ESSnet Big Data

WP 2 meeting - Gdansk 5-6 October 2017

Quality issues in *e-commerce*, *online job applications* and *social media presence* predictions

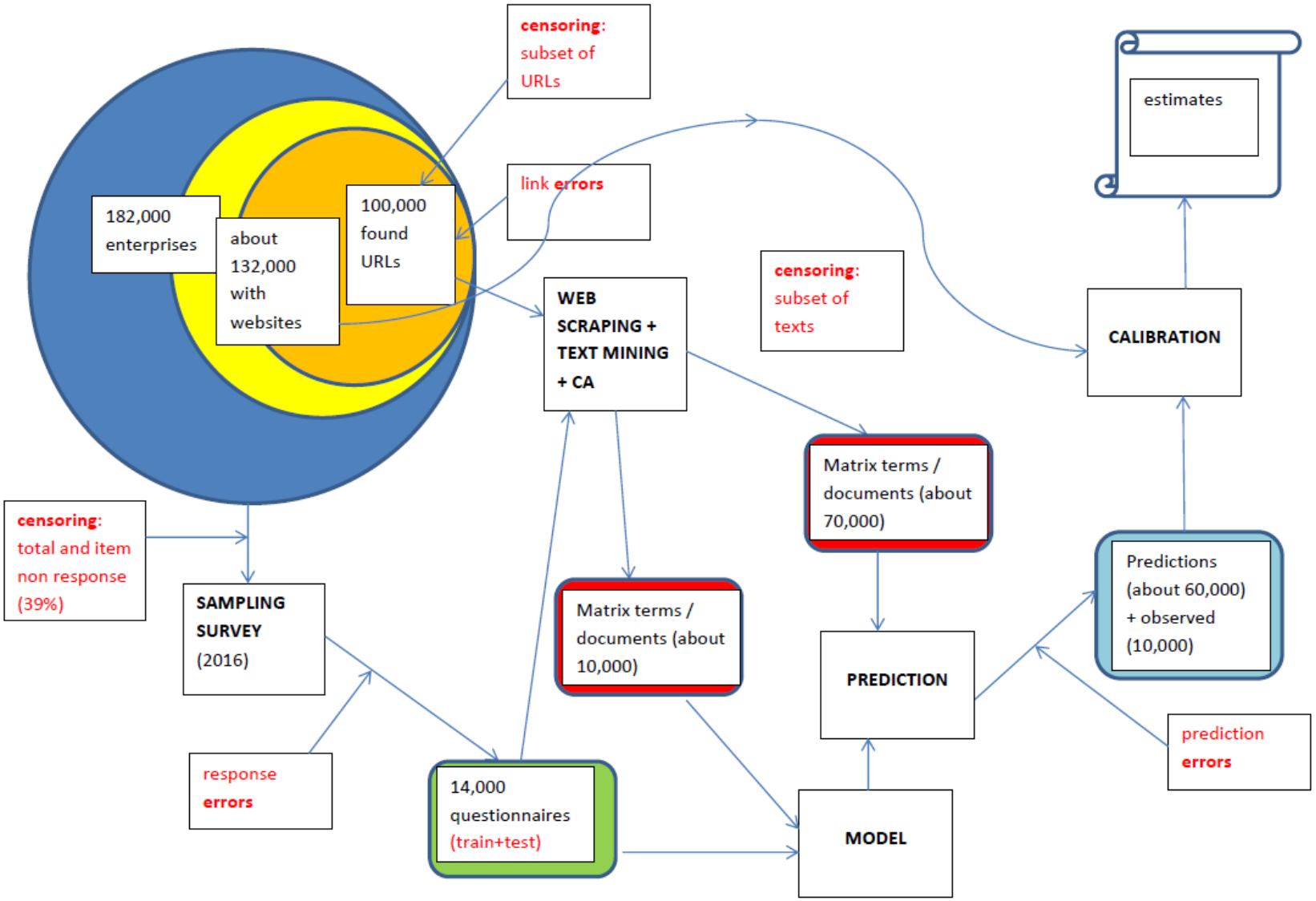
Giulio Barcaroli

Istat

Outline

- General description of Istat procedure for predicting enterprises websites characteristics at unit level
- Representativeness problem
- Production of aggregate estimates
- Comparison to current ICT survey estimates
- Preliminary evaluation of quality

General overview of the procedure



Quality of prediction at unit level (e-commerce)

Learner	Accuracy	Recall	Precision	F1-measure
Naïve Bayes	0.84	0.56	0.56	0.56
Logistic	**0.84**	**0.57**	**0.57**	**0.57**
Decision Tree	**0.87**	**0.64**	**0.64**	**0.64**
Neural Net	0.88	0.65	0.66	0.66
Bagging	**0.88**	**0.66**	**0.67**	**0.67**
SVM	0.90	0.62	0.76	0.68
Boosting	**0.90**	**0.71**	**0.71**	**0.71**
Random Forest	**0.90**	**0.73**	**0.73**	**0.73**

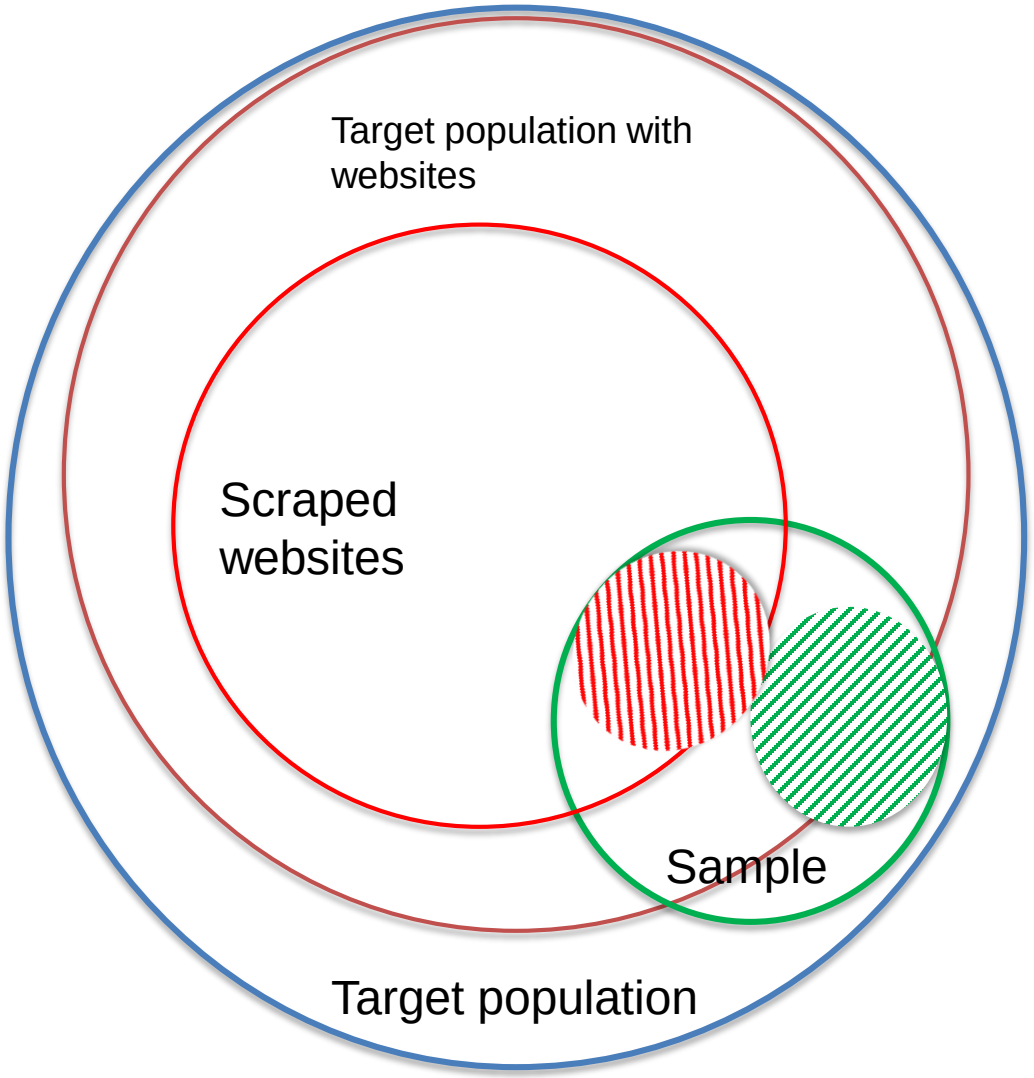
```
RANDOM FOREST
Estimate on train     0.2272634

Estimate on test      0.2268041

Confusion Matrix and Statistics

             observTest
predictedTest    0    1
            0 3288  462
            1  459  641

               Accuracy : 0.8101
                 95% CI : (0.7988, 0.8211)
    No Information Rate : 0.7726
    P-Value [Acc > NIR] : 1.109e-10

                  Kappa : 0.4591
 Mcnemar's Test P-Value : 0.9475

            Sensitivity : 0.5811
            Specificity : 0.8775
         Pos Pred Value : 0.5827
         Neg Pred Value : 0.8768
             Prevalence : 0.2274
         Detection Rate : 0.1322
   Detection Prevalence : 0.2268
      Balanced Accuracy : 0.7293

       'Positive' Class : 1

F1 measure          0.5819337
```

with response errors

```
RANDOM FOREST
Estimate on train     0.1817417

Estimate on test      0.1816653

Confusion Matrix and Statistics

             observTest
predictedTest    0    1
            0 1830  116
            1  116  316

               Accuracy : 0.9024
                 95% CI : (0.8898, 0.9141)
    No Information Rate : 0.8183
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6719
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7315
            Specificity : 0.9404
         Pos Pred Value : 0.7315
         Neg Pred Value : 0.9404
             Prevalence : 0.1817
         Detection Rate : 0.1329
   Detection Prevalence : 0.1817
      Balanced Accuracy : 0.8359

       'Positive' Class : 1

F1 measure          0.7314815
```

with no response error

Representativeness problem



- Target population: 182,000
- Target population with websites: 132,000
- Sample: 32,000
- Sample non-respondents: 13,000 (Respondents 19,000)
- Scraped websites (Big Data): 68,000
- Respondents with scraped websites: 10,000

(i) Adjustment for total non response in sample data

Models are fitted on the sample data.

Are these data representative of the whole population of interest?

Total non response is the result of a non ignorable process.

Models should be fitted taking into account also variables that can explain total non response, by:

- considering the same calibrated weights used in the current estimation procedure (in parametric models i.e. the logistic model);
- considering also those variables as explanatory variables (in non parametric models).

(ii) Adjustment for undercoverage of Big Data population

Actually, we are able to reach only a subset of the population of enterprises having a website: U' $\subset$ U (*undercoverage of target population by using Big Data*).

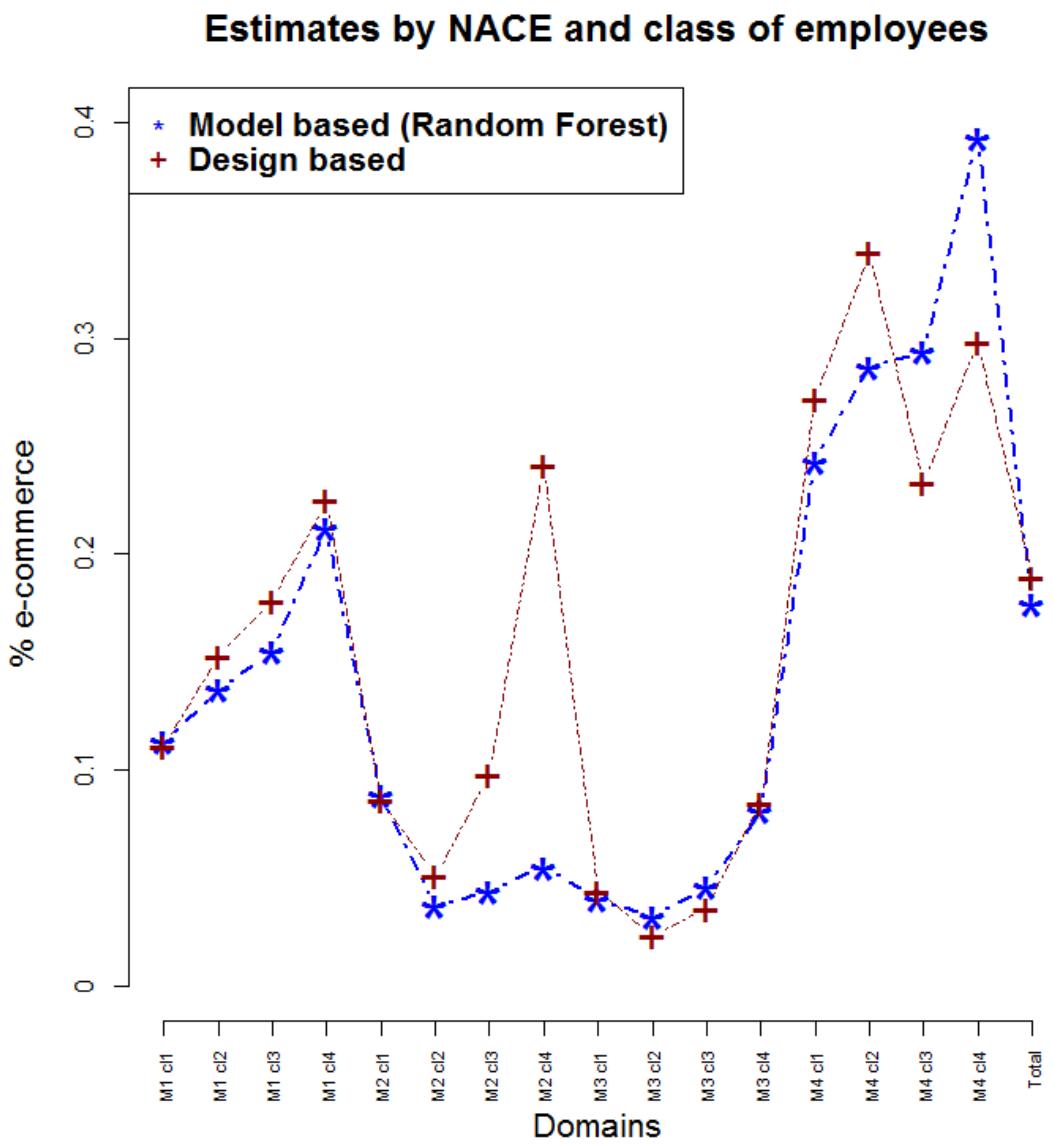
We can use calibration in order to let Big Data population represent the whole population of interest.

Therefore, the estimates are obtained by calibrating, using the same auxiliary variables used to calibrate sample observations (number of firms and the number of employees, according to the information contained in the Italian Business Register ASIA).

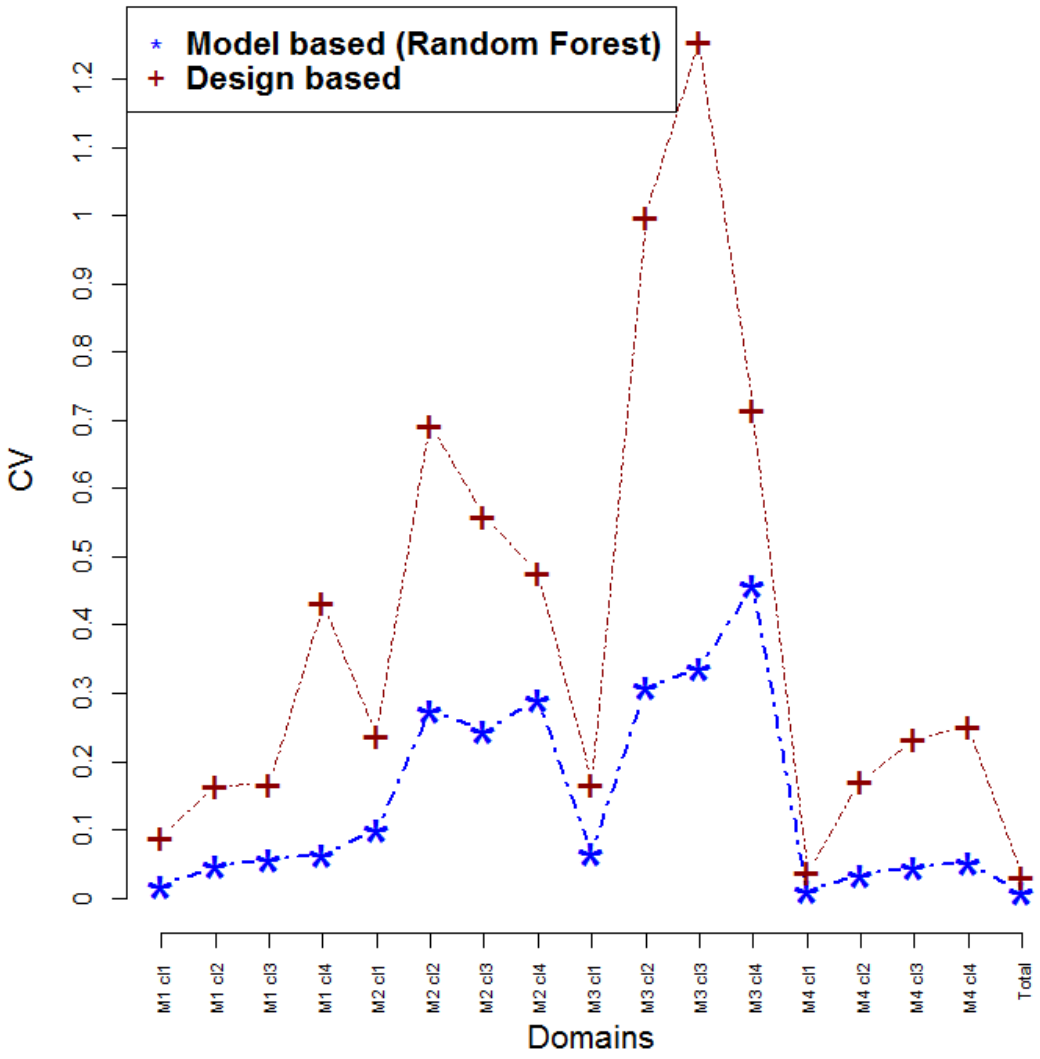
A comparison of estimates

For the 2016 round of the survey, a subset of related estimates have been compared to those obtained by the alternative model-based approach.

This subset includes the proportion of enterprises offering e-commerce in their websites, cross-classified by economic activity (NACE 1 digit, 4 values) and size (4 classes of number of employees.


Estimates by NACE and class of employees
Model based (Random Forest)
Design based
% e-commerce
0
0.1
0.2
0.3
0.4
M1 cl1
M1 cl2
M1 cl3
M1 cl4
M2 cl1
M2 cl2
M2 cl3
M2 cl4
M3 cl1
M3 cl2
M3 cl3
M3 cl4
M4 cl1
M4 cl2
M4 cl3
M4 cl4
Total
Domains

Coefficients of variation



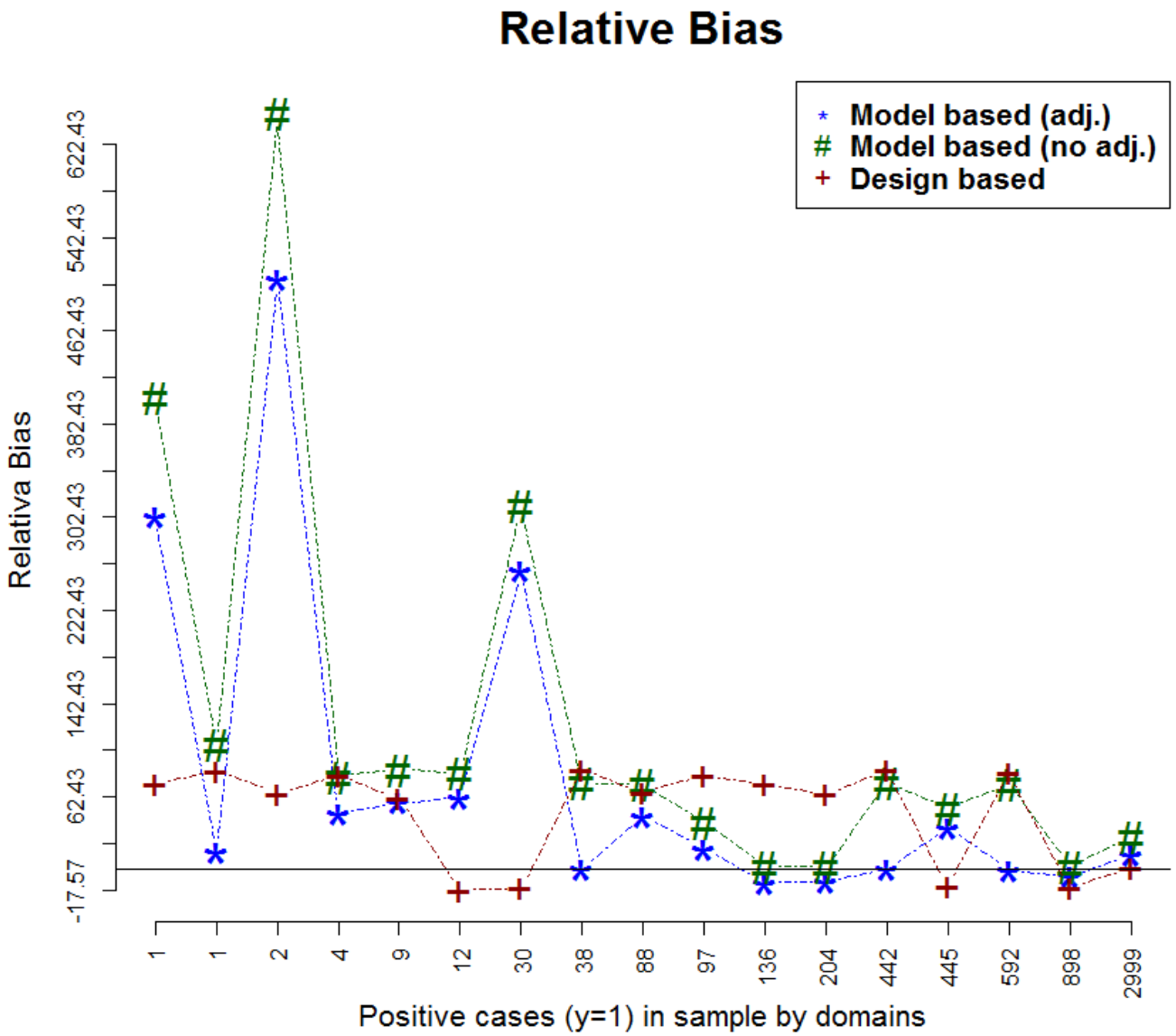
Simulation study

A **simulation study** has been conducted, on a very realistic basis: all the conditions related to the sampling survey and on the availability of Internet data have been re-created.
In particular, varying prediction performance of models have been considered, from the current one to an optimal one.

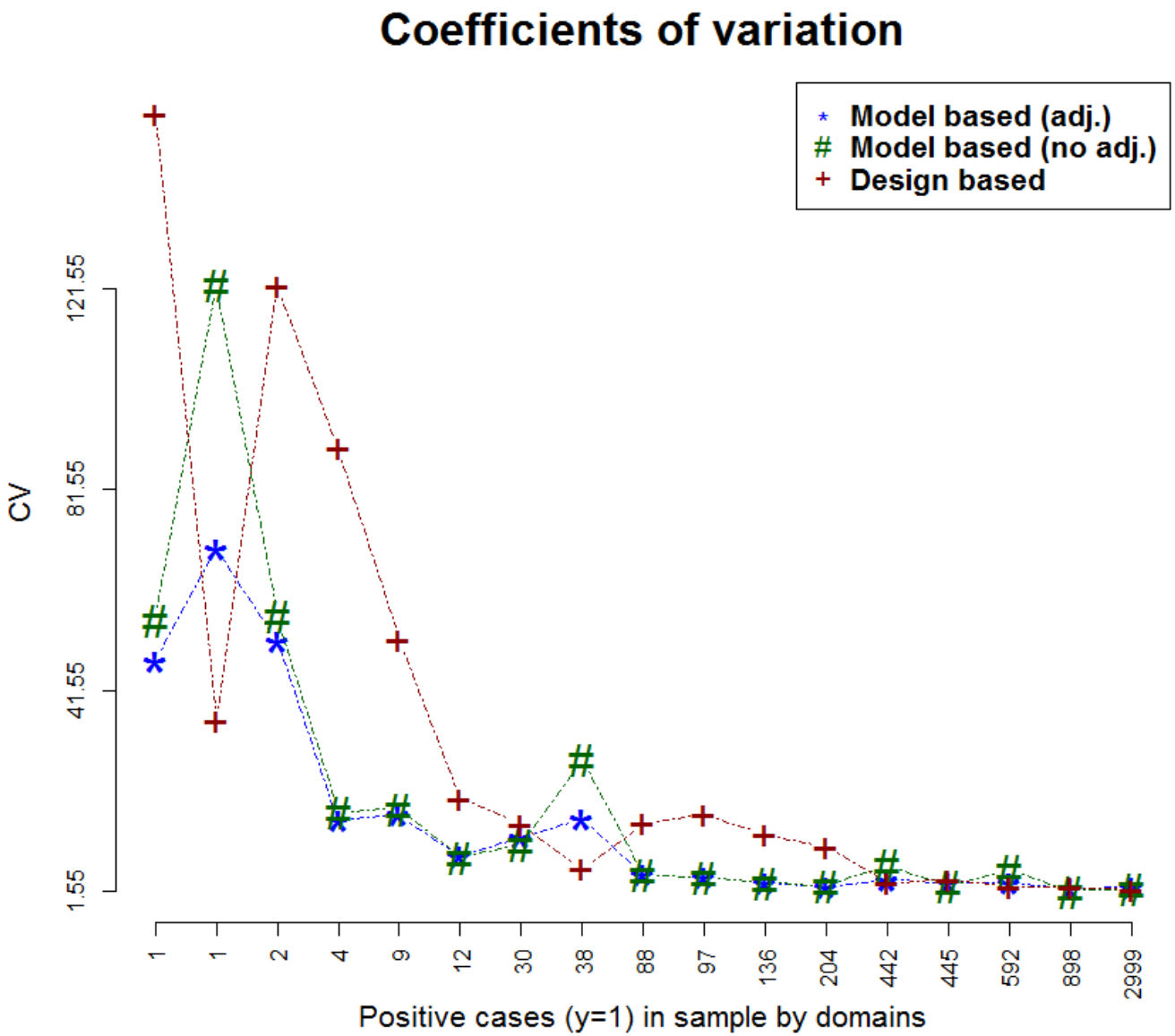
A **comparison between the quality of estimates** currently obtained by the **statistical survey**, and the ones obtainable by the use of **Internet data** has been carried out, based on the calculation of:

1. bias
2. variance
3. Mean Squared Error ($bias^2$+variance)

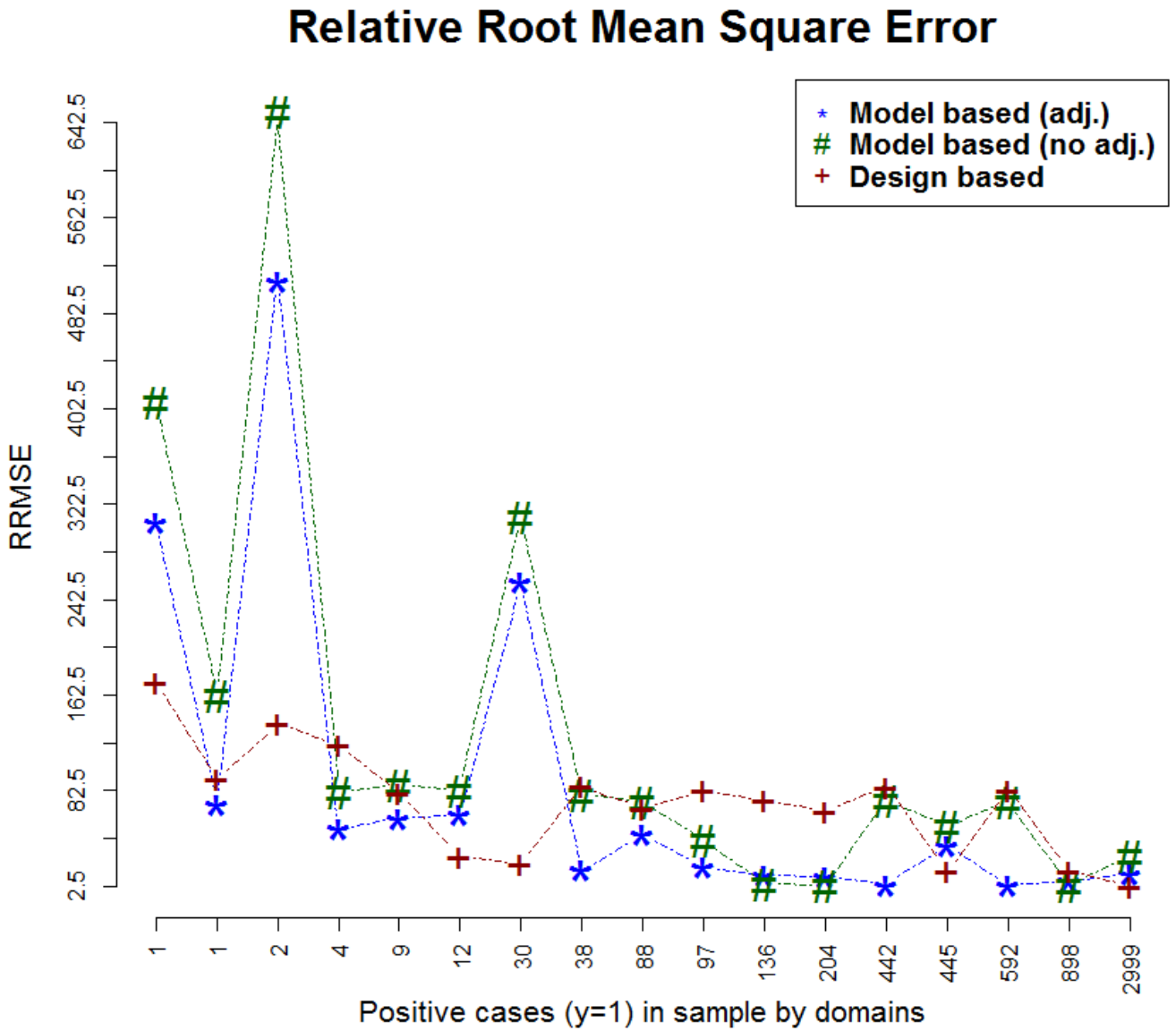
Simulation study



Simulation study



Simulation study



Simulation study

The simulation has shown that:

1. Comparing adjusted and non adjusted model based estimates, in terms of bias and MSE the former are always better than the latter
2. Comparing model based and design based estimates
   a. in terms of bias and variance, model based are often better than design based
   b. in terms of MSE, the adjusted model based estimates are better than design based with few exceptions.

These conclusions are true only *with an acceptable prediction performance of models (F1-measure)*.

Conclusions

The procedure based on the sequential application of web scraping / text mining / machine learning techniques allows to

- predict values at unit level
- produce estimates of aggregates

In both cases, the evaluation of the quality of the obtained results is necessary.

At unit level, whenever survey data and/or other sources data are available, the evaluation of the quality of the predictions is straightforward.

At aggregate level it is more complex, and requires the application of simulation techniques if we want to evaluate the different components of the mean square error.

GUS
SURS SSB
INE INSEE BNSI
CBS DESTATIS
INS ELSTAT SCB
STAT ISTAT
ONS

Thank you for your attention!

barcarol@istat.it