Machine learning techniques to forecast population using Eurostat data: an exploratory study

Europen Conference on Quality in Official Statistics
Q2018

Krakow, 27th-29th June 2018

Álvaro Gómez Losada and Néstor Duch-Brown

Joint Research Centre - Seville (Spain)

alvaro.gomez-losada@ec.europa.eu

Contents:

1. Introduction and aim of this exploratory study.

2. Source of data and input variables.

3. Machine learning methodology.

4. Results.

5. Conclusions: strengths and limitations.

1. Introduction

1. Population projections are accomplished using well-settled, widely-accepted statistical methodologies

2. Machine learning (ML) techniques have emerged as common forecasting approaches in multitude of study fields

3. Typically, ML requires huge amount of data to properly perform

4. Eurostat data are initially not suited to be analyzed by means of ML approaches due to their small size

5. This exploratory study aims to find out how well ML algorithms perform in forecasting population one-year in advance.

2. Source of data and input variables

Data were obtained using the *Eurostat* library *(Lahti, 2017)* from R (RCran, 2017). Target country: *Spain*.

Eurostat's tables and variables:

demo_pjan: population by sex and age (1-99). 2005 to 2016 –pop-
migr_imm8: immigration by sex and age (2004 to 2015) –imm-.
migr_emi2: emigration by sex and age (2004 to 2015) –emi-.
demo_r_frate2: fertility rate by age (2004 to 2015) -fer-.
demo_magec: death by age and sex (2004 to 2015) –dea-.
demo_nsinrt: first marriage rate by age and sex (2004 to 2015) –mar-.
demo_mlexpec: life expectancy by age and sex (2004 to 2015) –exp-.

3. Machine Learning methodology

R software (*https://rstudio.jrc.es/*)

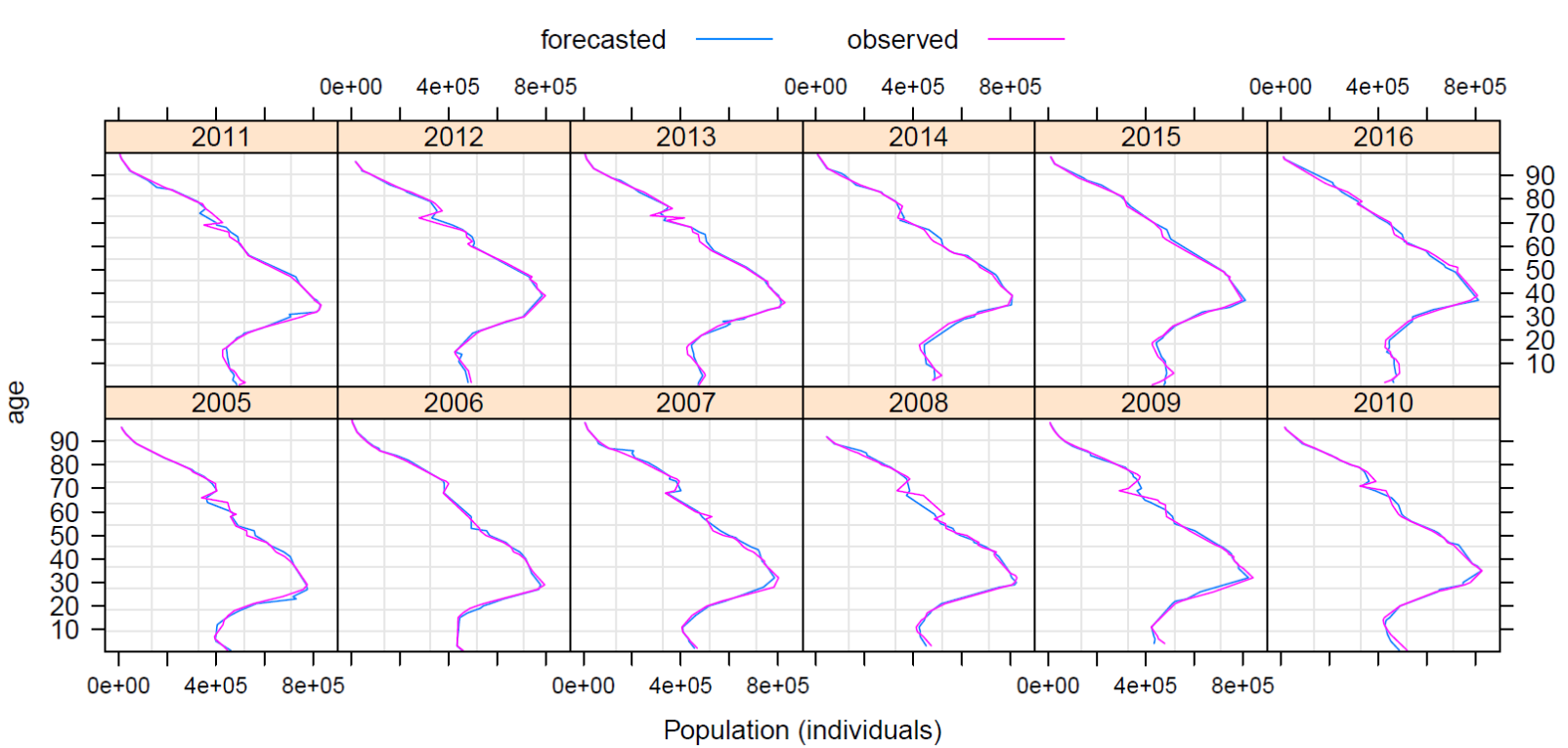
Objective: forecasting the population by age (by year of age)

Forecasting approach:

$$\hat{p}_{year} = \hat{f}(pop_{year-1}, imm_{year-1}, emi_{year-1}, fer_{year-1}, dea_{year-1}, mar_{year-1}, exp_{year-1})$$

Algorithm: random forest (Breiman, 2001).
Crossvalidation: 10 hold-out fractions, repeated 3 times

4. Results


forecasted
observed
2011
2012
2013
2014
2015
2016
2005
2006
2007
2008
2009
2010
0e+00
4e+05
8e+05
10
20
30
40
50
60
70
80
90
age
Population (individuals)

4. Results

Comparison with classical techniques:
Machine learning (ML) vs Arima & exponential smoothing (ES)

Units: population (individuals)

	ML	Arima	ES
RMSE	20318	81010	81247
MAE	14353	60596	62269

RMSE: root mean square deviation
MAE: mean absolute error

5. Conclusions

1. Good fit outperforming traditional techniques

2. Promising application of machine learning to traditional statistics

3. Small data sets feed the machine learning approach

4. Very linear relation among variables

5. Further developments will be addressed

Joint Research Centre –Seville

Álvaro Gómez Losada
alvaro.gomez-losada@ec.europa.eu

Néstor Duch-Brown
nestor.duch-brown@ec.europa.eu