An Stochastic Method to Forecast Mexican Migration

Víctor Manuel García-Guerrero

Centre of Demographic, Urban and Enviromental Studies, El Colegio de México, Mexico.
Camino al Ajusco No. 20 Col. Pedregal de Santa Teresa. C.P. 10740. Mexico City.

Abstract

In this paper a method to model and forecast net migration by age and sex is proposed. Such method is inspired in the previous work of Lee (1993) which at the same time is based on the well known Lee-Carter model. The method proposed allows to forecast migration consistently with the population dynamic in the sense that it gets sustainable migration levels. This sustainability is reached thanks to a logistic transformation of total net migration which allows to constrain it avoiding negative populations in a total population forecast. Such method is applied to recent Mexican net migration estimates by age and sex from 1960 to 2010 (SOMEDE, 2011). The predictive power of the model is tested applying the method to period 1960-1990 and forecasting from 1991 to 2010 and comparing the results with the estimates for such period.

Keywords:
Stochastic migration forecasts, demographic forecasting, Lee-Carter model, Mexican migration

---

Email address: `vmgarcia@colmex.mx` (Víctor Manuel García-Guerrero)

1. Introduction

In recent years Mexican migration to United States has decreased by several causes. According to Passel, "[t]he standstill appears to be the result of many factors, including the weakened U.S. job and housing construction markets, heightened border enforcement, a rise in deportations, the growing dangers associated with illegal border crossings, the long–term decline in Mexico's birth rates and broader economic conditions in Mexico" (Passel et al., 2012, p.6). These factors were unpredictable at the beginning of 21st century, when migration from Mexico to U.S. was at its highest levels: In 2000, the annual immigration from Mexico to U.S. was around 770 thousand people and in 2010 it drops down to 140 thousand per year approximately (Passel et al., 2012, p.8).

Nowadays, as well as in many other Latin American countries, the migration to Mexico is affected also by another unexpected factor called the "third wave" of immigrants from Spain. According to the Mexican National Institute of Migration at the beginning of 2012, more than one thousand of Spanish asked for permission to work in Mexican territory and in the last four years this permission was given to 7,630 Spanish.[1] The main factor that push this third wave is the strong effects the world economic crisis has had in countries like Spain, Greece and Portugal[2] Other collateral effect of the

[1]The first Spanish wave was between 1880 and 1930; in such period Mexico received around 30 thousand Spanish; the second wave was between 1939 and 1950 and in this period 25 thousand Spanish arrived to Mexico mainly exiled from the Franquist regime.

[2]The unemployment rate in Spain in 2012 was around 25% of economically active population (INE, 2013).

world economical crisis is that many migrants from Central America that try to cross Mexican territory to U.S. remain in Mexico waiting for better economical conditions to cross to U.S.

Thus, the migratory phenomenon is very complex and it depends of many exogenous and highly uncertain factors, whose nature can mainly be economic and labor.[3] The past population projections could not foresee the strong impact that the recent world economical crisis would have in demographic dynamics, mainly on the migration component.[4] That happened because the method in which they are based does not take into account such source of uncertainty.

There has been less effort to forecast migration stochastically as it has been for mortality and fertility. The reason for this is the lack of long reliable time series by single age and sex of migrants. Nevertheless, there are some very interesting approaches which can be divided in two: Those that forecast total migration amounts (Miller and Lee, 2004 and Bijak, 2011) and those that do it considering the age and sex structure (Hyndman and Booth, 2008). The first kind has been quite more developed than the second and the aims of such aproaches were different.[5] Although the works of Lee and Tuljapurkar (1994) and Hyndman and Booth (2008) persued the same goal: Forecast-

[3]The other important factor that increases migration's uncertainty is the criminalization of undocumented migration.

[4]Do it would be equivalent to forecast this world economical crisis.

[5]The main lack of the first approach is that if an estimation of migration the age patern is requiered, some assumptions about such age pattern should be made. Usually, the model proposed by Rogers and Castro (1982) is employed to solve this issue but one should to assume that such distribution will be the same for each forecasted year.

ing the total population by age and sex by means of the cohort-component method, the second is superior to the former in the sense that the authors forecasted migration by age and sex using a Box-Cox transformation (with $\lambda = 1$) of a generalization of LC model while the Lee and Tuljapurkar's approach just used the U.S. Census Bureau's migration projections –which are deterministic.

The issue with Hyndman and Booth's approach is that they do not constrain migration and for some contexts where outmigration is the predominant migration flow, it is possible to have negative populations, mainly with the lower prediction interval. Thus, the aim of this paper is to develop and apply a method to forecast net migration by age and sex that it forecasts net migration consistently with the hole population dynamic. This forecast assess the uncertainty related with net migration and it could be used as an input for total population forecasts –by age and sex also– in a consistent way.

2. Methodology

Following the ideas of Lee (1993), the historical levels of net migration by age and sex are a lineal function which depends of three unknown parameters: the *net migration index*, g_t, which is only time-dependent with $t \in [1, n]$; the age schedule of net migration, a_x for $x \in [0, \omega+]$, where $\omega+$ is the last open age group, and the migration *intensity* at each age, b_x. Then, the net migration by age –and sex– is defined as[6],

$$g_{x,t} = a_x + b_x g_t + \varepsilon_{x,t}, \tag{1}$$

this expression is based on the well-known Lee-Carter model (1992).

The parameter of age schedule, a_x, is calculated as the average of net migration along time,

$$a_x = \frac{\sum_{t=1}^{n} g_{x,t}}{n}, \tag{2}$$

for each age. Each entry of the vector g_t is estimated summing each column of the matrix $g_{x,t} - a_x$.[7] Mathematically,

$$g_t = \sum_{x=0}^{\omega+} (g_{x,t} - a_x), \tag{3}$$

for each $t \in [1, n]$.

Finally, to estimate b_x, the equation (1) is solved by ordinary least squares. It is,

$$b_x = \frac{\sum_{t=1}^{n} g_t (g_{x,t} - a_x)}{\sum_{t=1}^{n} g_t^2}, \tag{4}$$

for each $x \in [0, \omega+]$.[8]

[6]The superscript that denotes the sex is eliminated to avoid unnecessary notation, but it is clear that this model applies for each sex as it is shown later.

[7]Some other estimation procedures were applied (like the SVD and the proposed by Booth et al. (2002), Wilmoth (1993), Hyndman and Booth (2008) among others) but the here exposed gives the best fit for Mexican data. The statistic χ^2 and the coefficient of determination was the criteria to discriminate between each procedure (see Booth et al., 2002).

[8]To guarantee that equation (1) is unique, it is necessary to constraint it to $\sum_{x=0}^{\omega+} b_x = 1$ and $\sum_{t=1}^{n} g_t = 0$.

The total net migration (NM) for each year is the sum of net migration by age over all ages, that is,

$$NM_t = \sum_{x=0}^{\omega+} g_{x,t} = \sum_{x=0}^{\omega+} a_x + \sum_{x=0}^{\omega+} b_x g_t = A + g_t \tag{5}$$

where $A = \sum_{x=0}^{\omega+} a_x$ is the mean historical net migration. Then, a couple of prior constrains are incorporated by means of a transformation of NM.

$$z_t = \ln\left(\frac{NM_t - L}{U - NM_t},\right) \tag{6}$$

where L and U are prespecified lower and upper bounds of total net migration (see Alho, 1990, p.524) and z_t is called *modified net migration index*.[9] If the main historical migration flow has been outmigration, then $L = \min(NM_t)$ and $U = -\min(NM_t)$. If the main historical migration flow has been immigration, then $U = \max(NM_t)$ and $L = -\max(NM_t)$ or $L = 0$. If the historical minimum is quite far from the sustainable level of net migration a nonlinear optimization method should be used (like the Levenberg-Marquardt nonlinear least-squares algorithm found in the R-package *minpack.lm*) or, as a rule of thumb, it is possible to approach L to the next lower-integer ended in zero. Furthermore, if our knowledge about the constraints is limited "but rather leads to some probability distribution [...] estimation could be carried out using Bayesian methods" (Lee, 1993, p.192).

After the estimation process, the main task is modeling and forecasting the modified net migration index as a time series process using the Box and Jenkins approach (Chatfield, 1989).

[9]Once z_t is forecasted it is possible to get the net migration solving equation (6) for NM_t (a logistic transform) and then solving equation (5) for g_t and replacing the result in (1) to get the migration age shedule.

3. Application

3.1. Data sources

The most reliable recent estimation of the Mexican population is used to apply the method proposed (SOMEDE, 2011). In such estimation, the Mexican international migration is based in both Mexican and U.S. census from 1970 to 2010, Mexican households demographic surveys at national level and the American Community Survey (ACS) for several years. The immigration flows are extracted from Mexican data sources and the outmigration flows are extracted from U.S. data sources (the outmigration flow to other countries difference of U.S. is extracted from Mexican data sources but it is relatively small). Four kinds of migration flows are considered: The Mexicans who go to live to U.S., those no-born in Mexico who go from Mexico to U.S., those (Mexicans and non-Mexicans) who go to live to other country different of U.S. and the immigrants to Mexico (included the return flows) (SOMEDE, 2011, p.85).

Some procedures were used to fit, correct (mainly for digit preference) and smooth the information. The final result is presented by sex and single age yearly for period 1960-2010 as it is shown in figures 1 and 2 (SOMEDE, 2011, pp.86-89). According to such results, female and male net migration have mainly been affected by outmigration flow. In the whole period, the net migration has been negative. Just very recently (from 2005) we find positive figures in the first ages, mainly because return of families and children born in foreign countries (mainly in U.S.) from Mexican parents.

Figure 1 shows female net migration by single age yearly from 1960 to 2010. The net result of migration flows in such period started in a loss of

14,740 people in 1960 to its highest level a loss of 346,852 people in 2000 –i.e. it increases almost 24 times in 40 years. From 2001 to 2010 such loss of people has been decreasing because the motives previously described. In 2001 the loss of female population was of 251,350 people and in 2010 it decreases 14%, a loss of 35,195 females in such year. As it shows, the historical age-pattern is concentrated in children and working ages.

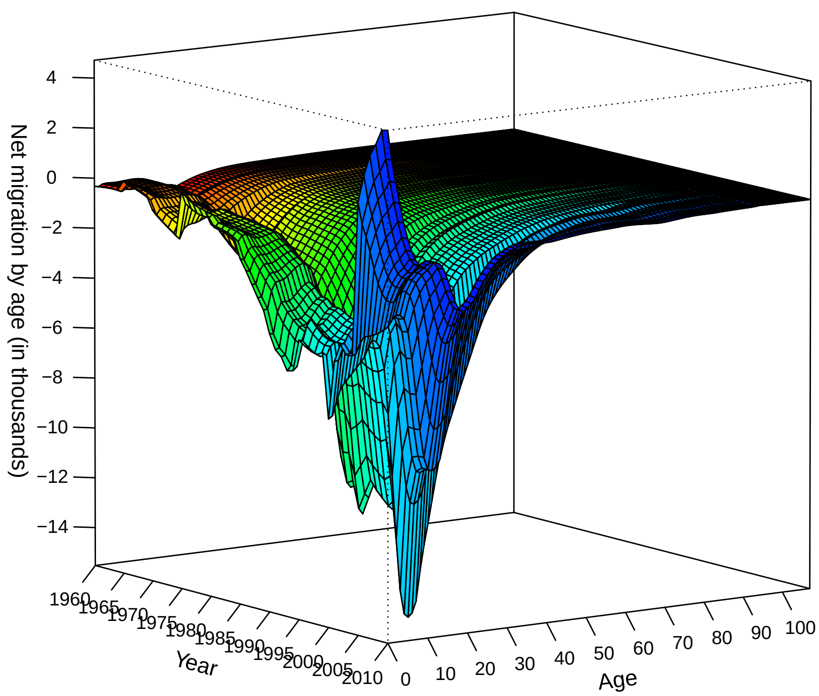

Figure 1: Female net migration, Mexico 1960-2010

By other way, Figure 2 shows male net migration by single age yearly from 1960 to 2010. The net result of migration flows in such period started in a loss of 13,704 people in 1960 to its highest level a loss of 424,347 people in 2000 –an increase of almost 31 times in 40 years.

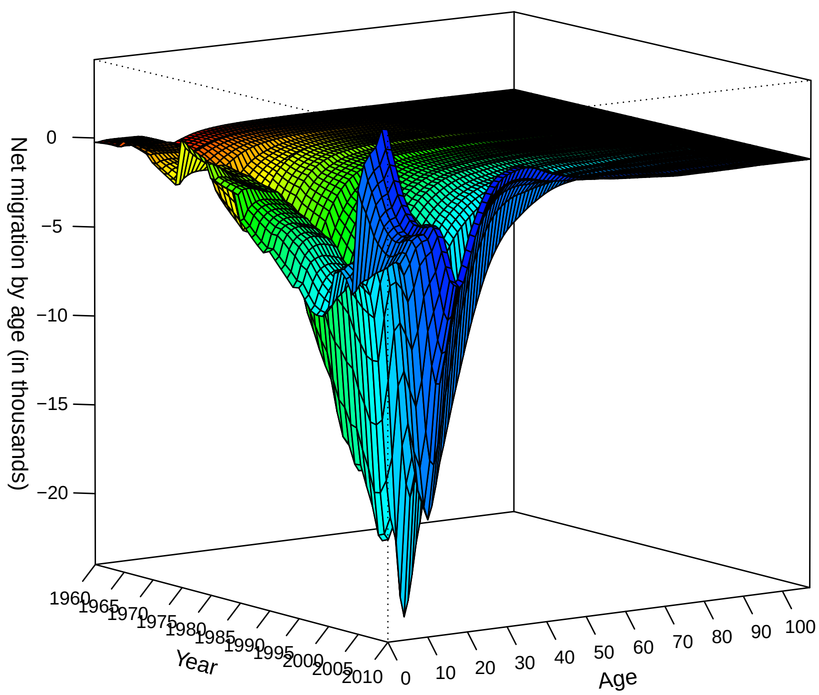

Figure 2: Male net migration, Mexico 1960-2010

From 2001 to 2010 such loss of people has been decreasing because the motives previously described. In 2001 the loss of male population was of

305,475 people and in 2010 it decreases 18%, a loss of 5,654 males in such year –6 times fewer than women. As women net migration, the historical age-pattern of male net migration is concentrated in children and working ages.

3.2. Results

Following the method described before, the parameter that determine the trend, age patern and intensity of net migration are shown in figures 3 and 4. The coefficient of determination is $R^2 = 0.94$ for females and $R^2 = 0.92$ for males.

In both cases, male and female, the shape of parameter a_x is almost the same that the proposed by Rogers and Castro (1982), although the method is quite deferent. Such age patern indicates that migration has an intrinsic labor character and because familiar reunífication, it is intensified in the first ages of life. Moreover, the intensity parameter indicates that net migration is strong in labor ages. Finally, the trend parameter indicates that such trend is descendant but it is quite hard to state that such trend will continue falling down continuously. That is because of the high volatility of migration phenomenon and, as was stated before, many collateral factors and policies.

For model (6), as sustenaibility limit levels the predefined bounds are $L = -500,000$ and $U = 500,000$ for males and $L = -400,000$ and $U = 400,000$ for females. The last bound refers to the maximum level of migration that mexican demographic dynamic could support.

After estimation procedure described in equations (2), (3), (4), (5) and (6), the modified net migration index is modeled and forecasted as a time series process. For males, the best fit was a random walk ARIMA(0,1,0),

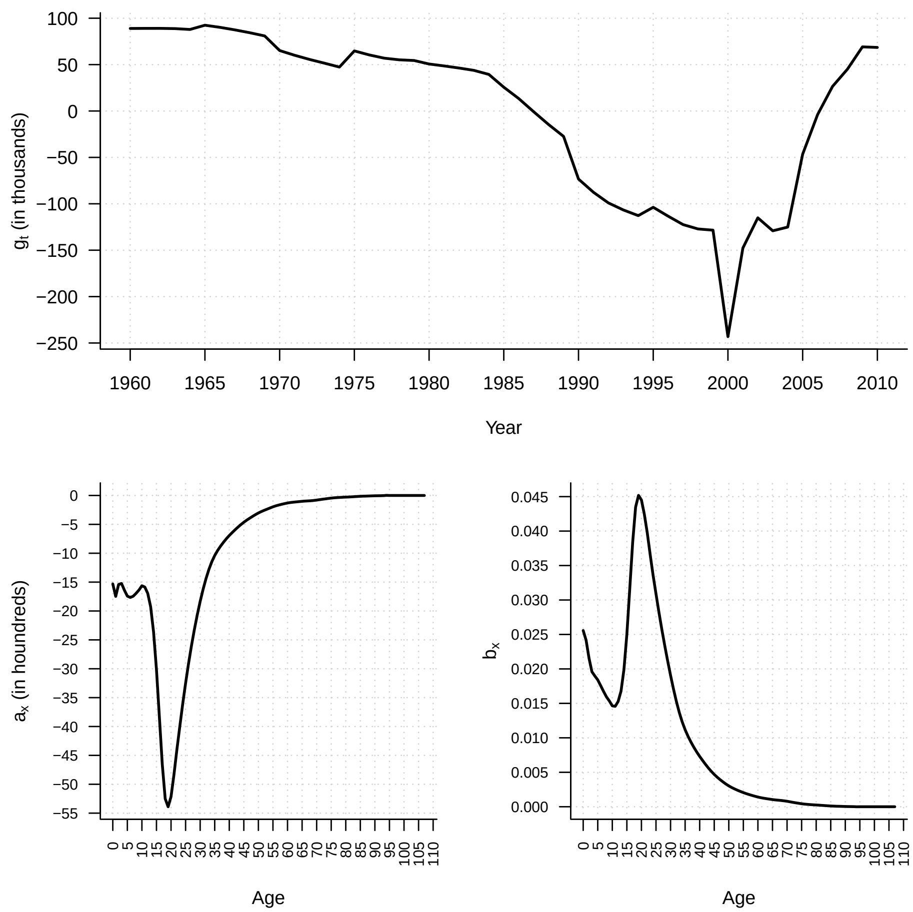

Figure 3: The three parameters estimated according to equation (1), Mexico, females.

that is

$$z_t = z_{t-1} + e_t \tag{7}$$

and the variance estimated is $\hat{\sigma}^2_{e_t} = 0.0617$. For females, the best fit was an integrated moving average process ARIMA(0,1,1), that is

$$z_t \;=\; z_{t-1} + e_t + \theta_1 e_{t-1}$$

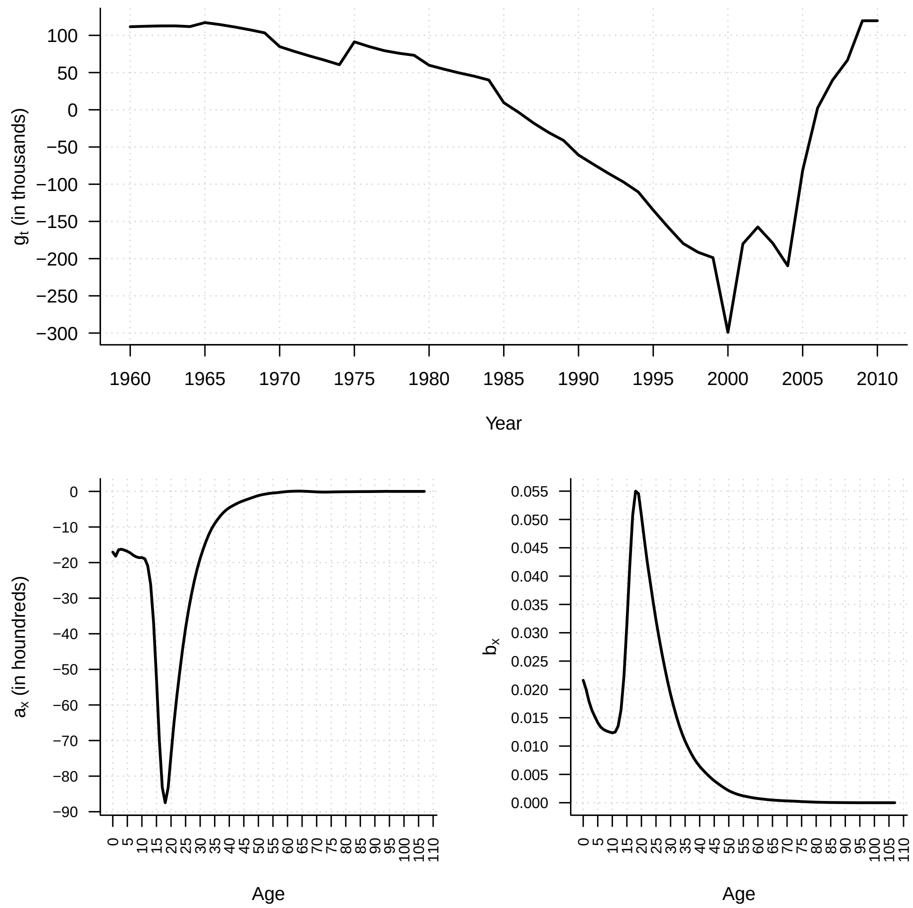

Figure 4: The three parameters estimated according to equation (1), Mexico, males.

$$= \quad z_{t-1} + e_t - \underset{(0.1226)}{0.2783} e_{t-1} \tag{8}$$

and the variance estimated is $\hat{\sigma}^2_{e_t} = 0.06784$. This model implies that in the median, the last estimated value remains constant along the projection horizon. That is because the model errors are independent and identically distributed random variables, its distribution is assumed normal with zero mean and the variance mentioned before. In figures 5 and 6 the female and male net migration estimated and forecasted are shown respectively with its prediction intervals of 95% of confidence. The gray lines represents a sample of the 10 thousand simulation realized and blue lines are the percentiles 2.5, 50 and 97.5.

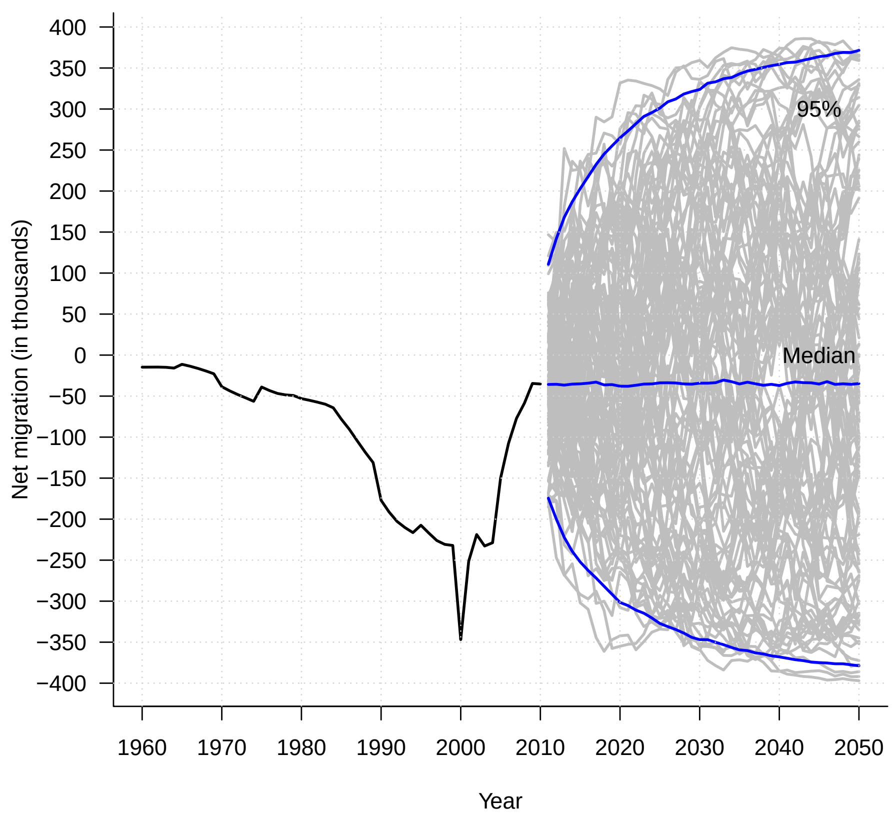

Figure 5: Net migration estimated 1960-2010 and forecasted 2011-2050, females, Mexico.

Note that in both cases, the prediction intervals are wider than one could expect, but it only reflects the high degree of uncertainty under Mexican net migration is. By other side, as it was noted before, in the median the net migration remains constant at the last level estimated.

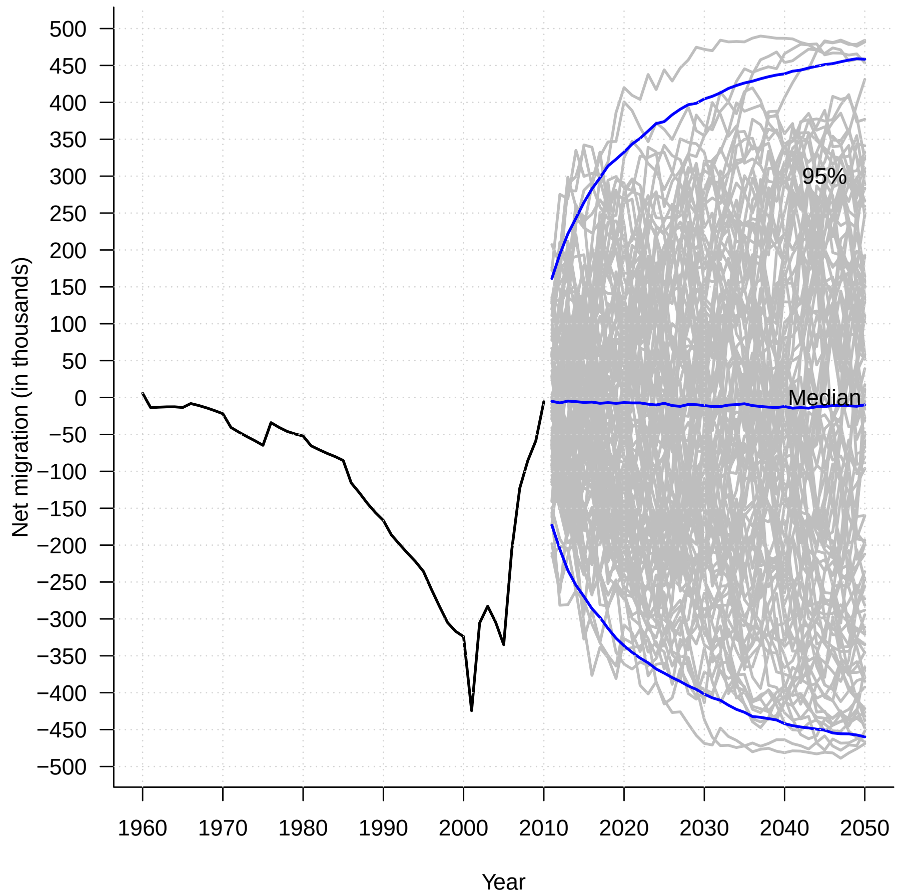

Figure 6: Net migration estimated 1960-2010 and forecasted 2011-2050, males, Mexico.

The difference between female and male net migration median forecast is that the latest remains constant closer to zero than the first. Other difference is that male net migration prediction interval is wider than female's. This means that, in males case, the probability that it will be positive or negative is equivalent to toss a coin.

Now, keeping each simulation and going back in the method, a final result is the net migration by single-age. Calculating the respective percentiles to state the predictive interval (PI) of 95 and 67% of probability (orange and yellow shades, respectively in Figure 7). Thus, as it shown in the figure, the PI's width is consistent with what it is expected: It is wider for males than for females and it is also wider when forecasted year is far away from base year.

Other characteristic of the single age net migration forecast is that in the median remains negative numbers –an exception are the very first ages. Thus, it is likely that in the future the coin will be quite more charged to outflow migration. By another hand, note that as time goes by, PIs start to become symmetrical, but at the beginning the PIs are quite biased to negative net migration flows.

3.3. Testing the predictive power

To test the predictive power of the method it was replied to estimate the period 1960-1990 and forecasting the period 1991-2010. The forecast obtained is compared with the net migration estimates for such period. With this test it is possible to assess the forecasting capability of the model for short, mean and long horizon.

After estimation procedure described in equations (2), (3), (4), (5) and (6), the modified net migration index is modeled and forecasted as a time series process. The coefficient of determination is $R^2 = 0.98$ for females and $R^2 = 0.92$ for males.

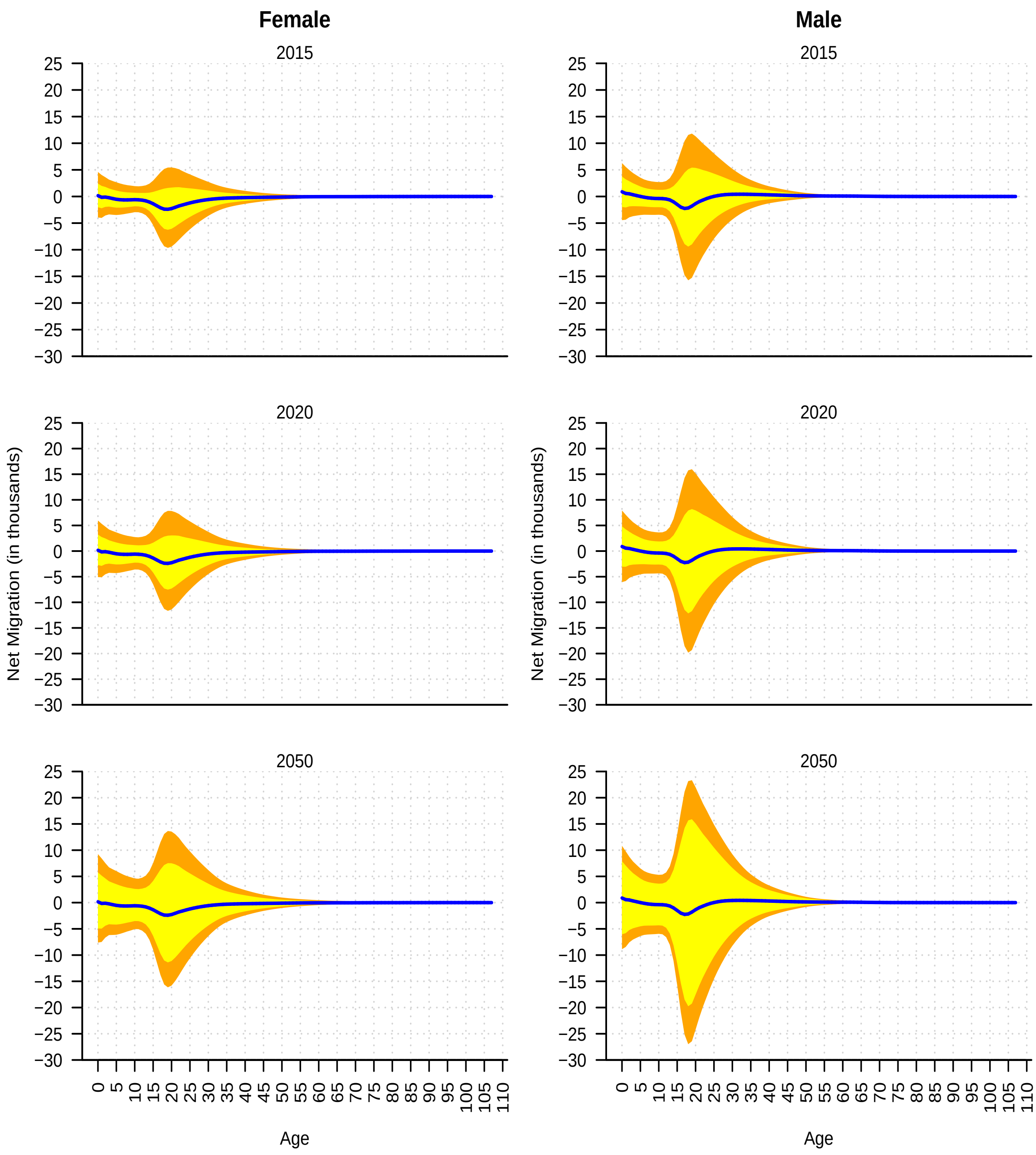
Female
Male
2015
2020
2050
Net Migration (in thousands)
Age
25
20
15
10
5
0
−5
−10
−15
−20
−25
−30
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110

Figure 7:

For males, the best fit was an autoregresive integrated moving averages process ARIMA(1,1,1), that is

$$\begin{aligned} z_t &= z_{t-1} + \phi_1(z_{t-1} - z_{t-2}) + e_t + \theta_1 e_{t-1} \\ &= z_{t-1} + \underset{(0.0741)}{0.9565}(z_{t-1} - z_{t-2}) + e_t - \underset{(0.1834)}{0.7201} e_{t-1} \end{aligned} \tag{9}$$

and the variance estimated is $\hat{\sigma}^2_{e_t} = 0.001541$. For females, the best fit was an integrated autoregresive process ARIMA(1,1,0), that is

$$\begin{aligned} z_t &= z_{t-1} + \phi_1(z_{t-1} - z_{t-2}) + e_t \\ &= z_{t-1} + \underset{(0.1727)}{0.8237}(z_{t-1} - z_{t-2}) + e_t \end{aligned} \tag{10}$$

and the variance estimated is $\hat{\sigma}^2_{e_t} = 0.0025$.

Figures 8 and 9 show the estimated level of net migration and the prediction interval of 95% for females and males, respectively. In these figures, it is clear that females model is better than males'. But in general terms it is possible to state that the model forecasts well in short, medium and long terms.

In the case of female net migration, in short term the estimated values are over the upper prediction limit but just for around five thousand people. At the end of the forecasting interval the estimations are over again the upper prediction interval but just by few houndreds people.

In the case of male net migration, the model looks like less able to forecast rare events. A preliminary proof of that is that both the highest historical level and the lowest are outside de 95% prediction interval (representing the

Figure 8: Prediction power test for female net migration forecast model. Mexico 1990-2010

20% of forecasted years). So it is possible that those rare events were completely out of sight of the model and because Mexican migration is stronger in males than females the uncertainty is higher in them. It means that males are exposed to more rare events than females –at least until the information we have.

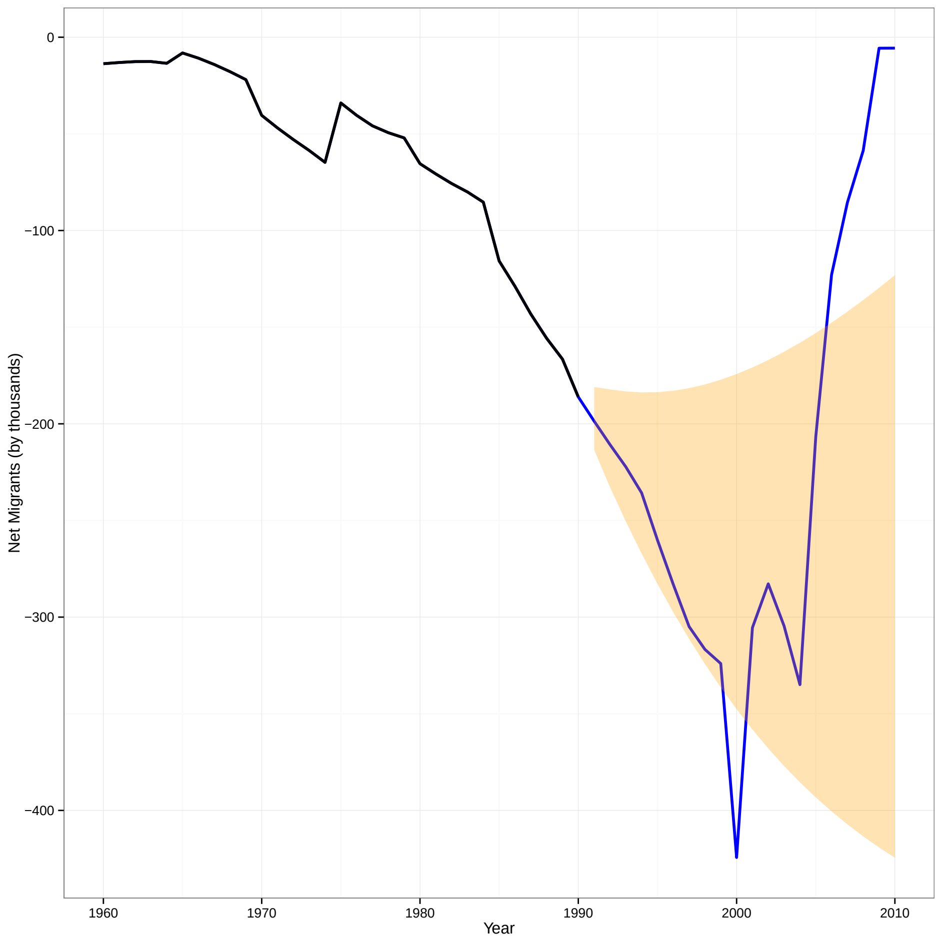

Figure 9: Prediction power test for male net migration forecast model. Mexico 1990-2010

4. Discussion and final remarks

The model here proposed is inspired by the previous work of Lee (1993) and it is based in a logistic function. It allows to constrain the main function – the total net migration– to certain bounds. This is a key to forecast migration consistently with the population dynamic in the sense that if migration is forecasted, regardless its impact on the entire population dynamic –according to the balancing equation of population change– then it is possible to get negative figures, which is senseless. That means the model here proposed allows to get sustainable migration levels for the entire demographic dynamic.

One of the most interesting results is that the best time series model for migration is a differentiated moving average random process –i.e. an ARIMA$(0, d, q)$. It means that, in average, it is better to forecast net migration without change than assuming some future fluctuation –thinking in a deterministic way. For a moving average process with one difference –without a drift–, the trend is determined by its order –q– so after q-years in the future, all the uncertainty will be concentrated in a purely random process with zero-mean. It means that after q-years the trend is constant along the forecasting period (Chatfield, 1989, pp.31-35).

The uncertainty associated to Mexican migratory phenomenon is pretty high but for sure the method here proposed helps to assess the whole uncertainty of population dynamic. After testing the predictive power of the method we conclude that it models and forecasts very well the net migration. Some improvements could be made, for example, using a bayesian approach to model and forecast the modified migration index, mainly those parameters where it is possible to state prior probability distributions, the upper and lower bounds of logarithmic transformation. This model is also applicable to immigration and outmigration separately.

References

Alho, Juha (1990). “Stochastic Methods in Population Forecasting” *International Journal of Forecasting*, 6, 521-530.

Bijak, Jakub (2011). *Forecasting International Migration in Europe. A Bayesian View*, Springer Series on Demographic Methods and Population Analysis 24, Springer.

Booth, Heather, J. Maindonald and L. Smith (2011). “Applying Lee-Carter under conditions of variable mortality decline” *Population Studies: A Journal of Demography*, 56(3), 325-336

Chatfield, Christopher (1989). *The Analysis of Time Series*, Chapman and Hall, Ltd.

García-Guerrero, Víctor M. (2010). *Projections and Population Policies in Mexico. A Critical Review and Methodological Proposal* (in Spanish), PhD thesis on Population Studies, Centre of Demographic, Urban and Enviromental Studies, El Colegio de México, Mexico.

Hyndman, Rob and H. Booth (2008). “Stochastic population forecasts using functional data models for mortality, fertility and migration”, *International Journal of Forecasting*, 24(3), 323-342.

National Institute of Statisctics, Spain (2013). *Tasas de actividad, park y empleo, por sexo y distintos grupos de edad*, viewed February 26th, 2013 at http://www.ine.es

Lee, Ronald (1993). "Modeling and forecasting the time series of U.S. fertility: Age distribution, range and ultimate level", *International Journal of Forecasting*, Vol. 9, No. 2, pp. 187-202.

Lee, Ronald and S. Tuljapurkar (1994). "Stochastic population forecast for the United States: Beyond high, medium and low", *Journal of the American Statistical Association*, Vol. 89, No. 428 pp. 1175-1189.

Lee, Ronald and L. Carter (1992). "Modeling and forecasting U.S. mortality", *Journal of the American Statistical Association*, Vol. 87, No. 419, pp. 659-671.

Mexican Demographic Society, SOMEDE (2011). *Demographic estimates for Mexico and its states, 1990-2010*, unpublished.

Miller, Timothy and R. Lee (2004). "Report I. A Probabilistic Forecast of Net Migration to the United States" in Lee, Miller and Anderson *Stochastic Infinite Horizon Forecasts for Social Security and Related Studies, National Bureau of Economic Research*, Working Paper No. 10917

National Population Council, CONAPO (2008). *Projections of Mexico's population, its states, municipalities and localities, 2005-2050. Methodological document* (in Spanish), CONAPO, Mexico.

CONAPO (2003). *Projections of Mexico's population, its states, municipalities and localities, 2000-2050. Methodological document* (in Spanish), CONAPO, Mexico.

Passel, Jeffrey, D. Cohn and A. González-Barrera (2012). *Net Migration from*

Mexico Falls to Zero and Perhaps Less, Pew Hispanic Center, Pew Research Center, Washington D.C., U.S.A.

R Development Core Team (2005) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org

Rogers, Andrei and J. L. Castro (1982). “Patrones modelo de migración”, *Demografía y Economía*, 3(51), Vol. XVI, El Colegio de México, México.

Wilmoth, John, (1993). “Computational methods for fitting and extrapolating the Lee-Carter model of mortality change”, *Tecnical Report*, Department of Demography, University of California, Berkeley, U.S.A.