Use of dedicated business websites to enhance the statistical business register in the Netherlands

Sharing experiences

Arnout van Delden, Nick de Wolf, Naomi Schalken, Sander Scholtus, Olav ten Bosch and others; Feb 4-5 2025, Gdansk

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Icon from www.freepik.com; by zero_wing

Web Intelligence Network

Funded by the European Union

Introduction

Automatic use of information on websites to reduce manual labour for maintenance variables in a SBR (units, contact information, NACE)

Experiences by Statistics Netherlands:

1. Finding of URLs using data from an external company
2. Development of a model to predict NACE misclassifications

URL finding

Source	Population	Frequency	Linkage
Chamber of commerce	Registration of (new) legal units	Continuous	Legal unit ID number
ICT survey	Sample of enterprises	Yearly	Enterprise ID number
DataProvider	Dutch websites that are not blocked	Monthly	ID numbers, name, email address and so on

Third party DataProvider (DP) scrapes URLs (and contact information) in many countries and makes a selection of Dutch businesses

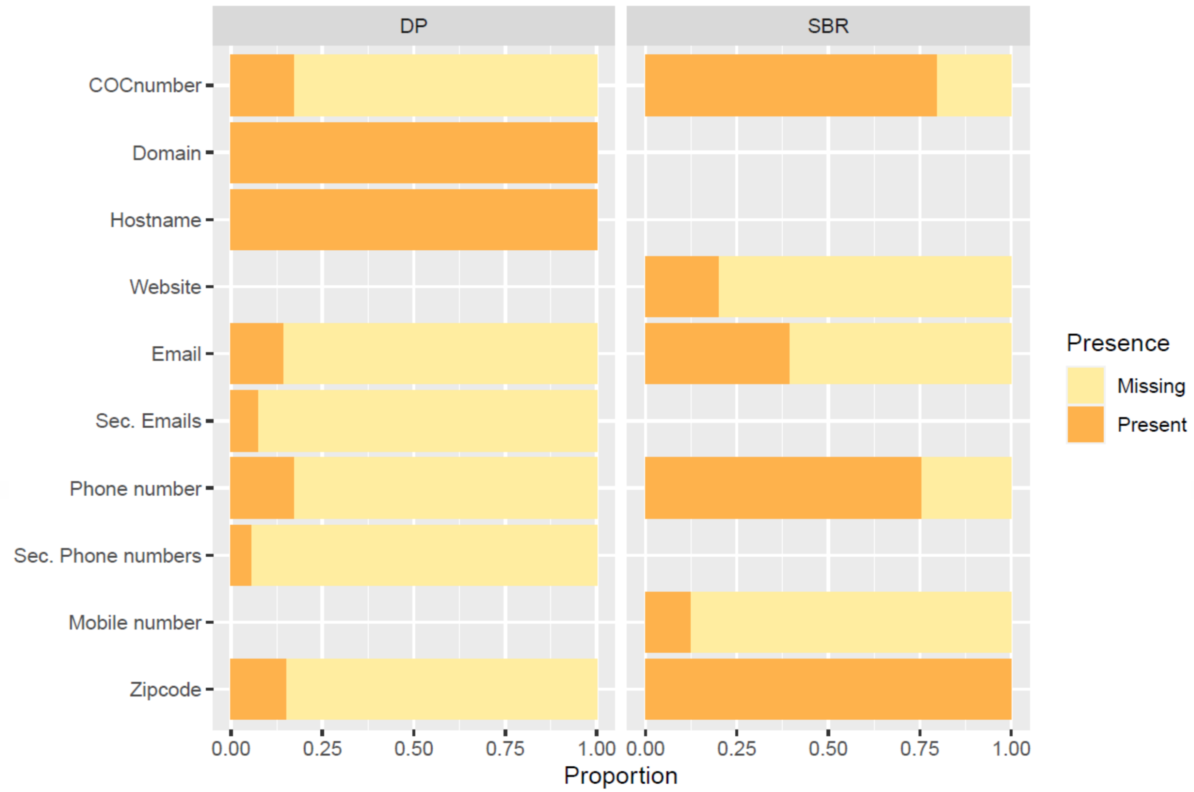
URLs collected by third parties are a potentially useful source for NSI's, but

- The collected URLs need to be linked to legal / statistical units in the SBR
- values of identifying variables need to be present in both sources

Web Intelligence Network

Funded by the European Union

URL finding: linkage of DP



URL finding: contribution of COC versus DP

Number of Legal Units in the SBR (Oct 2020)

Groups	URL from COC	URL from DP	DP URL-LU linkage probability						
			> 0%	≥ 10-50%	≥ 65%	≥ 75%	≥ 85	≥ 95%	100%
Total			4 630 836	4 630 836	4 630 836	4 630 836	4 630 836	4 630 836	4 630 836
Group A	+	+	700 973	670 528	656 672	644 217	635 936	424 151	389 165
Group B	-	+	671 011	213 781	123 765	29 265	1 109	1	1
Group C	+	-	221 605	252 050	265 906	278 361	286 642	498 427	533 413
Group D	-	-	3 037 247	3 494 477	3 584 493	3 678 993	3 707 149	3 708 257	3 708 257

Web Intelligence Network

Funded by the European Union

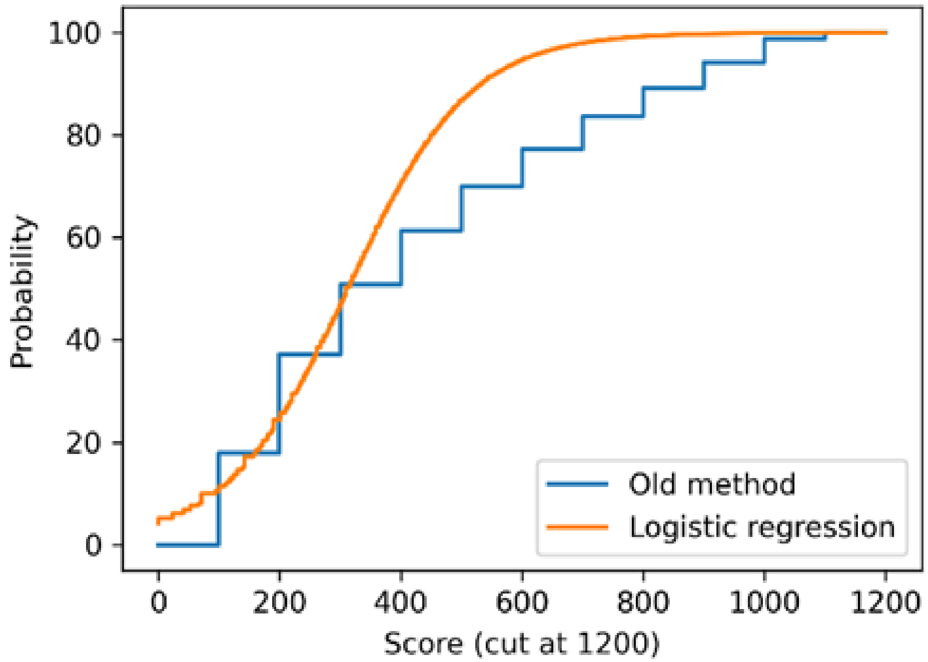
With websites scraped by third-parties:

- considerable effort is needed to build and maintain a probabilistic linkage function to link non-unique identifiers, or …
- limit the linkage to unique identifiers and accept fewer linkages

Web Intelligence Network

Funded by the European Union

URL finding linkage of DP



P_{LK} : linkage probability of pair {LU, URL} *k,* X_{kj}: agreement on variable *j of* pair *k:*,
score $S_k = \boldsymbol{X}_k\beta = \beta_0 + \beta_1 X_{k1} + \; ... + \beta_J X_{kJ}$

1 Old method: $P_{LK} = 47.5\ln(S_K) - 234$, with S_K by expert knowledge (sample 7 × 20 units)
2 Fitted LR model: logit(P_{LK}) = S_K (sample 4 × 400 units)
3 Updated LR model (more linkage variables, slightly more links)

URLs and legal units/statistics units are different unit types:

- 1:1, 1:n, n:1 and m:n linkages
- How to cope with multiple URLs per legal unit: primary/secondary URLs?

Web Intelligence Network

Funded by the European Union

Linkage between Legal Units (SBR) and URLs (oct 2020)

<table>
<tr><th></th><th></th><th># URLs</th><th></th><th></th></tr>
<tr><td># LUs</td><td>2+ (n)</td><td>1</td><td>0</td><td>Total</td></tr>
<tr><td>2+ (m)</td><td>4863</td><td>27935</td><td rowspan="2">3957354</td><td rowspan="2">4630836</td></tr>
<tr><td>1</td><td>111904</td><td>528780</td></tr>
<tr><td>0</td><td colspan="2">5057922</td><td>X</td><td>X</td></tr>
</table>

Finding NACE misclassifications

Sources of (mainly textual) information:

- scraped website texts (main page, about us page, … up to 10 pages)
- activity descriptions of registered legal units (chamber of commerce)
- Textual descriptions of establishments where the units are located

NACE Section R as a case study example, because

- was manually checked 2021-2022
- prone to misclassifications
- ‘challenge’: number of codes difficult to predict

With respect to classical ML models:

- The more knowledge-based the features are the better the performance of NACE predictions
- *adding* the knowledge-based features to a standard feature set without such selections slightly improves the performance

24 5-digit NACE codes in Section R

Algorithm		all	YAKE	IGFSS	D-words	all+YAKE	all + IGFSS	all + D-words
	number features	500	210	237	500	710	737	1000
SVM	macro avg	0,756	0.646	0.730	0.861	0.752	0.802	0.861
	weighted avg	0.755	0.649	0.732	0.862	0.753	0.802	0.862
	st.dev							
NB	macro avg	0.757	0.605	0.716	0.829	0.746	0.788	0.836
	weighted avg	0.756	0.608	0.716	0.830	0.745	0.786	0.837
	st.dev							

All: no selection

YAKE: general keyword selection (no NACE information)

IGFSS: select words pos. / neg. related to the predicted NACE code

D-words: descriptive words used by manual editors

Web Intelligence Network

Funded by the European Union

The quality of the ML predictions depend on:

- the quality of the available text data,
- the heterogeneity of the code

Lowest F1-scores

NACE code	Label	Size	F1-score (100, ≥75%)
90012	Production of live theatrical presentations, concerts, opera, dance and other productions	522	**0.08**
93299	Other recreation (no marina)	3602	**0.17**
90020	Services for performing arts	5102	**0.20**
90030	Writing and other creative arts	18895	**0.21**
90011	Performance of stage art	10291	**0.22**

Highest F1-scores

NACE code	Label	Size	F1-score (100, ≥75%)
93291	Marinas	251	**0.81**
91021	Museums	280	**0.78**
86912	Practice of physiotherapists	5206	**0.77**
96022	Beauty care, pedicures and manicures	13903	**0.74**
93121	Field football	135	**0.73**

Web Intelligence Network

Funded by the European Union

Model to capture just which codes in SBR are likely to be incorrect gives promising results, but

- Sensitive to unequal population sizes per code

24 5-digit NACE codes in Section R

	Full set	Test set	Test set	Test set	Test set
setting	Estimated prop. errors	True prop. errors	Estimated prop. errors	TPR	TNR
All data	0.064	0.190	0.294	0.284	0.703
Max 1000	0.067	0.190	0.143	0.459	0.941
Max 1000, suppl	0.065	0.199	0.138	0.433	0.935

TPR: units that are in reality misclassified that are identified as being misclassified

TNR: units that are in reality correct that are identified as correct

Thanks for your attention.

Let us exchange experiences among us.