Internet as a Data Source : ICT use of enterprises: web ordering, job advertising and presence on social media

Objective

A multi-source approach (based on a combined use of survey, administrative and BD sources) should allow to overcome usual limits of each single source, in particular those affecting Big Data.

This multi-source approach requires a shift in the paradigm of statistical inference. The traditional one followed by NSOs is usually based on design-based survey sampling theory and model-assisted inference. The new one (algorithmic-based inference) is derived by data science: the emphasis is on the exploration of all available data, seeking information that has not been extracted so far; models have to be evaluated no longer by their interpretability, but rather by their capability to correctly predict values at unit level, and to use them for estimating the parameters of interest.

Istat has experimented this new approach in order to obtain a subset of the estimates currently produced by the sampling "Survey on ICT usage and e-Commerce in Enterprises", yearly carried out by Istat and by the other member states in the EU. Target estimates of this survey include the characteristics of websites used by enterprises to present their business (for instance, if the website offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data are collected by means of traditional questionnaires.

An alternative way is to make use of Internet data, i.e. to collect data by accessing directly the websites, processing the collected texts to individuate relevant terms, and modelling the relationships between these terms and the characteristics we are interested to estimate. To do that, the sample of surveyed data plays the role of a training set useful to fit models that can be applied to the generality of enterprises owning a website. Administrative data (mainly contained in the Business Register) are used to cope with representativeness problems related to BD source. The sequential application of web scraping, text mining and machine learning techniques allows to obtain auxiliary variables suitable for applying a prediction approach and produce estimate that can be compared to the survey ones.

In terms of quality (accuracy), the impact of the new estimators is both positive (reduction of the variability and of the bias due to sampling variance, to total non-response and to measurement errors in the survey) and negative (model bias and variance). Whenever the quality of estimates obtained by means of this new approach reveals to be not lower than the ones produced by the traditional process, the former has to be preferred, as it allows not only to produce aggregate estimates, but also to predict individual values, useful for instance to enrich the information contained in registers.

Results achieved

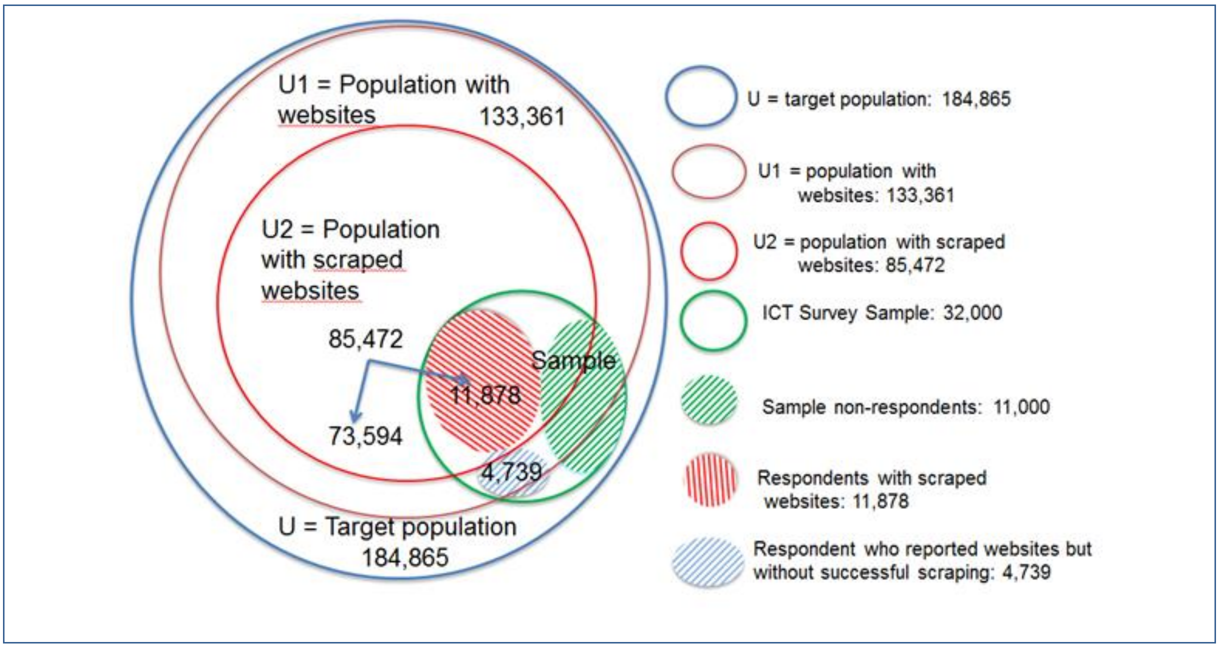
A complex procedure has been developed in order to:

1. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (URL retrieval);
2. access websites with available URL and scrape their content (web scraping);
3. process the content of the scraped websites in order to identify the best predictors for the target variables (text mining);

4. fit models (machine learning) in the subset of enterprises where both Internet data and survey data were available (considering survey data as the true values) and predict the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful.

The following Figure 1 reports the different subsets of the population of interest (enterprises with at least 10 persons employed operating in various economic activities of manufacture and non-financial services), involved in the overall procedure:

Figure 1 Subsets of the population of interest


U1 = Population with websites 133,361
U2 = Population with scraped websites
85,472
73,594
Sample
11,878
4,739
U = Target population 184,865
U = target population: 184,865
U1 = population with websites: 133,361
U2 = population with scraped websites: 85,472
ICT Survey Sample: 32,000
Sample non-respondents: 11,000
Respondents with scraped websites: 11,878
Respondent who reported websites but without successful scraping: 4,739

The "Survey on ICT usage and e-Commerce in Enterprises" produce on a yearly basis a set of estimates reporting rates of web-ordering, job advertising and presence on social media declared by enterprises that own or make use of websites. In particular enterprises are asked to answer to filter question about having own web site of Internet page. This filter question does not refer specifically to the ownership of the website, but to the use of a website by the enterprise to present its 'business'. It includes not only the existence of a website which is located on servers belonging to the enterprise or located at one of the enterprise's sites, but also third party websites (e.g. one of the group of enterprises to which it belongs i.e. website of the parent company or holding company). However, it does not include any presence of the enterprise on the web (for example the presence of the enterprise with e.g. its name or its contact information in online yellow pages are not included in this variable). Moreover enterprises on e-marketplaces where they have the possibility to advertise themselves, quote prices for ad hoc services etc. are not enterprises that are considered to have a website.

These estimates are available for the total population, and for different domains of interest, among which:

1. Cross-classification by Size Classes of persons employed (4) and Economic macro sectors (4) (16 different sub-domains);
2. Administrative Regions (21 different domains);
3. Detailed economic activities (26 domains).

Together with the current estimation method (*design based / model assisted*), alternative estimates have been calculated by adopting two different estimators: a *full model based* one and a *combined* one. The characteristics of the three different estimators are reported in the following table.

Table 1 **Estimators**

Estimator	Formula	Weighting	Description
Design based / model assisted	$\hat{Y} = \sum_{r} y_k w_k$	$\sum_{k=1}^{r} w_k = N_U$	w_k weights are obtained by calibration procedure of basic weights (inverse of inclusion probabilities) making use of known totals in the population in order to reduce the bias due to non-response and the variability due to sampling errors
Model based	$\hat{Y} = \sum_{U^2} \tilde{y}_k w'_k$	$\sum_{k=1}^{U^2} w'_k = N_{U^1}$	The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values $\tilde{y}_k$ for all units for which it was possible reach their websites (population U^2), calibrated in order to make them representative of all the population having websites (U^1).
Combined	$\hat{Y} = \sum_{U^2} \tilde{y}_k + \sum_{r^1} (\tilde{y}_k - y_k) w''_k + \sum_{r^2} y_k w'''_k$	$\sum_{k=1}^{r^1} w''_k = N_{U^2}$ and $\sum_{k=1}^{r^2} w'''_k = N_{U^1-U^2}$	Estimates are produced by summing three components: 1. the counting of predicted values in the subpopulation U^2 of units for which it was possible to scrape and process corresponding websites; 2. an adjustment based on the consideration of the differences between the r^1 reported values and the predicted values (expanded to the same subpopulation U^2); 3. the counting of observed values for the r^2 respondents that declared a website, that was not found nor scraped, expanded to the whole subpopulation $U^1 - U^2$.

Once computed, the 3 different sets of estimates can be compared. For instance, considering web-ordering the results are reported in Table 2. The first column indicates the domain for which the estimates are calculated. The absolute values of sample units, population, and websites offering web-ordering facilities are listed. Current design-based estimates together with lower and upper limits of corresponding confidence interval are reported. Finally, model based and combined estimates are shown (highlighted in red when they lay outside the design based confidence intervals).

Table 2 Web-ordering estimates comparison

DOMAIN		Design based estimate	Lower limit C.I.	Upper limit C.I.	Model based estimate	Combined estimate
Size class of persons employed						
cl1	from 10 to 49	14.57	13.32	15.83	15.22	13.8
cl2	from 50 to 99	15.96	13.83	18.08	16.23	15.1
cl3	from 100 to 249	17.91	16.04	19.78	17.71	17.38
cl4	from 250 and more	25.72	23.78	27.65	23.25	26.04
Economic macro sectors and size classes						
M1cl1	Manufacturing (C) 10-49	10.04	8.08	11.99	11.06	9.88
M1cl2	Manufacturing (C) 50-99	12.09	8.87	15.3	14.8	14.29
M1cl3	Manufacturing (C) 100-249	15.69	12.6	18.77	15.76	15.38
M1cl4	Manufacturing (C) 250+	24.18	21.06	27.3	22.65	21.09
M2cl1	Energy (D,E) 10-49	8.69	6.54	10.84	9.73	11.51
M2cl2	Energy (D,E) 50-99	10.5	5.98	15.03	11.55	9.73
M2cl3	Energy (D,E) 100-249	13.89	8.95	18.84	15.04	11.79
M2cl4	Energy (D,E) 250+	18.79	11.86	25.72	16.97	14.55
M3cl1	Construction (F) 10-49	2.92	2.03	3.81	5.54	5.02
M3cl2	Construction (F) 50-99	3.1	0.29	5.91	5.32	4.28
M3cl3	Construction (F) 100-249	2.05	0.3	3.81	5.19	5.19
M3cl4	Construction (F) 250+	8.12	1.09	15.16	10	8.75
M4cl1	Non-financial services 10-49	20.28	18.26	22.3	20.26	18.4
M4cl2	Non-financial services 50-99	21.76	18.36	25.16	19.36	17.68
M4cl3	Non-financial services 100-249	21.76	19.03	24.48	20.89	20.82
M4cl4	Non-financial services 250+	28.32	25.56	31.07	24.85	31.51
Nace economic activities						
naceict0	activities not included in ICT Sector (defined in terms of NACE as 261, 262, 263, 264, 268, 465, 582, 61, 62, 631, 951)	15.13	13.94	16.31	15.54	14.25
naceict1	activities included in ICT Sector	10.97	8.17	13.77	14.88	13.65
naceist01	manufacture of food products, beverages and tobacco products	19.4	12.86	25.94	17.04	14.82
naceist02	manufacture of textiles, apparel, leather and related products	16.05	9.2	22.91	13.85	11.93
naceist03	manufacture of wood and paper products, and printing	12.45	6.55	18.36	13.21	11.3
naceist04	manufacture of coke and refined petroleum products, of chemicals and chemical products, of basic pharmaceutical products and preparations, of rubber, plastic and of other non-metallic mineral products	10.44	6.85	14.02	11.78	11.73
naceist05	manufacture of basic metals and fabricated metal products, except machinery and equipment	5.94	3.02	8.85	7.65	7.25
naceist06	manufacture of computer, electronic and optical products	9.47	4.94	13.99	11.98	9.73
naceist07	manufacture of electrical equipment and of machinery and equipment n.e.c.	5.62	2.86	8.38	10.45	8.88
naceist08	manufacture of transport equipment	16.68	3.04	30.32	12.49	14.72
naceist09	manufacture of furniture, other manufacturing, and repair and installation of machinery and equipment	8.84	4.57	13.11	11.79	11.27
naceist10	electricity, gas steam, air conditioning supply, water supply, sewerage, waste management and remediation activities (d-e)	9.87	8.03	11.71	10.77	11.5
naceist11	construction	2.94	2.07	3.81	5.54	5
naceist12g	wholesale and retail trade and repair of motor vehicles and motorcycles	20.39	18.98	21.81	20.32	20.28
naceist15	transport and storage, except warehousing and support activities for transportation (h except 53)	14.16	6.57	21.75	11.47	10.9
naceist16	postal and courier activities	26.13	16.37	35.89	14.16	18.26
naceist17	accommodation	82.57	77.37	87.78	71.77	68.71
naceist18	food service activities	23.63	14.59	32.67	22.23	15.48

Table 2 *(continued)* Web-ordering estimates comparison

DOMAIN	Design	Lower	Upper	Model	Combined

		based	limit	limit	based	estimate
		estimate	C.I.	C.I.	estimate	
Nace economic activities						
naceist19	publishing activities	62	45.62	78.39	49.21	49.44
naceist20	motion picture, video and television programme production, sound recording	15.62	2.97	28.27	23.63	17.64
naceist21	telecommunications	21.45	13.71	29.19	20.44	20.8
naceist22	IT and other information services	8.82	5.65	11.99	12.89	12.45
naceist23	real estate activities	11.08	5.78	16.39	13.68	13.99
naceist24	professional, scientific and technical activities except veterinary activities	5.27	1.57	8.97	10.33	7.5
naceist25	administrative and support service activities except travel agency, tour operator and other reservation service and related activities (N except 79)	4.83	2.93	6.73	8.4	7.33
naceist26	travel agency, tour operator and other reservation service and related activities	44.2	31.8	56.59	44.19	47.71
Administrative Regions						
REG01	PIEMONTE	11.96	7.46	16.46	13.77	13.6
REG02	VALLE D'AOSTA	16.8	6.43	27.17	21.77	20.76
REG03	LOMBARDIA	11.76	10.42	13.1	14.38	13.07
REG05	VENETO	14.72	12.22	17.22	16.67	15.8
REG06	FRIULI-VENEZIA GIULIA	17.17	5.62	28.73	14.23	14.67
REG07	LIGURIA	11.39	5.96	16.83	14.86	12.02
REG08	EMILIA-ROMAGNA	12.63	9.89	15.36	15	14.9
REG09	TOSCANA	14.55	10.3	18.8	15.91	14.35
REG10	UMBRIA	24.23	20.35	28.1	16.34	15.43
REG11	MARCHE	20.37	7.51	33.23	16.58	14.04
REG12	LAZIO	16.62	12.47	20.77	16.02	13.79
REG13	ABRUZZO	17.41	9.08	25.74	13.87	14.23
REG14	MOLISE	14.06	4.08	24.03	12.41	15.17
REG15	CAMPANIA	15.87	10.82	20.91	14.4	14.33
REG16	PUGLIA	20.32	14.46	26.18	14.61	12.21
REG17	BASILICATA	12.02	4.34	19.7	13.78	8.34
REG18	CALABRIA	20.4	10.93	29.87	17.47	10.05
REG19	SICILIA	19.17	6.95	31.4	16.7	12.56
REG20	SARDEGNA	14	7.85	20.14	14.93	15.29
REG21	Provincia Autonoma Bolzano	31.43	24.93	37.92	29.38	26.64
REG22	Provincia Autonoma Trento	19.51	16.87	22.14	22.78	23.21
Total		14.97	13.81	16.13	15.51	14.22

For web-ordering estimates a graphical comparison is shown in Figure 2. The dashed lines define the area delimitated by the lower and upper limits of the confidence intervals calculated in correspondence of each design based estimate.

The same distributions are reported also for Job Advertisements (Figure 3) and Presence in Social Media (Figure 4).

Figure 2 Web-ordering estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)

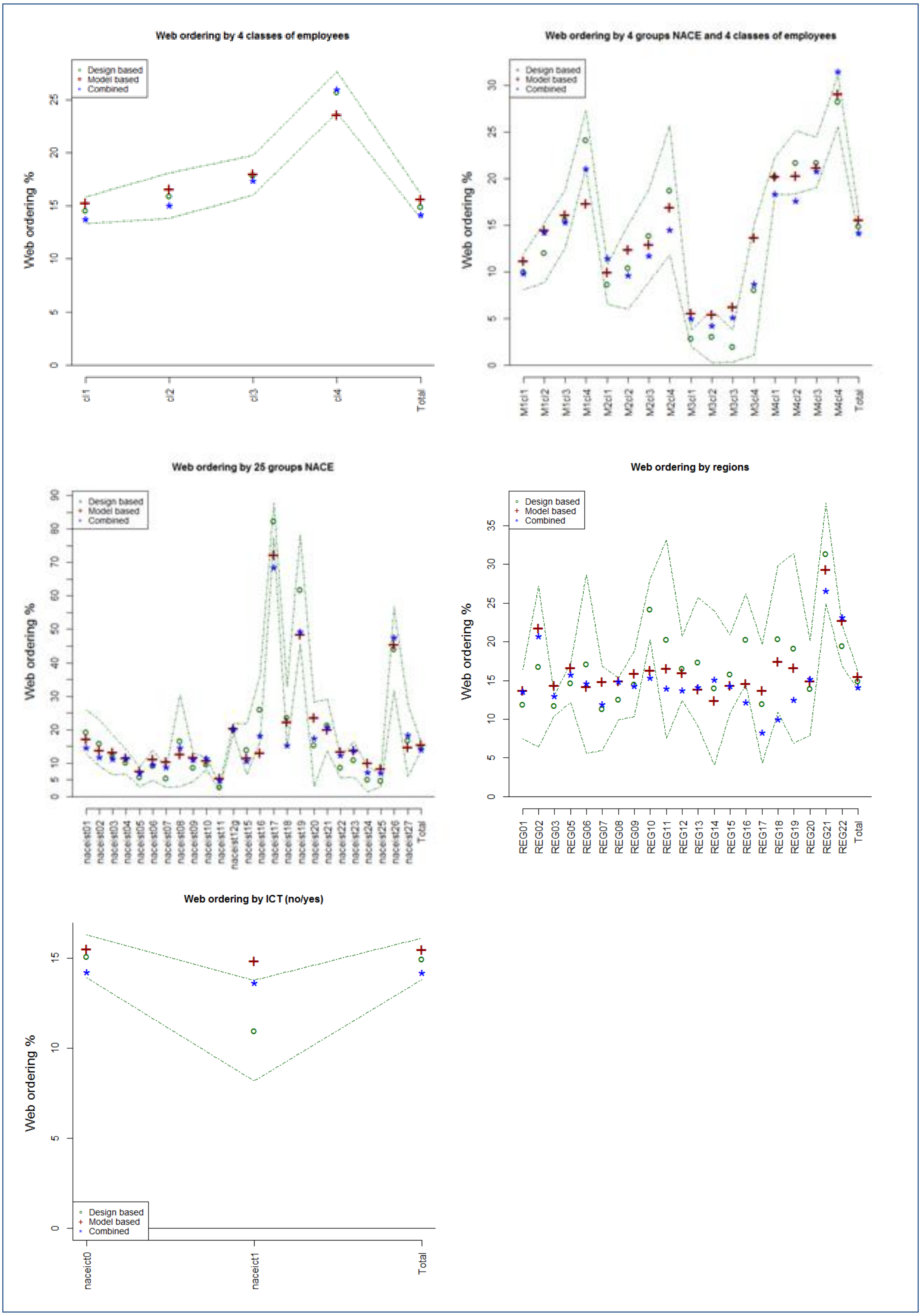


Figure 3 **Job advertisements estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)**

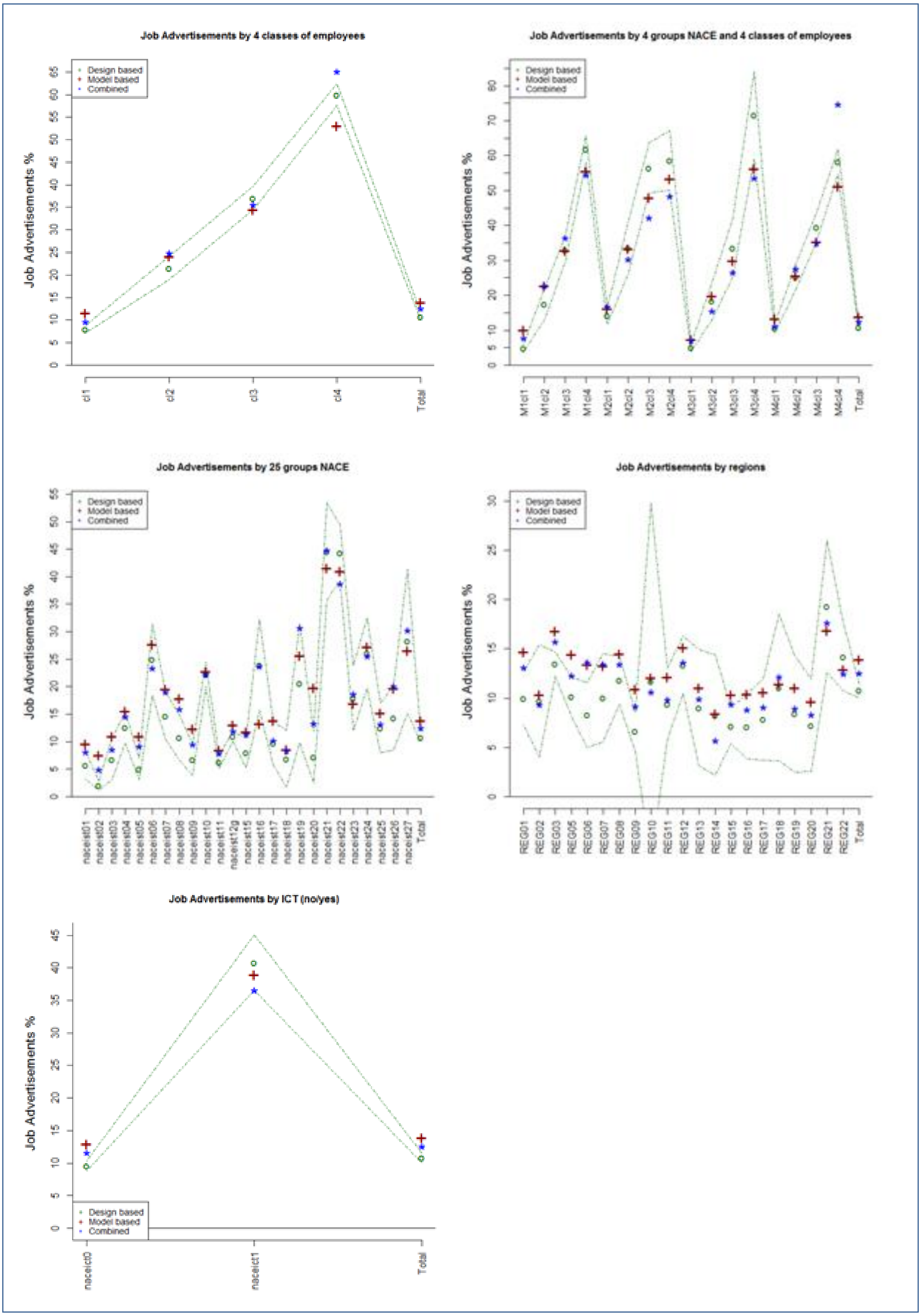
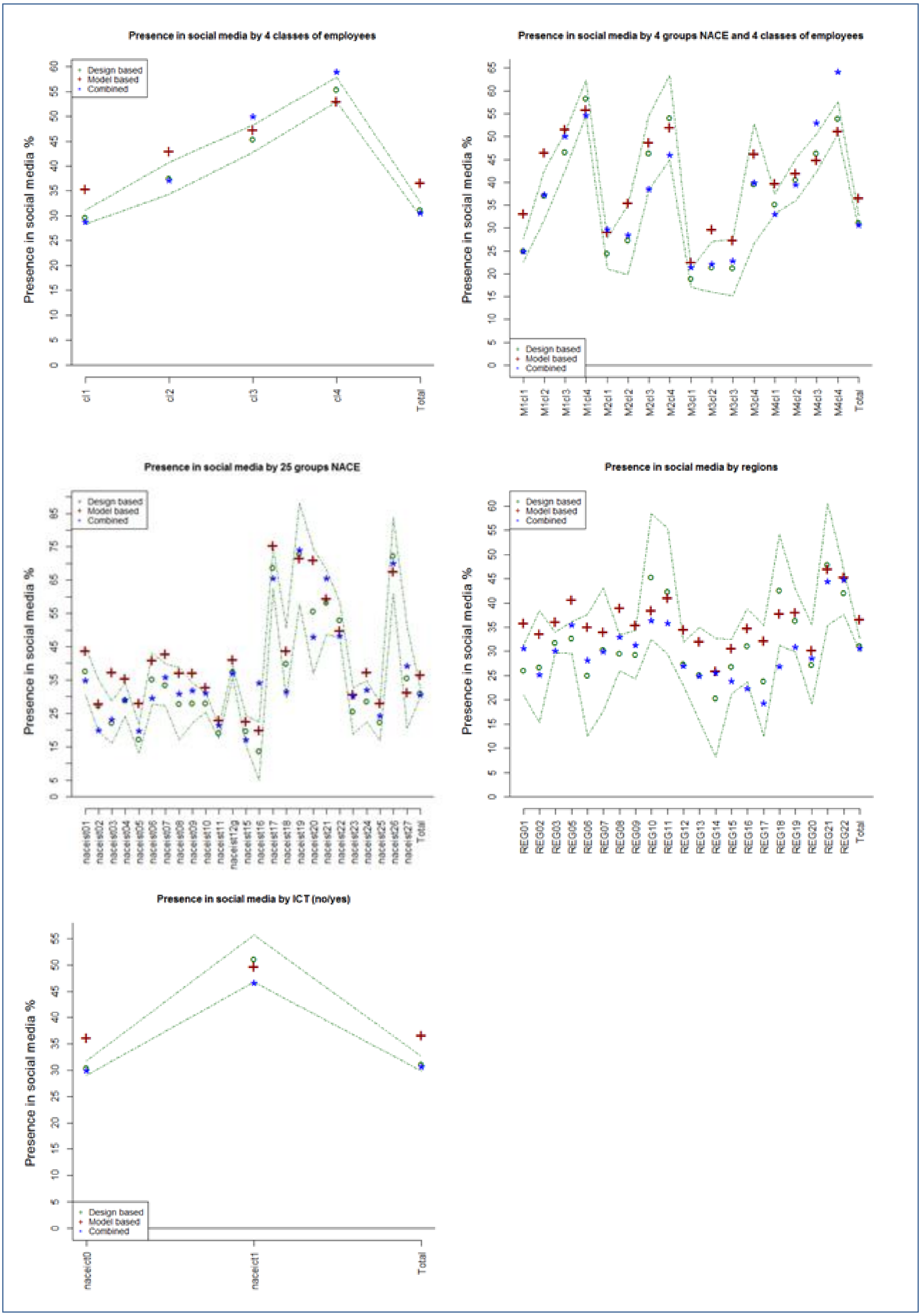


Figure 4 Presence in social media estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)



Presence in social media by 4 classes of employees

Presence in social media by 4 groups NACE and 4 classes of employees

Presence in social media by 25 groups NACE

Presence in social media by regions

Presence in social media by ICT (no/yes)

Lessons learnt

A first analysis of the estimates related to web-ordering, job-advertisements and presence in social media rates, obtained with the two alternative estimators, compared to the estimates produced by the official survey, allows some preliminary conclusions.

The three different sets are not incoherent. For instance, considering web-ordering the estimates for the total are well inside the confidence interval of the survey estimate, and this is the same for many values in the different domains.

Looking at coherence as one important dimension of quality, both combined estimates and full model based estimates can be considered as equally acceptable. But two considerations can be made:

1. the second component of the combined estimator is based on an assumption of perfect correctness of reported values, and considers predicted values as errors when they do not coincide with the reported ones. But controls have been carried out when fitting models, and in half of the cases in which predicted values were contradictory with reported ones, this was not due to model fault, but to response errors. So, this assumption does not always hold. In any case it would be advisable to deepen this phase also by returning to the respondents to verify if it is an error in response or if, for example, the model has evaluated the content of a site different from that one considered by the respondent;
2. if a medium-term aim is to make multi-annual frequency of the questions in the survey related to the websites characteristics (as Eurostat envisaged), then the combined estimator cannot be applied, as it relies on the current availability of reported values from the survey, and the full model based estimators remains the only alternative. In this case, there would be an issue in time series analysis due to problems in comparability between survey estimates and model based ones.

The main flaws of the model based estimator are in the presence of

- prediction errors;
- under-coverage of the population of enterprises owning websites, part of which has not been reached by web scraping.

As for the first, taking into consideration the presence of response errors in the test set, once eliminating them by manual inspection, the accuracy of the model predictions increases to more than acceptable levels (around 90% for web ordering, about the same for the other two variables), in any case comparable with the accuracy of survey data.

As for the second, pseudo-calibration allow to limit the bias, especially when the difference in the values of the parameters in the two sub-populations is not high, as it is the case.