_{\gamma_i}t}\}_{i=1}^{I} \end{aligned} \tag{41}$$

We will choose a noninformative prior pobability in the same manner as for the beta-binomial model presented in section 3.2. We hence set $\alpha_{l\tau}^{(i)} = 0.1, \forall i$, hence assigning a small probability for subcohorts without observations to make the model feasible.

3.5.1 The location partition problem

Equivalent to the cohort partition problem presented in the hierarchical beta-binomial model section for intrinsic demographic processes, there is an equivalent problem regarding finding a set of locations such that their subcohort probability parameters can be regarded as the same. Looking only at the location of interest if the simple answer, however there may be very few observations in such area and the prior probability may play a way to strong role. The prior probability is not completely uninformative. A typical example that is often considered is to predict age specific immigration counts. There are $\mathcal{J}+1 = 101$ age categories, corresponding to $I = 101$ subcohorts for any given location. The area is is newly built, and did not have any population nor immigration prior to now. Only using local information when inferring the distribution of the probability parameters for each age group will, without any observation, just yield a copy of the prior probability. In such cases, one may use observations of age distributions of other locations that can be thought of to be of the same type. This classification of locations into clusters of similar type can be done by expert knowledge, just as in the case of Beta-binomial model. Environmental factors such as the type of housing, distance to city center, income level of the area can be used to categorize the areas that make up the demographic data.

3.6 Graphical model overview

In the recent sections, we have specified a cascade of different demographic variables and their interdependence. To get an overview of the model structure, we will represent the conditional probability distributions with a Bayesian network. A Bayesian network is simply a decomposition of the joint probability distribution into conditional distributions, such that a large number of conditional independence assumption is used to simplify the expression [14][23]. A simple example of a graphical representation of four random variables, a, b, c and d. Now, consider the joint distribution decomposed into component in the following way:

$$p(a, b, c, d) = p(a|b, c, d)p(b|c, d)p(c|d)p(d) \tag{42}$$

Now, say that we happen to know that a is conditionally independent of c given b, that b is conditionally independent of d given c and that c is independent of d. Then, the expression simplifies to:

$$p(a, b, c, d) = p(a|b, d)p(b|c)p(c)p(d) \tag{43}$$

A graphical representation of this would correspond to figure 4. Each random variable is represented by a node and a directed edge is going from the random variables which the first random variable is conditioned on in (43) to itself.

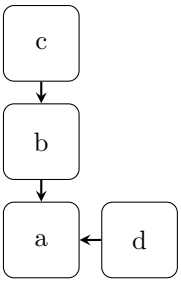

Figure 4: Graphical representation of the random variables decomposed in (43)

Let $\boldsymbol{N}_t$ denote a set of populations, $\boldsymbol{F}_t$ denote a set of birth counts, $\boldsymbol{M}_t$ denote a set of death counts, $\boldsymbol{Q}_t$ denote a set of emigration counts, $\boldsymbol{M}_{j=0,t}$ and $\boldsymbol{Q}_{j=0,t}$ denote corresponding sets for cohorts consisting exclusively of zero year olds, $\boldsymbol{P}_t^{(l)}$ denote a set of location specific immigration counts, $\boldsymbol{P}_t$ denote a set of immigration counts partitioned on any set of variables, $\boldsymbol{H}_t$ denote a set of variables indicating the number of new housing development and $\boldsymbol{\lambda}_t$ denote a set of immigration rates for some time point t within the stable period τ. Furthermore, let $\boldsymbol{\mu}_\tau^D, D \in \{F, M, Q\}$ denote sets of demographic rates for each demographic variable, $\boldsymbol{\beta}_\tau$ denote a set of regression parameters for the immigration rates, $\boldsymbol{\sigma}_\tau$ denote a set of parameters corresponding to σ in equation (20) and $\boldsymbol{\alpha}_\tau$ denote a set of Dirichlet parameters governing the subcohort distribution of immigration counts for the stable time period τ. Using this kind of decomposition, we can represent our model with figure 5, which neatly and compactly summarizes the structure of the model. We have indicated the entities given in the historic data by green and the latent parameter which are to be estimated by white. Hyperparameters are for simplicity not included in figure 5. The death and emigration

counts for cohorts consisting of zero year olds has has the number of new births as source populations (the number of trials in the binomial model) in contrast to the population of last year which is the case of these demographic variables for other age groups. This is why they are separated in the graphical model.

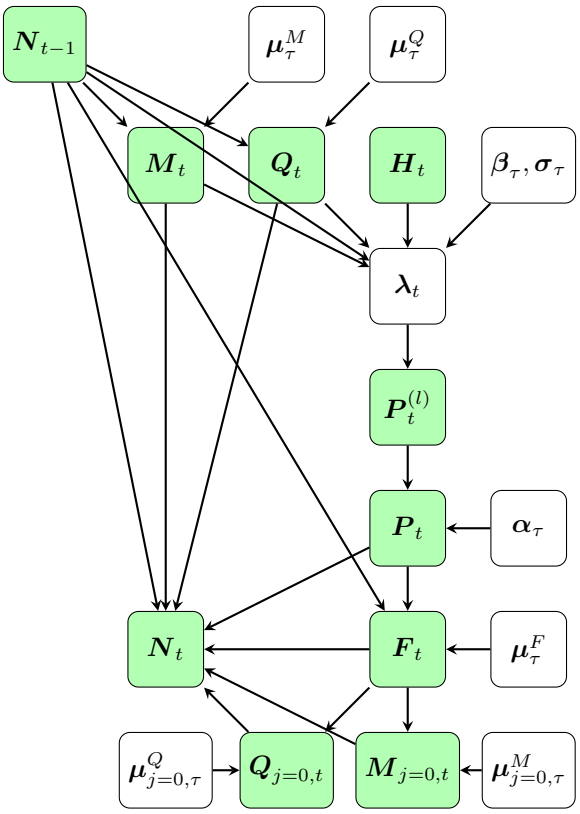

Figure 5: The graphical representation of the model framework represented in this study.

The graphical model of figure 5 can be considered a *causal Bayesian network* (as described in [23]) since each arrow is pointing from a node that can be considered to cause the other. Emigrations and death counts are caused by there being a population there to begin with, immigration counts are (generally) caused by available housing through either new developments or existing inhabitants either moving out or dying. Birth is in turn caused by the people living there or moving in there. And the emigration and death counts of new born children are caused by them being born. The populations of next year is caused by last years populations and the demographic processes that affects them. The interpretability of these causal relations are one of the main strength with the graphic representation of the model framework.

4 Application

4.1 Forecast generation

After having estimated all the parameters for the cohorts of interest, we will now lay out the procedures for generating forecasts. We will assume that the time frame in which we wish to establish the forecast lie within the stable time period that the parameters have been estimated within. By making that assumption, we can now consider the posterior probability distributions of the parameters as given. Let T denote the latest time in which we have access to demographic data and we now wish to make forecasts of times $T+1, T+2, ..., T+H$ for some time horizon H years into the future. These forecast will consist of predictive distributions of future population levels given current populaions and the historic data used to infer the posterior probability distributions of the parameters, or $p(\boldsymbol{N}_{T+1}, \boldsymbol{N}_{T+2}, ..., \boldsymbol{N}_{T+H} | \boldsymbol{N}_T, [historic \quad data])$. The forecasts are iteratively generated one year at the time using the procedures summarized in figure 5. Now that the posterior distributions of the parameters are given, the forecast generating process simplifies to a top-down simulation of each process, starting from the current population counts $\boldsymbol{N}_t$. Since there is no hope of finding closed for expression for the predictive distributions, we will generate a large amount of parallel samples for each process in figure X. Each sample from a parent enters the child generating process as a parameter. To illustrate the procedure, take some population such that the source cohort has a population of $N_{\gamma^* T} = 54$. The emigration rate for the source cohort has been infered to be $Beta(13, 140)$. We now sample a large number of realisations from the emigration rate distribution, for example $\mu^Q_{\gamma t} = 0.1316, 0.1169, 0.0707, 0.09118, ...$ When drawing emigration count samples from $Bin(N_{\gamma^* T}, \mu^Q_{\gamma T+1})$, use each sampled value for $\mu^Q_{\gamma t}$ for one emigration count sample.

$$Q_{\gamma T+1} \in Bin(54, 0.1316), Bin(54, 0.1169), Bin(54, 0.0707), Bin(54, 0.09118), ...$$

Sampling from this yields $Q_{\gamma T+1} = 4, 4, 8, 4, ...,$ where each sample represent some possible future outcome. In the next step, we draw the same number of samples from the other parentless nodes in figure 5 and keep on going until we reach the end of the graph, where we will end up with a nuber of samples from $N_{\gamma T+1}$ which is the target of the forecast. The Glivenko-Cantelli theorem stated in (36) ensures that the empirical distribution of $N_{\gamma T+1}$ will approach the true distribution for a large number of samples. The law of large number ensures that the sample mean approaches the true expected value $\mathbb{E}[N_{\gamma T+1}]$.

As the forecasts consist of predictive distributions, it is fairly easy to infer *credible intervals* of the approximations. The notion of credible intervals was introduced by Ward Edwards in [5] and it is defined for credible level $\alpha \in (0, 1)$ by an interval $\mathcal{I} = (\mathcal{I}_{inf}, \mathcal{I}_{sup})$ such that a share α of the density function of the random variable lies within the interval. That is, for the random variable of population $N_{\gamma T+1}$, we have for credible level α:

$$p(N_{\gamma T+1} \in \mathcal{I}) = \alpha \tag{44}$$

Multiple credible intervals that fulfills (44) exist for each credible level. In this study, we focus on central credible intervals, or intervals defined such that:

$$F_{N_{\gamma T+1}}(\mathcal{I}_{inf}) = \frac{1}{2} - \frac{\alpha}{2}, \quad F_{N_{\gamma T+1}}(\mathcal{I}_{sup}) = \frac{1}{2} + \frac{\alpha}{2} \tag{45}$$

Where $F_{N_{\gamma T+1}}$ denotes the cumulative distribution function of $N_{\gamma T+1}$. The credible intervals of the true posterior distribution of population levels will be approximated by the credible intervals of the empirical distribution.

Forecasting $\boldsymbol{N}_{T+2}$ involves an analogous procedure, where the starting population now is represented by the empirical distribution of $\boldsymbol{N}_{T+1}$ instead of $\boldsymbol{N}_T$. This iterative process is continued for as long as one wish, and more randomness is introduced in each time step. This leads to forecasts with larger variance for long time frames, which is a reasonable expectation. Note that the environmental variables used in determining the immigration rates are not forecast along with the demographic variables and the populations, and future values of these must be determined separately. This procedure can be done through expert estimations where each environmental variable is assigned some probability distribution for future times. We can also use some simple time series model (for example, ARIMA-models, see [22]) to estimate the typical volatility of the environmental variables. Note that in the application of the model framework used in this study, we will not make multiple year forecasts.

4.2 Data

The model specified in earlier sections was used to predict populations of different cohorts in Stockholm county, Sweden. The cohorts of interest were generated by dividing the population of Stockholm county into different partitions based on age groups and locations. We consider two different level of location partitions. One of the levels is corresponding to each municipality in Stockholm county. A map of this location partition is provided in figure 6. The units of the second level of location are called *base areas*, which are defined by *Tillväxt- och regionplaneförvaltningen*, which is the regional planning authority in Stockholm county. The base areas consists of areas of about one thousand inhabitants[6] and they are designed to represent as homogeneous populations as possible. Stockholm county consists of 26 municipalities and 1418 base areas. In figure 7, we show how Täby municipality has been divided into base area as an example of this.
The base areas used in this study are the ones defined in 2010. This is the finest partition of location available in our demographic data. The demographic data consists of population counts and demographic variables for each combination of age, gender and base area and for each year between 2000 and 2016. The demographic data has been collected from Swedish population registers by *Statistiska centralbyrån* (SCB) or *Statistics Sweden*, which is the governmental agency of statistics in Sweden. The population counts of each year denotes the population counts as of December 31:st of that year, and the demographic variables for each year denotes the number of events of that demographic process that occurred during the year. The effect of incorrectly registered and unregistered individuals will not be taken into account in this study. The forecast will hence denote the number of people registered as part of a given cohort, which may differ from

[6]The larges base area has a little less than ten thousand inhabitants and there are several base area with no population whatsoever.

Figure 6: Partition of Stockholm county into municipalities.

the de-facto population of the cohort. Unregistered international immigrants are assumed to account for the largest portion of unaccounted individuals. This will not pose a problem since the purpose of this study is to develop a tool for societal planing of areas such as schooling and taxation, which are generally based on the registered populations and not de-facto populations.
The historic data of new housing development has also been collected by SCB based on reports from the Municipalities. However, municipal authorities have flagged for discrepancy between the dates that the new housing is reported to have been completed and the actual date when people are moving in. This warning was confirmed by inspecting the data. Figure 8 shows two cases where spikes in housing development causes immigration counts to rise. In the right case, there is an almost perfect correlation with the number of new housing units and the immigration counts within the area. In the left case however, the spikes in new housing units lags behind the spike of immigration counts of 2010. This phenomenon of new housing units being delayed was found in multiple cases. This discrepancy can be taken into account by expanding the regression function to take future values of new housing units into account. To reduce the number of parameters for factors that are likely to have a small effect, we collapse all lagged parameters into one that takes the total number of housing units into account, regardless of type. Letting $H_{lt}^{(tot)}$ denote the total number of newly built housing unit in area l at time t, the modified version of (22) becomes:

$$g_\tau(\boldsymbol{X}_{lt}) = \beta_{0\tau} + \sum_{k=1}^{K} \beta_{k\tau} X_{lt}^k + \beta_\tau^{(lag)} H_{lt+1}^{(tot)} \tag{46}$$

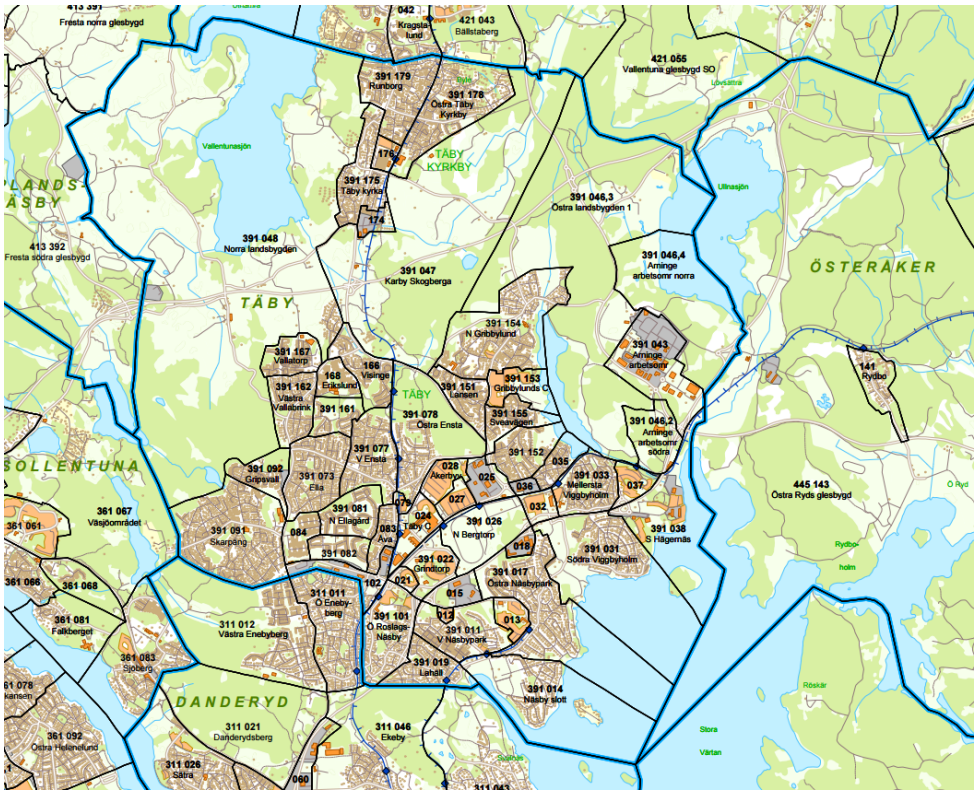

Figure 7: The partition of Täby municipality into base areas. The numbers denote their respective reference code. Maps are available on http://rufs.se/kartor/omradesdata/basomradeskartor-2010/

4.3 Model training

The model framework described in the previous sections has been used to make forecasts of different populations of Stockholm county. We have made forecasts for the total population of Stockholm county (denoted *Partition A*), for each age group in Stockholm county (denoted *Partition B*), for the total population of each municipality (denoted *Partition C*), for the population of each age group in each county (denoted *Partition D*), for total population each base area (denoted *Partition E*) and for the populations of each age group in each base area (denoted *Partition F*). There are 1 population in partition A, 101 populations in partition B (using the cap-age of 100 years old), 26 populations in partition C, 2626 populations in partition D, 1418 populations in partition E and a staggering 143218 cohorts in partition F that can be trained for each year.

The posterior distribution of demographic rates and other parameters were estimated using the procedures described in section 3. They were estimated using different training schemes and tested on historic data left out of the training part to analyze which specific model in the model framework that produces the best results with regard to different populations and demographic variables. For the intrinsic demographic variables, we used two homogeneous cohort assumptions

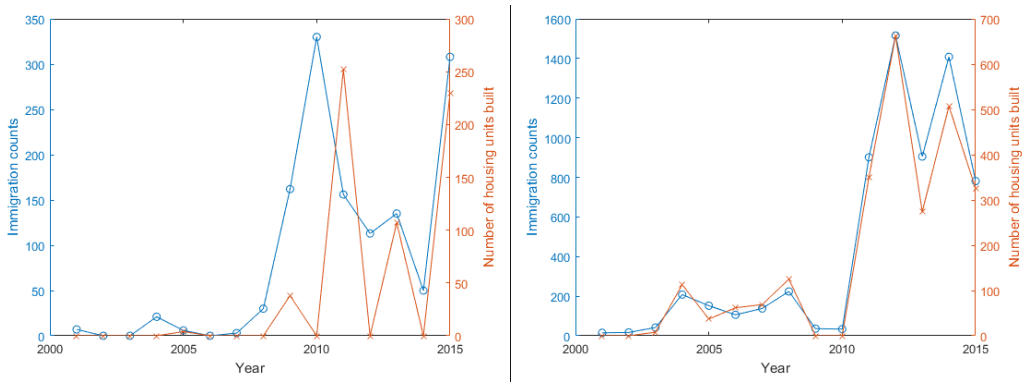

Figure 8: Area specific immigration counts and new housing units plotted alongside each other for two base areas. Note that the axes differ in value for each variable.

in the preliminary test rounds. We partitioned our cohorts on age and gender (denoted *Partition B**) and on age, gender and municipality (denoted *Partition D**). For the birth rates however, we will not train the parameters on gender of the children. This is due to the fact that we often lack data for the gender of new born children. We will instead only estimate $\mu^F_{jl\tau}$ for each age of mother $j = 14, ..., 49$ and location in the stable time period. We will then simulate the gender of the child assume using the well established 51.4% probability of a new-born baby being a boy in Sweden. Furthermore, due to the lack of data regarding the age of the mothers of zero year olds that is moving into an area, we will not use the approximation of total birth counts stated in (11). Since the location cohorts used to estimate the birth parameters are large enough and not mainly made up by new housing development, this negligence of immigrant births should not affect the predictive power of the birth rate estimates. In other words, we will simply be using (8) to model birth in this application of the model framework.

For the estimation of within location subcohort immigration estimation given total immigration to the area described in section 3.5, we define the entire Stockholm county (Partition A) as an homogeneous area in one model and each municipality (Partition C) as a homogeneous area in another model. Hence, for each intrinsic demographic variable and the subcohort immigration probabilities, we will fit models on a finer and a coarser homogeneous cohort assumption. The finer models (partitions D* and C) will in general have fewer observations to base its inference on compared to the coarser models (partitions B* and A). This will lead to parameter estimates with higher variance and broader credible intervals, which is demonstrated in figure 2. The strength of using a cohort partition finer than the population at interest for training the parameters is that if the populations consists of several homogeneous cohorts and the the proportions of these homogeneous cohorts shift, the finer model will capture this whereas the coarser model will not. There may be motivation to use a homogeneous cohort assumptions that is coarser than the population of interest if there are too few historic observations of a certain demographic process

within the cohort of interest. In such situation, the parameter estimations will be very sensitive to single observations and may overfit the data. It may also be motivated by missing data, which is the case for partition F in this study. We simply don't have any available data of demographic variables in partition F. We do however have populations, which is why forecasts will be tested on these to see how good the coarser model will perform on such small populations.

For the location specific immigration model, we will try two different regression models. Recall the regression function of the negbin-II-model described in section 2.4 (stable time period index is dropped for convenience):

$$g(\boldsymbol{X}_{lt}) = \beta_0 + \beta_{Qt}Q_{lt} + \beta_{Mt}M_{lt} + \sum_{k=1}^{K_H} \beta_{H_k} H_{lt}^k + \beta_H^{(lag)} H_{lt+1}^{(tot)} \tag{47}$$

The current population of the location N_{lt-1} was dropped out of the model due to preliminary model analysis showed that including it did not significantly affect the results if the other variable were included. $H_{lt}^k, \quad k = 1, ..., K_H$ denotes the values of the K_H different number of variable related to new housing. The difference in the two models is which housing categories that the newly developed housing in an area is divided into. The new housing variables are either partitioned into components representing the number of new built apartments (ap), terraced houses (th) and detached houses (dh) $\boldsymbol{H}_{lt} = [H_{lt}^{ap}, H_{lt}^{th}, H_{lt}^{dh}]$ (hereby denoted *model 1*) or partitioned after type and number of rooms (kitchens excluded) for apartments $\boldsymbol{H}_{lt} = [H_{lt}^{ap1}, H_{lt}^{ap2}, ..., H_{lt}^{ap5}, H_{lt}^{th}, H_{lt}^{dh}]$ (hereby denoted *Model 2*). Note that the H^{ap5}-variable captures newly built housing with five or more rooms. Apartments reported as having zero rooms generally denotes student housing with shared kitchens. The location specific immigration model will for each housing parameter schema be trained using data points generated by location partition partition E (resulting in 1418 data points for each year) to obtain posterior distributions of parameters.
A few exploratory results on the estimated demographic rates of historic data are presented in figure 9. The "high" lines represent the upper bounds of the 90 percent credible intervals and the "low" lines represent the lower bounds of the 90 percent credible intervals for each demographic rate presented. These graphs show that demographic rates may significantly differ for certain genders, age groups and areas which is motivating the partition of homogeneous cohorts. The graph on the left show some slight trends and random variance between years which may question the validity of the stable time period assumptions.

4.4 Validation procedure

Any selection we can make in the model (such as altering the homogeneous cohort partition assumption, or choosing different variables to include in our immigration negbin-II-model) will hereby be referred to as a *training schema*. After the parameters have been trained using the schemas presented in the last section, we wish to establish the schema that systematically produce the best results for each demographic variable and for each cohort of interest. We will use the historic data of demographic variables to make one year forecasts of each demographic variable for each year. We will for each (feasible) year train a set of model on historic data up till the year before. Using the demographic states

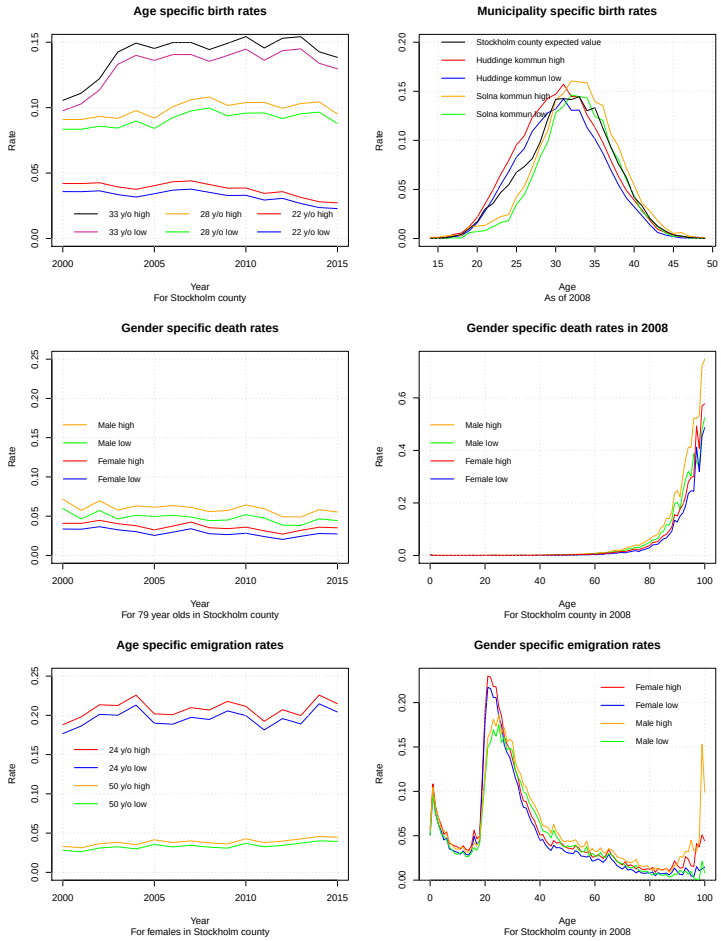

Figure 9: 90% credible intervals of the posterior distribution of the intrinsic demographic rates, plotted against time and different age group for a selection of cases.

of the year before, we will make forecasts of demographic variables of that year. We can then compare the forecasts with the true values in order to determine which schema is best suited for a particular situation. This kind of rolling leave one out validation approach of time series data has been suggested in [18].

For the tree-based parameters, the cohort partition used for the training may either be finer, identical, coarser or non-overlapping with the cohort partition that the forecasts are made of. Coarser partitions will in general produce tighter confidence bounds than finer approaches due to a higher number of observations per parameter. If the training partition is partitioned coarser than the cohorts we are forecasting and the homogeneous cohort assumption erroneous, we should expect erroneous forecasts as well. Another hypothesis to be tested is that partitions finer than the population of interest should be able to make better forecasts, as they capture the structural difference of the demographic

tendencies of the individuals within the cohort of interest and as the population composition changes in an area, the stable time assumption will no longer hold.

4.4.1 Regression diagnostics

The regression model of the immigration counts stated in (12), (20) and (22) are base on a number of distributional assumptions. In (21), we get the variance of the immigration counts given the covariates. Since the variance is invariant to a change in location parameter, we also have:

$$Var(P_{lt} - g(\boldsymbol{X}_{lt})) = g(\boldsymbol{X}_{lt}) + \frac{g(\boldsymbol{X}_{lt})^2}{\sigma} \tag{48}$$

The conditioning on the covariates and σ is omitted for convenience. Now consider the expected value of the residual squared:

$$\begin{aligned} &Var(g(\boldsymbol{X}_{lt}) - P_{lt}) = \mathbb{E}[(g(\boldsymbol{X}_{lt}) - P_{lt})^2] - \mathbb{E}[g(\boldsymbol{X}_{lt}) - P_{lt}]^2 \\ &= \mathbb{E}[(g(\boldsymbol{X}_{lt}) - P_{lt})^2] - (\mathbb{E}[g(\boldsymbol{X}_{lt})] - \mathbb{E}[P_{lt}])^2 = \mathbb{E}[(g(\boldsymbol{X}_{lt}) - P_{lt})^2] - (P_{lt} - P_{lt})^2 \\ &= \mathbb{E}[(g(\boldsymbol{X}_{lt}) - P_{lt})^2] \end{aligned} \tag{49}$$

Hence, from (48) and (49) we get :

$$\mathbb{E}[(g(\boldsymbol{X}_{lt}) - P_{lt})^2] = g(\boldsymbol{X}_{lt}) + \frac{g(\boldsymbol{X}_{lt})^2}{\sigma} \tag{50}$$

This is the expected distribution of the residuals squared as a function of their expected value $g(\boldsymbol{X}_{lt})$. As a diagnostics tool, we will plot the residuals squared as a function of the predicted value for the immigration counts and graphically inspect weather the residuals squared value can reasonably have been generated from a random variable with the expected value given by (49). Another assumption is that the expected value of the residuals should be zero, independently on the predicted value $g(\boldsymbol{X}_{lt})$. Hence, plotting the residuals as a function of predicted value will be helpful tools to investigate weather the regression model is misspecified or not.

4.4.2 Accuracy metrics

A few accuracy metrics will be used to evaluate the performance of the model, of which one is the distribution of error terms as a function of various variable. This is to investigate weather the model framework systematically over-or underestimates the populations for some specific groups. We will in particular look at the distribution of error terms for different ages and different cohort sizes to uncover any bias that the models create. We will also look at the distributional assumptions for the predictive distributions. For any credible level α, we should expect around that share of observation lying within the corresponding credible interval. If too few observations fall within the interval, the model is overconfident in its distributional assumptions, and if too many observations fall within the credible interval, the model does not produce forecasts that are confident enough. This metric will also be applied to different age groups and cohort sizes in order to evaluate if the distributional assumptions of the models hold

particularly well for some specific populations. The mean sum of squared error is used to compare different models of the same population and is not meant to be used as a tool to compare model strength for different populations.

5 Results

In this section, we will summarize the performances of the models we have tested according to the procedures specified in the previous section. To summarize the results, there was almost always at least one training schema for each demographic variable and for each type of target population that performed what could be considered well. A well performing model is unbiased, ie the residuals of the test data points seems to be generated from a zero mean random variable. A well performing model is also a model that correctly estimates the built in uncertainties of the demographic processes. Hence, around 90 percent of the observed outcomes should fall within the 90 percent credible interval that is given for each forecast. The model framework tend to perform worse on cohorts not divided into age categories, where the typical problem is that the predictive distribution of the demographic variables in these cases have lower variance than the true distributions. This means that the built in beta binomial variance does not fully explain the variances of demographic variables in these populations. We also notice a general tendency of the model framework to overestimate the number of deaths, especially for large cohorts.

5.1 Intrinsic demographic variable forecasts

We applied the algorithm to make next-year forecast simulations using data up to time t for each past time data available. The results are summarized in table 1, where the model was tested for each homogeneous cohort assumption schema on all cohorts in the partition presented in the column "Cohorts". The MSE-column denotes the mean squared residual of the predicted demographic variable compared to the true outcome. The in cred.int. column denotes the share of observed demographic variables that fall within the 90 percent credible rate of the predictive distribution in each case. Figures 14 to 43 of Appendix B shows the residual and the average share of observations within the 90 percent credible rate for the predictive distribution plotted against cohort size and age (in the cases where cohorts are partition on age) for each row in table 1, which serves to further illustrate the goodness of each schema. In figures 10 to 12, we have plotted the total demographic counts in Stockholm county for each demographic variable with the one year forecasts for each year, credible intervals included. These graphs clearly illustrate the over confidence of the model framework for cohort A, and by observing figures 38 to 43 depicting the performance of the model framework for cohort C disaggregated by cohort sizes, we see that this overestimation of certainty tend to affect larger cohorts.

Using the populations at interest as the homogeneous cohort assumptions worked well in the cases where they could be tested. With the exception for an unusually bad ability to accurately estimate the variance of emigration counts of people around 80 years of age for partition D (see figure 36), the demographic variables

Process	Cohorts	Homogeneous Cohorts	MSE	In cred.int.
Birth	A	B*	512449	0.500
Birth	A	D*	545089	0.500
Death	A	B*	154396	0.500
Death	A	D*	2045300	0
Emigration	A	B*	19177152	0
Emigration	A	D*	15488950	0.200
Birth	B	B*	2431	0.825
Birth	B	D*	2450	0.817
Death	B	B*	298	0.898
Death	B	D*	532	0.654
Emigration	B	B*	8190	0.738
Emigration	B	D*	9576	0.515
Birth	C	B*	18983	0.577
Birth	C	D*	6431	0.842
Death	C	B*	2131	0.627
Death	C	D*	5196	0.434
Emigration	C	B*	2648899	0.196
Emigration	C	D*	200188	0.480
Birth	D	B*	298	0.780
Birth	D	D*	67	0.923
Death	D	B*	6.69	0.959
Death	D	D*	10.1	0.946
Emigration	D	B*	1478	0.745
Emigration	D	D*	161	0.869

Table 1: Summary of results

for all cohorts that were trained using identical homogeneous cohort assumptions yielded unbiased results and correctly estimated the uncertainty of the forecast. Using a finer homogeneous cohort assumption than the cohorts of interest yielded equivalent results in the case of birth counts, but biased results for the other demographic variables. This was clearly evident for deaths and also in the case of partition A, where using an even finer partition (D*) gave more biased results than using a coarser (B*). This is probably due to the already present general bias to the death variable which is exaggerated when using finer partitions. The adaptive stable time period algorithm which was described in section 3.2.1 have also contributed to the bias, since it actively searches for rare instances of a demographic event. Looking at figures 21 and 35, we see that it is the small cohorts for which the events are rare that the forecasting algorithm overestimates the occurrences systematically. This attempt to mimic the heuristics of the forecasting procedures used for rare-event forecasting clearly must be redefined for future studies of the subject.
Using a coarser homogeneous cohort assumption than the cohorts of interest have yielded mixed results. Applying the B* schema for cohorts partitioned by D gave comparable results with D* for the death process, but yielded overconfident forecasts for emigration and extremely bad forecasts for births. The bias of these demographic variables seem to grow with cohort size (see fig 27 and

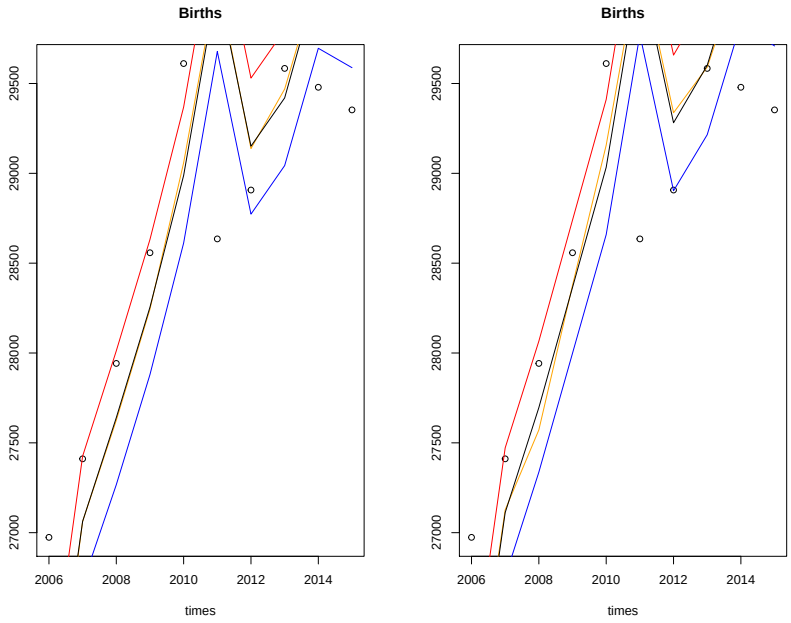

Figure 10: Predicted total birth counts in Stockholm county versus the outcome. The left figure depicts forecasts with parameters from training schema B* and the right figure depicts forecasts with parameters from training schema D*. The circles are the true outcomes, the black line denotes the expected value of the predictive distribution, the yellow line denotes the mode of the predictive distribution, the blue line denotes the lower bound and the red line denotes the upper bound of the 90% credible interval

35). In the case of cohort partition C where one training schema is strictly finer (partition D*) and one is finer in one dimension and coarser in another (partition B* is finer in age and coarser in location), the different training schemas performed differently well on different demographic variables. For emigration, the B*-schema was biased and the D*-schema, even though the forecasts were overconfident, they lacked bias. Hence, the D*-schema is the preferred one for emigration and partition C. Regarding deaths, the bias was conversely present in the D*-schema but not int the B*-schema.

5.2 Immigration forecasts

5.2.1 Outliers

Before we trained the Metropolis-Hastings algorithm, we performed preliminary linear regressions with the sole purpose of identifying high leverage points in the model using cooks distance. We found two particular base areas whose data points were putting extraordinary leverage on the model. Both cases were characterized by times of extraordinary large net migration without any contributing factors present. Looking closer into these cases, we find that the areas are the base areas with reference numbers 2242310 and 2230460, roughly cov-

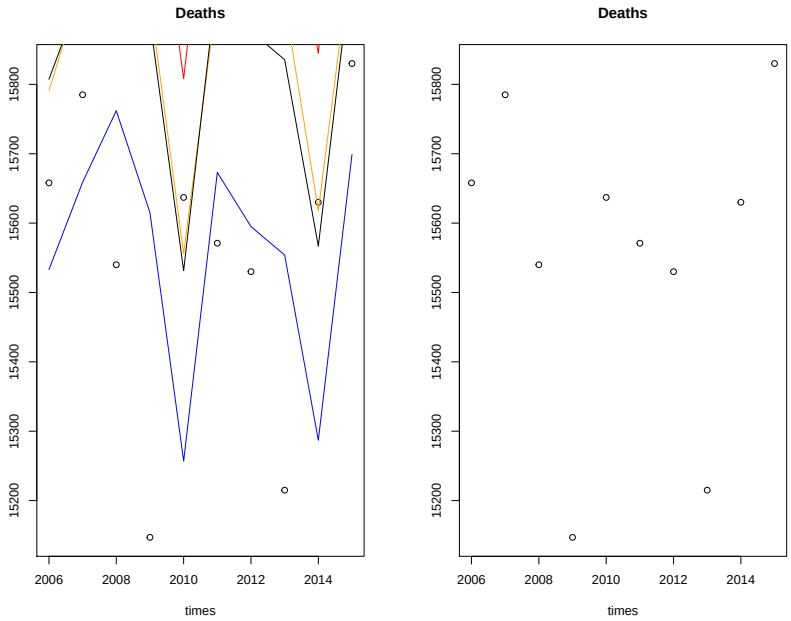

Figure 11: Predicted total death counts in Stockholm county versus the outcome. The left figure depicts forecasts with parameters from training schema B* and the right figure depicts forecasts with parameters from training schema D*. The circles are the true outcomes, the black line denotes the expected value of the predictive distribution, the yellow line denotes the mode of the predictive distribution, the blue line denotes the lower bound and the red line denotes the upper bound of the 90% credible interval. Note that the predictions are not even present on the right graph, since they all fall outside the scope of the y-axis. The predictions are above the actual values for each time in the right graph.

ering the areas of Kista centrum and Råcksta. In the Kista case, a large chunk of student housing was built in 2002 and the subsequent immigration came in subsequent years. Hence, we would need to specify further lag-components to account for this effect. In the Råcksta case, a large former office building (Vattenfallshuset) were converted into housing. This was however not reported as newly built housing in our data, and there were hence available housing lacking. We removed all data points associated with these areas from the model.

5.2.2 Diagnosis of Metropolis-Hasting algorithm

The Metropolis Hastings algorithm was trained on location specific immigration counts for both models with the parameters specified in table is summarized in tables 2 for model 1 and in 3 for model 2. The columns x_0 represent the initial value of the Markov chain for each parameter. As stated in section 3.4, the symmetric random walk proposal kernel was used with $x^* = x_{i-1} + \epsilon$ for generating proposals x^* (where x is a vector consisting of each parameter to be estimated $[\beta_0, \beta_M, ..., \beta_H^{(dh)}, \sigma]$) and where $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for some covariance

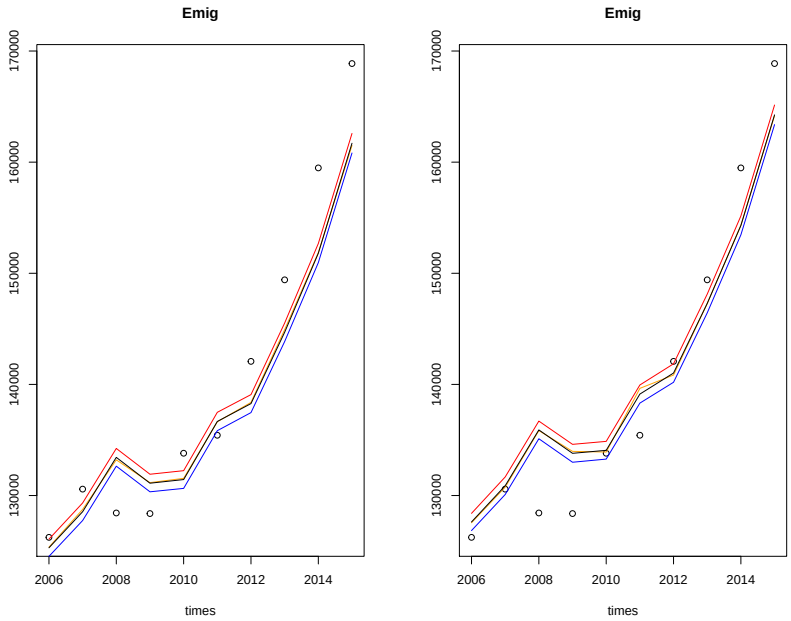

Figure 12: Predicted total emigration counts in Stockholm county versus the outcome. The left figure depicts forecasts with parameters from training schema B* and the right figure depicts forecasts with parameters from training schema D*. The circles are the true outcomes, the black line denotes the expected value of the predictive distribution, the yellow line denotes the mode of the predictive distribution, the blue line denotes the lower bound and the red line denotes the upper bound of the 90% credible interval

matrix $\boldsymbol{\Sigma}$. Fast mixing was obtained using a diagonal covariance matrix $\Sigma = \boldsymbol{I}\boldsymbol{s}$ where $\boldsymbol{I}$ denotes the identity matrix and $\boldsymbol{s} = [s_1, s_2, ..., s_K]$ is a vector containing component-wise variances for each parameter. These component-wise variances s_k had to be tuned in to obtain fast mixing and the column s_k shows the proposal density variance used in the algorithm used to generate the empirical posterior distribution. The column *a.r.* denotes the acceptance rate for each parameter that was obtained when training the model on data between 2005 and 2010. The column *p.v.* denotes the sample mean of the posterior empirical distribution and the column *s.d.* denotes the standard deviation when training the data between. In appendix B, there are plots, histograms and autocorrelation functions for the Markov chain Monte Carlo samples for each parameter and model, trained on data from between 2005 and 2010.

We found a negative correlation in both models between the parameter for emigration counts and the parameter for death counts in both model 1 and model 2. This is probably due to the fact that they are correlated in the data and show a similar effect on immigration counts. A full presentation of the the sample correlation between the posterior parameters distributions for model 1

Parameter	x_0	s_k	a.r.	p.v	s.d.
β_0	1	0.15	0.2440	0.9402	0.03033
β_M	1	0.19	0.2478	0.9658	0.04620
β_Q	1	0.016	0.2382	0.9616	0.004139
$\beta_H^{(ap)}$	1.5	0.25	0.2342	1.464	0.04801
$\beta_H^{(th)}$	1.8	1.3	0.2758	1.758	0.2903
$\beta_H^{(dh)}$	2.7	0.5	0.2344	2.482	0.0.1040
$\beta_H^{(lag)}$	0.3	0.1	0.2286	0.2820	0.02008
σ	18.5	3	0.1980	18.23	0.4842

Table 2: Table of cohort partitions used for training and forecasting.

respectively 2 when trained on data from between 2005 and 2010 can be found in table 4 in Appendix B. The emigration count parameter and the death count parameter all have an expected value close to one. This indicates that for each person dying or moving out of an area, we can expect one person moving in. For the new housing developments, we have for model 1, about 1.5 expected immigrant per newly built apartment, 1.8 expected immigrant per terraced housing unit and 2.7 expected immigrant per detached home. The model was trained for each year between 2006 and 2014 and the left side of figure X shows the expected value of each parameter plotted against the year after the last year that the data was trained on. For each year, we trained the data on immigration data from six years back. For model 2, we observed some interesting behaviour in the results, which are summarized in table 3. The posterior distribution of the intercept, death parameter, emigration parameter, terraced homes, detached homes, lag and sigma parameters were strikingly similar between the models. The expected value of 1.75 immigrants per newly built apartment in model 1 was replaced by an array of expected parameter values for model 2. Interesting results were that the expected number of people moving in per newly build two room apartment were smaller than the expected number of people moving in per newly built one room apartment. We also found that we can expect fewer people to move in per newly built 5 room or more apartment than per newly built four or three room apartment. A large part of this can be explained by lacking variation in the data (i.e multicollinearity). The number of three room apartments built in a base area during a year can be predicted by the number of two room apartments built using linear regression with an adjusted R-squared value of 0.7221, meaning that there are few cases where the effect of each variable can be tested individually. This is also evident in the strong negative correlation of -0.69 between the posterior distribution of these parameter, as presented in table 4. This may be problematic if such cases were to appear in the future, and a recommended respecification of the model would be to aggregate the variables for two room apartments and three room apartments into one variable representing the total counts.

In figure 13 to the left, we have plotted the (training) residuals $P_{lt} - g(\boldsymbol{X}_{lt})$ against the predicted value $g(\boldsymbol{X}_{lt})$ and it is clear that the model does not have a systematic off prediction based on predicted size. In figure 13 to the middle is a

Parameter	x_0	s_k	a.r.	p.v	s.d.
β_0	1	0.16	0.2276	0.9413	0.03014
β_M	1	0.18	0.2432	0.9708	0.04543
β_Q	1	0.02	0.1924	0.9613	0.004001
$\beta_H^{(ap0)}$	0.6	0.5	0.2790	0.5921	0.1194
$\beta_H^{(ap1)}$	0.9	1.8	0.2434	0.7525	0.4182
$\beta_H^{(ap2)}$	0.6	0.75	0.2094	0.6001	0.2254
$\beta_H^{(ap3)}$	2	0.9	0.2060	2.280	0.295
$\beta_H^{(ap4)}$	2	1	0.2838	2.380	0.3004
$\beta_H^{(ap5)}$	1.5	2.5	0.3040	1.180	0.8323
$\beta_H^{(th)}$	1.8	1.3	0.2754	1.810	0.3144
$\beta_H^{(dh)}$	2.5	0.5	0.2252	2.462	0.09661
$\beta_H^{(lag)}$	0.4	0.1	0.2368	0.2764	0.01888
σ	19	2	0.2866	18.51	0.4769

Table 3: Table of cohort partitions used for training and forecasting.

histogram of the residual. The residuals are close to zero mean and have slightly heavier tails that a normal distributed random variables sharing the first two moments. In figure 13 to the right, we have plotted the residual squared against the predicted values $g(\boldsymbol{X}_{lt})$. The expected value of these points as a function of $\boldsymbol{X}_{lt}$ and σ is given by (50) and we have plotted this function for the expected value of the sigma fitted in model 1 in green and the expected value of the sigma fitted in model 2 in red. For data points having high predicted immigration, the distribution of points follow the expected value line rather poorly. This might be an indication of model misspecification. We will in this study however focus on the predictive power of the model. In general, the models performed similarly on training data. The mean of the absolute value of the residuals were 21.42 for model 1 and 21.25 for model 2.

5.2.3 Predictions

Due to time constraints, the immigration model have not been systematically tested on immigration counts in the same manner as the intrinsic demographic variables have. The regression model works well on training data and assuming that it has not overfitted the training data, it should be expected to work on test data as well.

6 Discussion

The model was successfully applied to one year forecasts of demographic variables, and tested on a large set of points. The hierarchic beta-binomial approach seems to be able to produce unbiased forecasts of all demographic variables for the entire range of populations consistently for the time period of 2006 to 2015 on which the framework was tested, given that the correct homogeneous cohort assumptions are made. These seem to differ between the demographic variables and the size of the populations which we are trying to predict, but a general

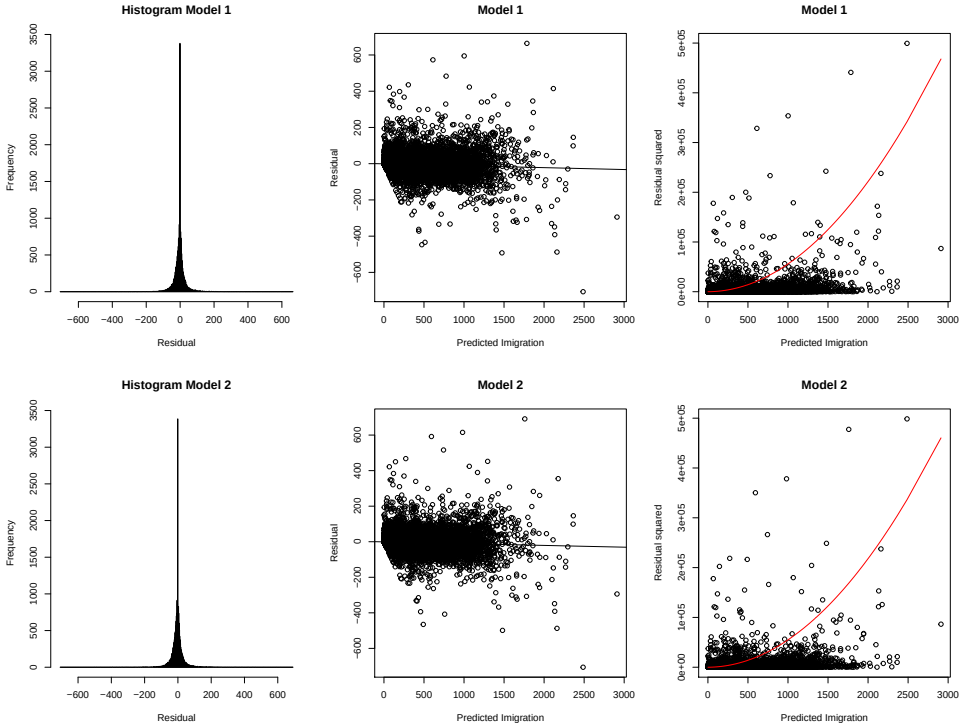

Figure 13: Residual diagnostics of both models on training data

pattern that emerged was that for births and emigration, we need to specify the age and gender specific demographic rates separately for each municipality. For deaths however, the rates obtained from Stockholm county produced accurate predictive distributions of deaths counts, meaning that the death rates can be concluded to be the same between municipalities. Not choosing an excessively fine homogeneous cohort assumption is helpful when considering the computational feasibility of implementing these models. Multiple year ancestral simulations is going to involve a huge network of entangled probability distribution of various demographic variables and population counts and as the homogeneous cohort assumptions becomes finer, the computational requirements increases.

The tendency of the models to overestimate death counts may be attributed to the general trend of declining death rates for all age groups [25]. The time dependent one year estimations of death rates from Figure 9 shows a slightly declining trend for the age group that was sampled and the long term decline of death rates is a well established fact [7]. This demonstrate the need to include trends in the model when making long term forecasts. By observing long term time series data of the demographic rates, one can identify the historic variation by fitting an ARIMA-model to the expected demographic rates. Even if there is no clear trend, one can still estimate the variation of the demographic rates from time to time, and the additional uncertainty of future demographic rates can be included. This is exactly what [7] does with mortality rates, and [24]

expands this framework to all demographic processes.

The model of immigration counts produced unbiased estimates on the training data. Plotting the residuals squared against the predicted values in figure 13 did however reveal a possibility of misspecification, as the squared residuals did not seem to have the predicted expected value specified in (50). The actual residuals seemed to be smaller than expected for large predicted immigration counts and slightly higher for smaller predicted immigration counts. This could take the effect of skewing the predictive distributions to have too high variance for high predicted immigration counts (leading to an predictive distribution with unnecessarily wide credible bounds) and too confident estimations for cases with low predicted immigration counts. In further studies, one could consider the generalization of the negbin-II-model, which is the negbin-X-model as described in [16]. In this model family, the residual dependence of the right term in (50) is no longer limited to a quadratic relation but can be have any positive real number X as exponent. This extra degree of freedom will make it easier for the red line in the rightmost graphs of figure 13 to follow the actual data. Another interesting part of the immigration count model is the notion of "null-area", i.e. areas where the value of all covariates are zero. The most commonly estimated value of the intercept is around one, and hence, an area without existing emigration, deaths and newly built housing should expect one person moving in. Observing the actual outcomes of immigration counts in these null-areas reveals that the vast majority of these areas indeed have no immigration, and the model hence does a poor job of estimating the immigration counts of null areas in its current form. We hence recommend further studies to utilize a *Zero-inflated* model, more on which can be read in [16].

References

[1] J. Graunt. *Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality.* John Martyn and James Allestry, London. 1662.

[2] J. B. S. Haldane. "The precision of observed values of small frequencies". In: *Biometrika* 35 (1948), pp. 297–300.

[3] N. et al. Metropolis. "Equation of State Calculations by Fast Computing Machines". In: *Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[4] H. G. Tucker. "A Generalization of the Glivenko-Cantelli Theorem". In: *The Annals of Mathematical Statistics* 30.3 (Sept. 1959), pp. 828–830.

[5] W. et al. Edwards. "Bayesian statistical inference in psychological research". In: *Psychological review* 70 (1963), pp. 193–242.

[6] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[7] R. D. Lee and L. R. Carter. "Modeling and Forecasting U. S. Mortality". In: *Journal of the American Statistical Association* 87.419 (1992), pp. 659–671.

[8] M. K. Cowels and B. P. Carlin. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *Journal of the American Statistical Association* 91.434 (June 1996), pp. 883–904.

[9] W. R. Gelman A. Gilks and G. O. Roberts. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms". In: *The Annals of Applied Probabilit* 7.1 (1997), pp. 110–120.

[10] P. Preston S. Heuveline and M. Guillot. *Demography*. Oxford: Blackwell, 2001.

[11] G. Castella and R. L. Berger. *Statistical Inference*. Second edition. Duxbury, Thompson Learning, 2002.

[12] D. Q. Keilman N. Pham and A. Hetland. "Why population forecasts should be probabilistic - illustrated by the case of Norway". In: *Demographic Research* 6.15 (2002), pp. 409–454.

[13] C. P. Robert and G. Castella. *Monte Carlo Statistical Methods*. Second edition. Springer, 2004.

[14] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] W. Reichmund and S. Sarferaz. "Bayesian Demographic Modeling and Forecasting: An Application to U.S. Mortality". In: *SFB 649 Discussion Paper 2008-052* (2008).

[16] R. Winkelman. *Econometric Analysis of Count Data*. Fifth edition. Springer, 2008.

[17] J. Smits and C. Monden. "Twinning across the Developing World". In: *PLoS ONE Public Library of Science* 6.9 (2011).

[18] L. Torgo. *Data Mining With R - learning with case studies*. CRC Press, 2011.

[19] P. M. Lee. *Bayesian statistics: an introduction*. Fourth edition. Oxford University Press, 2012.

[20] J. R. Bryant and P. J. Graham. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources". In: *Bayesian Analysis* 8.3 (2013), pp. 591–622.

[21] K. B. Newbold. *Population Geography*. Second edition. Rowman & Litlefield, 2014.

[22] G. M. et al. Box G. E. P. Jenkins. *Time Series Analysis, Forcasting and Control*. Fifth edition. Wiley, 2015.

[23] L. E. Sucar. *Probabilistic Graphical Models, Principles and Applications*. Springer, 2015.

[24] A. et al. Wiśniowski. "Bayesian Population Forecasting: Extending the Lee-Carter Method". In: *Demography* 52 (2015), pp. 1035–1059.

[25] Stefan Lundgren. *Sveriges framtida befolkning 2016–2060/The future population of Sweden 2016–2060*. Tech. rep. Statistiska centralbyrån/Statistics Sweden, 2016.

Appendices

A Further plots and results

In this section, we have appended the plots that does not fit in the results-section but is however an important part of the results. Figure 14 to 43 depicts the test residuals of single year intrinsic demographic variable forecasts and the share of observations falling within the 90 percent credible interval. The red line on the right diagrams of each figure represents the average number of observations for the value of the variable specified in the x-axis lable. The points in the left figure simply plots the residuals. Table 4 shows the correlations of the joint posterior distribution of the parameters. Figure 44 to 54 depicts the Markov-chain stages for each parameter as well as the sample autocorrelation function and a histogram of the posterior distribution.

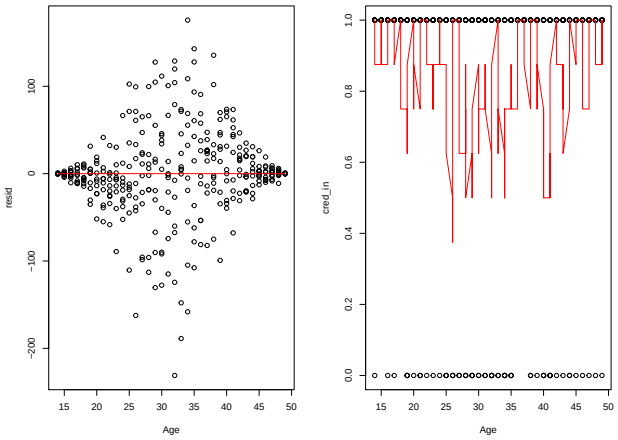

Figure 14: Births, Partition B, Schema B*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

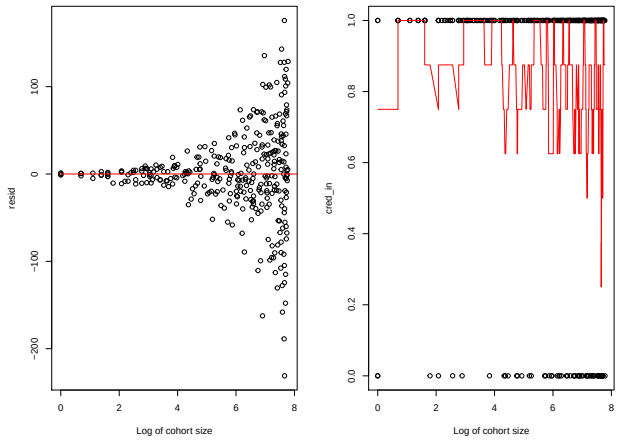

Figure 15: Births, Partition B, Schema B*: Residuals as function of number of births in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

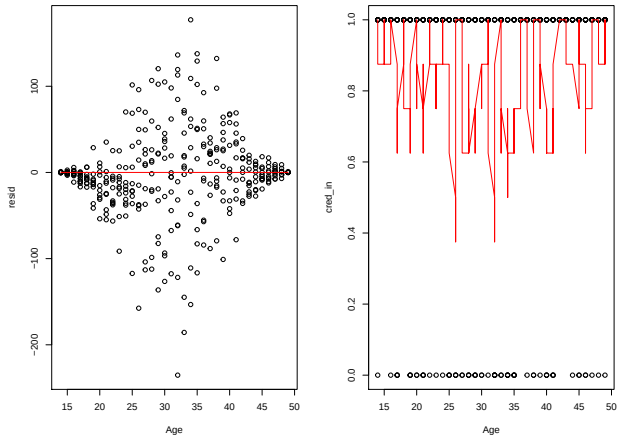

Figure 16: Births, Partition B, Schema D*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

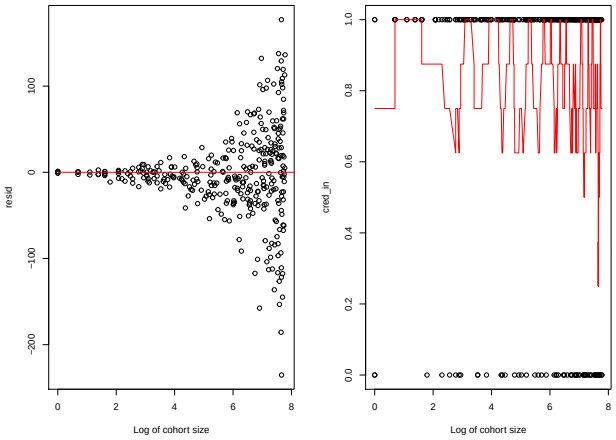

Figure 17: Births, Partition B, Schema B*: Residuals as function of number of births in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

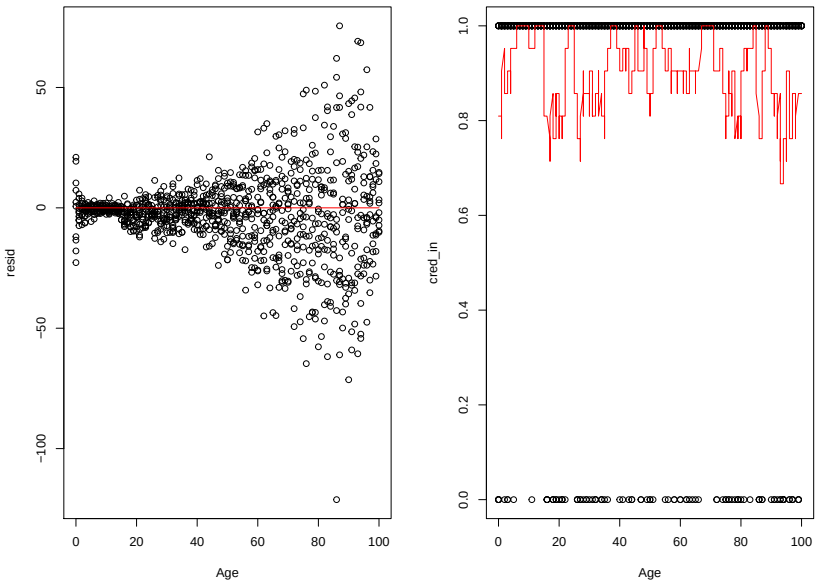

Figure 18: Deaths, Partition B, Schema B*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

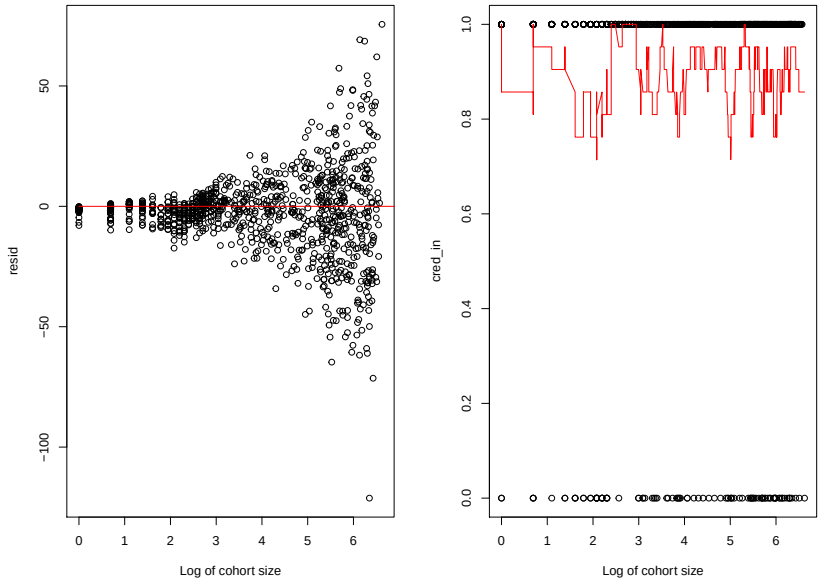

Figure 19: Deaths, Partition B, Schema B*: Residuals as function of number of deaths in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

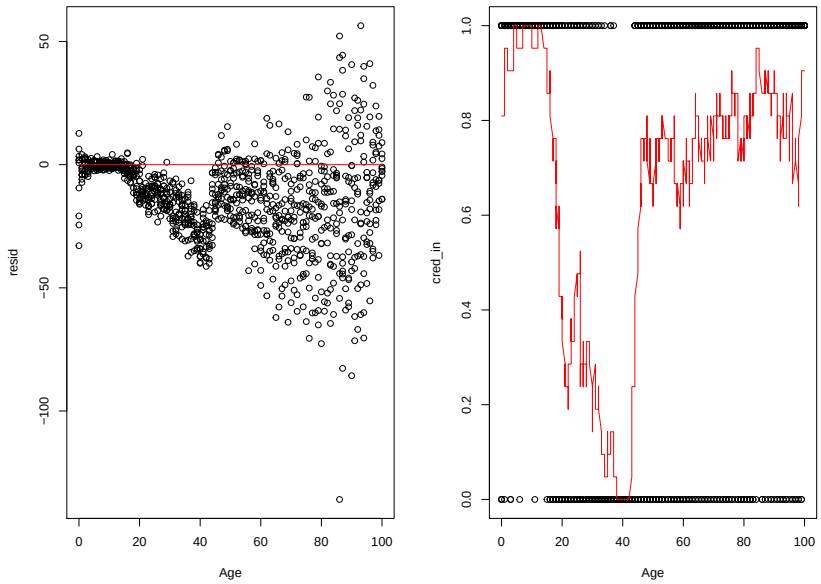

Figure 20: Deaths, Partition B, Schema D*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

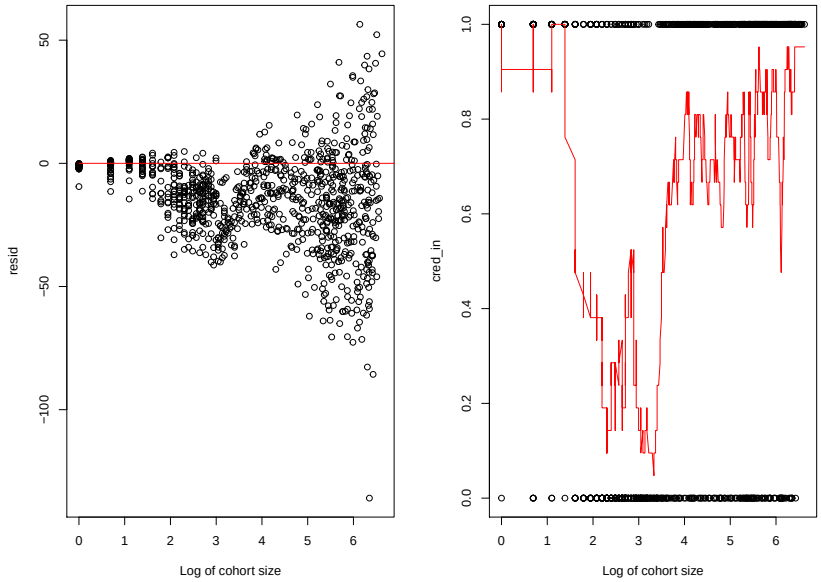

Figure 21: Deaths, Partition B, Schema D*: Residuals as function of number of deaths in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

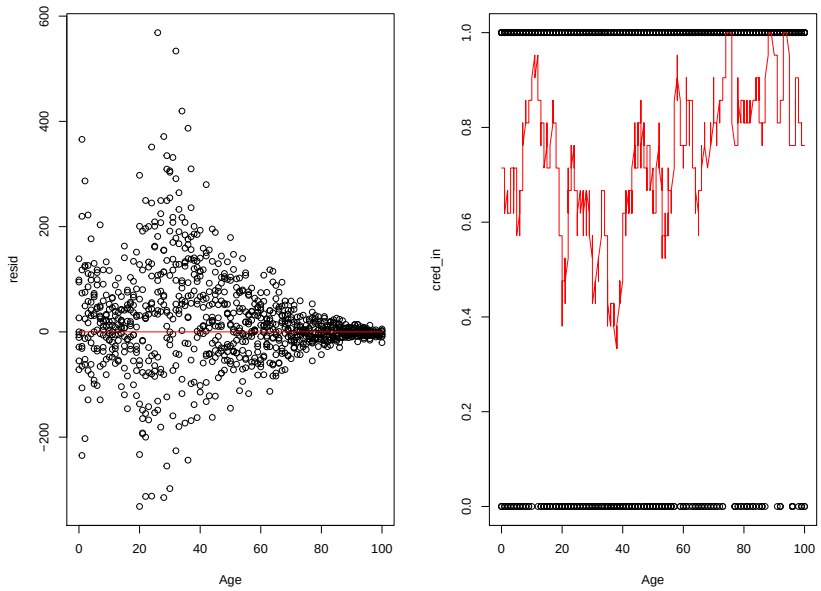

Figure 22: Emigrations, Partition B, Schema B*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

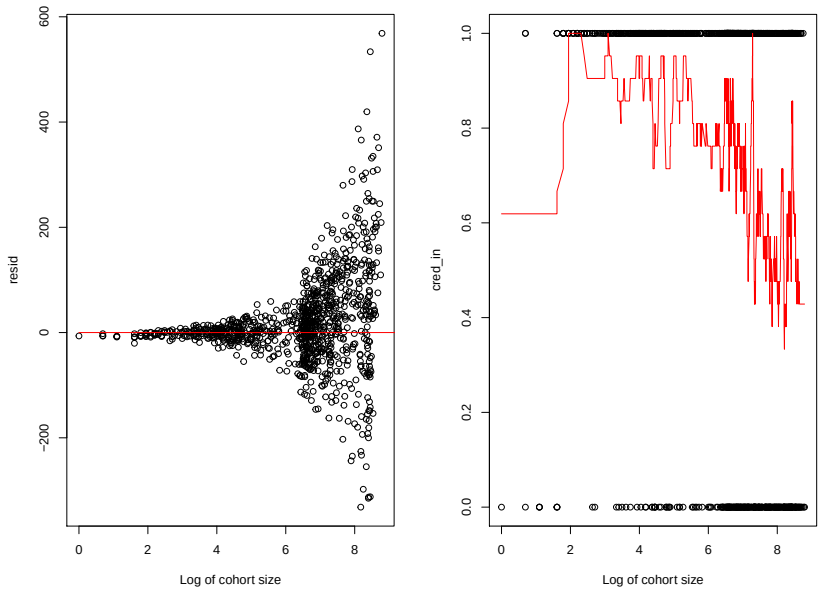

Figure 23: Emigrations, Partition B, Schema B*: Residuals as function of number of emigrations in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

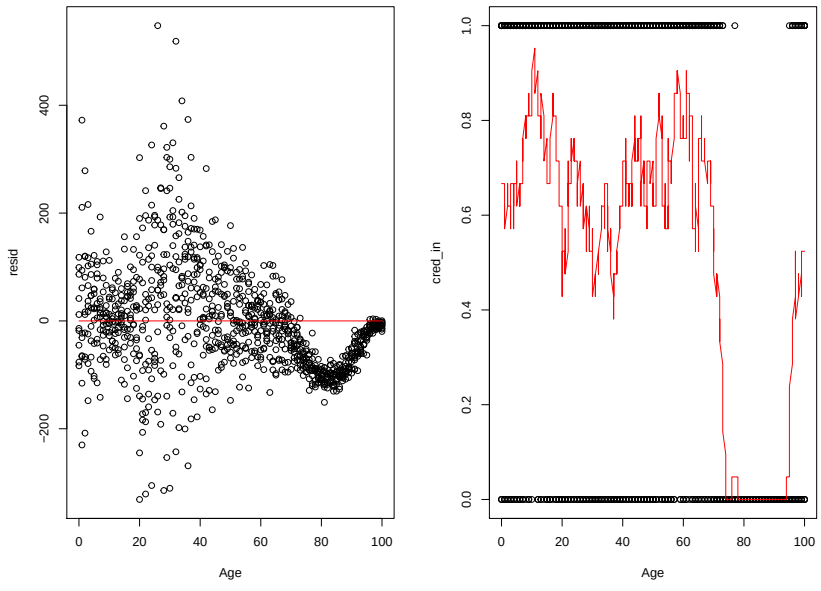

Figure 24: Emigrations, Partition B, Schema D*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

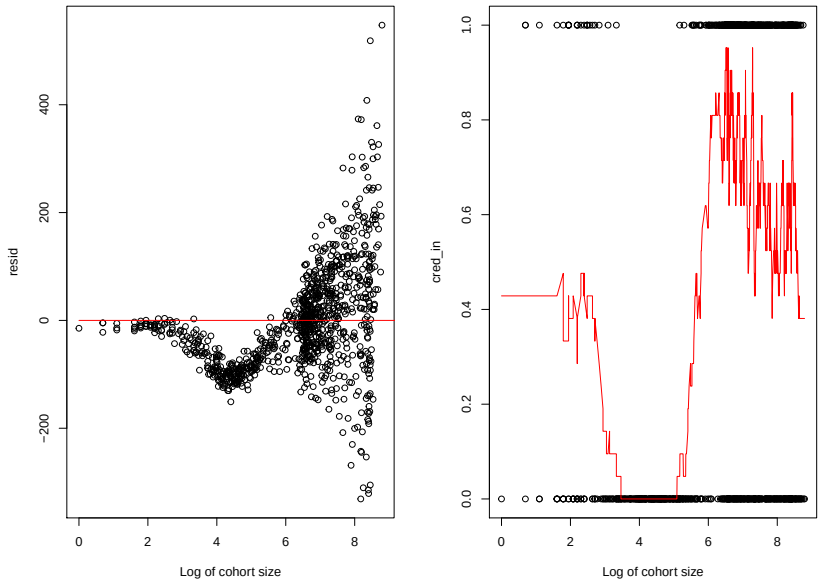

Figure 25: Emigrations, Partition B, Schema D*: Residuals as function of number of emigrations in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

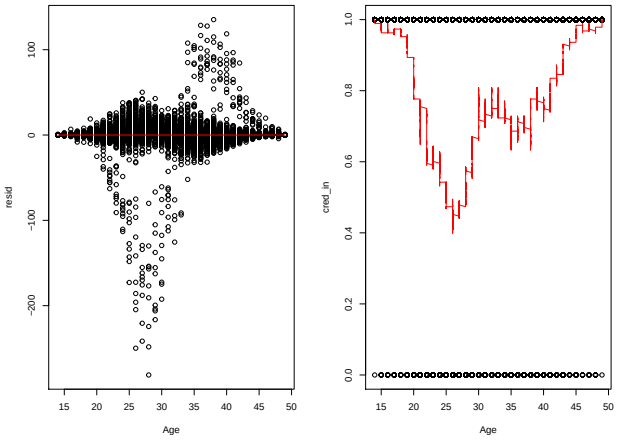

Figure 26: Births, Partition D, Schema B*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

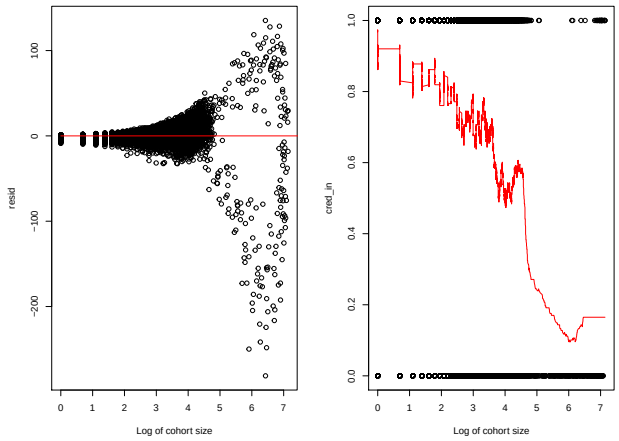

Figure 27: Births, Partition D, Schema B*: Residuals as function of number of births in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

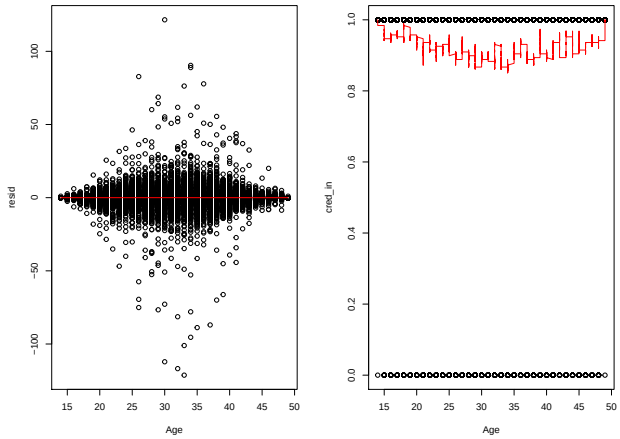

Figure 28: Births, Partition D, Schema D*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

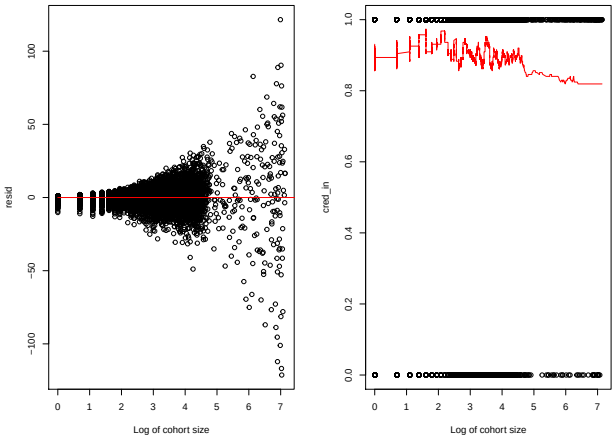

Figure 29: Births, Partition D, Schema D*:Residuals as function of number of births in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

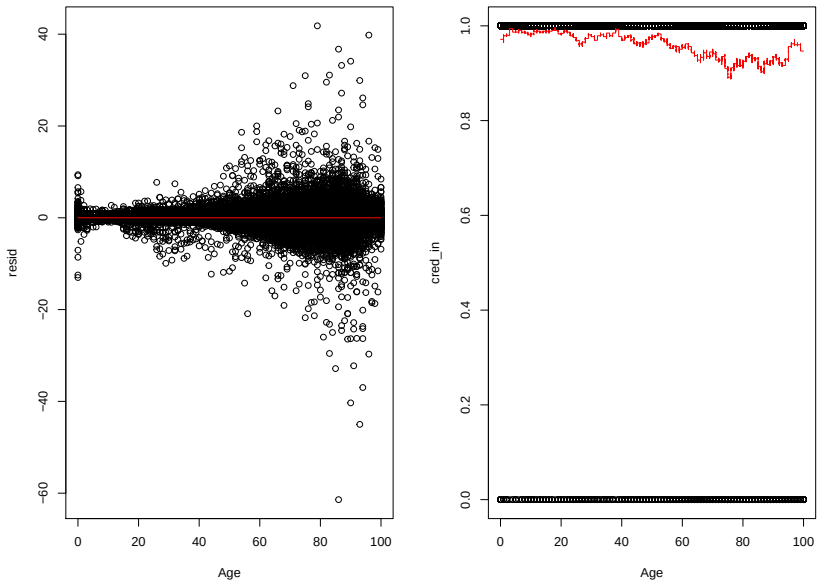

Figure 30: Deaths, Partition D, Schema B*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

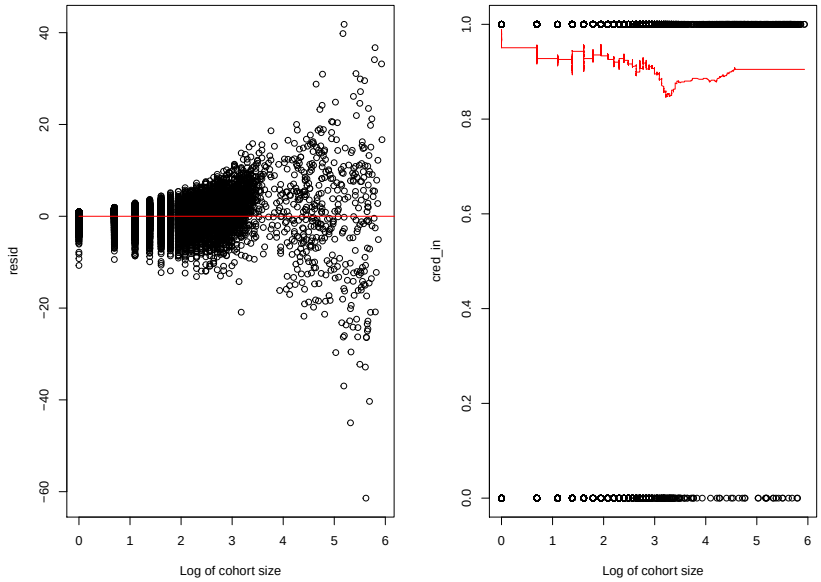

Figure 31: Deaths, Partition D, Schema B*: Residuals as function of number of deaths in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

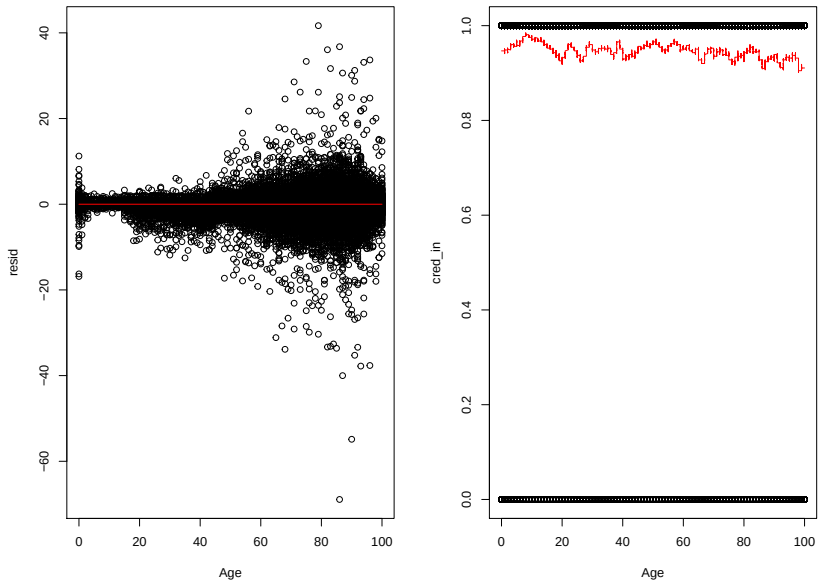

Figure 32: Deaths, Partition D, Schema D*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

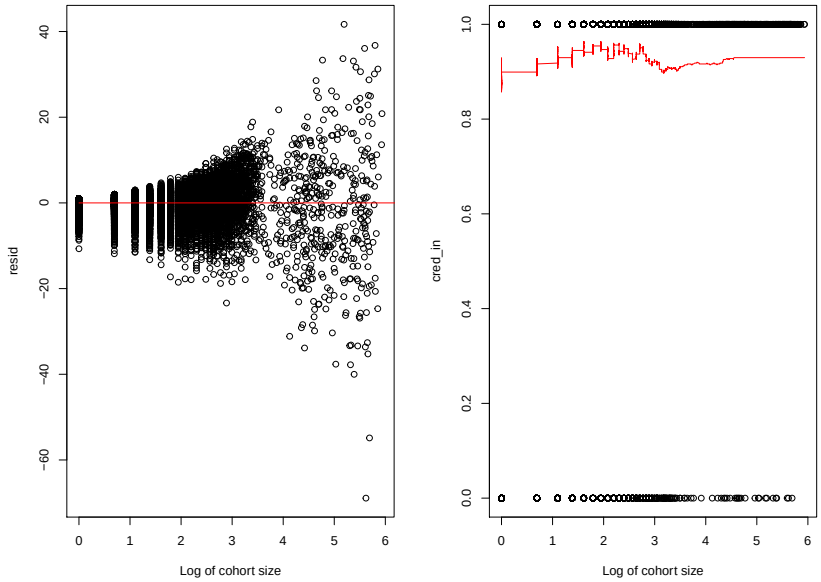

Figure 33: Deaths, Partition D, Schema D*: Residuals as function of number of deaths in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

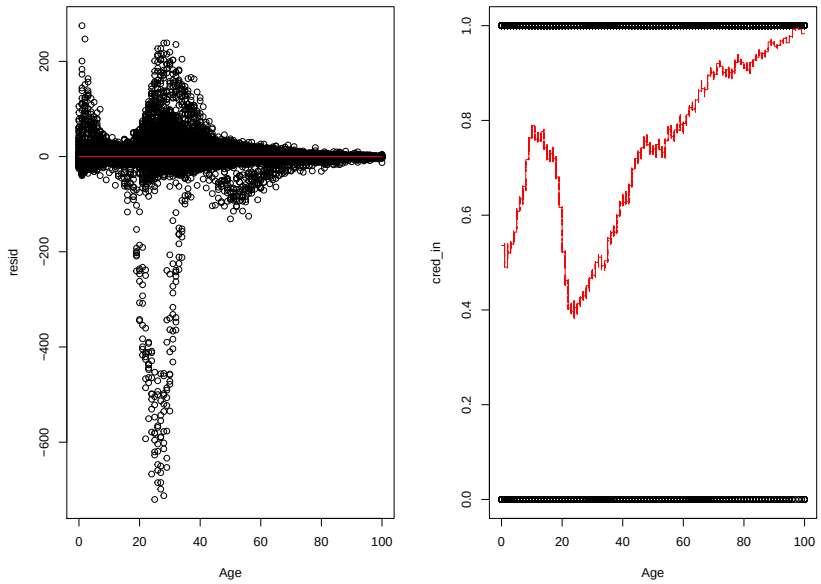

Figure 34: Emigrations, Partition D, Schema B*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

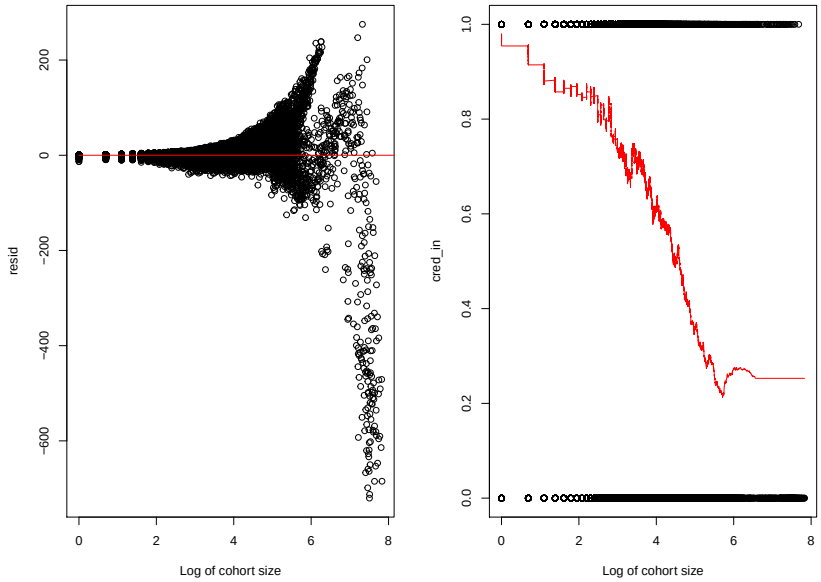

Figure 35: Emigrations, Partition D, Schema B*: Residuals as function of number of emigrations in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

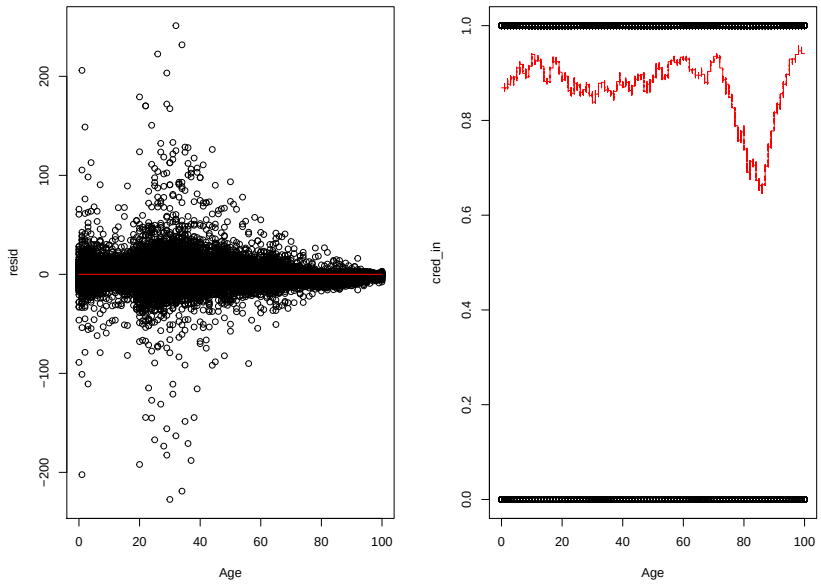

Figure 36: Emigrations, Partition D, Schema D*: Residuals as function of age to the left, the share of observations that fall within 90% credible interval as function of age to the right

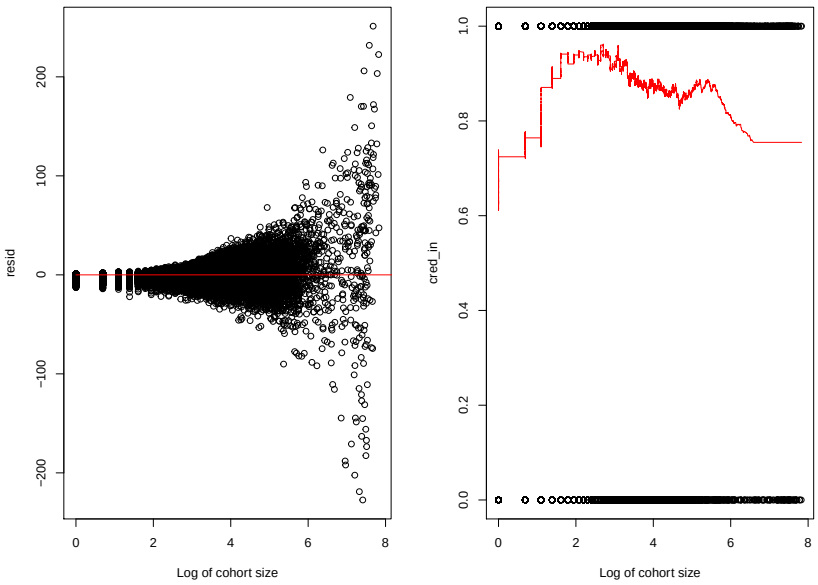

Figure 37: Emigrations, Partition D, Schema D*: Residuals as function of number of emigrations in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

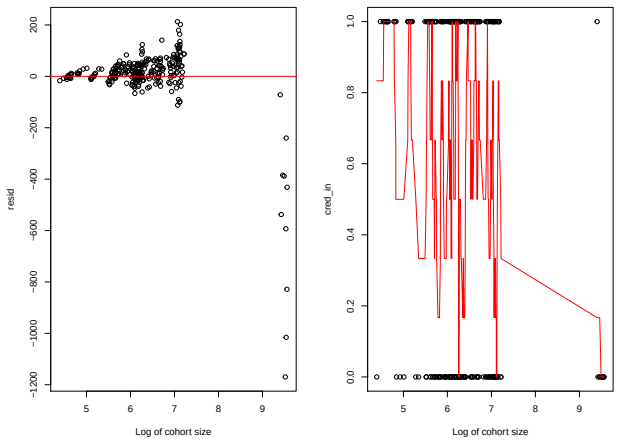

Figure 38: Births, Partition C, Schema B*: Residuals as function of number of births in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

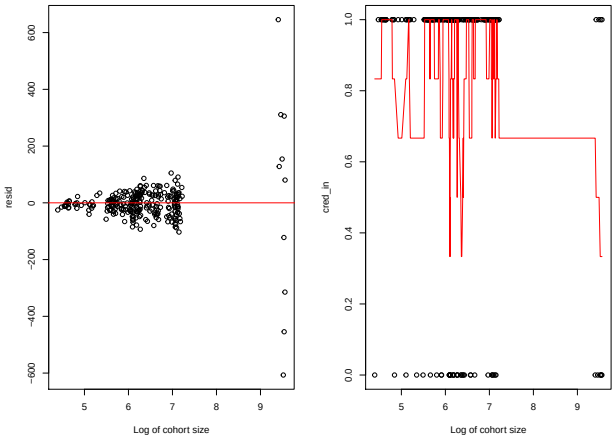

Figure 39: Births, Partition C, Schema D*: Residuals as function of number of births in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

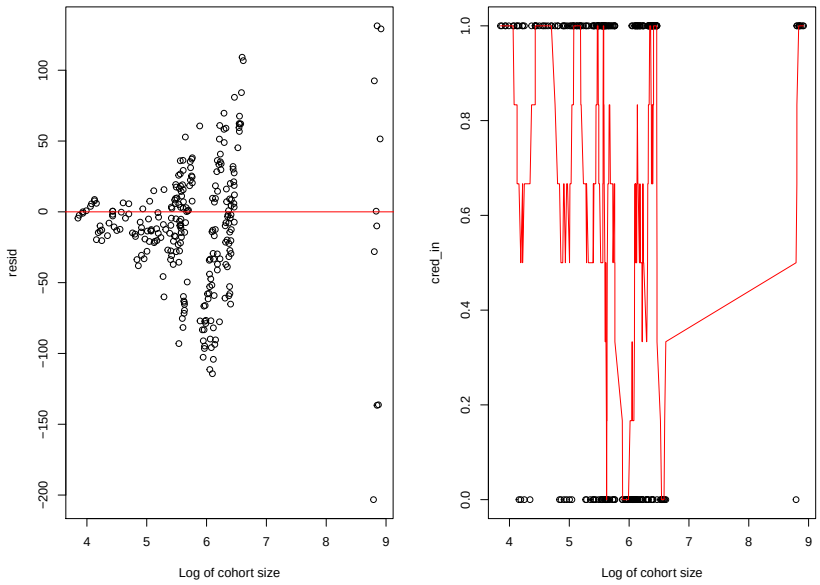

Figure 40: Deaths, Partition C, Schema B*: Residuals as function of number of deaths in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

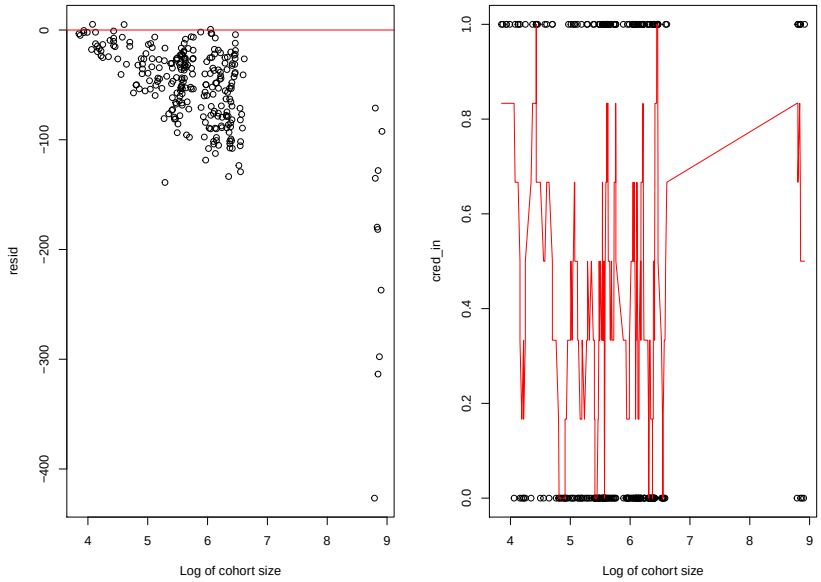

Figure 41: Deaths, Partition C, Schema D*: Residuals as function of number of deaths in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

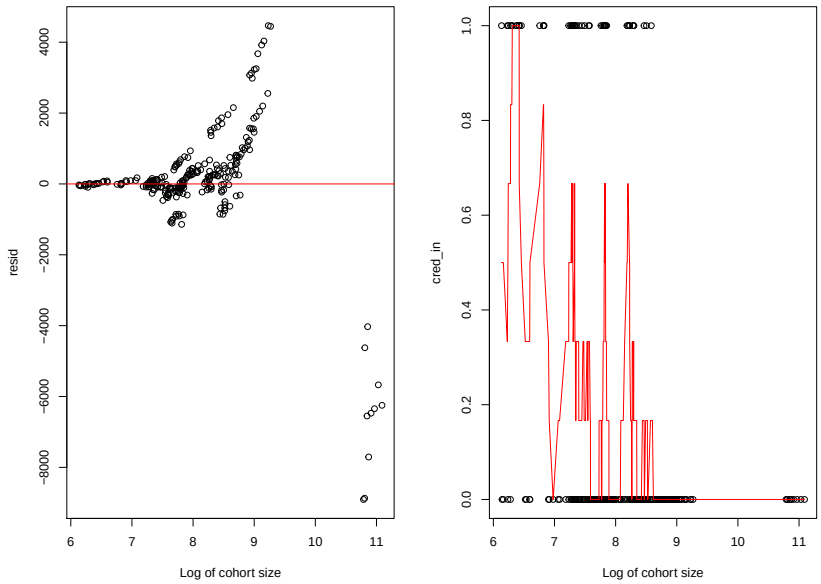

Figure 42: Emigration, Partition C, Schema B*: Residuals as function of number of emigrations in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

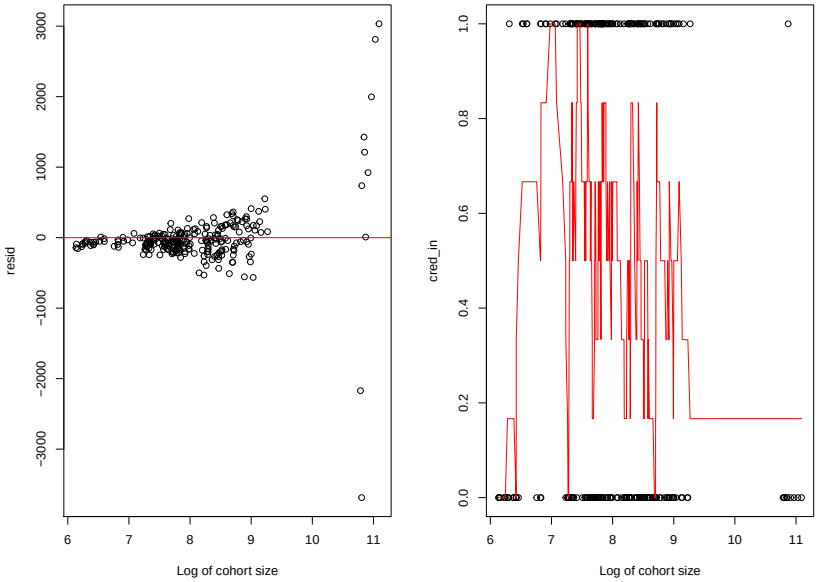

Figure 43: Emigration, Partition C, Schema D*: Residuals as function of number of emigrations in cohort to the left, the share of observations that fall within 90% credible interval as function of number of births in cohort to the right

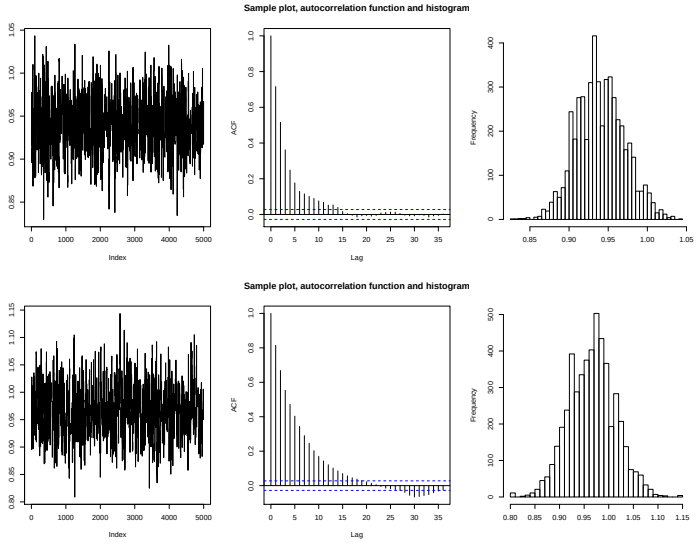

Figure 44: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_0 for Model 1, Lower - β_M for Model 1

	β_0	β_M	β_Q	$\beta_H^{(ap)}$	$\beta_H^{(th)}$	$\beta_H^{(dh)}$	$\beta_H^{(lag)}$	σ
β_0	1	-0.022	-0.093	0.0095	0.037	-0.0073	-0.034	-0.011
β_M	0.032	1	**-0.55**	0.0050	0.030	-0.017	-0.0097	-0.0075
β_Q	**-0.11**	**-0.53**	1	**-0.11**	0.0088	**-0.16**	-0.086	-0.053
$\beta_H^{(ap)}$				1	0.0076	0.028	**-0.13**	0.017
$\beta_H^{(th)}$	0.0030	0.049	-0.043		1	-0.065	-0.060	-0.033
$\beta_H^{(dh)}$	-0.025	0.014	**-0.15**		**-0.10**	1	**-0.14**	0.029
$\beta_H^{(lag)}$	0.044	0.0088	**-0.13**		-0.025	**-0.12**	1	0.013
σ	0.0054	0.021	-0.050		-0.021	-0.010	0.029	1
$\beta_H^{(ap0)}$	-0.023	-0.036	-0.0086		-0.021	-0.0066	-0.0040	0.013
$\beta_H^{(ap1)}$	0.015	0.047	-0.050		-0.021	0.039	-0.049	-0.016
$\beta_H^{(ap2)}$	0.0058	-0.019	0.026		0.030	0.034	-0.016	0.0041
$\beta_H^{(ap3)}$	0.013	0.017	-0.042		-0.024	-0.058	-0.033	0.023
$\beta_H^{(ap4)}$	-0.065	-0.052	0.024		0.0060	0.014	-0.049	-0.025
$\beta_H^{(ap5)}$	-0.0032	0.053	-0.050		0.023	-0.0074	0.042	0.023

	$\beta_H^{(ap0)}$	$\beta_H^{(ap1)}$	$\beta_H^{(ap2)}$	$\beta_H^{(ap3)}$	$\beta_H^{(ap4)}$	$\beta_H^{(ap5)}$
$\beta_H^{(ap0)}$	1					
$\beta_H^{(ap1)}$	-0.0081	1				
$\beta_H^{(ap2)}$	-0.066	0.010	1			
$\beta_H^{(ap3)}$	0.032	**-0.28**	**-0.69**	1		
$\beta_H^{(ap4)}$	-0.010	0.036	-0.019	**-0.40**	1	
$\beta_H^{(ap5)}$	0.016	-0.0039	-0.012	-0.061	**-0.26**	1

Table 4: Correlations between marginal posterior distribution for each parameter. The upper-right side of the diagonal are the correlations of model 1 and the lower-left side of the diagonal are the correlations of model 2. Both are trained on data from 2005 to 2010. The correlations smaller than -0.1 have been bolded.

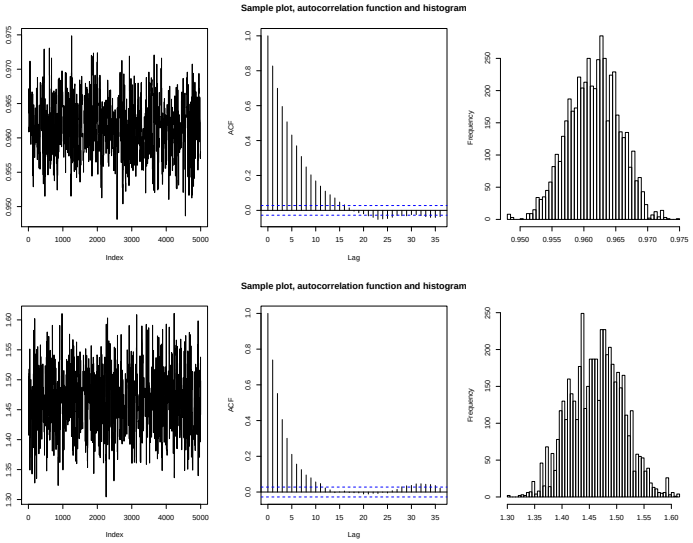

Figure 45: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_Q for Model 1, Lower - β_H^{ap} for Model 1

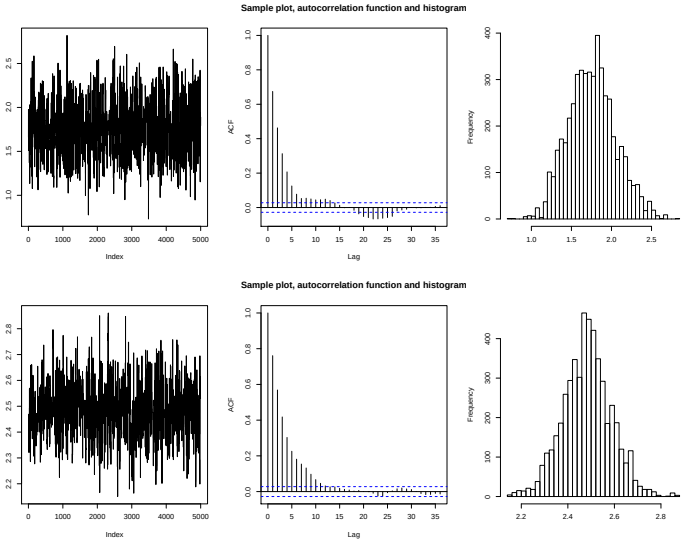

Figure 46: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_H^{th} for Model 1, Lower - β_H^{dh} for Model 1

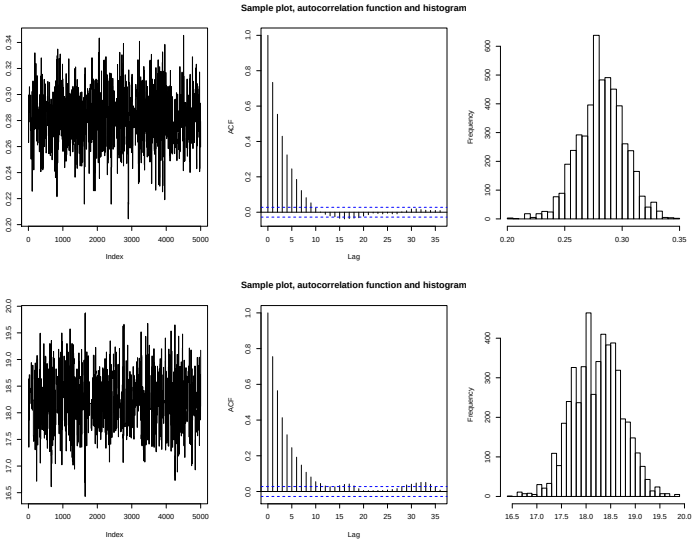

Figure 47: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_H^{lag} for Model 1, Lower - σ for Model 1

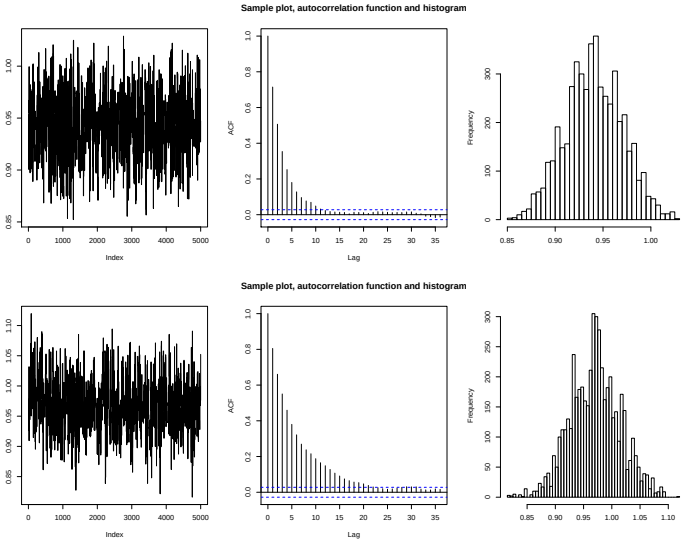

Figure 48: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_0 for Model 2, Lower - β_M for Model 2

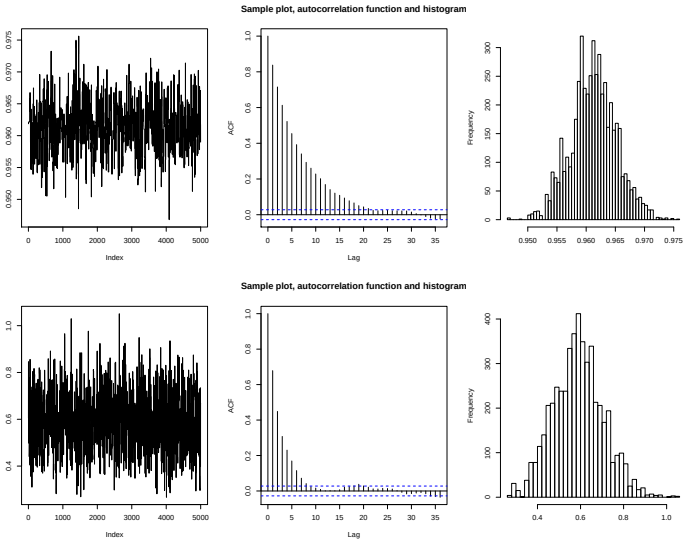

Figure 49: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_Q for Model 2, Lower - β_H^{ap0} for Model 2

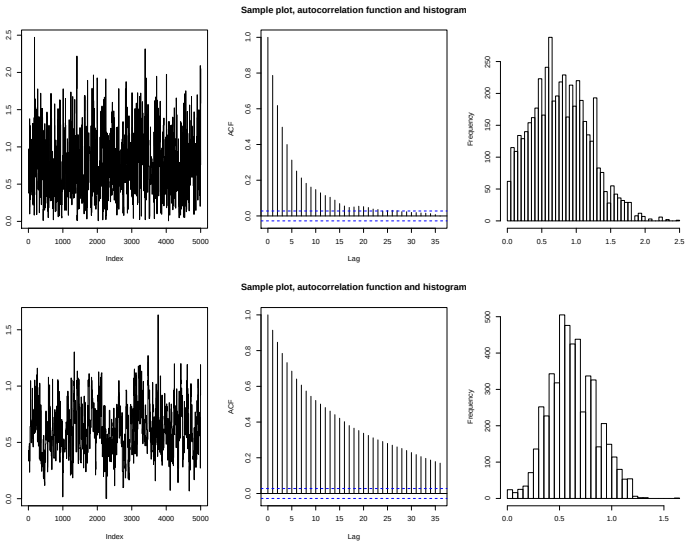

Figure 50: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_H^{ap1} for Model 2, Lower - β_H^{ap2} for Model 2

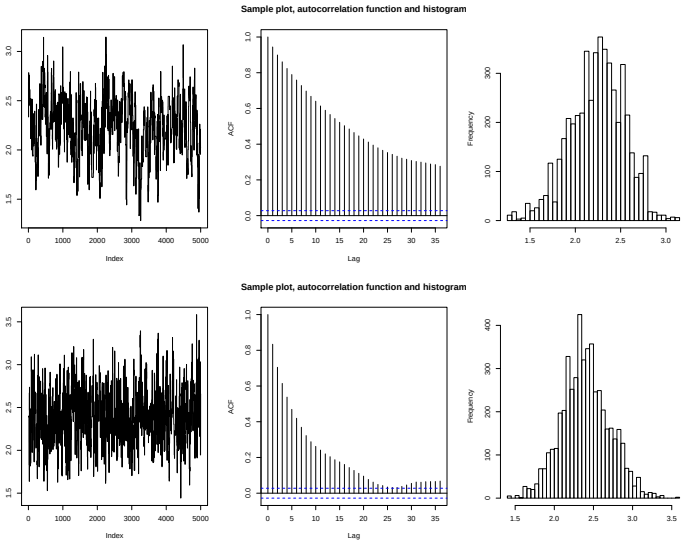

Figure 51: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_H^{ap3} for Model 2, Lower - β_H^{ap4} for Model 2

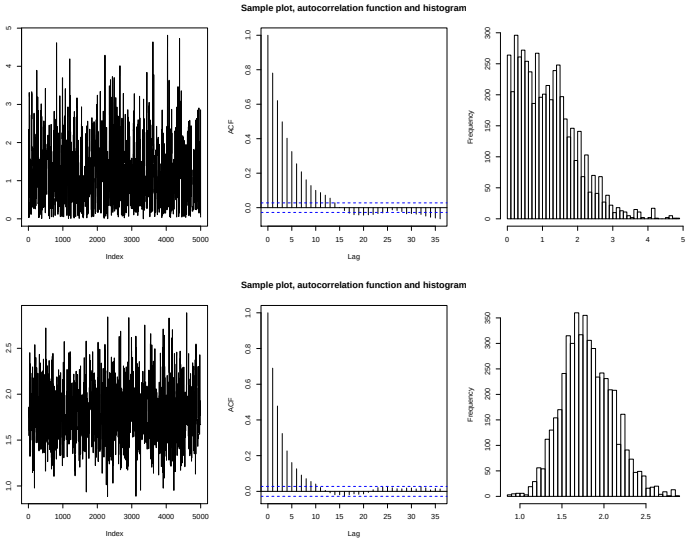

Figure 52: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_H^{ap5} for Model 2, Lower - β_H^{th} for Model 2

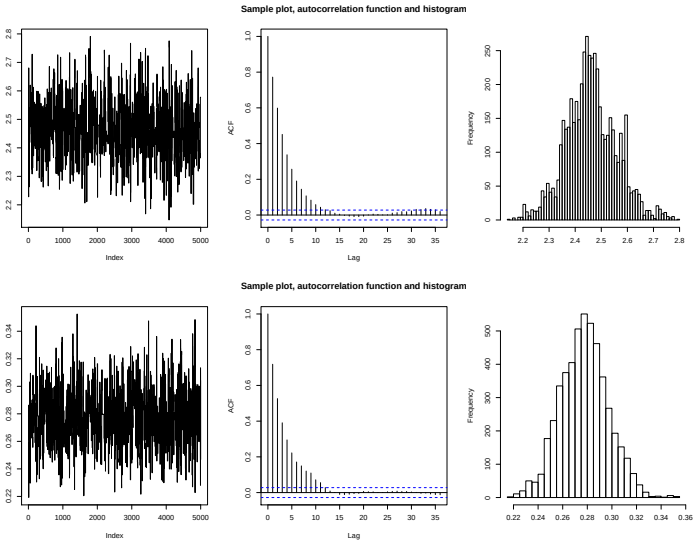

Figure 53: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, Upper - β_H^{dh} for Model 2, Lower - β_H^{lag} for Model 2

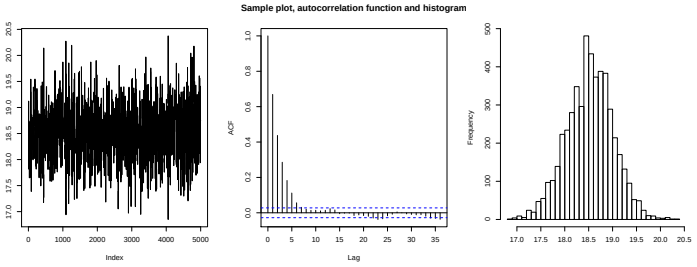

Figure 54: Left - Plot of MCMC-samples, Middle - Autocorrelation function, Right - Histogram of marginal empirical distribution, σ for Model 2