A Bayesian Semiparametric Approach for Trend-Seasonal Interaction: an Application to Migration Forecasts

Alice Milivinti*and Giacomo Benini

Université de Genève, Bd du Pont d'Arve 40, CH - 1211 Genève

April 1, 2019

Abstract

The current paper models complex trend-seasonal interactions within a Bayesian framework. The contribution devides in two parts. First, it proves, via a set of simulations, that a semiparametric specification of the interplay between the seasonal cycle and the global time trend outperforms parametric and nonparametric alternatives when the seasonal behavior is represented by Fourier series of order bigger than one. Second, the paper uses a Bayesian framework to forecast Swiss immigration merging the simulations' outcome with a set of priors derived from alternative hypothesis about the future number of incomers. The result is an effective symbiosis between Bayesian probability and semiparametric flexibility able to reconcile past observations with unprecedented expectations.

Keywords: Trend-Seasonal Interaction, Bayesian Forecast, Semiparametric, Immigration.
JEL Classification: C11, C14, C52, C53, J11.
Corresponding Author: alice.milivinti@unine.ch

*The research leading to these results has received funding from the Swiss National Science Foundation in the context of the NCCR on the move.

1 Introduction

In a world characterized by low fertility and increasing life expectancy, human mobility has gained prominence in driving population change. Any forward-looking policy should be designed around its forecast. However, a common belief around its non-repeatability and non-traceability has left most migration projections dependent upon deterministic methods until the mid 1990s. Nevertheless, being unquestionable that forecasting is not only about modelling, but also about quantifying uncertainty, these naïve approaches were progressively abandoned and substituted by standard probabilistic techniques like time series analysis (Lee & Tuljapurkar, 1994; De Beer, 1997; Keilman, Pham, Hetland, et al., 2002; Wilson & Bell, 2004; Raymer, Abel, & Rogers, 2012) and generalized linear regressions (Schmidt & Fertig, 2000; Alvarez-Plata, Brücker, & Siliverstovs, 2003; Cohen, Roig, Reuman, & GoGwilt, 2008; Cappelen, Skjerpen, & Tønnessen, 2015).

Even though this transition represented a significant leap forward in statistical rigor (Lutz & Goldstein, 2004), it was not immune from criticism. In particular, the substitution of a subjective approach with a set of data-driven methods has created a series of specifications unable to reconcile the expectations of the researchers with the information contained in the data. This major shortcoming represents a significant threat to the precision of out-of-sample forecasts, especially in the case of unique episodes. Suppose, for example, that a country signs a treaty which guarantees the free movement of people from neighbouring nations. Any prediction which ignores such unprecedented event would most probably produce large errors.

A possible way to prevent this type of mistakes is to adopt a Bayesian prospective (Bijak & Wiśniowski, 2010; Bijak, 2010; Billari, Graziani, & Melilli, 2014; Azose & Raftery, 2015; Azose, Ševčíková, & Raftery, 2016) able to condition the future uncertainty on past information and future intuitions, conjugating the "*correspondence to the observed reality*" with "*the awareness of multiple perspectives*" (Gelman & Hennig, 2017). This more convoluted practice is implemented at the expenses of a closed-form solution and the corresponding reliance on numerically intensive algorithms which require an *a priori* hypothesis about the prior distribution of the structural parameters and an *a posteriori* sampling of their probability distribution. Frequently, such a computationally intense way to manage the data is compensated assuming a linear relation between the dependent and the independent variables. This simplification fastens the speed of the algorithm's rate-of-convergence and delivers easy-to-interpret estimates (Blake & Mumtaz, 2012). Nonetheless, a rigid linear specification might fail to describe the migration process ignoring the complexity of global and seasonal trends, as well as, the anomalies related to the interdependency across migration flows.

The present paper explores the possibility of a non-linear interaction between long and short run migration within a Bayesian framework. Empirically, the global tendency is defined as a time trend and the seasonality as a sum of Fourier series. Being particularly malleable, the latter can conveniently model multiple seasonal spikes by simply increasing its order. These two components become the arguments of an unknown bivariate smooth function, which relaxes the hypothesis that trend and seasonality evolve independently (Koopman & Lee, 2009; Hindrayanto, Jacobs, & Osborn, 2014). The nonparametric nature of the interaction does not impose a rigid structure to the trend-seasonal co-movements, returning an Additive Model (AMI). At the same time the Bayesian prospective reconciles the AMI with the long standing demographic tradition of including experts' opinions into the predictions through the use of informative priors.

In order to test the statistical properties of the model we run a set of simulations. The AMI out-

performs, in terms of predictive accuracy, the alternatives, especially if the seasonal component has more than one spike. Given its good performance, we train the AMI on Swiss monthly data to test its capacity to predict the number of arrivals. Like in the simulation case, the AMI predicts the left-out data better than all the other specifications, suggesting, both for long and short run predictions, a non negligible amount of non-linear trend-seasonal interaction. Consequently, we set up altervative forecast scenarios for Swiss immigration flows until 2023 by making an active use of informative priors. Finally, we extend the model to a longitudinal analysis, producing age specific forecasts. This last case reduces the impact of the interaction, while still returning high degrees of non-linearity.

2 Methodological Framework

2.1 Model Construction

The basic model of the paper decomposes the log-transformation of the number of monthly arrivals, $\log Y_t = y_t$, into a trend and a seasonal component,

$$y_t = \text{trend}_t + \text{seasonality}_t + \text{e}_t, \quad t = 1, 2, ..., T. \tag{1}$$

The first element of equation (1) is the global direction of the series, defined as t/T, while the second one is the seasonal cycle obtained by averaging the "de-trended" series for each month over all periods. The random part of the series, e_t, is obtained removing the sum $(\text{trend}_t + \text{seasonality}_t)$ from the original time series.
The seasonal part of the regression can be parsimoniously modeled using a harmonic series. Recalling Fourier theorem, a periodic function, with period $p = 2\pi/\varpi$, can be rewritten using the cosine function $\cos(\varpi(t))$. For example, an annual seasonal pattern ($p = 12$) has a frequency $\varpi = \pi/6$. All the same, a model which includes only a cosine function would assume by default that the annual peak is the first period of the season, normally January. In order to generalize this result, and allow the model to shift the initial spike, it is sufficient to add a phase shifter, such that

$$\text{seasonality}_t = a_t \cos(\varpi(t) + \theta), \tag{2}$$

where θ is the magnitude of the shift. Given that equation (2) can be rewritten as the sum of cosine and sine, it is possible to express the whole seasonality as

$$\text{seasonality}_t = a_t \cos(\varpi(t)) + b_t \sin(\varpi(t)). \tag{3}$$

Equation (3) reproduces a cyclic behaviour with a single peak. In order to allow for multiple spikes it is necessary to add further sine and cosine terms up to the desired order. For example the N-th seasonality would be

$$\text{seasonality}_t = a_t \cos(\varpi(t)) + b_t \sin(\varpi(t)) + ... + v_t \cos(\varpi(Nt)) + z_t \sin(\varpi(Nt)), \tag{4}$$

where $[a_t, b_t, ..., v_t, z_t]$ define discretionary amplitudes. Substituting equation (4) into the funda-

mental model, allows us to rewrite (1) as a Linear Model with No Interaction (LMNI),

$$y_t = \beta_0 + \beta_1 \text{trend}_t + \sum_{i=1}^{N} \beta_{2i}(\cos_{it} + \sin_{it}) + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \tag{5}$$

where, we simply sum the trend and the cyclical components using $\cos_{it} = \cos(\varpi(it))$ and $\sin_{it} = \sin(\varpi(it))$.

Regression (5) allows for multiple peaks of a seasonal component which does not need to be in January. Furthermore, it accounts for the possibility that the unconditional mean of the weakly stationary process y_t is different from zero and equal to β_0. However, it leaves unchanged the fundamental assumption that y_t can be expressed as a combination of short-run waves and of a long-run component. Nonetheless, this does not always hold since seasonal times series often display a cycle amplitude which varies for different stages of the trend. Said differently, there might be an interaction between seasonality and trend, which transforms equation (5) into a Linear Model with Interaction (LMI),

$$y_t = \beta_0 + \beta_1 \text{trend}_t + \sum_{i=1}^{N} \beta_{2i}(\cos_{it} + \sin_{it}) + \sum_{i=1}^{N} \beta_{3i} \text{trend}_t * (\sin_{it} + \cos_{it}) + \epsilon_t. \tag{6}$$

A frequentist estimation of equation (6) presents two major shortcomings. First, in contrast to a long standing tradition in demographic studies, it does not allow to introduce a judgmental component to the empirical analysis. Second, the interaction can impact on the number of incomers only linearly. This last restriction can be particularly limiting since there is no empirical evidence supporting the idea that the interaction is linear or of any other specific functional form like $\exp(\text{trend}_t * (\sin_{it} + \cos_{it}))$ (Koopman & Lee, 2009). The combination of these two limitations might translates into estimates far away in probability from the true unobserved parameters. In order to see why, let us consider the following example. A researcher collects monthly migration inflows data to a small open economy well integrated in the global division of labour. An explanatory analysis of the time series suggests that, historically, there is an interaction between the evolution of the trend and the amplitude of the seasonality. In particular, the interaction changes for different stages of the global business cycle, which is expected to expand in the upcoming months.
A short-run forecast based on a frequentist interpretation of equation (5) would ignore the trend-cycle interaction, while equation (6) would only allows for a very specific impact of $(\text{trend}, \cos + \sin)$ on y. A Bayesian Additive Model with Interaction (AMI),

$$y_t = \beta_0 + f_1(\text{trend}_t) + \sum_{i=1}^{N} f_{2i}(\cos_{it} + \sin_{it}) + \sum_{i=1}^{N} f_{3i}(\text{trend}_t, \cos_{it} + \sin_{it}) + \epsilon_t, \tag{7}$$

generalizes the previous expressions allowing the impact of all the explanatory variables to be non-linear while including the expected distribution of the model's parameter β_0 and functions $[f_1 \quad f_{21} \quad \sum_i f_{2i} \quad \sum_i f_{3i}]$. The flexible form of $f_{3i}(.)$ allows to compute complex and potentially non-linear interactions between trend and seasonality, which might be driven by the business cycle. Furthermore, a prior about the distribution of the explanatory variables can incorporate experts' expectations about the increasing variability brought about by an unprecedented event such as the completion of a commercial agreement or the burst of a war potentially improving the model's

performance.
In order to see these two properties working in practice it is necessary to introduce the estimation technique used to fit (7). Using a first order Fourier series with interaction,

$$y_t = \beta_0 + f_1(\text{trend}_t) + f_{21}(\cos_{1t} + \sin_{1t}) + f_{31}(\text{trend}_t, \cos_{1t} + \sin_{1t}) + \epsilon_t, \tag{8}$$

we illustrate the estimation process step-by-step. The first one is to choose which nonparametric technique to implement. Among the different options, thin plate regression splines tend to outperform, at least in finite samples, more traditional alternatives, such as kernels, marginal integration and local polynomial approximation (Wood, 2003). This method requires to choose a base for each unknown function. For example, the nonparametric terms of equation (8) can be re-expressed as

$$f_1(\cdot) = \sum_{k=1}^{K_1} \beta_{1k} b_{1k}(\text{trend}_t) \tag{9}$$

$$f_{21}(\cdot) = \sum_{k=1}^{K_{21}} \beta_{21k} b_{21k}(\cos_{1t} + \sin_{1t}) \tag{10}$$

$$f_{31}(\cdot) = \sum_{k=1}^{K_{31}} \beta_{31k} b_{31k}(\text{trend}_t, \cos_{1t} + \sin_{1t}), \tag{11}$$

where $(\beta_{1k}, \beta_{21k}, \beta_{31k})$ are unknown vectors of parameters, $(b_{1k}(.), b_{21k}(.), b_{31k}(.))$ are basis functions and $k = [1, 2, ..., K_1]$, $k = [1, 2, ..., K_{21}]$ and $k = [1, 2, ..., K_{31}]$ are, respectively, the number of knots used to fit $f_1(.)$, $f_{21}(.)$ and $f_{31}(.)$. Estimating equation (8) using (9), (10) and (11) usually leads to an identification problem. A possible wayout is to center each smooth, such that, $1^T \tilde{\mathbf{X}}_k \tilde{\beta}_k = 0$, where

$$\tilde{\mathbf{X}}_k = [\tilde{\mathbf{X}}_{K_1} : \tilde{\mathbf{X}}_{K_{21}} : \tilde{\mathbf{X}}_{K_{31}}] \quad \text{and} \quad \tilde{\beta}_k = [\tilde{\beta}_{K_1} : \tilde{\beta}_{K_{21}} : \tilde{\beta}_{K_{31}}],$$

with $\tilde{\mathbf{X}}_{K_1} = [b_{11}(.), b_{12}(.), ..., b_{1K_1}(.)]$, $\tilde{\mathbf{X}}_{K_{21}} = [b_{211}(.), b_{212}(.), ..., b_{21K_{21}}(.)]$, $\tilde{\mathbf{X}}_{K_{31}} = [b_{311}(.), b_{312}(.), ..., b_{31K_{31}}(.)]$, $\tilde{\beta}_{K_1} = [\beta_{11}, \beta_{12}, ..., \beta_{1K_1}]^T$, $\tilde{\beta}_{K_{21}} = [\beta_{211}, \beta_{212}, ..., \beta_{21K_{21}}]^T$ and $\tilde{\beta}_{K_{31}} = [\beta_{311}, \beta_{312}, ..., \beta_{31K_{31}}]^T$. To ensure $1^T \tilde{\mathbf{X}}_k \tilde{\beta}_k = 0$, $\tilde{\mathbf{X}}_k \tilde{\beta}_k$ is reparametrized using a single Householder matrix $\mathbf{Z}$, such that

$$\mathbf{X}_k = \tilde{\mathbf{X}}_k \mathbf{Z} \quad \text{and} \quad \tilde{\beta}_k = \mathbf{Z}\beta_k.$$

Once the matrices have been centered the expected value of equation (8),

$$\mathbb{E}[y_t | \text{trend}_t, \cos_{1t} + \sin_{1t}] = \mathbf{X}_{kt}\beta, \tag{12}$$

with $\beta^T = [\beta_0^T, \beta_{K1}^T, \beta_{K21}^T, \beta_{K31}^T]$ can be estimated using a standard likelihood function. However, if the number of knots is large enough, specification (9), (10) and (11) would probably overfit the data. Therefore, thin plate regression splines replace the standard likelihood with a penalized one,

$$l_p(\beta) = l(\beta) - \frac{1}{2} \sum_k \lambda_k \beta^T \mathbf{S} \beta, \tag{13}$$

where, λ is an unknown smoothing parameter and $\mathbf{S}$ is a linear modification of a penalty, which measures the wiggliness of $f_1(\cdot)$, $f_{21}(\cdot)$ and $f_{31}(\cdot)$ as a quadratic form in the coefficients of the function. Likelihood 13 can be insert into a standard Bayesian procedure, setting the prior distributions of β and λ, computing the likelihoods and sampling from the posterior distributions. Further mathematical details are available in the Additional Material.

2.2 Simulations and Priors' Selection

Equations (5), (6) and (7) are estimated using the Stan language provided by the `brms` package of the statistical software R (Bürkner, 2017). The first step to calculate the regressions is to set the priors. Given that, we want to test the properties of the different models via a set of statistical simulations, we opt for weakly informative priors in order to validate the predictive power of the model[1]. In particular, for the parametric part, we use a multivariate uniform distribution defined between minus infinity and plus infinity without correlation among the betas, such that $\beta = [\beta_0, \beta_1, \sum_i \beta_{2i}, \sum_i \beta_{3i}]$, is distributed as

$$\beta \sim U(-\infty, +\infty, \mathbf{0}). \tag{14}$$

The nonparametric part of equation (7) requires a prior on the distribution of the explanatory variables. This reflects the need to make an assumption about every point in some continuous input space to which we associate a particular statistical process. The most common choice is to assume that all the arguments are normally distributed random variables, so that the functions $f_1(.), f_{21}(.), f_{31}(.)$ follow a Standardized Gaussian Process (GP),

$$f_1(\cdot) \sim GP(\cdot|0,1) \quad f_{21}(\cdot) \sim GP(\cdot|0,1) \quad f_{31}(\cdot) \sim GP(\cdot|0,1), \tag{15}$$

and, consequently, the splines' coefficients are normally distributed

$$\begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{1K_1} \\ \beta_{211} \\ \vdots \\ \beta_{21K_{21}} \\ \beta_{31} \\ \vdots \\ \beta_{31K_{31}} \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & \ddots & \dots & \dots & \dots & \dots & 0 \\ \vdots & \dots & \dots & \ddots & \dots & \dots & \dots & \vdots \\ \vdots & \dots & \dots & \dots & \ddots & \dots & \dots & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \ddots & \dots & \vdots \\ 0 & 0 & \dots & \dots & \dots & \dots & 1 & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} \right]. \tag{16}$$

In the same way the smoothing parameter λ follows a Normal distribution, $\lambda \sim N(0, \sigma_\lambda^2)$. The set of priors is completed by the distributions of error's variance term and of the variance of λ. Following the suggestion of Håvard (2016), we propose two standardized half t-student with three

[1]The default priors of the R-package `brms` are employed

degrees of freedom,

$$\sigma_\epsilon \sim t_3(0,1) \qquad \sigma_\lambda \sim t_3(0,1). \tag{17}$$

In practice, the sampling is run on four Markov chains and repeated for 4000 iterations, a reasonable number given the sampling efficacy granted by a Hamiltonian Monte Carlo Sampler, which uses the No-U-Turn Sampler (NUTS) (Neal et al., 2011)[2].

In order to test the model we construct three simulated data generating processes containing a first order Fourier seasonality for $t = 1, .., 360$;

$$y_t^{sim_1} = 3 + 0.4\text{trend}_t + 30\cos\left(\frac{\pi}{6}x_t\right) + 60\sin\left(\frac{\pi}{6}x_t\right) + \epsilon_t,$$
$$y_t^{sim_2} = 3 + 0.4\text{trend}_t + 30\cos\left(\frac{\pi}{6}x_t\right) + 60\sin\left(\frac{\pi}{6}x_t\right) + 0.8\text{trend}_t\left(\cos\left(\frac{\pi}{6}x_t\right) + \sin\left(\frac{\pi}{6}x_t\right)\right) + \epsilon_t,$$
$$y_t^{sim_3} = 3 + 0.4\text{trend}_t + 30\cos\left(\frac{\pi}{6}x_t\right) + 60\sin\left(\frac{\pi}{6}x_t\right) + 0.8\text{trend}_t \exp\left(\left(\cos\left(\frac{\pi}{6}x_t\right) + \sin\left(\frac{\pi}{6}x_t\right)\right)\right) + \epsilon_t.$$

sim_1 presents no trend-seasonal interaction, sim_2 a linear interaction and sim_3 an interaction where seasonality is not linear with respect to y_t. Both in sim_2 and sim_3, the interaction's coefficient is kept small to contain the eventual advantages of the AMI. In all three regressions x_t are drawn from a standardized normal distribution $x_t \sim N(0,1)$, the trend is the cumulative sum of a linear sum of those draws, $\text{trend}_t = \sum_{t=1}^{360}[\sum_{j=1}^{360} x_{tj} + x_t]$, and $\epsilon_t \sim N(0, 0.4)$.
In order to test the properties of the AMI we compare its performances with the ones of a linear model without (LMNI) and with interaction (LMI) and an additive model without interaction (AMNI),

$$y_t = \beta_0 + f_1(\text{trend}_t) + \sum_{i=1}^{N} f_{21}(\cos_{it} + \sin_{it}) + \epsilon_t. \tag{18}$$

To make the comparison independent form any type of structure we further introduce, as a limit case, a purely nonparametric model (NPM),

$$y_t = \beta_0 + f\left(\text{trend}_t, \sum_{i=1}^{N}(\cos_{it} + \sin_{it})\right) + \epsilon_t, \tag{19}$$

which is the most flexible option.

Three standard measures have been chosen to compare the *ex post* average forecast with the observed values, the root mean squared forecast error (RMSFE)[3], the mean absolute percentage error (MAPE)[4] (Hyndman & Koehler, 2006), as well as the coverage of the 95% and 90% prediction

[2]It is interesting to notice that the NUTS-Sampler, used by the software, does not require any special behaviour for conjugate priors, which much impact the priors' choice (Hoffman & Gelman, 2014). For more details see Stan Development Team (2015) and Hoffman and Gelman (2014).

[3]The root-mean-square error (or root-mean-square deviation) is a measure of the differences between the values predicted by a model or an estimator ($\hat{\theta}$) and the values observed (θ). It is computed using the formula: $\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}$.

[4]The mean absolute percentage error (or mean absolute percentage deviation) is a measure of prediction accuracy of a forecasting method, which expresses accuracy as a percentage and is defined by the formula: $\text{MAPE}(\hat{\theta}) = E\left(\left|\frac{\theta - \hat{\theta}}{\theta}\right|\right)$.

intervals (Azose & Raftery, 2015)[5].

Table 1 shows that for sim_1, which does not include an interaction term, the LMNI does slightly better than the LMI, but that the AMI delivers more accurate results with respect to the AMNI. For sim_2 there are clearly two clusters in terms of RMSFE and MAPE. As expected the LMNI and the AMNI have a poor performance compared to the interaction models. However, the absence of a significant difference between the LMI, the AMI and the NPM, suggests that, once the interaction is introduced, the gain in using a nonparametric or a semiparametric strategy in presence of a linear interaction is trivial. None of the models is particularly accurate in capturing the uncertainty related to observations far away from the mean since the coverage intervals falls between 45% and 55% for the 95% CI and between 36% and 45% for the 90% CI. sim_3 reverses this last finding returning preciser credible intervals, which range from 76% to 92.5% at the 95% CI and from 76% to 86% for the 90% CI. The non-linear interaction highlights the benefits of a flexible functional form. Nevertheless, there is no significant difference between the NPM and the AMI. Therefore, unless the curse of dimensionality implies a sub-optimal choice of knots, it might be better to simply put all the explanatory variables into a single unspecified function.
Based on the previous simulation exercise we introduce a second order Fourier seasonality,

$$
\begin{aligned}
y_t^{sim_4} &= 12 + 0.1\text{trend}_t + 20\cos\left(\frac{\pi}{6}x_t\right) + 48\sin\left(\frac{\pi}{6}x_t\right) + 24\cos\left(\frac{\pi}{3}x_t\right) + 8\sin\left(\frac{\pi}{3}x_t\right) + \epsilon_t,\\
y_t^{sim_5} &= 12 + 0.1\text{trend}_t + 20\cos\left(\frac{\pi}{6}x_t\right) + 48\sin\left(\frac{\pi}{6}x_t\right) + 24\cos\left(\frac{\pi}{3}x_t\right) + 8\sin\left(\frac{\pi}{3}x_t\right) +\\
&\quad + 0.8\text{trend}_t\left(\cos\left(\frac{\pi}{6}x_t\right) + \sin\left(\frac{\pi}{6}x_t\right) + \cos\left(\frac{\pi}{3}x_t\right) + \sin\left(\frac{\pi}{3}x_t\right)\right) + \epsilon_t,\\
y_t^{sim_6} &= 12 + 0.1\text{trend}_t + 20\cos\left(\frac{\pi}{6}x_t\right) + 48\sin\left(\frac{\pi}{6}x_t\right) + 24\cos\left(\frac{\pi}{3}x_t\right) + 8\sin\left(\frac{\pi}{3}x_t\right) +\\
&\quad + 0.8\text{trend}_t \exp\left(\left(\cos\left(\frac{\pi}{6}x_t\right) + \sin\left(\frac{\pi}{6}x_t\right) + \cos\left(\frac{\pi}{3}x_t\right) + \sin\left(\frac{\pi}{3}x_t\right)\right)\right) + \epsilon_t.
\end{aligned}
$$

As expected in sim_4, the models without interaction outperform the ones with interaction. The AMNI improves the LMNI and it substantially increases the coverage capacity shifting it from 80% to 94% for the 95% CI and from 74% to 86% for the 90% CI. When in sim_5 the interaction is introduced the behaviour of the different models is comparable to the one presented in Table 1. In other words, the models with interaction outperform the ones without interaction in terms of RMSFE, MAPE and coverage intervals, with the NPM being the most accurate among them. To the contrary, sim_6 reports more remarkable differences between the models. The AMI does not only outperform all the parametric alternatives and the AMNI, but also the NPM with a final improvement of the RMSFE of the 14%, see Table 2. This last outcome shows that equation (7) gives a non negligible advantage, compared to the other specifications, if different parts of the seasonality interact with the trend, especially in the coverage capacity of the credible intervals. A detailed visualization of the results is provided in section 2 of the Additional Material.

[5]The interval coverage is a measure of prediction accuracy, which computes the actual coverage percentage of the prediction intervals on hold-out samples. Therefore, the larger the coverage, the better the model. For example, a value of 50 at a 95% level means that 50% of the observations fall into the 95% credible intervals.

Table 1: 1st order Fourier Simulations: Root Mean Square Forecast Error (RMSFE) and Mean Average Percentage Error (MAPE) and prediction interval coverage for the Bayesian Linear Model with no Interaction (LMNI), the Additive Model with no Interaction (AMNI), the Linear Model with Interaction (LMI), the Additive Model with Interactions (AMI) and the Nonparametric Model (NPM).

	sim_1				sim_2				sim_3			
	RMSFE	MAPE	95%	90%	RMSFE	MAPE	95%	90%	RMSFE	MAPE	95%	90%
LMNI	76	0.06	64	60	6441	3.12	45	36	127144	0.94	92.5	86
AMNI	89	0.06	52.5	50	6461	2.82	49	40	105791	0.90	90	79
LMI	77	0.06	65	60	3139	0.72	55	36	121757	0.88	82.5	76
AMI	61	0.05	66	55	3363	0.77	50	44	91830	0.72	76	76
NPM	93	0.07	49	45	3354	0.75	50	45	92478	0.76	76	76

Table 2: 2nd order Fourier Simulations: Root Mean Square Forecast Error (RMSFE) and Mean Average Percentage Error (MAPE) and prediction interval coverage for the Bayesian Linear Model with no Interaction (LMNI), the Additive Model with no Interaction (AMNI), the Linear Model with Interaction (LMI), the Additive Model with Interactions (AMI) and the Nonparametric Model (NPM).

	sim_4				sim_5				sim_6			
	RMSFE	MAPE	95%	90%	RMSFE	MAPE	95%	90%	RMSFE	MAPE	95%	90%
LMNI	30	0.10	80	74	1462	1.17	59	49	498	0.23	60	51
AMNI	28	0.09	94	86	1426	1.28	57.5	51	412	0.19	92.5	81
LMI	36	0.13	80	75	847	1.46	55	51	345	0.14	69	59
AMI	30	0.19	87.5	76	815	1.39	60	40	212	0.12	90	87
NPM	36	0.13	85	75	250	0.12	60	49	246	0.17	25	21

3 Time Series Analysis of Migration Data

3.1 Swiss Immigration Data 1981-2013

We take advantage of the monthly released Swiss migration data to elaborate an empirical implementation of the model. In Switzerland the migratory inflows data stem from the Swiss Federal Statistical Office (SFSO). The sources used for the immigration of the non-Swiss population come from PETRA (Statistics for the Resident Population of Foreign Nationality), for the period 1981-2009, and from STATPOP (Population and Households Statistics) for the subsequent years.
The combination of different counts creates a minor discrepancy. PETRA reports the month and the year of the residence permit emission, while STATPOP records the date of movement of the migrants. Since people arriving in Switzerland can obtain their *Ausländerausweis* (Residential Permit) few months after their arrival, profiting of legal arrangements like the tourist allowances, trivial incongruities could emerge. The present study accounts for the year 2010 people who arrived in 2010, but who obtained their permit in 2011. Due to this choice, inconsistencies can be detected between the gross immigration data used in this paper and the ones reported by the SFSO. The magnitude of the disparities amounts to about 9.000-10.000 migrants per year for the period 2011-2013. This procedure might be imperfect, since from 1981 to 2010 we consider the date in which the permit was obtained while from 2011 to 2013 the arrival data. However, we think that this solution is better than the withdrawing of all the observations for which the arrival month was not available.
The resulting data format is a time series, which has two peculiarities compared to datasets normally used to produce immigration forecasts. On one side, it uses monthly, rather than annual, data. This choice is grounded in the idea to capture, otherwise unobservable, short-term fluctuations (Disney et al., 2015), while avoiding the inconsistencies which characterize the long term forecasts based on annual observations (Hajnal, 1955; Taeuber et al., 1969). On the other side, the forecasts are performed over the aggregate time series rather than over its single age components. This decision has several upsides. The models set-up costs are reduced to a minimum, any "sum-back" process is avoided and all the typical problems related to the forecasting of aggregate values starting from disaggregated ones are avoided (Bermingham & DAgostino, 2014; Hendry & Hubrich, 2006). Furthermore, given the invariance of the age frequencies in our sample, see Figure 1, in order to obtain the age profile of the migrants, it is sufficient to multiply the aggregate outcome by the age frequencies. A disaggregated age prediction is performed in section 4 in order to show the versatility of our approach and its applicability to longitudinal datasets.

3.2 Model Validation

The previous sections point out how the construction of a predictive model is performed at two stages: the analysis of the observed data and the incorporation of new available information. Hereafter we focus on both stages, first by undertaking a descriptive analysis of the immigration time series and then by elaborating suitable priors. The starting point is to decompose the logarithm of the number of monthly arrivals into a time and a seasonal trend, as in equation (1), see Figure 2.

Given that the best trade trade-off between parsimony and accuracy is achieved with a 2nd

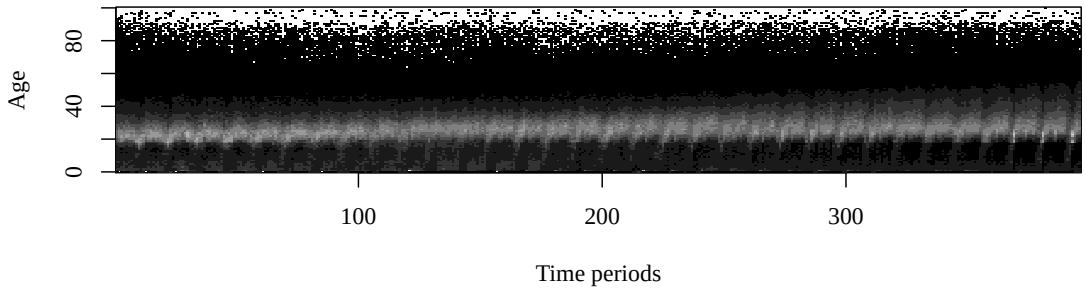

Figure 1: Image plot showing migration frequencies, where dark gray indicates low frequency and light gray high frequency, by age from 01.01.1981 to 01.12.2013 (T=396).

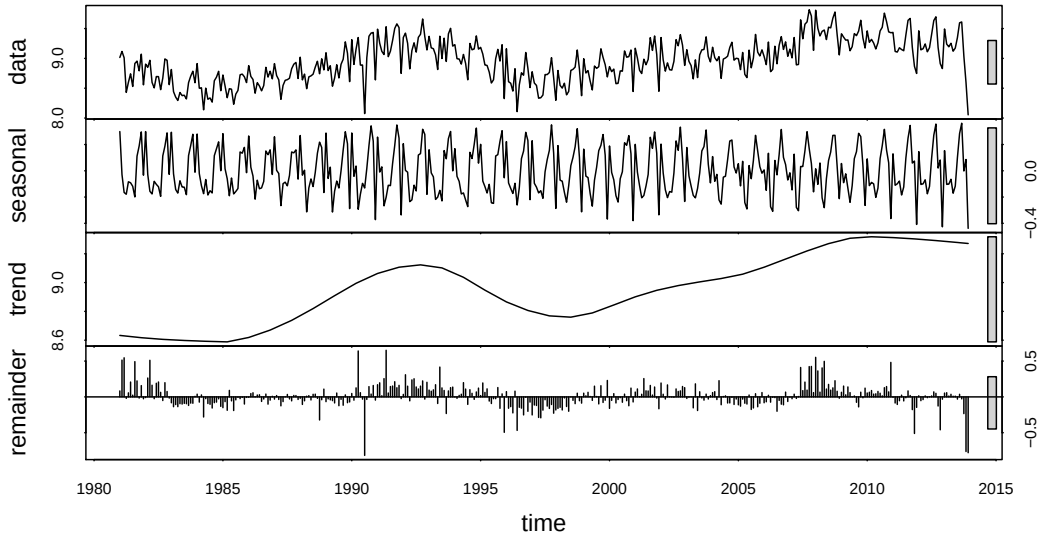

Figure 2: Swiss Immigrant Flows Decomposition obtained using a locally weighted scatter-plot smoother (LOESS curve). From the top to the bottom, the plots show: observed data, global trend, seasonal trend and random noise.

order Fourier seasonality the reference model with interaction becomes,

$$y_t = \beta_0 + \beta_1 \text{trend}_t + \sum_{i=1}^{2} \beta_{2i}(\cos_{it} + \sin_{it}) + \sum_{i=1}^{2} \beta_{3i} \text{trend}_t * (\cos_{it} + \sin_{it}) + \epsilon_t. \tag{20}$$

For the coefficients of equation (20) we use as prior a multivariate normal distribution,

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{21} \\ \beta_{22} \\ \beta_{31} \\ \beta_{32} \end{bmatrix} \sim N \left[\begin{pmatrix} 9 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 \end{pmatrix} \right], \tag{21}$$

where the mean of the intercept has been set to 9, which is the average value of the logarithm of the immigration for the observed periods, while the standard deviation is 0.5 since we are confident that the mean can be neither lower than 7.4 nor higher than 10.6. The mean of the other coefficients is centered on 0 with a standard deviation of 0.5. This choice imposes the means of the βs of global trend, sine, cosine and trend-(sin+cos) interaction to be between -2 and 2. Note that the absence of covariance between the different coefficients allows to vectorize the prior and speed up the convergence rate of the algorithm. The same prior is used for the only coefficient of the equivalent additive model,

$$y_t = \beta_0 + f_1(\text{trend}_t) + \sum_{i=1}^{2} f_{2i}(\cos_{it} + \sin_{it}) + \sum_{i=1}^{2} f_{3i}(\text{trend}_t, \cos_{it} + \sin_{it}) + \epsilon_t, \tag{22}$$

such that $\beta_0 \sim N(9, 0.5)$. The coefficients which identify the splines are assumed to be normally distributed with variances equal to 1, while the priors of the standard deviation of the error and of λ are set to half Cauchy, see Håvard (2016),

$$\sigma_\epsilon \sim HC(0, 2) \qquad \sigma_\lambda \sim HC(0, 2). \tag{23}$$

Therefore, the coefficients' priors should be considered as empirical informative, while the variances' ones weakly informative, see Figure 3.

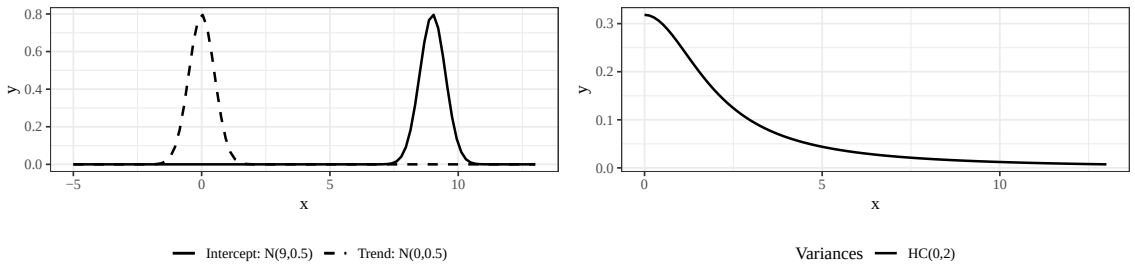

Figure 3: Representation of the priors for the coefficients and the variances.

Given the aim of our analysis, we need to check which of the specifications presented in section 2 produces the best forecasts. The first step in this direction requires to split the data into two subsets. The *training data* (from January 1981 till December 2003) are used to fit the model, and the *test data* (from January 2004 till December 2013), are used to predict the immigration flows. Once this first comparison has been done, it is possible to check the forecast's accuracy. Table 3

Table 3: 2nd order Fourier models on Swiss Immigration Aggregated Data for Short and Long Run Predictions.

	Short Run Predictions				Long Run Predictions			
	RMSFE	MAPE	95%	90%	RMSFE	MAPE	95%	90%
LMNI	0.324	0.029	94	88	0.293	0.025	98	94
AMNI	0.281	0.023	98	94	0.281	0.024	98	94
LMI	0.324	0.029	94	88	0.287	0.025	99	96
AMI	0.260	0.022	89	94	0.270	0.023	99	95.5
NPM	0.586	0.060	50	39	1.572	0.183	21	17

reports three standard measures to compare the *ex post* average forecast with the observed values.

According to the RMSFE and the MAPE the AMI emerges as the most accurate model followed by the AMNI, the two linear models, LMNI and LMI, and far behind the NPM. Also in terms of Credible Intervals (CI) the AMI and the AMNI achieve the best coverage with a 98% for the 95% CI and a 94% for the 90% CI. The poor performance of the NPM seems driven by the usual overfitting of non-additive models. Such specification returns a decreasing trend, while the observed one is increasing. The results' comparison between AMNI and NPM suggests that, in the Swiss data, the necessity of modelling the trend is stronger than the one of the interaction. Nevertheless, a certain degree of non-linear trend-seasonal interplay is still present making the AMI the best predictive model. Figure 4 visually portrays the gains quantitatively evaluated in Table 3. More detailed results are given in the Additional Material

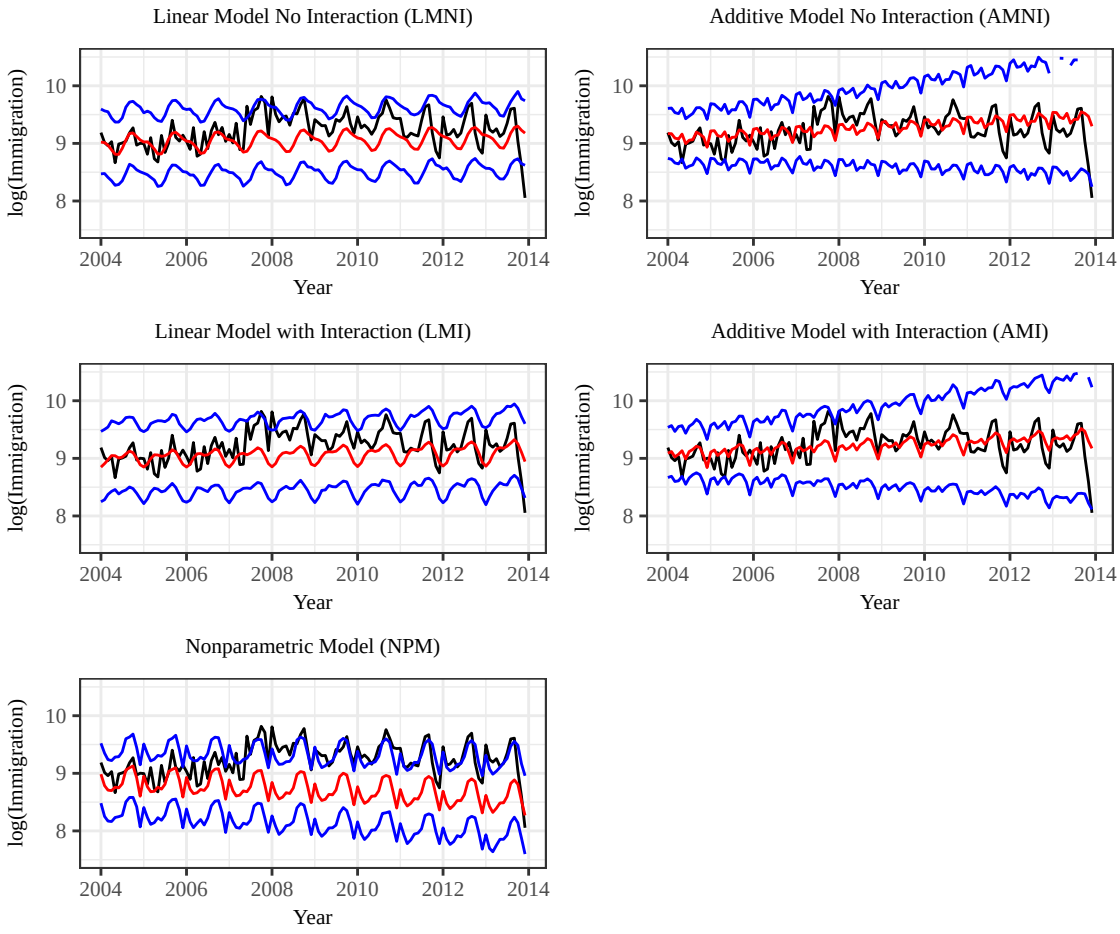

Figure 4: Aggregated forecasts 2004-2013 with 95% prediction credible interval from the posterior predictive distribution.

Nonetheless, the short time horizon chosen to implement the forecasts might influence the performance results' in favor of the semiparametric models. In fact, the smooth trend term is computed using a piece-wise regression where the *optimal* number of knots is obtained from the training, not the testing data. Therefore, in presence of a highly non-linear trend, the number of knots might be quite high. While this could be of no particular problem in the short run, when the probability of maintaining a trend close to the one of the last periods is realistic, in the long run this might be more problematic. Hence, a safer choice is either to use a low number of knots or to substitute $f_1(\cdot)$ with a parametric component in order to stabilize the semiparametric tendency to overfit out-of-sample forecasts, such that the AMNI becomes,

$$y_t = \beta_0 + \beta_1 \text{trend}_t + \sum_{i=1}^{2} f_{2i}(\cos_{it} + \sin_{it}) + \epsilon_t, \tag{24}$$

while the AMI is,

$$y_t = \beta_0 + \beta_1 \text{trend}_t + \sum_{i=1}^{2} f_{2i}(\cos_{it} + \sin_{it}) + \sum_{i=1}^{2} f_{3i}(\text{trend}_t, \cos_{it} + \sin_{it}) + \epsilon_t, \tag{25}$$

and the NPM

$$y_t = \beta_0 + \beta_1 \text{trend}_t + f\left(\text{trend}_t, \sum_{i=1}^{2}(\cos_{it} + \sin_{it})\right) + \epsilon_t. \tag{26}$$

The results show that the long term predictions highly benefit from a linear trend reducing the RMSFE by 600% and the MAPE by 900%. Note that the addition of the parametric term β_1 is strictly related to the proposed estimation technique, i.e. the thin plate regression splines. If, for example, $f_1(.)$ would have been estimated using a local polynomial approximation the required stabilization would have been embodied in the estimation methodology. The results for the forecast period 1998-2013 mainly reflect the ones of the shorter term, with a slight improvement of the LMI over the LMNI. In general, both time horizons show a premium for the AMI specification. It is also interesting to notice how the NPM still overfits the data regardless the additional linear trend term.

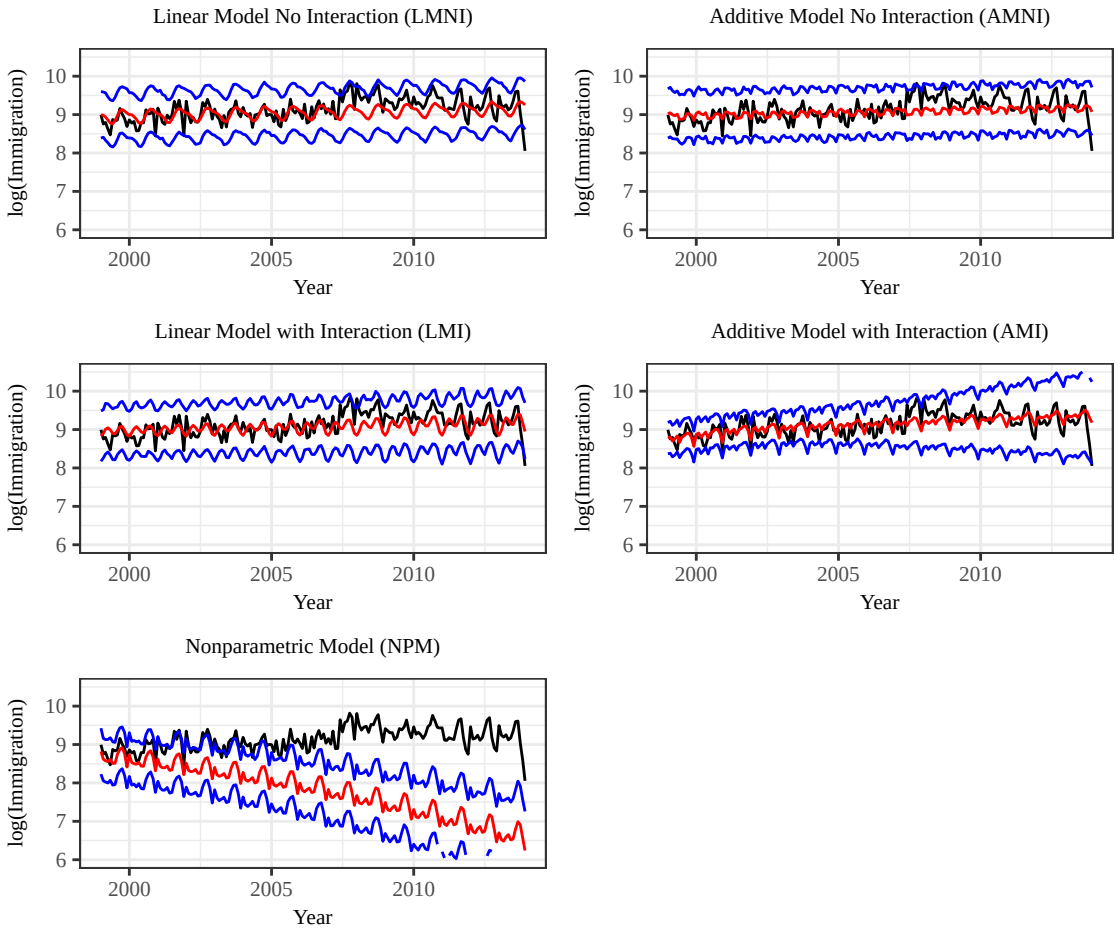

Figure 5: Forecasts 1998-2013 with 95% prediction credible interval from the posterior predictive distribution.

A closer look to the reminders plotted at the bottom of Figure 2 may suggest the presence of autocorrelation among the residuals. Potentially, such autocorrelation might persist also in the residuals of the LMNI, AMNI, LMI, AMI and NPM models. Hence, we try to re-estimate all the models, both for the short and the long run, by allowing their residuals to be autocorrelated of order 1 (AR(1)):

$$\epsilon_t = \alpha\epsilon_{t-1} + \xi_t, \quad -1 < \alpha < 1, \quad \xi_t \sim N(0, \sigma_\xi^2). \tag{27}$$

Table 4 confirms an improvement in prediction accuracy with a reduction of the RMSFE of 11%, of the MAPE of 33% and an extra coverage of the 95% and 90% credible intervals of respectively of 1.3% and 15%. The amelioration can also be traced in Figures 6 and 7. In general models' performances are in line with the one of Table 3. However, the comparative advantage of the AMI with respect to the LMI, while being improved in the long run, is reduced for the short run. A possible explanation can be found in the origin of the autocorrelation among errors. For example, if there are logarithmic or exponential terms in the data generatin process, the LMs would most probably generate autocorrelated errors. Therefore, equation 20 benefits from the introduction

Table 4: 2nd order Fourier models on Swiss Immigration Aggregated Data for Short and Long Run Predictions with AR(1) errors.

	Short Run Predictions				Long Run Predictions			
	RMSFE	MAPE	95%	90%	RMSFE	MAPE	95%	90%
LMNI	0.255	0.022	96	94	0.251	0.022	95	94
AMNI	0.264	0.021	97.5	94	0.247	0.020	97	93
LMI	0.248	0.021	96	93	0.252	0.021	97	94
AMI	0.247	0.021	98	95	0.232	0.020	100	99
NPM	0.271	0.026	94	88	0.550	0.053	70	53

of (27). To the contrary, the AMs, by construction, tend to solve autocorrelation generated by functional form misspecifications, gaining less from the inclusion of an AR(1) component. Figure 8 shows how the LMI reports stronger evidence of residuals dependence than the AMI[6].

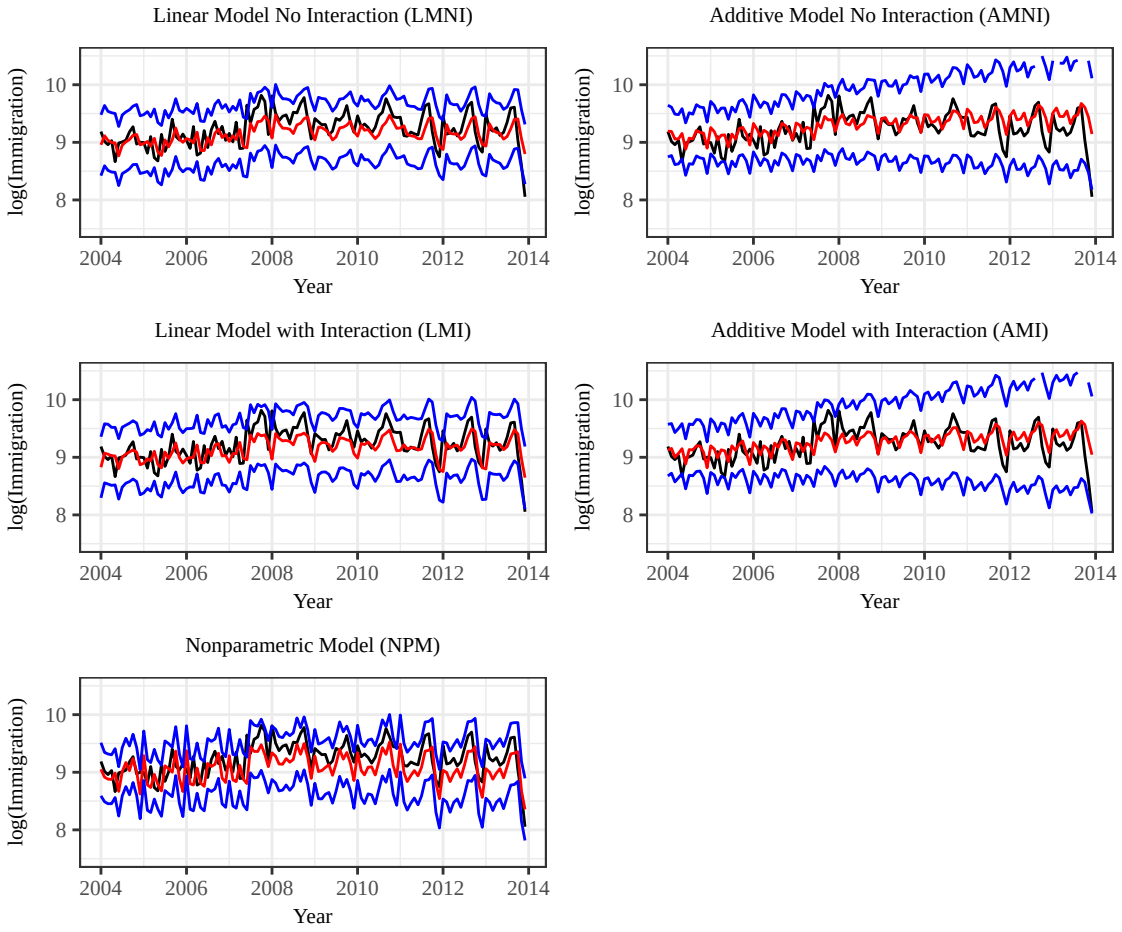

Figure 6: Aggregated forecasts 2004-2013 with 95% prediction credible interval from the posterior predictive distribution for models with autocorrelated residuals of order one (AR(1)).

[6]A Box-Pierce test statistic for examining the null hypothesis of independence of the residuals is rejected with a p-value $<$ 2.2e-16 for the LMI and a p-value $=$ 0.0022 for the AMI.

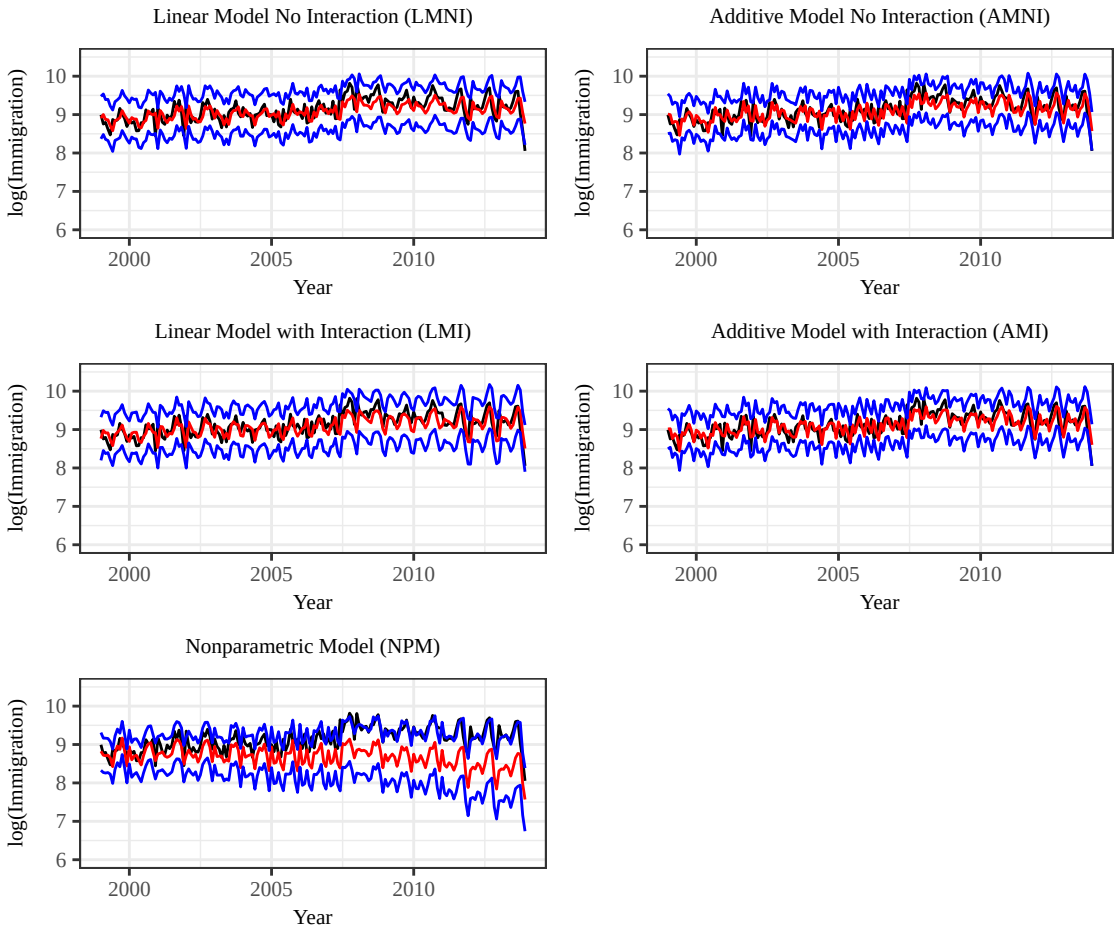

Figure 7: Forecasts 1998-2013 with 95% prediction credible interval from the posterior predictive distribution for models with autocorrelated residuals of order one (AR(1)).

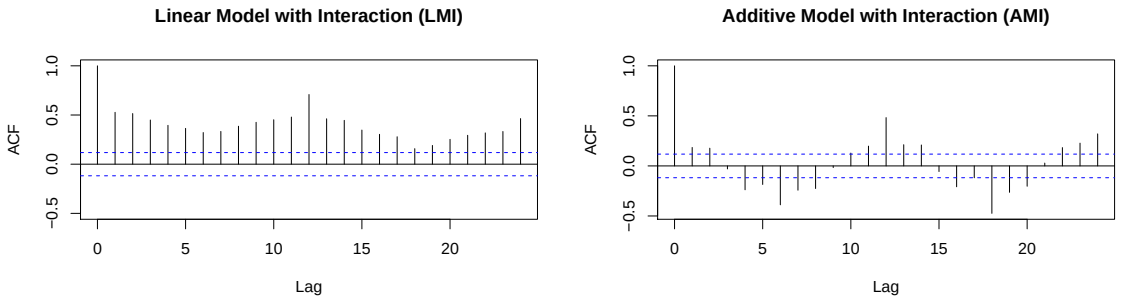

Figure 8: Estimates of the autocovariance or autocorrelation function for the linear model with interaction (left plot) and for the additive model with interaction (right plot).

While the previous sections have shown how the semiparametric models tend to outperform the other alternatives both with simulated and historical data, we now illustrate how robust these results are with respect to the setting of different priors.

The first alternative is an uninformative prior defined over $\mathbb{R}$ for β combined again with two stan-

dardized half t-student with three degrees of freedom for σ_ϵ and σ_λ (Bürkner, 2017). A second possibility is to use the horseshoe hierarchical shrinkage prior with parameter 1 for β combined again with two standardized half t-student for the variances of ϵ and λ. The outcomes confirm that the AMI is robust to these alternative specifications, see Figure 9.

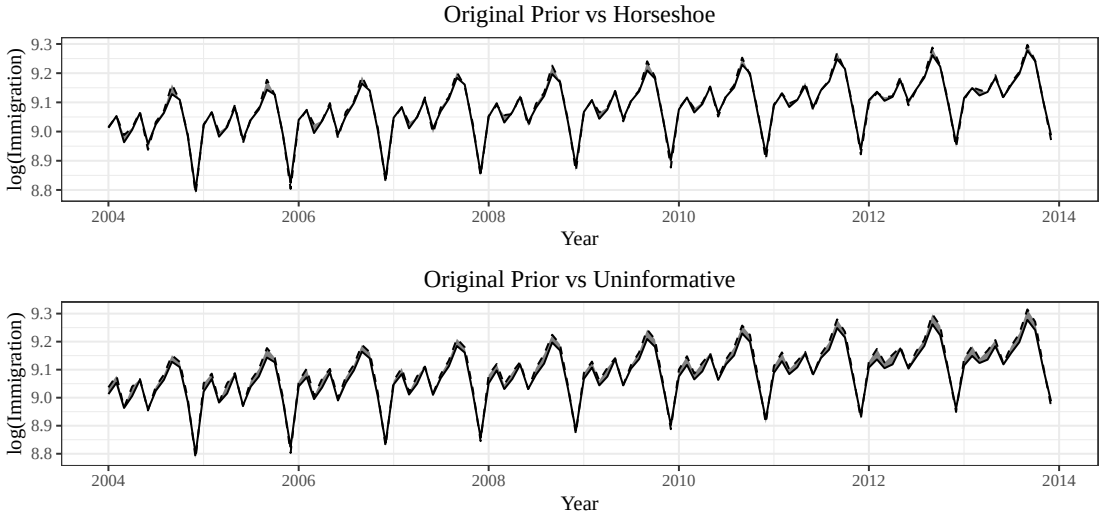

Figure 9: Priors' robustness comparison for the predictions over the period 2004-2014. The black line denotes the forecasts obtained with the original prior distributions used in the analysis. The red lines are the forecasts obtained with comparative priors, the horseshoe in the top and the uninformative in the bottom plot. The difference is the blue area.

So far we have shown how the Bayesian Additive Model with Interaction outperforms the alternatives in predicting the behaviour of the test data starting from the fit of the training data. However, our argument in favour of Bayesian statistics was rooted in its capacity to introduce informative believes through the choice of the prior distributions. Therefore, we propose, in the next Subsection, an illustrative example which directly shows the advantages of adopting different priors in the predictions.

3.3 Forecast Exercise

In this subsection we fit the AMI on the Swiss data for all the available years (1981-2013) and we try to forecast immigration flows until 2023. Since the priors are used, at the same time, to estimate the model and, indirectly, to implement the forecasts, their distributions link historical knowledge with future expectations. Therefore, if migration is foreseen to look like the past, flat priors with relatively high variances should be chosen. To the contrary, if migration is anticipated to change, informative priors, with smaller variances around the expected means should be selected. The latter case, however, may give room to divergent transitions within the sampling.
For this study we set up three scenarios. The first pictures the sentiment of a future migration in line with its historical average. Therefore, it relies on weakly informative priors. For example, the intercept's prior takes the same values as in the model validation exercise with a mean of 9 and a standard deviation of 0.5. In the same way, the trend's prior is centered around the posterior obtained from section 3.2, but with a wider variance to convey a minimal impact on its posterior

distribution. The second scenario is a middle story line, which reflects the possibility of a small shock. In this case, the trend's prior is centered around 1.5, rather than 0, with a variance of 0.2, which suggests an increased return of the trend. The third is a scene, which mirrors the expectation of a more evident structural break on the trend's historical impact on y_t. This last case is achieved by increasing the expected mean up to 2, while shrinking its variance to 0.1.

In order to make sure that the difference between the three scenarios is only about the researcher's expectations about structural changes we do not modify the priors for the standard deviation of the error term (σ_ξ) and of the smoothing parameter (σ_λ) assuming that in every case they are distributed as Half-Cauchy with a scale parameter of 2, like in section 3.2. Instead, we play mostly with the priors of the error's autocorrelation term (α) assuming increasing path dependency coherently following the discussion on the trend's priors. All the distributions are described in Table 5 and portrayed in Figure 10.

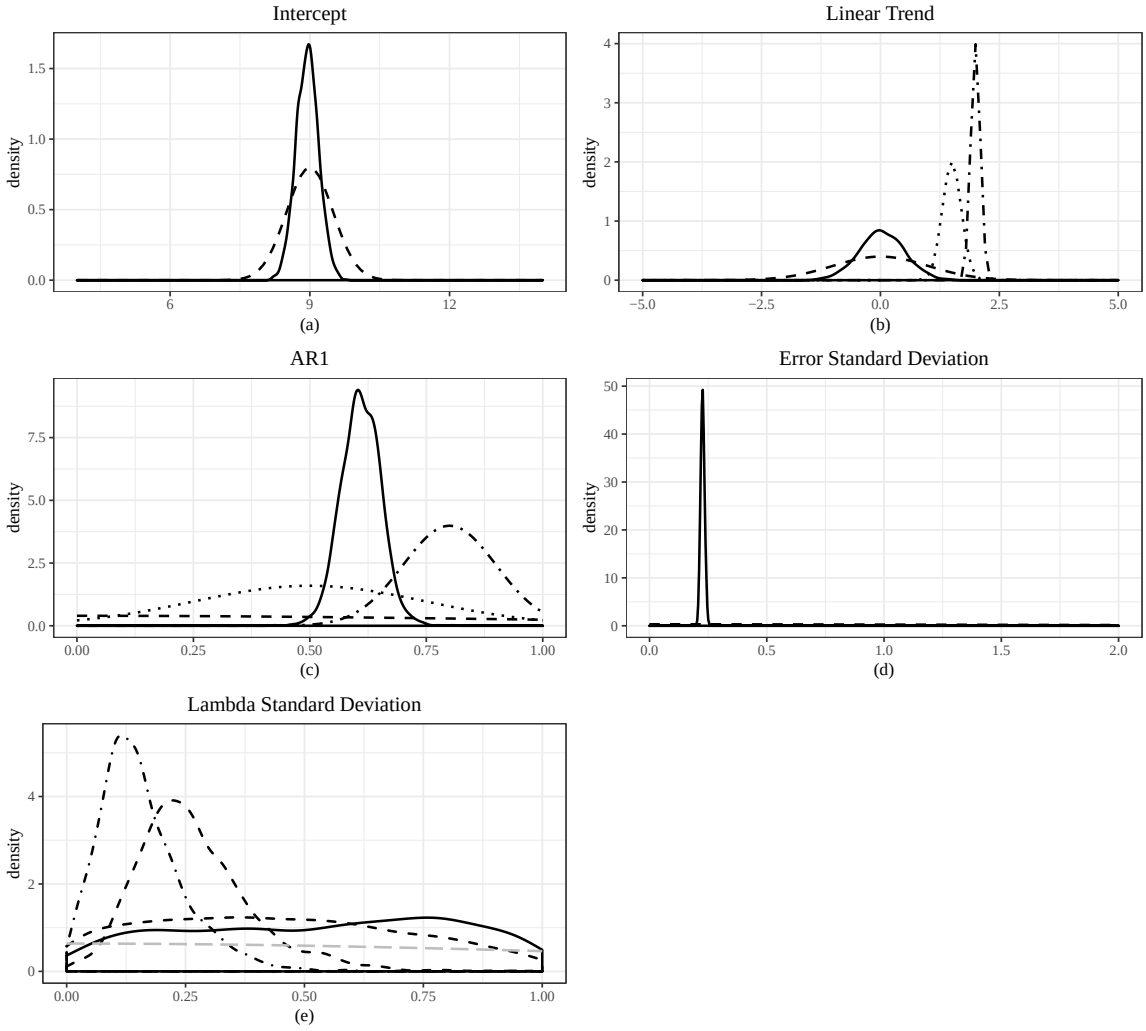

Figure 10: The plots compare the posterior distributions obtained from the AMI with autoregressive errors of order one for the period 1981-2013 (solid line) with the priors chosen to set the different forecast scenarios (dashed and dotted lines). From top to bottom and left to right the graphs show the distributions for the intercept, the linear trend's coefficient, the error autoregressive coefficient (AR1), the error's standard deviation and the standard deviation of the smoothing parameter.

Table 5: Prior Distributions for the Scenario Analysis

	Historical Scenario	Middle Scenario	High Scenario
β_0	N(9;0.5)	N(9;0.5)	N(9;0.5)
β_1	N(0;1)	N(1.5;0.2)	N(2;0.1)
α	N(0;1)	N(0.5;0.25)	N(0.8;0.1)
σ_ξ	HC(0;2)	HC(0;2)	HC(0;2)
σ_λ	HC(0;2)	HC(0;2)	HC(0;2)

For each scenario we model the global trend in three different ways. The choice stems from

the reasoning already presented in the previous sections, where we outlined the potential problems raising from a smooth trend with a considerable number of knots for long term forecasts. Thus, the three models include as alternatives a linear trend ($\beta_1 \text{trend}_t$) and two smooth trends with respectively four, $f_1(\text{trend}_t, \text{k} = 4)$, and six knots, $f_1(\text{trend}_t, \text{k} = 6)$. The difference in the results is more visible for the trend fitted on six knots, especially in the case of historical and middle scenarios, where the global trend is down-warding. In general when the trend is nonparametrically computed with k=6, the average future immigration is approximately 40% lower than in the linear case for the historical and the middle scenario.

In the historical scenario, the median number of immigrants rises from its historic value of 99,070 people per year to 147,646 for the linear trend, to 206,151 for the smooth trend with k=4 and shrinks to 82,701 for the smooth trend with k=6, showing respectively an increase of 49% and 68% and a decrease of 17%. In the middle scenario the median number of immigrants becomes 148,313 for the linear trend, 185,752 for the smoothing trend with k=4 and 109,584 for k=6 corresponding respectively to a 50%, 87% and 11% increase. Finally, the high scenario finds a growth rate of 53% for the linear trend, of 168% for the smoothing trend with k=4 and of 136% for k=6. Figure 11 visualizes the results. Note that all the growth rates would be down-sized if we would look at the median immigration of the last 10 and 5 years rather than the one from 1981 till 2013.

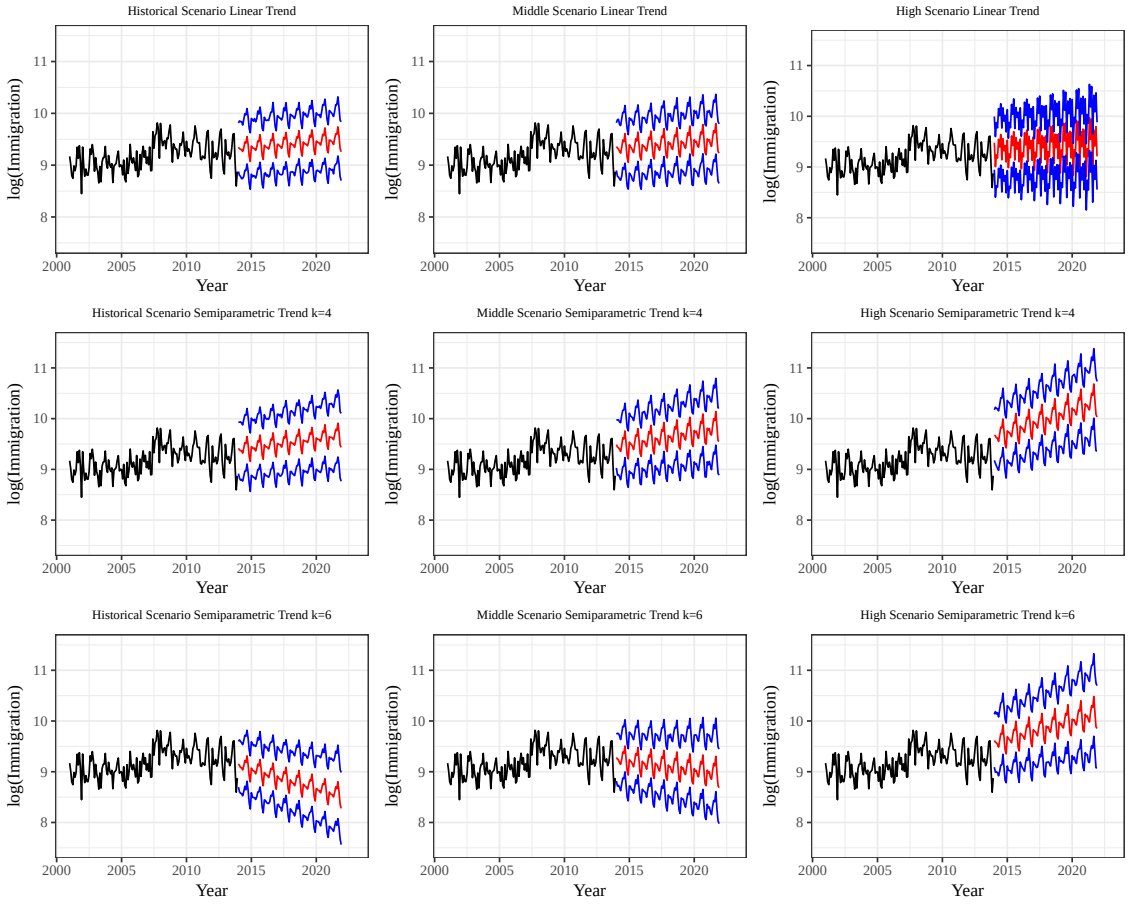

Figure 11: Forecasts 2013-2021 with 95% prediction credible interval from the posterior predictive distribution for three different priors.

As a general remark from our forecast exercise, practitioners should be careful in treating the global trend either parametrically or nonparametrically, as well as, in the degree of volatility conveyed by the priors. In general a large number of knots requires more coefficients (one for each knots interval) and, consequently, a loss in degrees of freedom. Said differently, including an excessive number of knots reduces the degrees of freedom available to estimate the parameters' variability, potentially having a negative impact on the forecast's quality. Furthermore, since the paper stresses the benefits of modelling semiparametrically the trend-seasonal interaction, estimating the trend in parsimonious ways allows to devote a larger number of degrees of freedom to fit the interaction without worsening the prediction accuracy.

4 Longitudinal Analysis of Age Categorized Data

4.1 Model Validation

To illustrate the adaptability of our model to a disaggregated problem we employ longitudinal data categorizing the monthly number of arrivals by age. Further distinctions by gender or nationality can, at any rate, be implemented, but they are not considered here.

Table 6: 2nd order Fourier models on Swiss Immigration Disaggregated Data: Root Mean Square Forecast Error (RMSFE) and Mean Average Percentage Error (MAPE) and prediction interval coverage

	RMSFE	MAPE	95%	90%
LMNI	1.165	0.314	95	81
AMNI	0.662	0.245	89	79.5
LMI	1.165	0.314	95	81
AMI	0.662	0.246	90	79
NPM	0.547	0.330	91	83

Splitting our data into different ages allows us to check the persistence of non-linearity and trend-cycle interactions. We add Age as an explanatory variable to the models estimated for the time series. The latter is introduced as a parametric term, $\beta_4 Age_{age,t}$, in LMNI and LMI, as a smooth function $f_4(Age_{age,t})$ in AMNI and AMI (Dodd et al., 2018), while for the the nonparametric model (NPM), we add it to the main smoothing function f. To estimate the Bayesian models we adopt the same priors as in the aggregate exercise, with in addition $\beta_4 \sim N(0, 0.5)$.

The results in Table 6 show an increase in uncertainty produced by the disaggregation. On average the RMSFE increases by 130% and the MAPE by 770%. The improvements given by the use of semiparametric and nonparametric models are reinforced, dropping the RMSFE from 1.16 (LMNI, LMI) to 0.66 (AMNI, AMI) and to 0.54 (NPM). However, the relative forecast accuracy measure, the mean absolute percentage error, portrays a different picture. The AMNI and the AMI achieve the most precise predictions with a MAPE of 0.24, followed by the LMNI and the LMI with a MAPE of 0.31 and lastly by the NPM with a MAPE of 0.33. The difference reflects the specific peculiarities of the absolute (RMSFE) vs the relative (MAPE) accuracy measures. In our case the NPM is doing better than the alternatives in minimizing the big error generated by the predictions of outlier observations, i.e. low RMSFE, but it produces a poorer performance, on average, when predicting smaller values, i.e. high MAPE. On the other hand, AMNI and AMI minimize the errors between, let say 0.11 and 0.12 rater than the ones between 0.81 and 0.82. In light of such considerations, the semiparametric models seem a more cautious choice over the nonparametric one.

In general, the disaggregation, while it does not generate any significant linearization, confirming the better performance of the semiparametric models, it minimizes the importance of the trend-seasonal interaction, which is reflected in the absence of a significant difference between the models with and without it.

4.2 Forecast Exercise

The previous section validates the additive models as the preferable options to predict future immigration flows by age. This section presents a forecast exercise on the Swiss data for the period 2014-2021, which replicates the one in section 3.3, for the longitudinal case. Even though the difference between the AMNI and the AMI is not particularly relevant, we use the AMI, not only to keep a certain degree of consistency with section 3.3, but also because the AMI has a faster rate of convergence. Using the same priors as in the time series analysis, see Table 5, we check the

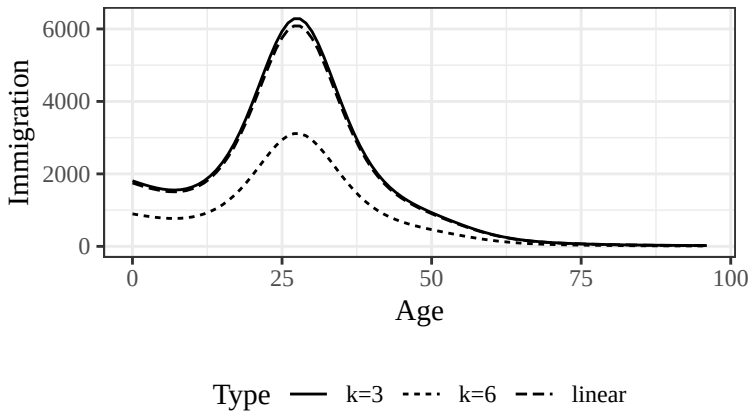

Figure 12: Effect of a different trend specification on the predictions of age-specific immigration forecast with on the y-axis the average yearly number of immigrants for the predicted period (2014-2021) and on the x-axis the age. The blue line denotes the results obtained with a linear trend, the red one with a smooth trend with 3 knots (k=4) and the green one with a smooth trend with 6 knots (k=6).

performances of different trends: linear, smooth with three (k=4) and six (k=6) knots. Figure 12 reports the results under the low volatility scenario averaged over all the forecast periods. While the linear and the smooth trend with three knots have a very similar behaviour, the smoothest trend suffers from the same down-ward trend as in the aggregate case portrayed in Figure 4. Due to the non significant difference between the blue (linear) and the red (k=4) line, we use the former to produce the disaggregated forecast scenario analysis.

The results are depicted in Figure 13. In all the three cases, the expected amount of immigration by age resembles a normal distribution centered around 33, with a third moment bigger than zero. All the scenarios roughly maintain the same immigrant population age structure as the historical data. Nevertheless, as volatility increases, the degree of smoothness decreases. The same is true as the forecast horizon augments.

A final remark that needs to be made when considering disaggregated forecasts is how to manage the "pooling-back" when the final interest is to know the future of Swiss migration as a whole. In fact, stacking the average forecasts by age implies also stacking the credibility intervals which might be difficult to handle and risk to degenerate in an uninformative explosion of uncertainty. In light of the results, while disaggregation is always a possibility, the age pooling seems a safer choice, which guarantees forecast accuracy, as well as stability.

5 Conclusion

Migration gained the reputation of being an unpredictable component of population change (Pijpers, 2008; Bijak & Wiśniowski, 2010). The current paper tries to show how merging Bayesian statistics with semiparametric methods can help to handle the uncertainty surrounding the number of future incomers.

The core of the research lies in the choice to consider migration as a seasonal, rather than an annual, phenomenon and to exploit the monthly frequency of the output to deal with eventual

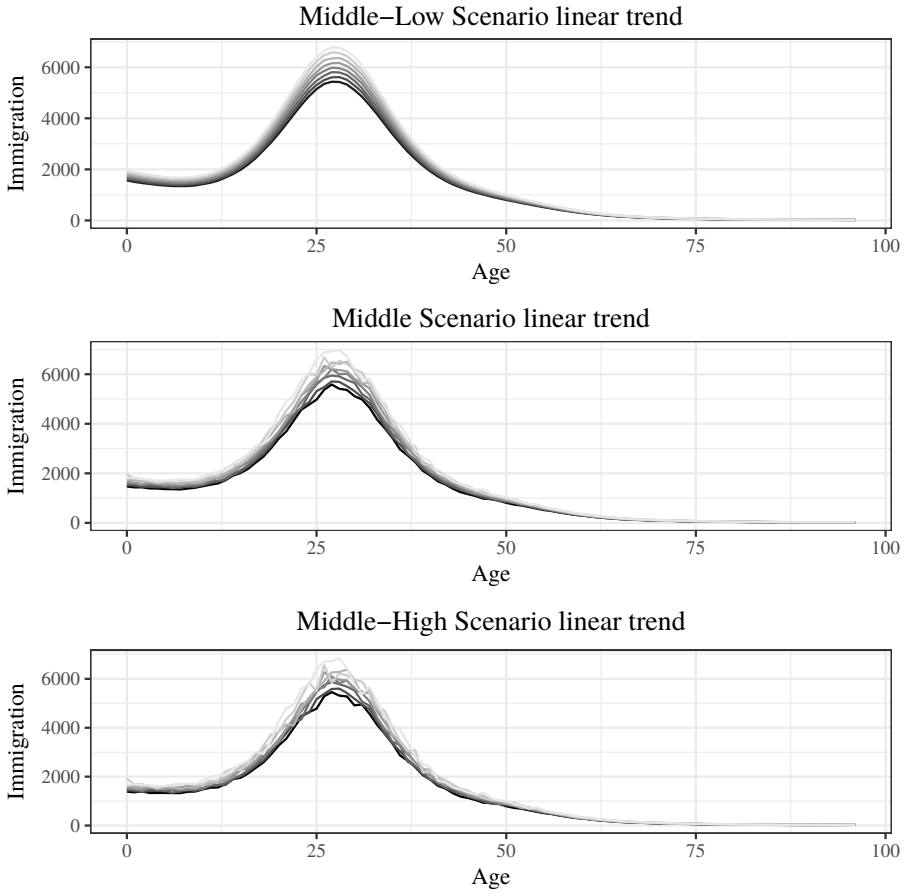

Figure 13: Disaggregated monthly forecasts averaged by year over the period 2014 (orange) - 2021 (pink) for three different scenarios from the posterior predictive distribution.

trend-seasonal interactions. Even if such focus limits the methodology's application to countries which dispose of high frequency data, it also potentially opens new perspectives for analyzing new migration trends, which show high seasonality, like the ones of the recent refugee crisis (European Union Publications Office, 2017).

The message our results deliver is mainly twofold. On one side, the semiparametric models can represent an appealing alternative in presence of non-linear trend-cycle interactions. On the other, a Bayesian prospective can be proactively embraced through the choice of informative prior distributions to build forecast scenarios accounting for unprecedented events. The latter adds the possibility to condition the forecast on a set of macroeconomic projections (Bijak, 2010).

Despite the model's choice belongs to the researcher discretionality, we have a few recommendations for future users in light of our investigation. Semiparametric models can be preferable in case of foreseen growing volatility since their flexibility can be fully exploited. However, they tend to exhibit increasing instability when dealing with long forecast horizons due to the usual overfitting of nonparametric models in out-of-sample performances. In such cases, a Bayesian perspective can help by setting boundaries on the prior distribution of the structural coefficients.

References

Alvarez-Plata, P., Brücker, H., & Siliverstovs, B. (2003). *Potential migration from central and eastern europe into the eu-15: An update*. European Commission, Directorate-General for Employment and Social Affairs, Unit A. 1.

Azose, J. J., & Raftery, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography*, *52*(5), 1627–1650.

Azose, J. J., Ševčíková, H., & Raftery, A. E. (2016). Probabilistic population projections with migration uncertainty. *Proceedings of the National Academy of Sciences*, *113*(23), 6460–6465.

Bermingham, C., & DAgostino, A. (2014). Understanding and forecasting aggregate and disaggregate price dynamics. *Empirical Economics*, *46*(2), 765–788.

Bijak, J. (2010). *Forecasting international migration in europe: A bayesian view* (Vol. 24). Springer Science & Business Media.

Bijak, J., & Wiśniowski, A. (2010). Bayesian forecasting of immigration to selected european countries by using expert knowledge. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(4), 775–796.

Billari, F., Graziani, R., & Melilli, E. (2014). Stochastic population forecasting based on combinations of expert evaluations within the bayesian paradigm. *Demography*, *51*(5), 1933–1954. doi: 10.1007/s13524-014-0318-5

Blake, A., & Mumtaz, H. (2012). Applied bayesian econometrics for central bankers.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software, Articles*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Cappelen, Å., Skjerpen, T., & Tønnessen, M. (2015). Forecasting immigration in official population projections using an econometric model. *International Migration Review*, *49*(4), 945–980.

Cohen, J. E., Roig, M., Reuman, D. C., & GoGwilt, C. (2008). International migration beyond gravity: A statistical model for use in population projections. *Proceedings of the National Academy of Sciences*, *105*(40), 15269–15274.

De Beer, J. (1997). The effect of uncertainty of migration on national population forecasts: The case of the netherlands. *Journal of Official Statistics-Stockholm*, *13*, 227–244.

Disney, G., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2015). *Evaluation of existing migration forecasting methods and models* (Tech. Rep.). ESRC Centre for Population Change, University of Southampton.

Dodd, E., Forster, J. J., Bijak, J., & Smith, P. W. (2018). Smoothing mortality data: the english life tables, 2010–2012. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

European Union Publications Office. (2017). *Quantitative assessment of asylum-related migration: a survey of methodology.* (Tech. Rep.). Office of the European Union.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society, Series A*.

Hajnal, J. (1955). The prospects for population forecasts. *Journal of the American Statistical Association*, *50*(270), 309–322.

Håvard, R. (2016). *Stan Prior Choice Recommendations.* `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`. (Accessed: 2016-11-15)

Hendry, D. F., & Hubrich, K. S. (2006). Forecasting economic aggregates by disaggregates.

Hindrayanto, I., Jacobs, J., & Osborn, D. (2014). On trend-cycle-seasonal interactions.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679–688.

Keilman, N., Pham, D. Q., Hetland, A., et al. (2002). Why population forecasts should be probabilistic–illustrated by the case of norway. *Demographic Research*, *6*(15), 409–454.

Koopman, S. J., & Lee, K. M. (2009). Seasonality with trend and cycle interactions in unobserved components models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *58*(4), 427–448.

Lee, R., & Tuljapurkar, S. (1994). Stochastic population forecasts for the united states: Beyond high, medium, and low. *Journal of the American Statistical Association*, *89*(428), 1175–1189.

Lutz, W., & Goldstein, J. (2004). Introduction: How to deal with uncertainty in population forecasting? *International Statistical Review*, *72*(1), 1–4. doi: 10.1111/j.1751-5823.2004.tb00219.x

Neal, R. M., et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*(11).

Pijpers, R. (2008). Problematising the orderlyaesthetic assumptions of forecasts of east–west migration in the european union. *Environment and Planning A*, *40*(1), 174–188.

Raymer, J., Abel, G. J., & Rogers, A. (2012). Does specification matter? experiments with simple multiregional probabilistic population projections. *Environment and Planning A*, *44*(11), 2664–2686.

Schmidt, C. M., & Fertig, M. (2000). Aggregate-level migration studies as a tool for forecasting future migration streams. *IZA Discussion paper No. 183*.

Stan Development Team. (2015). Stan modeling language user's guide and reference manual, version 2.10.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Taeuber, C., Keyfitz, N., & Flieger, W. (1969). *Introduction to the Mathematics of Population.*

(Vol. 34) (No. 3). American Sociological Review. doi: 10.2307/2092532

Wilson, T., & Bell, M. (2004). Australia's uncertain demographic future. *Demographic Research*, *11*, 195–234.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114.