Producing Turnover Indices in Industry at INE Spain

Editing, validation and imputation

Workshop on Turnover Indices

Muscat, 22-24 May 2017

elena.rosa.perez@ine.es

Introduction

One of the difficulties in providing high-quality statistical output arises from the fact that the data sources contain **errors** that may **influence** the **estimates**.

It has been estimated that NSIs spend approximately **20%-40% of their resources** on editing and imputing data.

Statistical data editing refers to detecting and correcting errors .

It is not necessary to correct all data in every detail. In order to obtain data of sufficiently high quality, it is usually enough to remove only the most influential errors.

"Over-editing": correcting errors that do not have a noticeable impact on the ultimately published figures.

Introduction

Small errors in individual records are acceptable:

- small errors in individual records tend to cancel out when aggregated;
- there will always be a sampling error in the published figures, even when all collected data are completely correct.

The main four principles of the E&I:

- The data quality at the beginning and at the end of the E&I process must be assessed;
- The E&I process has to be designed and executed in a way that allows for control of the process;
- The data quality at the end of the process should satisfy the needs of the users;
- The process should be as simple, cheap and fast as possible.

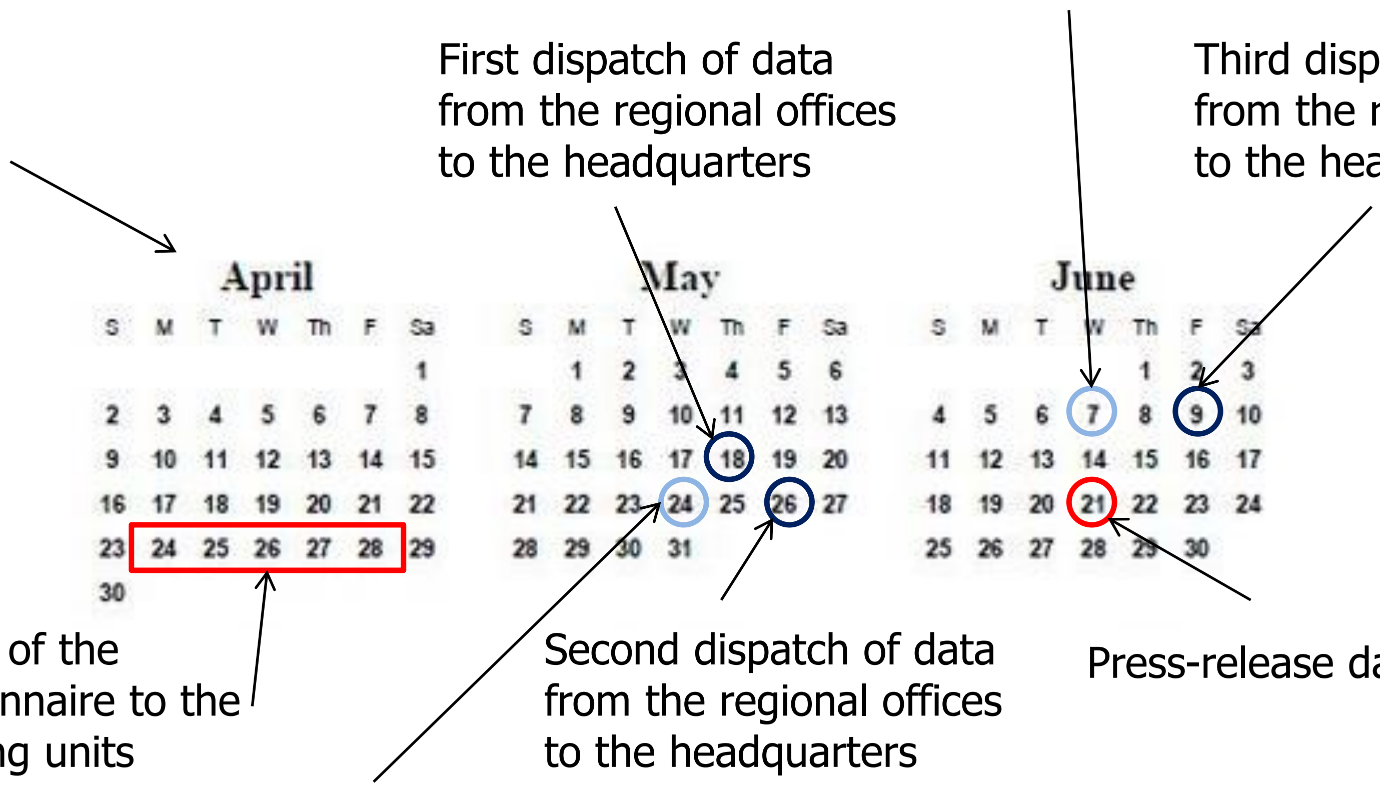
Data collection dates

IPI Press-release day

Reference month

First dispatch of data
from the regional offices
to the headquarters

Third dispatch of data
from the regional offices
to the headquarters



April

S	M	T	W	Th	F	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

May

S	M	T	W	Th	F	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

June

S	M	T	W	Th	F	Sa
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

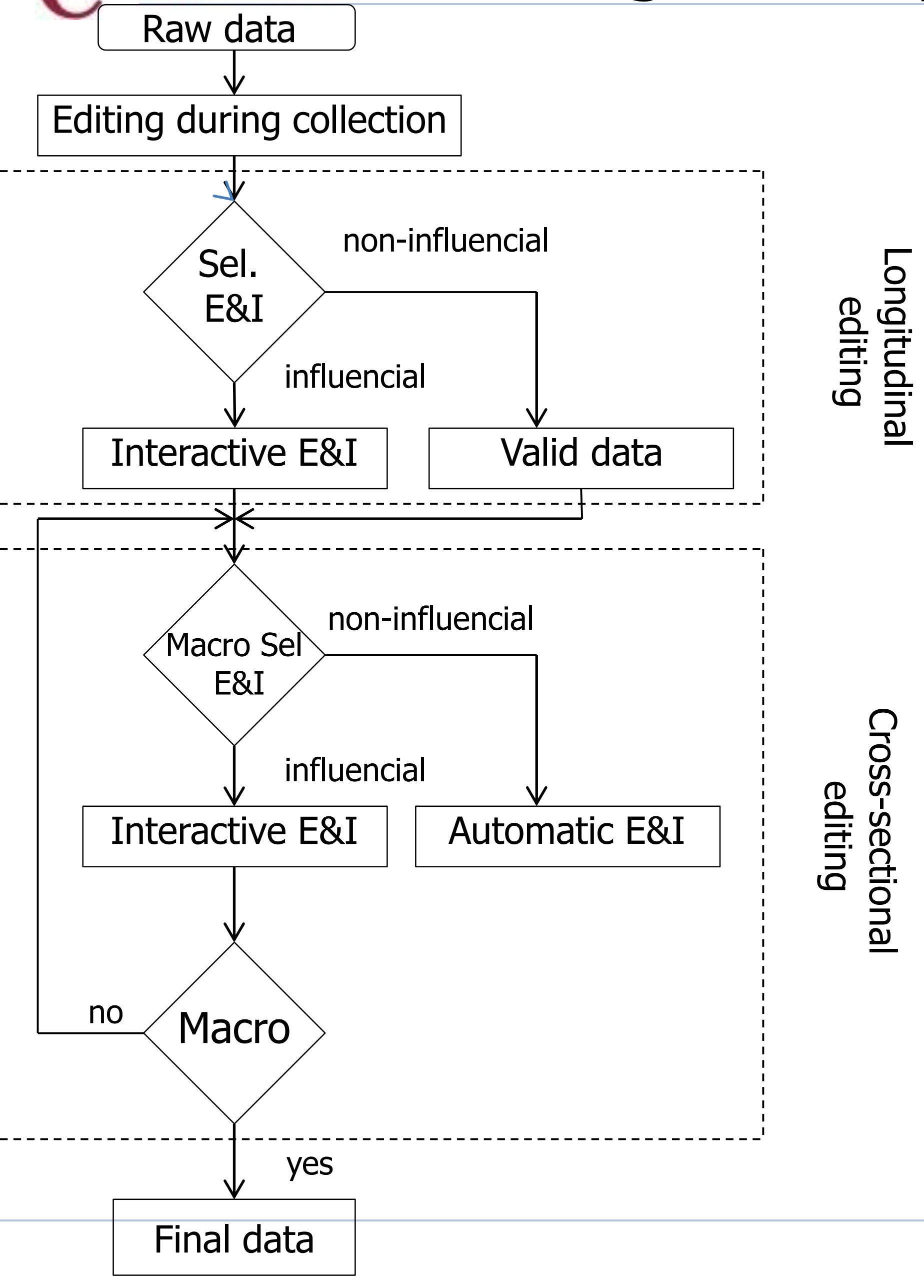
Mailing of the
questionnaire to the
sampling units

Second dispatch of data
from the regional offices
to the headquarters

Press-release day

IPRI Press-release day

Editing and imputation strategy



The E&I strategy at INE Spain in ITI is the following:

- Editing during data collection: the e-questionnaire contains hard and soft edits;

- Interactive editing at the regional offices: including recontact with the respondents and editing of paper questionnaires;

- Macro editing at the central office: questionnaires flagged are again subjected to interactive editing.

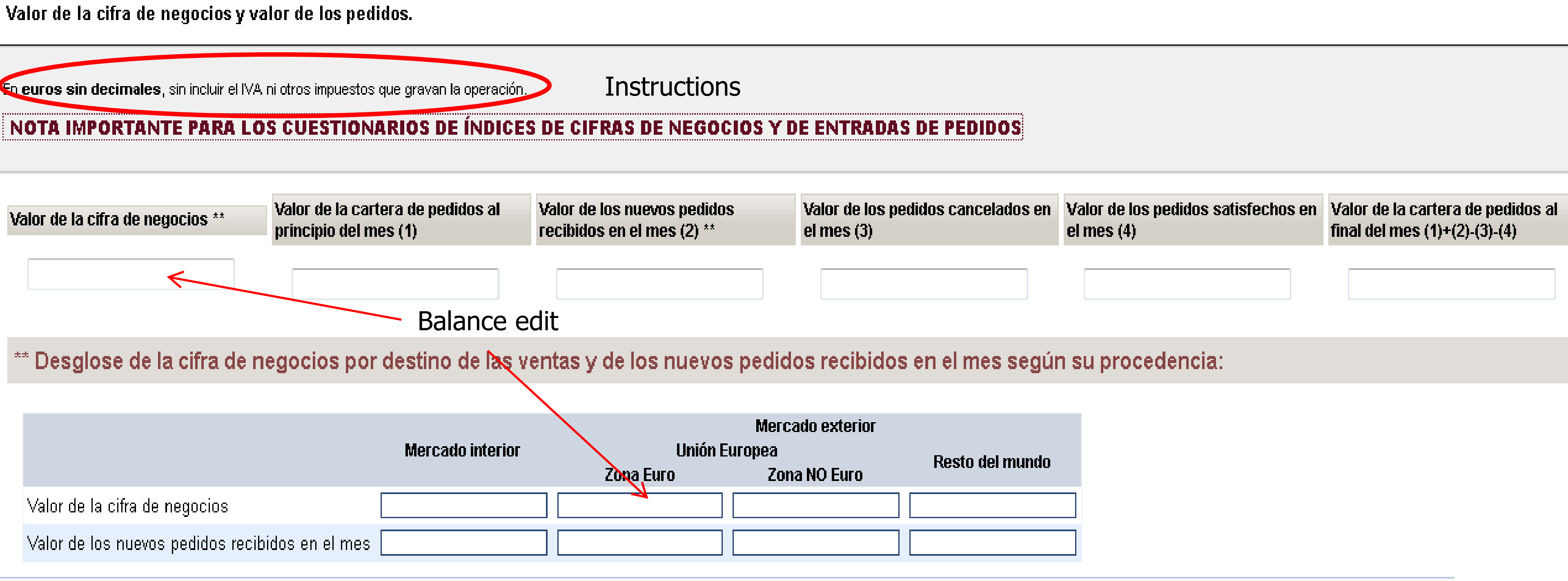
Each stage comprises a set of check controls for the whole sample.

Editing during data collection. e-Q

Unity measurement error that occurs if the respondent reports in euros when it was required to report in thousands of euros, or viceversa. Turnover must be provided in euros without decimals.

Rage restrictions: There is no rage restriction in turnover

Balance edits: Total turnover = ∑ market turnover



Valor de la cifra de negocios y valor de los pedidos.

En **euros sin decimales**, sin incluir el IVA ni otros impuestos que gravan la operación.

Instructions

NOTA IMPORTANTE PARA LOS CUESTIONARIOS DE ÍNDICES DE CIFRAS DE NEGOCIOS Y DE ENTRADAS DE PEDIDOS

Valor de la cifra de negocios **	Valor de la cartera de pedidos al principio del mes (1)	Valor de los nuevos pedidos recibidos en el mes (2) **	Valor de los pedidos cancelados en el mes (3)	Valor de los pedidos satisfechos en el mes (4)	Valor de la cartera de pedidos al final del mes (1)+(2)-(3)-(4)

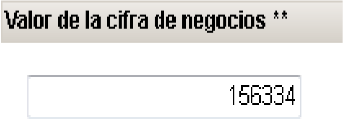
Balance edit

** Desglose de la cifra de negocios por destino de las ventas y de los nuevos pedidos recibidos en el mes según su procedencia:

		Mercado exterior		
	Mercado interior	Unión Europea		Resto del mundo
		Zona Euro	Zona NO Euro	
Valor de la cifra de negocios				
Valor de los nuevos pedidos recibidos en el mes				

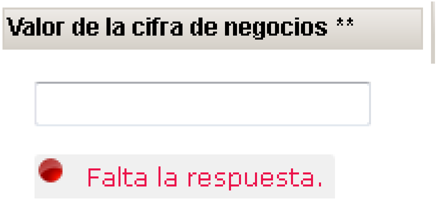
Editing during data collection. e-Q

Missing data are not allowed



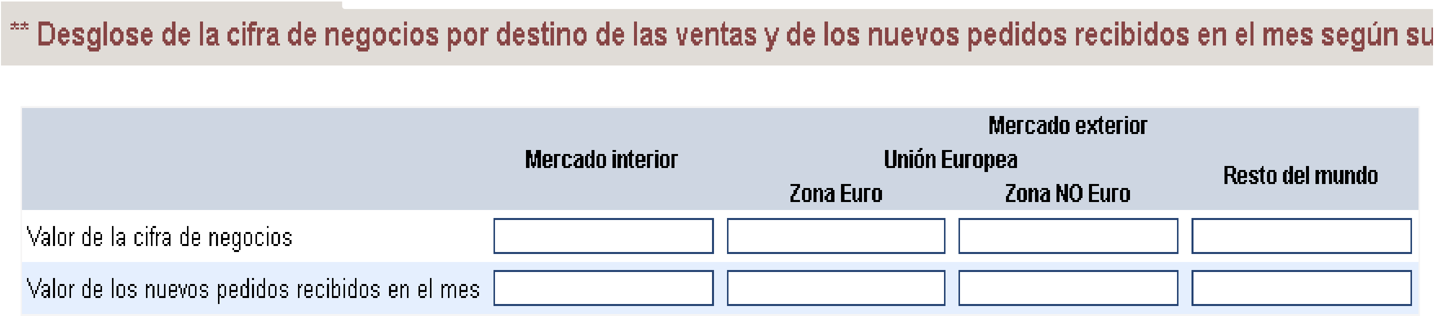
Valor de la cifra de negocios **

156334



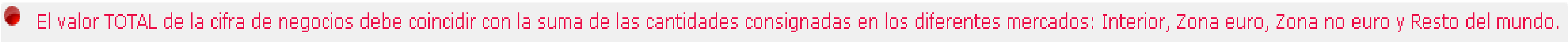
Valor de la cifra de negocios **

● Falta la respuesta.



** Desglose de la cifra de negocios por destino de las ventas y de los nuevos pedidos recibidos en el mes según su

	Mercado interior	Mercado exterior		
		Unión Europea		Resto del mundo
		Zona Euro	Zona NO Euro	
Valor de la cifra de negocios				
Valor de los nuevos pedidos recibidos en el mes				



● El valor TOTAL de la cifra de negocios debe coincidir con la suma de las cantidades consignadas en los diferentes mercados: Interior, Zona euro, Zona no euro y Resto del mundo.

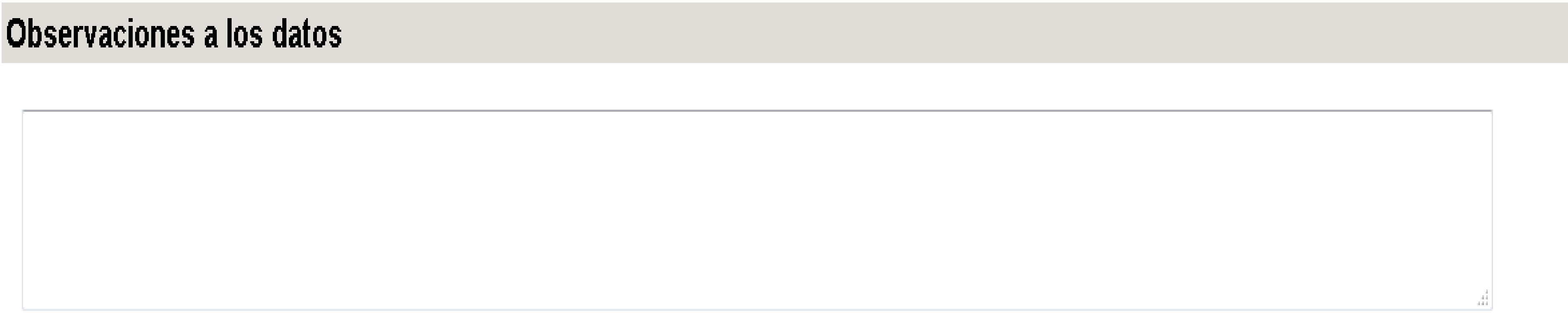
The same message appears if the markets turnover do not add up the total.

A questionnaire is flagged if:

- the total turnover for the current time period equals the total turnover of the preceding time period;
- the total turnover is 0.

Editing during data collection. e-Q

Interval-distance control checks: for the **total turnover** value of each respondent a validation interval is assigned. If the distance (from reported value to the interval) is greater than certain threshold (that can be changed), the questionnaire is flagged; it is not, otherwise.



Observaciones a los datos

If the questionnaire is flagged, then comments have to be included explaining the reason of the flag. Without any comments, the questionnaire will not be sent.

Editing during data collection

In the **regional offices** all questionnaires (on-line and paper) are recorded/downloaded in an internal application. This application is used to **edit paper questionnaires** (same edits as electronic questionnaires).

Interval-distance control checks for total turnover are carried out one more time. The intervals in this step are broader.

The regional staff also get in **contact** with the respondents if the don't send back the questionnaire filled in. Finally, if the establishment does not answer the questionnaire, this staff is in charge of **fining** them.

They send to the central office all the doubts that emerge and we are in permanent contact with them: to ask for non received questionnaires of important units, to clarify doubts of high/low data.

Selective editing

Selective editing applies interactive editing to a well-chosen subset of the **records** (the **most influential**). This way limited time and resources available for interactive editing are allocated to those that have most effect on the quality of the final estimates.

In selective editing, a score is calculated for each record, expressing the relevance of the potential error(s) in the record. This score is used to prioritize units.

The selective editing is an editing method that perfectly adapts to the ITI: skewness of the distribution of Industrial businesses size, cut-off sampling.

Selective editing. Longitudinal phase

Interval-distance controls: They are based on the construction of a validation **interval** (using ARIMA predictions), a **distance** (which is a score function) and a **threshold** for the reference period t.

This threshold (used to flag the questionnaires) is obtained using:

- the distance between the final edited values and their corresponding validation intervals for the preceding period $t-1$ for each unit k;
- the quantile over the distribution of distances for each dissemination domain.

These type of edits are called **longitudinal** because we use the longitudinal **information** for the total turnover of **each unit** in the sample.

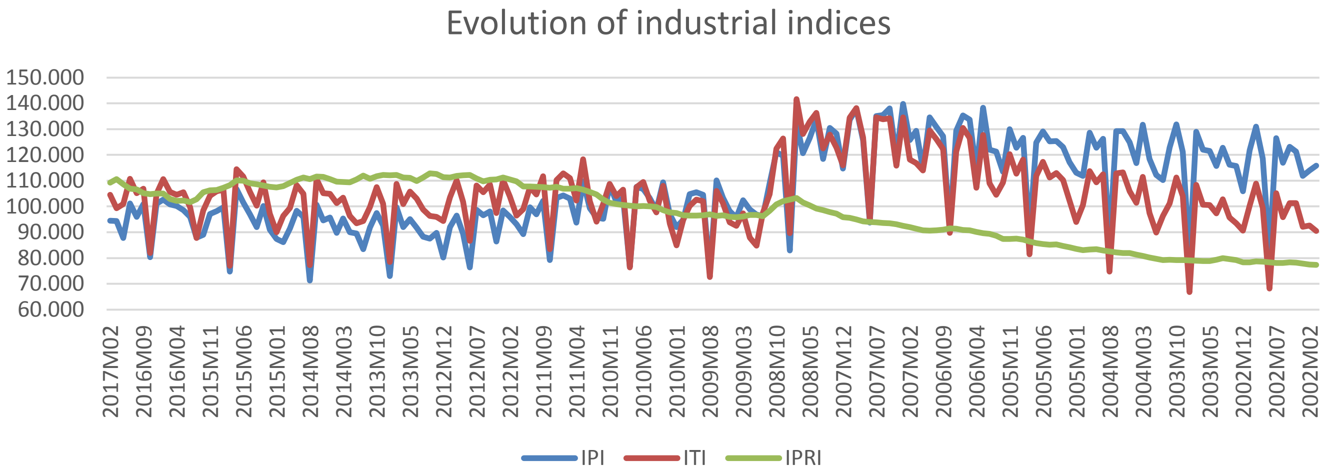
Selective editing. Cross-sectional phase

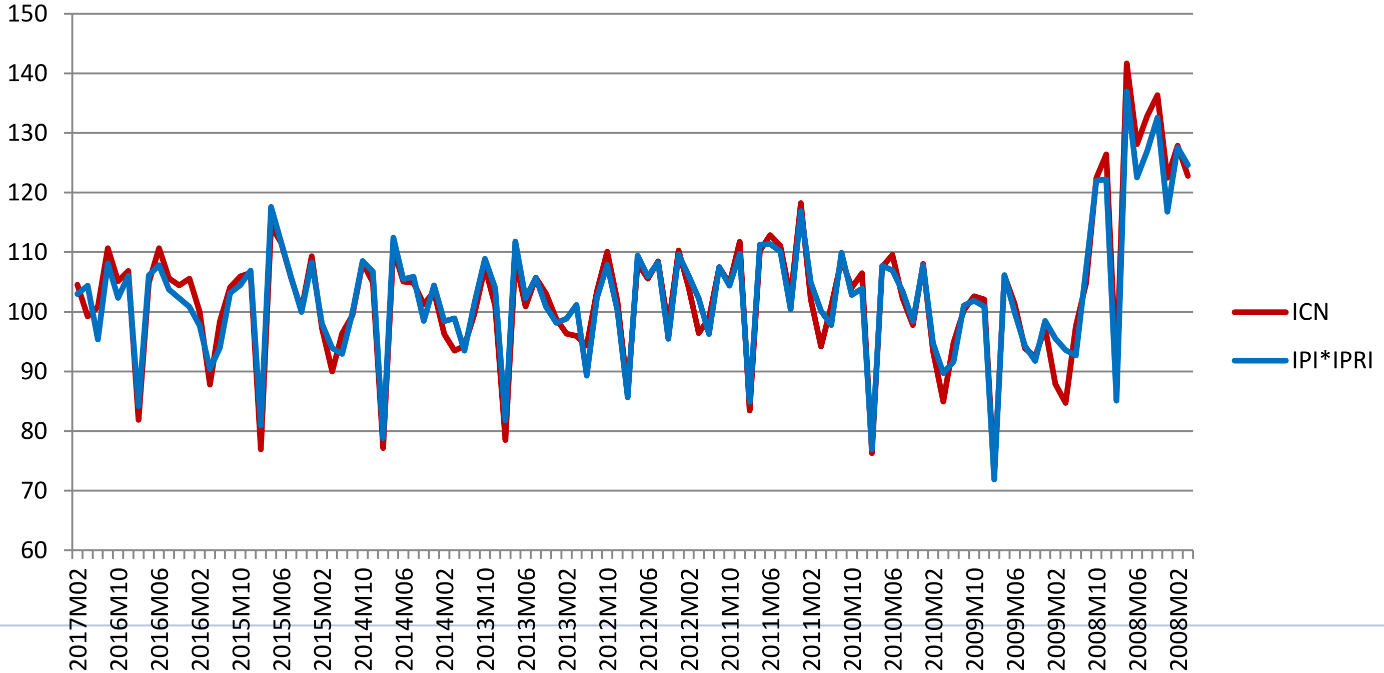
After the **first** and the **second** data dispatches a **cross-sectional** editing takes place. In each phase 100 questionnaires are flagged and sent back to the regional offices for further interactive editing. The edited data will be received in the following dispatch.

It is called **cross-sectional** because we are using the data of the **reference month** of the **whole sample**.

A prediction of the value of each unit is obtained using regression with the data of the reference month received in the dispatch. This prediction value is compared with the reported value and the units are scored (and prioritized) according to the distance.

Macroediting





As we have seen, IPRI and IPI are disseminated before ITI, and they can be helpful in the macroediting phase.

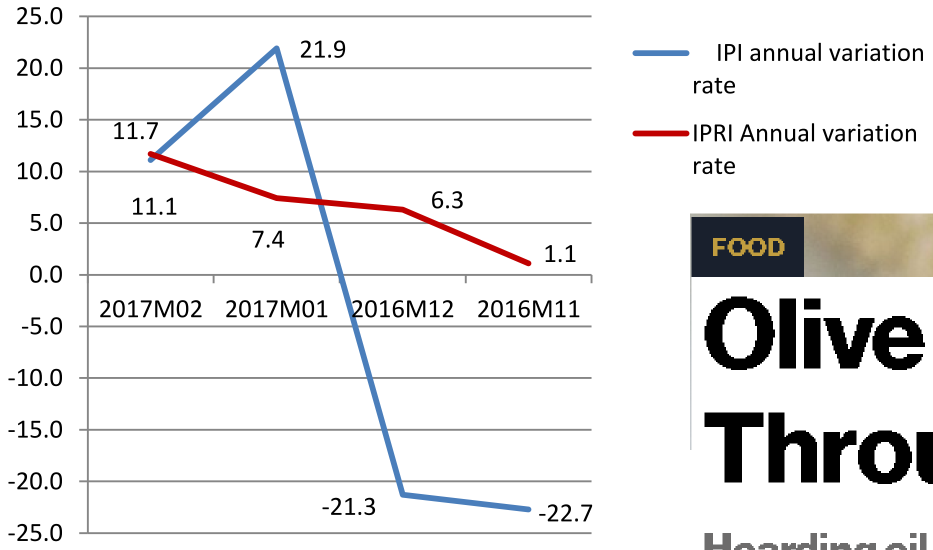
IPI measures the monthly development of productive activity of industrial branches, while IPRI measures the monthly development of the price of products manufactured and sold in the domestic market.

So IPI*IPRI trend should be "similar" to ITI's.

Macroediting

Monthly we obtain files containing the unit whose annual rate at microdata level are above 100% and under -70%. In some cases we can find that in some division/subdivision there is a general increase/decrease. IPI and IPRI data can useful to ascertain it.

For example: Olive oil in the last months in Spain. In November and December 2016 lot of units had negative rates, whereas in January and February 2017 the annual rate were in some cases above 100%.



More information can be obtain from the news in TV or in papers

FOOD

Olive Oil Prices Are Going Through the Roof

Hoarding oil of a new kind after terrible harvests in Italy, Spain and Greece.

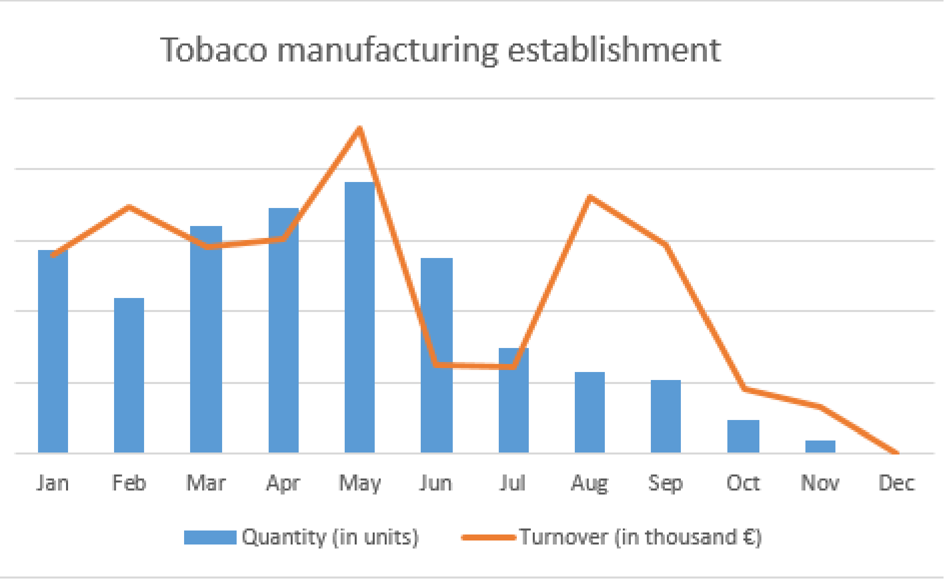
by **Agnieszka De Sousa** and **Richard Vines**
9 de febrero de 2017 14:15 CET

Macroediting

As we have already seen, the IPI sample and the ITI sample are really similar, and it can be used in the editing phase.

Once we detect that there is some division/subdivision with a rate that seems suspicious we take a look at the units included and if any has a very low/high value we can **compare** the **ITI microdata** with **IPI microdata**.

In many cases an increase/decrease in the turnover is reflected in the production and viceversa.



Imputation

The data of the establishments not sending the questionnaire are imputed automatically during the process. The imputation method consists in multiplying the data of the previous month by the monthly rate of the stratum.

The imputation method is not reliable, so units with imputed data are flagged if the monthly rate is higher than 100%. Suspicious units having a big impact in the indices are requested to be sent in the following data dispatch.

Establishment with imputed data are flagged in order to list these units in an easier way or to obtain quality indicators, like the imputation rate.

The imputation problem is further complicated owing to the existence of constraints in the form of edits that have to be satisfied by the data.

Bibliography

- De Waal T., Pannekoek J., and Scholtus S. (2007), *Handbook of Statistical Data Editing and Imputation,* Willey, N.Y.
- EDIMBUS. 2007. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. ISTAT; CBS; SFSO; EUROSTAT.
- INE Spain selective editing in ITI: http://www.ine.es/ss/Satellite?L=es_ES&c=INEDocTrabajo_C&cid=1259945892315&p=1254735839320&pagename=MetodologiaYEstandares%2FINELayout
- INE Spain Industrial Products Survey: http://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736149053&menu=ultiDatos&idp=1254735576715

Any question??

INSTITUTO NACIONAL DE ESTADISTICA INSTITUTO NACIONAL DE ESTADISTICA

INE