Correcting for population overestimates by using statistical classification methods

Violeta Calian, Margherita Zuppardo
Statistics Iceland

Keywords: population register, census, overestimates, supervised machine learning, unsupervised machine learning.

INTRODUCTION

Register based census and population statistics may overestimate the population living in a country at a reference date. This is due to people leaving the country without being de-registered, such as short or medium term migrants who leave for a different destination or return to their home country. The effect of this over-coverage may create biases in statistical estimates of demographic or social characteristics e.g. age distributions, fertility or mortality rates, migration flows, employment and education profiles. The adjustment methods adopted by official statistics institutes in order to correct for population over estimation are neither harmonized nor well known across countries.

Statistics Iceland employed for the purpose of 2011 census estimates a logistic regression model, fitted on past de-registration data enriched with financial, social, household attributes observed at successive points in time. Statistics Estonia has developed yet another approach, based on a residency index [1] built as a function of binary variables describing education, health care, social support, employment (signs of life) measures and calibrated on training data of certain outcome. Statistics Sweden used a scoring method [2], based on tracing changes in registers concerning characteristics related to education, income, migration, civil status or residency and even measuring the impact of the over-coverage on mortality and fertility estimates.

In this paper, we propose a systematic approach to the problem, based on rigorous statistical methods of classification, which belong, in the language of machine learning, to both the domain of supervised and unsupervised learning. We start by identifying the problem of over-coverage with a classification problem, where the status of any individual may be either in or out of the country and the variables which may be employed in order to predict this status are continuous (e.g. income, taxes, time since last change in population register) or categorical (e.g. status in employment). This creates the basis for using a rich spectrum of statistical models, given available training data, but also raises new questions such as the criteria for model selection and performance evaluation or the importance and selection of variables to be used for prediction. These issues are addressed in the following section while illustrative examples are described in section 3.

METHODS

Choosing the best solution for a classification problem depends strongly on the data structure. This is why we are developing a highly automated and flexible system which consists of the following steps: (i) exploratory analysis and data pre-processing (ii) unsupervised machine learning and (iii) supervised machine learning model fitting and testing. Our R-code [3] is open source, will be shared on public repositories like *github* and relies heavily on efficient and reliable packages like *caret* (classification and regression training) [4] both these tools making the tasks enumerated above very straightforward. The main purpose is to test multiple methods on any given data set, by using a generic function with a *set of methods* as argument and performance evaluating measures as well as comparisons as output.

The data set used for training and testing was built by using information regarding presence/absence of individuals from the SILC survey (EU statistics on income and living conditions) combined with register data regarding employment status, income and taxes, education level, changes in civil and residency status, family composition, previous migration events. Most often, classifiers such as neural networks need scaling and indeed, this and recoding were the most typical pre-processing steps in this project.

The exploratory stage follows standard statistical analysis regarding uni- and multi- variate distributions, correlations, outliers, distributional differences and it is performed by employing R-packages like *dataExplorer*[1], *tidyverse*[2] and *LaplaceDemon*[3].

Using unsupervised learners such as clustering is an informative complement to both model based methods and exploratory stages. Such a machine learning method (we used the *pam* function of the *cluster*[4] R-package) partitions the data set based on dissimilarity of patterns and may improve the models built in the next stage.

The spectrum of supervised learners is vast. We are in the process of testing and adding several most robust methods of this type to our automated system. Neural networks [5], classification trees and random forests were first considered, as illustrated in the next section and more experiments are work in progress.

Evaluation of classifiers' performance is a topic on its own. We only mention here measure such sensitivity, specificity, confusion matrices, accuracy rates and their confidence intervals, meaningful hypothesis testing concerning accuracy and information rates. Most of these measures and tests, as well as predictor importance evaluation are implemented in the *caret* package.

RESULTS

In order to illustrate our method, we show here the results of steps (i)-(iii) and methods described above for a simplified data set (only foreign citizens and few attributes), containing several variables which describe the class (presence or absence in country) as depending on income (yearly salary or pension) and changes in population register at two successive points in time, age, time since arrival in Iceland.

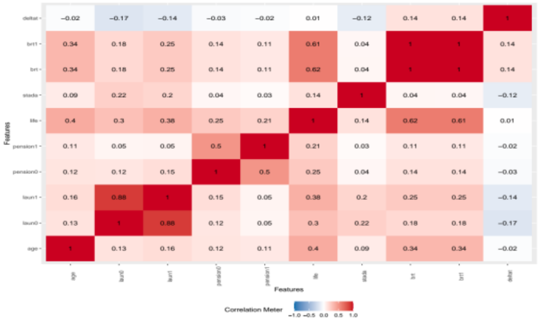
The correlation structure of this data set shows expected auto-correlations (values at time t1 and time t0 of income related variables, laun0, laun1 and pension0, pension1) or anti-correlations, such as between the absence in country and employed status or observed changes in registers, length of stay in Iceland. Cross-correlations are also confirmed, e.g. between income, status in employment and presence in country.

[1] https://cran.r-project.org/web/packages/DataExplorer/DataExplorer.pdf

[2] https://www.tidyverse.org/

[3] https://cran.r-project.org/web/packages/LaplacesDemon/index.html

[4] https://cran.r-project.org/web/packages/cluster/cluster.pdf



	age	laun0	laun1	pension0	pension1	life	stada	brt	brt1	deltat
deltat	−0.02	−0.17	−0.14	−0.03	−0.02	0.01	−0.12	0.14	0.14	1
brt1	0.34	0.18	0.25	0.14	0.11	0.61	0.04	1	1	0.14
brt	0.34	0.18	0.25	0.14	0.11	0.62	0.04	1	1	0.14
stada	0.09	0.22	0.2	0.04	0.03	0.14	1	0.04	0.04	−0.12
life	0.4	0.3	0.38	0.25	0.21	1	0.14	0.62	0.61	0.01
pension1	0.11	0.05	0.05	0.5	1	0.21	0.03	0.11	0.11	−0.02
pension0	0.12	0.12	0.15	1	0.5	0.25	0.04	0.14	0.14	−0.03
laun1	0.16	0.88	1	0.15	0.05	0.38	0.2	0.25	0.25	−0.14
laun0	0.13	1	0.88	0.12	0.05	0.3	0.22	0.18	0.18	−0.17
age	1	0.13	0.16	0.12	0.11	0.4	0.09	0.34	0.34	−0.02

Figure 1. Correlation meter for the simple dataset

The clustering algorithm partitioning around medoids (*pam*) confirms a good separation between groups which are / are not in the country, as shown in Figure 2 for the age and salary dimensions.

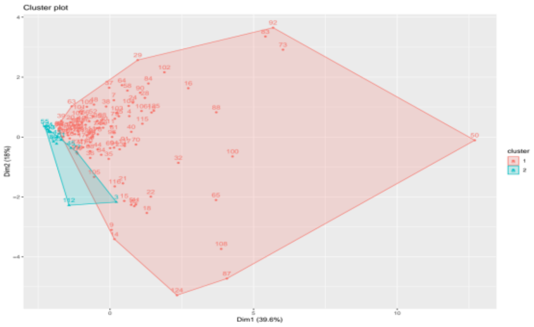


Figure 2. Result of *pam* clustering algorithm

One of the most popular supervised learning methods is the classification tree method, due to the easy interpretability of results. In Figure 3 we have a detailed confirmation of the fact that the salary, both in the most recent year (laun0) and the previous one (laun1), the time since last changes in register, the time spent in Iceland and the age, when compared with given thresholds and in a given order, may predict the presence in the country.

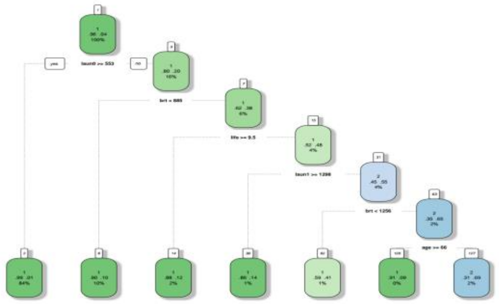


Figure 3. Classification tree diagram

The next example of a classifier is a neural network [5]. Figure 4 represents a network with two layers of five and three nodes and nonlinear output with good convergence and (cross entropy) error levels.

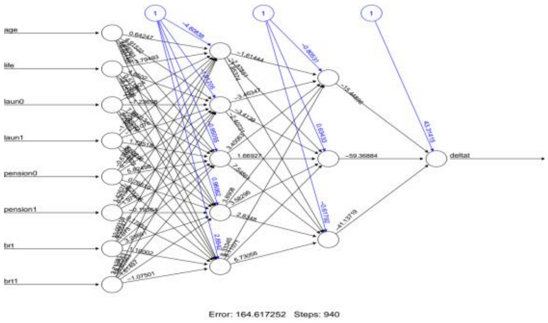


Figure 4. Neural network diagram showing trained synaptic weights

More models are being tested and a systematic and compact comparison of multiple models is under construction.

CONCLUSIONS

The goal of the present paper is to offer a solution to the population over estimation problem, based on its statistical formulation as a classification problem.

In addition, we propose a practical implementation of this solution, using multiple models and methods, systematic tests and evaluation of results, with minimum open source R-code and employing standard, trusted R-packages.

REFERENCES

[1] E. Maasing, E.-M. Tiit, M. Vahi, Residency index – a tool for measuring the population size, Acta et Commentationes Universitatis Tartuensis De Mathematica (2017), 129-139.

[2] A. Monti, S. Drefahl, E. Mussino, J. Härkönen, Over-coverage in population registers leads to bias in demographic estimates, Population Studies (2019), 1-19.

[3] R Core Team (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software (2008) 1-26.

[5] B. Venables, B. Ripley, VR: Bundle of MASS, class, nnet, spatial. R package version 7.2-42 (1999) URL http://CRAN.R-project.org/package=VR.