WP 4.4
16 October 2013

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint Eurostat/UNECE Work Session on Demographic Projections
organised in cooperation with Istat
(29-31 October 2013, Rome, Italy)

Item 4 – Assumptions on future migration

Dynamical models for migration projections

Violeta Calian, Statistics Iceland

Abstract

All model-based forecasts are subject to uncertainties: data and measurement method, model, parameters for any given model, future uncertainty due to effects of unobserved type of events/shocks. In addition, in order to give valid point estimates and prediction intervals for the forecasted values, one can not rely on classical regression type models due to the auto-correlated and non-stationary character of the time series involved in migration processes.

We have used dynamical, or auto-regressive distributed lag (*ARDL*) models as a solution to this problem. This approach does not treat auto-correlation and non-stationarity as nuissance phenomena but includes them into the model. The dependent variable at time “t” is modelled as a function of its own values at different time laggs and of the values of several simultaneous or lagged predictor variables.

We have thus obtained short time predictions for the number of immigrants/emmigrants of Icelandic and foreign citizenships as functions of several time series predictors: unemployment, change in GDP values, number of graduating students and dummy variables mirroring the EEA resizing in time and the Icelandic economic boom which ended in 2008. The time series we used for fitting the models are about 45 year time steps long and we produced a forecast for the next 5 years in 2011, 2012 and this year. The results are a good fit to the true values, even as point estimates, although our confidence intervals are rather large.

Our method could be further improved by using vector autoregressive distributed lag models, since all components of our migration process are correlated, but one has to evaluate carefully how realistic this is in terms of time series lengths and dimension of the vector space.

1. Introduction and formulation of the problem

Migration forecasting is an important part of population projections and may also have a significant impact on economical decision making. Statistics Iceland, as most statistical offices, has traditionally reported predictions for medium and low / high variants of migration flows, based on deterministic models. These models had always involved economic and social variables, but did not take into account the time series properties of the data and did not provide prediction intervals for the forecasted variants. Nor could they give reliable answers during high economic instability.

Starting with the year 2011, we have improved our methods. Based on a careful data analysis we found that, in order to be able to give valid forecasts, the main statistical problem to be solved was to deal correctly with the auto-correlated and non-stationary character of the time series data, for both the independent and dependent variables. We proved instead that these time series are first-difference stationary, making auto-regressive distributed lag (*ARDL*) models legitimate candidates for inference, see Pesaran (1995, 1999). This is a necessary but not sufficient condition, see Johansen (2010), for un-biased and consistent point estimates and independent and identically distributed residuals. Choosing the structure and the order of the ARDL model by a consistent model selection criterion is a crucial step, too.

Dynamical economical models for population projections are gradually attracting some well deserved attention, as discussed for example in Brunborg (2010) and already suggested by Keilman (2002). We prove here that they can also capture the very diverse evolution of the migration components (emmigrating/immigrating women/men/Icelandic/foreign citizens) under strong oscillations of the economic background and thus can be used for short term predictions.

In this paper: (i) we explain why we need to use dynamical models, (ii) what are the mathematical conditions for a reliable statistical inferrence and whether they are fulfilled by our data and models, (iii) we show how our models for all migration components are built and how they perform.

2. Data and notations

The source of migration demographic data is the Icelandic National Register, which contains information on migration events since 1961 and is updated on a continuous basis, as opposed to once a year, since 1986. As showed by Statistics Iceland in WP1 (2010) and WP2 (2010), the estimated values of migration flows are reasonably accurate, although the short term migration has an influence on the accuracy of the emmigration figures. This effect is mainly due to de-registration lag effects but is well measured and stable in time. The net migration numbers are not affected by this phenomenon in a significant way.

The data concerning unemployment rates, gross domestic product and their short term forecast is provided by the department of national accounts and public finances of Statistics Iceland.

The number of graduating students and its predicted values for the next few years is provided by the department of education of Statistics Iceland.

In the next sections, we will use the following notation: y_1, y_2 - the number of Icelandic immigrants/emmigrants, men; y_3 , y_4 - the number of Icelandic immigrants/emmigrants, women; y_7, y_8 - the number of immigrants/emmigrants, women of foreign citizenship; y_5, y_6 - the number of foreign immigrants/emmigrants, men, x_4 - the unemployment rate; x_8 - a measure of GDP, x_5, x_6 - the number of graduating students, men and women respectively; *boom* - a dummy variable coupled to the Icelandic economic boom, reflecting also temporary changes in the registration process; *eea* – a dummy variable which mirrors the re-sizing of the *EEA*.

All these (ten) time series of 42 years length were tested for: (i) stationarity , by using augumented Dickey-Fuller and Kwiatkowski-Philips_Schmidt_Shin (KPSS) and (ii) auto-correlation of first and higher order, by using Durbin-Watson and Breusch-Gofrey tests. We found (see figure 1 for illustration) that they are first difference stationary or I(1).

3. Statistical models and short term forecast

Statistics Iceland must provide short and long term predictions for all the 8 components of migration y_1 to y_8, and for the net migration values, once a year.

Predicting the net migration numbers can be done in two ways: (i) by building an ARDL model for the time series which is the algebraic sum of the emigration and immigration series or (ii) by combining the results of eight ARDL models of all migration time series which enter the definition of the net migration.

The last method has the advantages that the estimates of the net migration are manifestly equal to the sum of the components‘ estimates and that it displays the widely different influence of economic evolution on the behaviour of various migration groups. But it also has the dissadvantage that the confidence intervals are rather broad and one has to expertly take into account the multiple (auto-) correlations (intra/) betweeen components. The former method has to deal only with autocorrelations and usual model parameter errors and gives narrower confidence intervals, but the values of the net migration estimates would not (in general) match the sums of components perfectly. It is most adequated when one needs only the net migration to be reported or used for further calculations.

We have built, in a consistent and parcimonious way, the following ARDL models:

$$y_1(t) \sim y_1(t-1) + x_4(t) + x_4(t-1) + y_2(t-1)$$

$$y_2(t) \sim y_2(t-1) + y_2(t-2) + x_5(t-2)$$

$$y_3(t) \sim y_3(t-1) + x_4(t) + x_8(t) + x_4(t-1) + x_8(t-1)$$

$$y_4(t) \sim y_4(t-1) + y_4(t-2) + x_6(t-2)$$

$$y_7(t) \sim y_7(t-1) + x_4(t) + x_8(t) + boom(t) + eea(t) + x_4(t-1) + x_8(t-1)$$

$$y_8(t) \sim y_8(t-1) + y_7(t) + y_7(t-1) + x_4(t) + x_8(t) + x_4(t-1) + x_8(t-1)$$

A ***note of caution*** is in order here:

the interpretation and model diagnostics when using ARDL is very different from the classical so-called *static* models. Collinearity, short and long term effects are key ingredients which have to be treated appropriately. One way to apply the classical notions is to transform the dynamical model into the equivalent error correction model (ECM). We actually did build both ARDL and ECM in most cases.

Our calculation of prediction intervals is correct and optimal in some sense, see Pessaran (1997), for each dynamical model y_1 to y_8, and it requiered significant additional work for the net migration when not analysed as a time series on its own but as a sum of correlated time series. We have to note also that y_5 and y_6 were not modeled separately but we used instead the empirically verified correlation between men and women migration numbers and the results of models y_7, y_8.

We applied several tests in order to establish the goodness of fit and the behaviour of residuals for all our models:

(i) Augumented Dickey – Fuller and KPSS tests for stationarity of residuals.
(ii) Box – Ljung and Durbin – Watson tests for autocorrelation, the latter not reliable for ARDL residuals but aplicable to individual predictor seies.
(iii) Auto-correlation function and partial auto-correlation functions calculations.
(iv) Breusch-Godfrey test for higher order serial correlation which also performs correctly for ARDL models.
(v) rainbow tests for the quality of fit.
(vi) standard residuals' normality checks (q-q plots, histograms) and Jacque – Bera tests.

As shown in *Appendix 1*, the results of the tests confirm the assumption that the models can be used for valid inference. All these tests have to be interpreted with great care and flexibility, too, since most are themselves based on some models and null hypotheses rejections can be caused by more than one reason.

4. Results and conclusions

The net migration is projected to be positive for the next 5 years (see Figure 2a), slowly increasing if the economic factors evolve according to our forecast. The confidence intervals do not exclude zero net migration. They reflect both the model uncertainty and more ad-hoc measures of the regressor prediction errors. Our past predictions gave a good fit to the subsequently recorded values, both as point estimates and as degree of confidence of the prediction intervals which included the true values.

The economic factors have a strong effect on the migration rates, as illustrated by the predicted values of migration under different scenarios which are created by modifying the GDP growth values, see Figure 2b.

Modeling separately the eight migration components implies that we assumed independence. This is however not the case and an ideal solution to this issue would be to use vector autoregressive distributed lag models. This in turn can be done only after a careful examination of the dimension of the problem in terms of time series lenghts and number of parameters estimated on their joint distribution. Otherwise, uniquely modeling the net migration is a safer although not as comprehensive alternative.

References

Alho, J.M. (1997): Scenarios, uncertainty and conditional forecasts of the world population, *Journal of Royal Statistical Society* A 160, Part 1, 71-85.

Brunborg, H. And Cappelen, A. (2010): Forecasting migration flows to and from Norway using an economic model, *Joint Eurostat/UNECE work session on demographic projections*, Lisbon, Portugal, 28-20 April 2010.

Johansen, S. (1996): Likelihood-based inference in cointegrated vector autoregressive models, *Oxford University Press*, Oxford.

Johansen, S. (2010): The analysis of nonstationary time series using regression, correlation ans cointegration with an application to annual mean temperature and sea level, *Discussion Papers*, no.10-27, Department of economics, University of Copenhagen.

Keilman, N., Pham, D. Q., Hetland, A. (2002): Why population forecasts should be probabilistic - illustrated by the case of Norway, *Demographic Research*, Vol. 6, Article 15.

Pesaran, M. H. (1999): An autoregressive distributed lag modelling approach to cointegration analysis, *Symposyum of the Norwegian Academy of Science and Letters*, Oslo, 3-5 March 1995, and DAE Working Paper Series No. 9514. Department of Econometrics, University of Cambridge, 1999.

Pesaran, M.H., Shin, Y. and Smith, R.J. (1996): Testing the existence of long-run relationships, *DAE Working Paper Series No. 9622*, Department of Econometrics, University ofCambridge.

Pesaran, M.H., Shin, Y. and Smith, R.J. (2001): Bounds testing approaches to the analysis of level relationships, *J. Applied Econometrics*, 16: 289-326.

WP1 Statistics Iceland (2010): Long-term and short term migration in Iceland - Analysis of estimation methods of Statistics Iceland, *Working paper 13 at the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses*, The Hague, The Netherlands, 10-11 May 2010.

WP2 Statistics Iceland (2010): The population statistics in Iceland - Analysis of population estimation methods of Statistics Iceland, *Working paper 6 at the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses*, The Hague, The Netherlands, 10-11 May 2010.

List of figures and appendices:

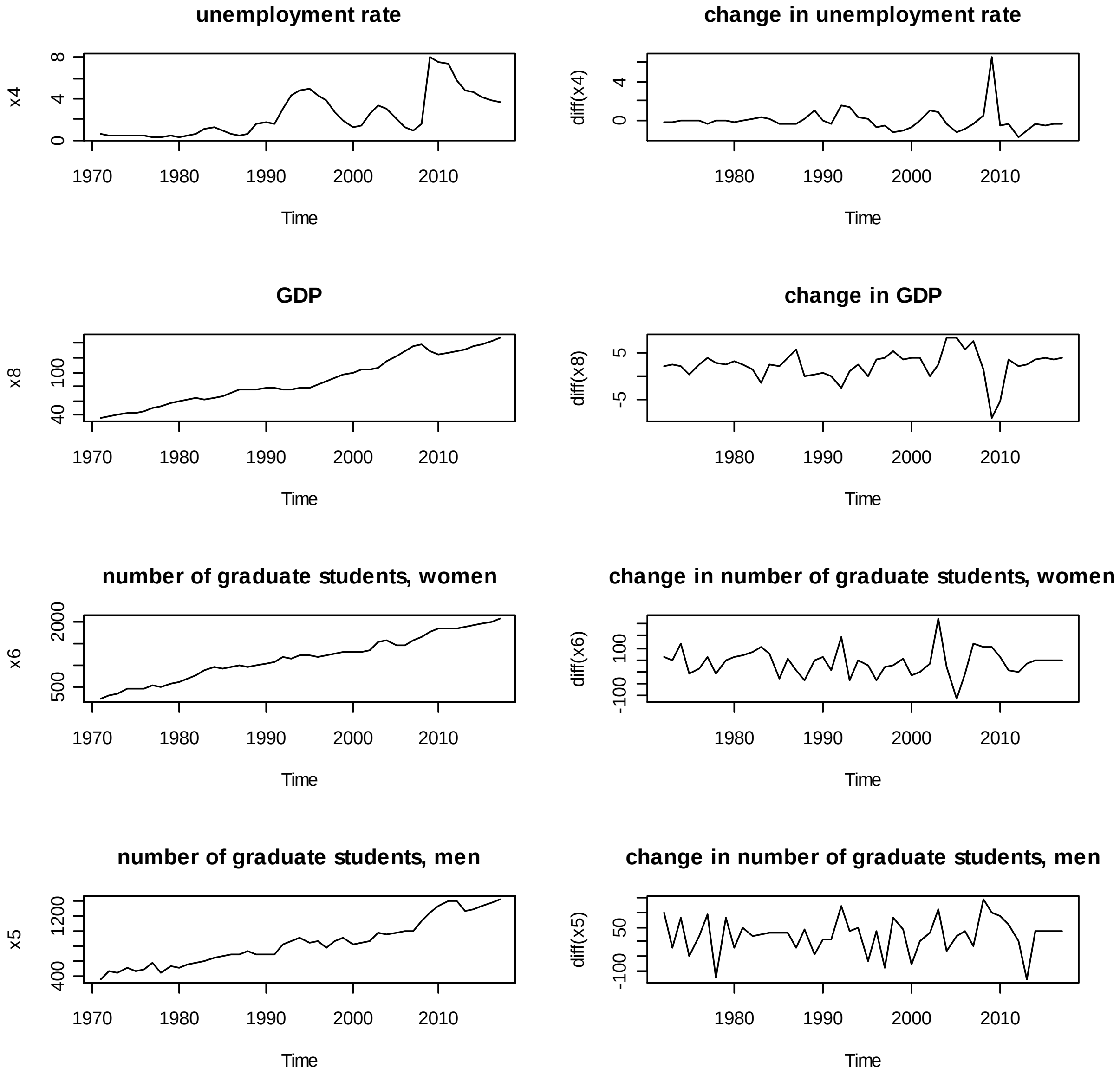

Figure 1-a : Regressor time series and their first order differences, displaying first difference stationarity, confirmed by Augumented Dickey-Fuller tests.

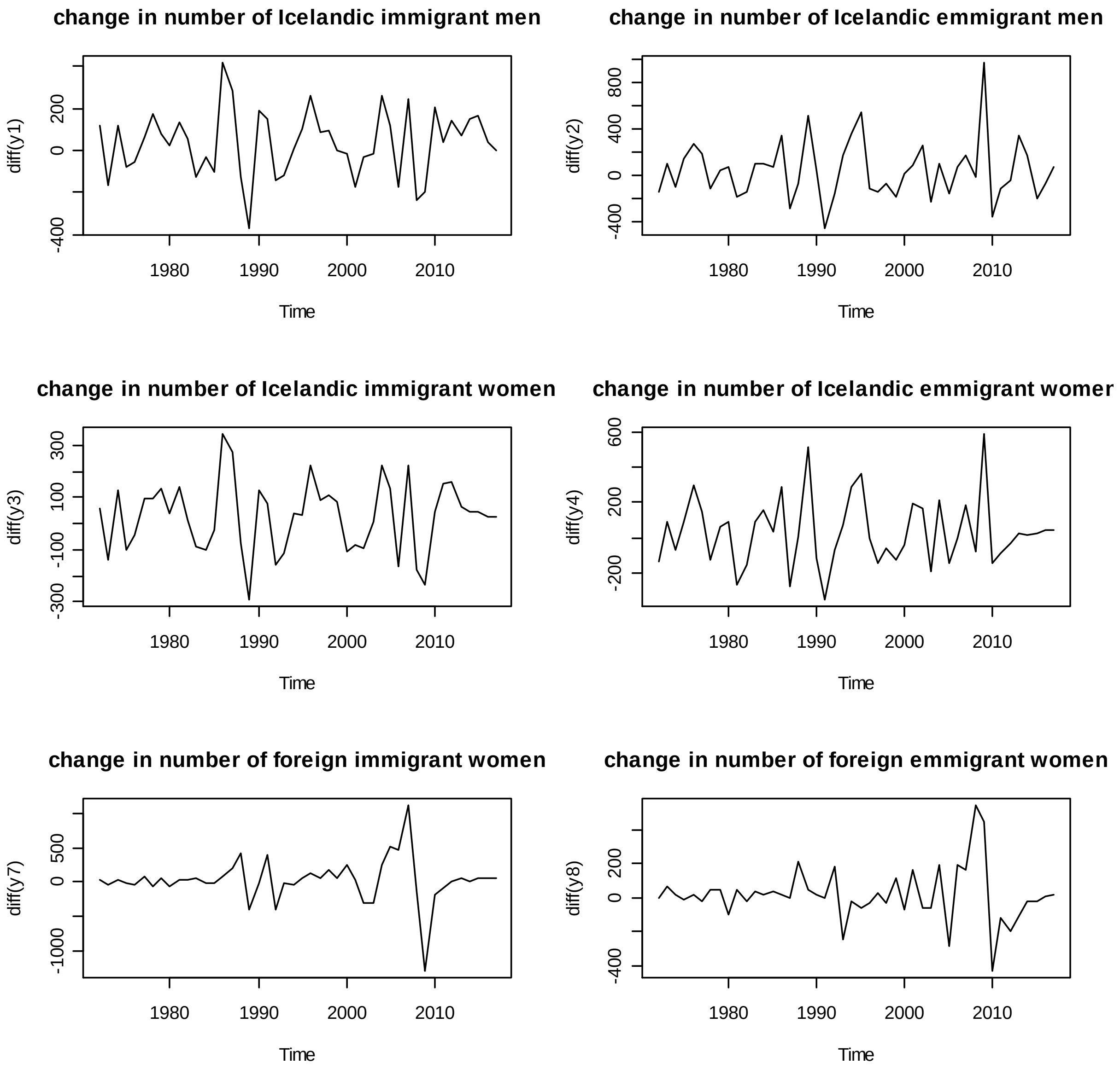

Figure 1-b: Independent time series displaying first difference stationarity, confirmed by augumented Dickey-Fuller tests.

Appendix 1: Models‘ goodness of fit and behaviour of residuals

A) Stationarity of residuals: the KPSS tests do *not* reject the hypothesis of stationarity of residuals‘ distributions for any of our models (all p-values greater than 0.1). Supporting this conclusion, augumented Dickey-Fuller tests reject *non*-stationarity (all p-values smaller than 0.01)

B) Normality of residuals: Jacques-Bera tests do *not* reject normality for any of the model residuals‘ distributions. The p – values of the tests are: 0.18, 0.12, 0.93, 0.06, 0.39, 0.35, respectively. These values reflect extremely well the general aspect of the empirircal distributions (histograms not shown).

C) Autocorrelation of residuals: Box-Ljung tests do *not* reject the hypothesis of random residuals. Same conclusion is supported by the direct calculation of autocorrelation for residuals shown in Figure A1:

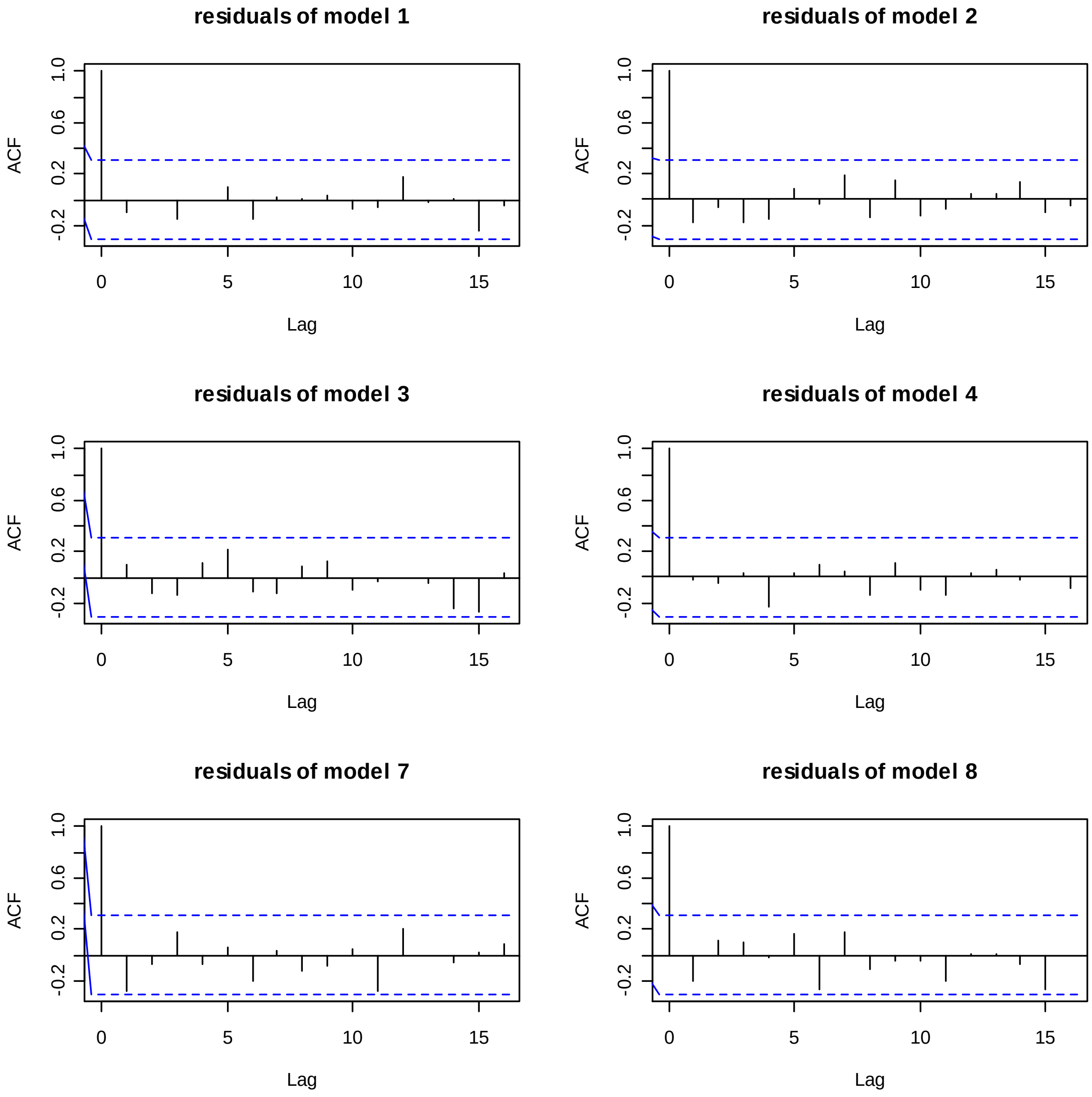

Figure A1: Values of residuals‘ autocorrelation of the economic models.

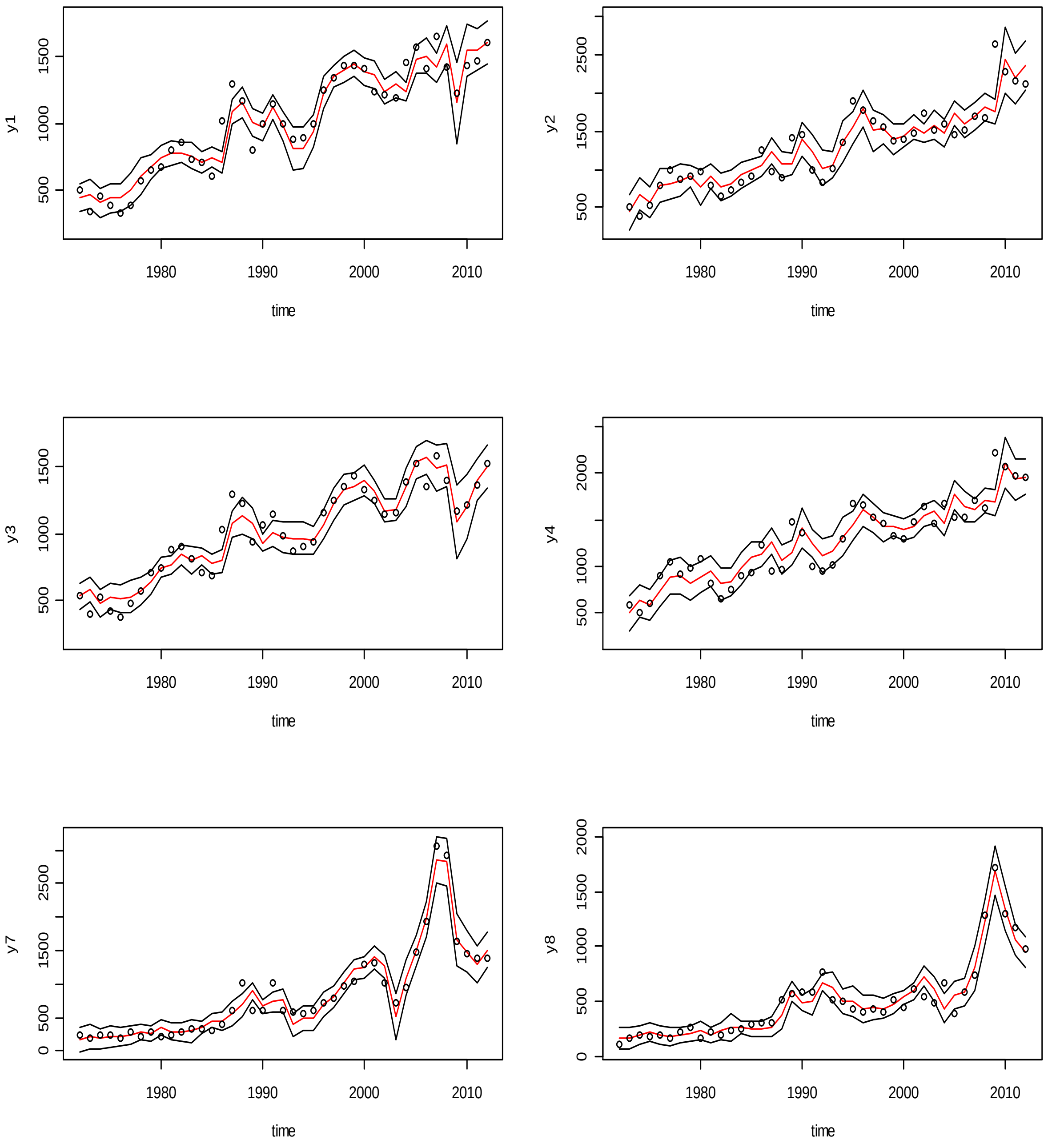

Figure A-2: Migration components: true (dots) and model estimated (continuous red lines) values, confidence intervals (continuous lines).

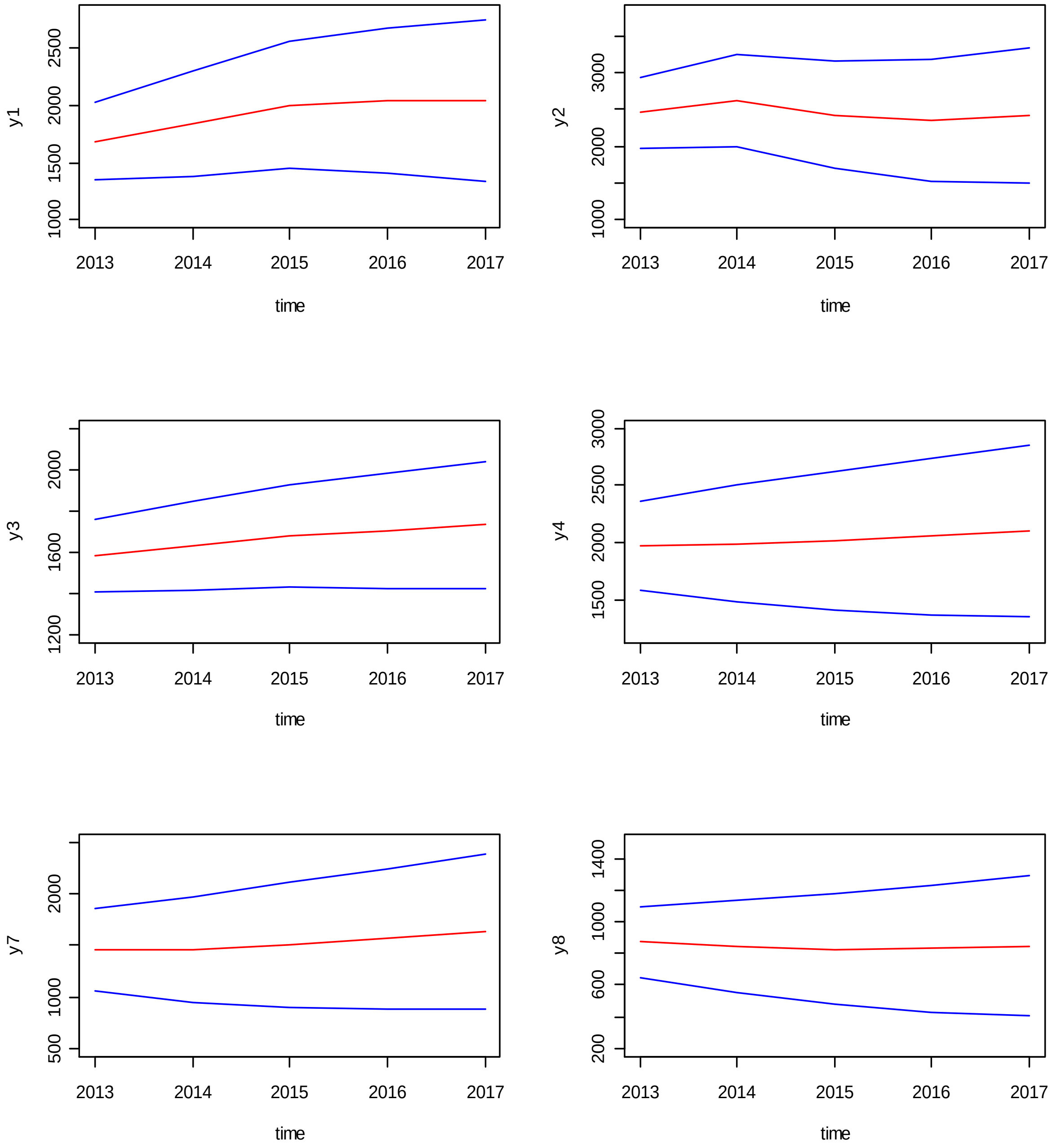

Figure 2-a: Short term projections (red lines) and prediction intervals (blue lines) for the migration components.

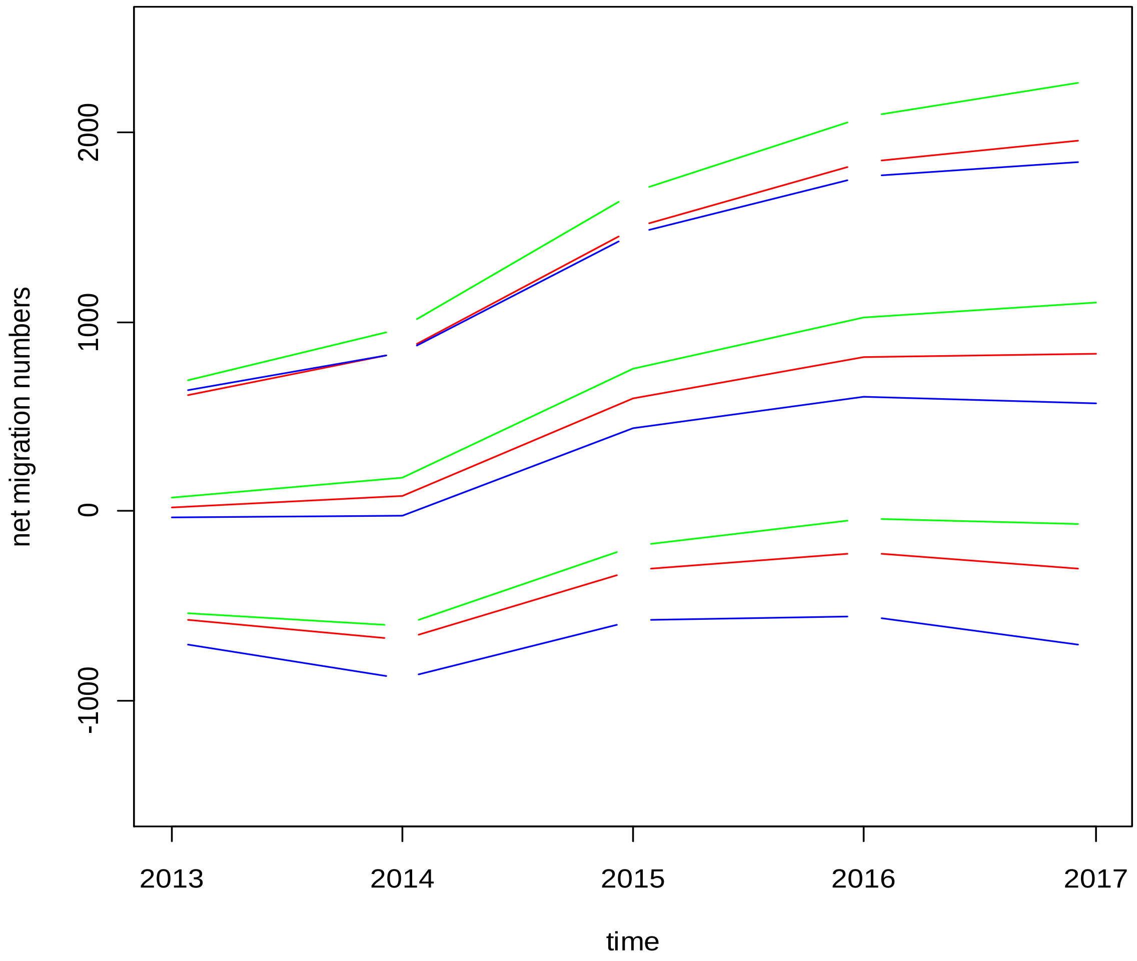

Figure 2-b: Net migration numbers: predicted values (continuous lines) and confidence intervals (dash lines), for various economic scenarios (current economic forecast in red, higher GDP growth in green, lower GDP growth in blue) .