The Italian new survey on Enterprises Final Consumption of Energy Products (COEN) – 2011: an innovative editing procedure

Giovanni Seri – Roberta Varriale – Ugo Guarnera
Italian National Statistical Institute (Istat)
{seri,varriale,guarnera}@istat.it

1. Introduction

Regulation (EC) No 1099/2008, successively amended by the Commission Regulation (EU) n.844/2010 (Regulation in the following), establishes a common framework for the production, transmission, evaluation and dissemination of comparable energy statistics in the Union. While energy statistics have traditionally been focused on both energy supply and fossil energies, the Regulation promotes, for the coming years, an increasing knowledge and monitoring of final energy consumption and renewable energy. To the aim of the Regulation, enterprises final energy consumption refers to the total energy consumption in industry, transport and commercial and public services.

The Italian National Statistical Institute (Istat) - supported by the Ministry of economic development (MISE) and the Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA) - planned, designed and performed a statistical survey addressed to final energy users in order to integrate data sources on energy products and their aggregates defined by the Regulation[1]. The survey on Enterprises Final Consumption of Energy Products (COEN) was conducted in 2012 (reference year 2011). COEN can be considered as a sort of "pilot" survey from both the contents and the technical/methodological point of views (a similar survey was carried out by Istat only in 1999).

This work aims at describing the main characteristics of the COEN survey (Paragraph 2) focusing, in particular, on the innovative procedure of data checking/editing (Paragraph 3) combining the use of selective editing, mixture modeling and deterministic procedures. In the last paragraph some conclusions and final remarks are given.

2. The survey on Enterprises Final Consumption of Energy Products (COEN)

Up to 2010 (data reference year), information on the expenditure of enterprises on energy products was annually collected by PRODCOM (PRODuction COMmunautaire) survey. Starting from 2011, energy statistics will be provided by COEN survey. The aim of this survey is indicated by the Regulation as the estimation of enterprises final consumption (not the expenditure, as in PRODCOM) of energy products. More specifically, COEN survey aims at estimating the enterprises consumption of a list of products belonging to the following categories:

- Electricity and heat
- Natural gas
- Oil and petroleum products
- Solid fossil fuels and manufactured gases
- Renewable energy and energy from waste.

The Regulation considers the 'final consumption' of the enterprises belonging to specific economic activities according to the NACE 2007 classification (the Energy Sector and the Supply and Transformation Sector are not investigated). Table 1 reports (first column) aggregates required by the Regulation for each energy product.

[1] Note that the household consumption is addressed by a specific survey.

Keeping into account that a large proportion of Italian enterprises have a very small size in terms of number of employees, the population investigated by COEN survey includes only enterprises with at least 3 employees, thus adopting an analogous scheme to that adopted by PRODCOM survey. The coverage of the whole population in terms of number of enterprises, number of employees and turnover are shown, for each aggregate required by the Regulation, in Table 1, columns 2, 3 and 4, respectively. We are confident that the threshold of 3 employees allows a better outcome of the survey by reducing the size of the population under investigation and consequently increasing the expected precision of the estimates, maintaining at the same time almost all the potential information, at least in terms of number of employees and turnover. Moreover, it usually happens that the lower is the enterprise size, the lower is the response rate (see, for example, [2]). Finally, auxiliary sources of information may be used in order to estimate final energy consumption of enterprises with less than 3 employees. Just to give an example, an auxiliary source of information about the survey objective is the administrative data stemming from the public company TERNA that should provide almost complete set of information on electricity [3].

The Italian business register ASIA 2011 (the sampling frame) counts around 4.5M units while the population under investigation is less than 1.1M units: Industry 33% (energy industries excluded), Transport 3%, Public and commercial services (Services) 65%.

Table 1. *Coverage in terms of number of enterprises (ent), number of employees (emp) and turnover (turn) of the surveyed over the whole population; response rate (resp) by NACE aggregates.*

NACE aggregates	% ent	% emp	% turn	% resp
Iron and Steel	74	99	100	45.4
Chemical (including Petrochemical)	69	99	92	47.1
Non-Ferrous Metals	68	98	99	45.2
Non-Metallic Minerals	54	94	98	38.8
Transport Equipment	65	99	99	39.6
Machinery	59	95	97	44.1
Mining and Quarrying	58	94	94	33.2
Food, Beverages and Tobacco	56	92	97	38.9
Pulp, Paper and Printing	50	92	97	43.2
Wood and Wood Products	35	80	92	35.9
Construction:	26	70	73	30.4
Textile and Leather	51	92	96	34.0
Not Elsewhere Specified – Industry	41	90	97	40.7
Industry Sector - Total	**36**	**86**	**92**	**40.7**
Rail	89	100	100	57.1
Domestic Navigation	34	96	97	37.0
Road	27	82	86	30.3
Pipeline Transport	36	100	96	33.3
Aviation	67	100	99	30.9
Transport Sector - Total	**34**	**84**	**91**	**30.9**
Commercial and Public Service	21	71	84	36.8
Other Sector – Total	**21**	**71**	**84**	**36.8**
Total	**27**	**77**	**88**	**37.1**

The sampling design was based on a one stage stratified simple random sample with a sample size of around 40k units. As a stratification criteria, the NACE regroupings defined by the Regulation (estimate domains), the level 1 of the NUTS classification and the number of persons employed size class were considered. All the enterprises belonging to the bigger size classes were included in the

sample. Size class threshold varied among NACE domains in order to preserve the Industry sector (Constructions excluded) and small domains.

Regarding data capturing, COEN has been designed and conducted completely as a web survey. Both analytical and observational unit was the enterprise but optionally data could be reported for local units. In order to facilitate the respondents, different unit measures (those relevant for each product) were available in the questionnaire. Expenses, even if non mandatory information, were asked in thousands of euro.

COEN, as a first edition of a business survey, experienced a significant response rate, greater than 37% (see Table 1 column 5 for response rates detailed for NACE domains), corresponding to more than 15k enterprises respondents. As expected, response rate are higher in the Industry sector (excluded Constructions) and for bigger size classes.

3. The data editing procedure

The data editing strategy was differentiated for class of products, depending on the number of observations and the available auxiliary information. Generally, we used a combination of: selective editing of influential errors based on contamination models, mixture modeling to detect unity measure errors and deterministic procedures to remove other systematic errors. In particular, when the number of observations was suitable (Electricity, Gas, Diesel oil and Heat), we detected unity measure errors using an approach based on mixture models [5]. After this treatment, only for Electricity and Gas (the two products with the highest number of observations) we identified influential errors through a contamination model, implemented with the R package Selemix [4]. We used deterministic procedures for the other products.

In the following, we report the results obtained over the set of the 12968 observations collected by October 2012 and used to produce preliminary estimates for Electricity consumption. The consumption of Electricity was a mandatory data: to register zero consumption respondents were invited to provide a motivation before going on with filling in the questionnaire.

The data editing procedure was followed separately for the three main sector: Industry, Transport and Services. Table 2 shows the response pattern on Electricity consumption (Quantity and Expenses). Note that the information on the Expenses for Electricity consumption was collected to be used as an auxiliary variable in the editing process. In particular, 11688 respondents out of 12968 filled in both the Quantity and the relative Expenses; in the Industry sector, 4406 out of 5805 enterprises also responded to the PRODCOM survey in 2010.

Table 2. *Number of respondents for Electricity, Quantity and Expenses, and number of respondents to the PRODCOM survey 2010 by sector of economic activity*

	Respondents (10 Oct 2012)	Zero consumption	Quantity only	Expenses only	Both Quantity and Expenses	PRODCOM 2010
Industry	6117	75	157	80	5805	4406
Transport	1243	264	70	28	881	
Services	5608	160	245	201	5002	
Total	12968	499	472	309	11688	

In the following, in order to exemplify the editing procedure, we focus on the Industry sector, where the procedure was more complex and complete because of the presence of the auxiliary information from the PRODCOM survey 2010.

The detection of unity measure errors based on mixture models was applied for the 5805 units responding to both Quantity and Expenses. In the COEN survey, enterprises could indicate the Quantity in kilowatt (kW), megawatt (MW) or gigawatt (GW), and had to indicate the Expenses in

thousands of euros. After a descriptive analysis (see Figure 1.a), we identified 5 clusters of possible unity measure error pattern: 1. no errors in both Quantity and Expenses, 2. Expenses expressed in euros instead of thousands of euros, 3. Quantity expressed in MW or GW instead of kW and MW, respectively, 4. Quantity expressed in kW or MW instead of MW and GW, respectively, 5. a combination of error types 2 and 4. Once a specific type of error is identified, the corresponding correction is obvious. Thus, the main difficulty is to classify the units according to their error pattern. To this aim, we used a model-based approach using mixture models. According to this approach, each mixture component corresponds to a population group associated with a particular error pattern. Each unit has been classified (and therefore errors have been edited) according to the highest posterior probability of group membership.

The procedure has been applied separately to the units with auxiliary information from the PRODCOM survey 2010 (1399) and to the units without auxiliary information (4406). Figure 1.b and Table 3 (column 1) show the results of the process. It is worth noting that, as expected, clusters 1 (no correction required), and 2 (expenses in euro instead of thousands of euro) contain more than 90% of the whole set of observations.

Figure 1. *Clusters of units by consumption of Electricity: Quantity and relative Expenses. Industry sector (units without auxiliary information from the PRODCOM survey 2010), log scale.*

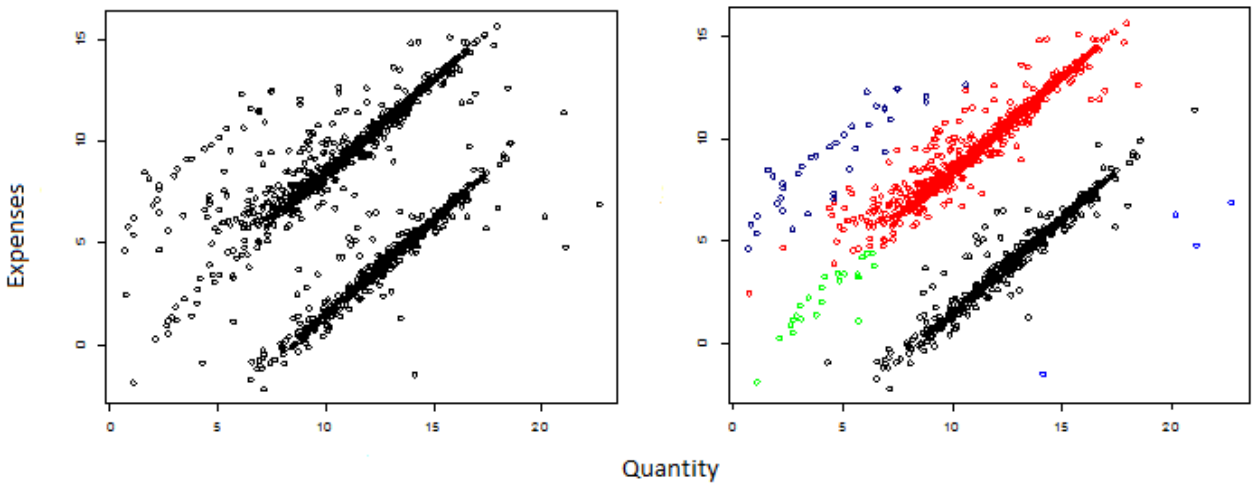

Table 3. *Unit measure error pattern: respondents for Electricity consumption and respondents to the Prodcom survey 2010 (Yes and No). Industry sector.*

		PRODCOM	
Cluster	**Number of units**	**Yes**	**No**
1	3343	572	2771
2	2092	763	1329
3	32	4	28
4	247	23	224
5	91	37	54
Total	**5805**	**1399**	**4406**

After the grouping process, observations with auxiliary information from PRODCOM survey were treated with Selemix. It is important to note that in the contamination model also the number of employees was used as auxiliary information. The result in the Industry sector was 80 (out of 4406) observations marked as ‘influential’ and corrected.

Respondents to both Quantity and Expenses without auxiliary information from PRODCOM survey were checked deterministically using the interval stemming from a set of 'good observations': 43 observations were corrected. The set of 11239 'good observations' was identified by first applying the grouping process to all units in the Industry sector having information on both Quantity and Expenses, and choosing the units with unit price (for kWh) lower than a 'reasonable' threshold (0.9 euros). On the base of this set of observations, intervals of 'acceptance' of Electricity consumption were defined taking into account NACE domains and size classes.

Also the units responding to Quantity only (157) were checked deterministically using the interval stemming from the set of 'good observations', while those responding to Expenses only (80) were first imputed using the median unit price per combination of NACE domains and size classes and then checked deterministically using the interval from the set of 'good observations'.

For the Transport and Service sectors, without auxiliary information from the PRODCOM survey 2010, a similar combination of selective editing, mixture modeling and deterministic procedures was used.

For all sectors and all products except Electricity, Gas, Diesel oil and Heat, the editing process was fully deterministic based on the following steps:

(a) unit price belongs to a given interval predefined;
(b) unit price belongs to a given interval predefined after expenses are divided by 1000;
(c) large enterprises with too low consumption are marked and clerically checked;
(d) respondent to Expenses only were imputed using the median unit price in a set of 'similar' observations.

As a final step, units mainly 'contributing' to the final estimates were clerically checked. Moreover, estimates were compared to value stemming from external sources and discussed with expert in the field.

4. Conclusions

COEN survey can be considered the first experimental edition of a survey on energetic statistics, given that previously only once, in 1999, a similar survey was conducted [1,6,7]. Through the experience of this first 'pilot' edition of the survey, Istat is planning to conduct a steady survey in the future. Obviously, at the moment, not all the decisions about a steady survey are already taken and the information described in the work is provided only related to the one shot survey carried out in 2012.

The results obtained through the survey are quite good with respect to both collaboration with enterprises, response rate and evaluation of the subject matter experts. Despite of this, the analysis of the raw and treated data and the editing procedure highlighted some features of the survey that need to be revised in the future. Just to give some examples, only one unit of measure will be used for each product, and no decimals will be required. Questions on the products will be sorted depending on how commonly they are used by enterprises, and the definition of some products not commonly used will be clarified. Furthermore, some auxiliary information that has been used only in the deterministic phase of the editing procedure, will be used also in the planning phase, to define the survey design.

References

1. Istat: Gli acquisti di prodotti energetici delle imprese industriali, Anno 2009, Tavole dati Istat (2012).
   http://www.istat.it/it/archivio/49962
2. Casciano M.C., Cirianni A., De Giorgi V., Di Francescantonio T., Mazzilli A., Luzi O., Oropallo F., Rinaldi M., Santi E., Seri G., Siesto G.: Utilizzo delle fonti amministrative nella rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni. Istat Working Papers N. 7/2011 (2011)
   http://www.istat.it/it/files/2011/10/Istat-Working-Papers-n.-7-2011.pdf
   http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
3. Terna S.p.a.: *Dati Statistici sull'energia elettrica in Italia. Annuario* (2010)
   http://www.terna.it/default/Home/SISTEMA_ELETTRICO/statistiche/dati_statistici.aspx
4. Di Zio M., Guarnera U.: Selemix: an R package for selective editing via contamination models. In *Proceedings of Statistics Canada Symposium,* 2011, Ottawa (2011)
5. Di Zio M., Guarnera U., Luzi O.: Editing Systematic Unity Measure Errors Through Mixture Modelling. *Survey Methodology*, 31, 1, 53-63 (2005)
6. Ballin M., Iorio G., Mercanti A., Perrella G., Poggi A.: Indagine sugli impieghi delle fonti energetiche nel settore Industria in Italia - Anno 1999. ENEA Ente per le Nuove Tecnologie, l'Energia e l'Ambiente - Serie RT/STUDI/2001 (2001)
7. Iorio G., Perrella G., Ballin M.: Indagine sugli impieghi delle fonti energetiche nel settore Terziario in Italia - Anno 1999. ENEA Ente per le Nuove Tecnologie, l'Energia e l'Ambiente - Serie RT/STUDI/2001 (2001)