THE INCORPORATION OF PROCESSES TO IMPROVE THE INDEX NUMBERS PRODUCTION QUALITY:

THE BASQUE COUNTRY EXPERIENCE[1]

Cristina Prado, Lourdes Llorens, Marina Ayestarán

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

[1] Paper submitted at The International Conference on Quality in Official Statistics, May 2001, Stockholm.

Index

Chapter 1

Introduction

The aim of this paper is to put forward the modifications being implemented on the index generation process taking into account the following features:

Firstly and fundamental: EUSTAT is going towards an integrated information system. It is worth pointing out that the system advantages lie in taking the information in and outflows and using them in the most efficient way possible.

Secondly: It takes into account the knowledge built in along time. The know-how of doing a statistical operation. The goal has been both to model automatically the analyst knowledge and to make the information checks easier.

Thirdly: The system implementation has made clear the feedback between several statistical operations. It has served as a proof of what it has to come in the future in relation with an integrated information system.

Graph 1 describes in a simple how the system is trying to work to integrate the economical statistics within.

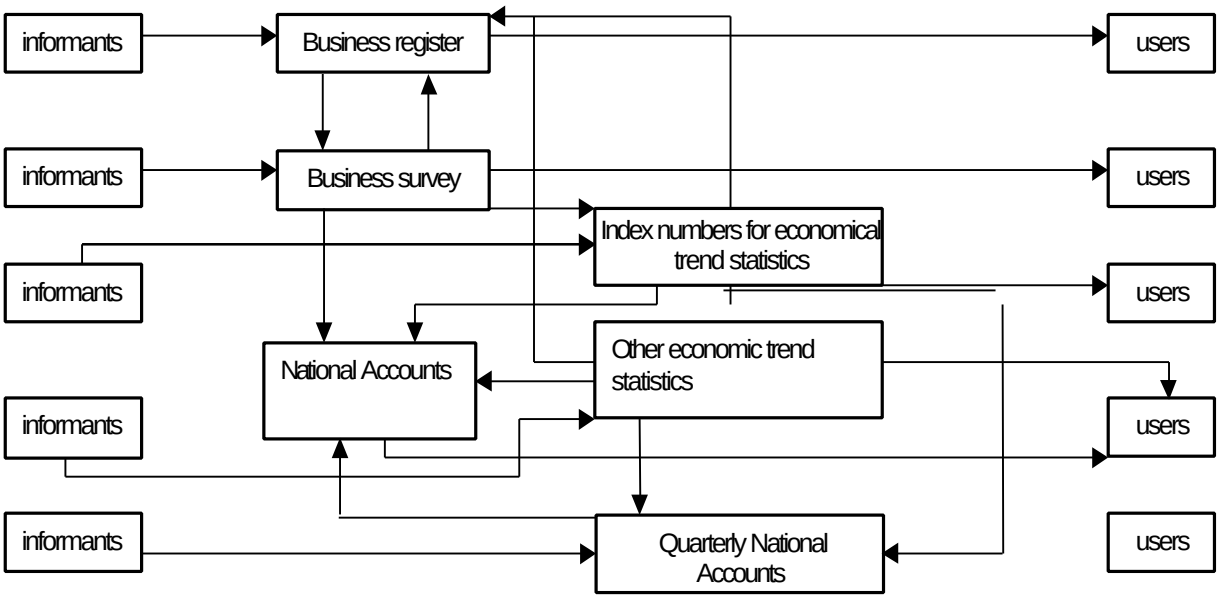
The main informational bases are our informants. Those come from different sources: administrative registers and business surveys.

The business register is the most important administrative base to select samples both for trend and structural surveys. The register is mainly fed by administrative sources but also collects information from the mentioned surveys.

All the statistics generated are going to be integrated so the information collected is used efficiently and timelessly. Up to now information is downloaded to the business register once a year. The final goal is to make automatic the updating of the information. This would generate, at least, three types of consequences: data would more efficiently collected, it would decrease the informant burden and results would be available before and with better quality.

In the following diagram it can be easily seen the existence of the feedback so then the integration and automation of the circuits will make the system efficient.

Graph 1 Economic statistics integrated information system.

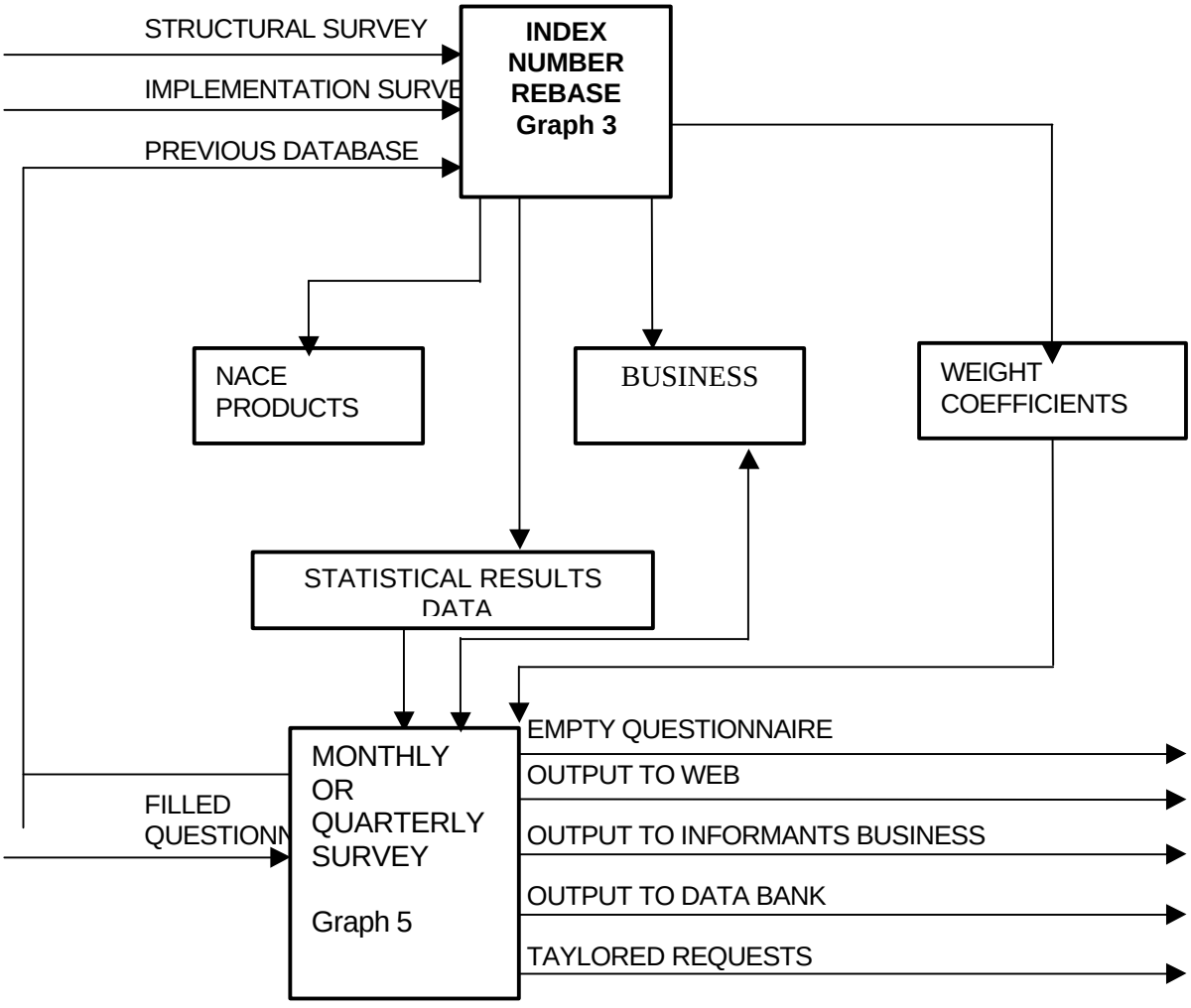


Chapter 2

The statistical operation process based on index numbers

We now refer to the index number production linked to the economical statistics. How this works can be seen in graph 2.

Graph 2.Statistical operation process based on index numbers general.



In general all the statistical operations based on index numbers consist of two clearly differentiated phases.

The first phase is the implementation of the operation. Under such a phase we could be under two cases, either the starting off of the operation or the rebase needed to reveal economical structural changes. In the first case a previous database would not be available. The diagram shows the relationship between the structural and trend operations in such a way that the results of the first ones affect the implementation of the latter ones.

The second phase refers to the periodical generation of index under its monthly or quarterly character.

Graph 2 shows the tasks to be done in order to update and generate the index. Among the activities there are three to point out since they are essential to make the index obtained valid and representative of the reality is intending to put forward. Such tasks require the analyst ability and knowledge of the sector.

1. Rebase. This change implies the analysis of a lot of information and the inclusion of new enterprises or establishments. We will explain further down how experience built in during many years has been incorporated in an automated way. In this task as well as in many others it is difficult to substitute the expertise analysis, but the experts can be helped in their work.
2. Data validation: Firms provide information that must be validated with consistent criteria and must be done as quickly as possible. The work incorporates the experience built in through out the years in making index numbers.
3. The imputation and treatment of no response: Trend statistics are subject to calendar so the imputation and treatment of no response is vital to generate valid indexes. The know-how incorporation in an automated way is the core of the whole question

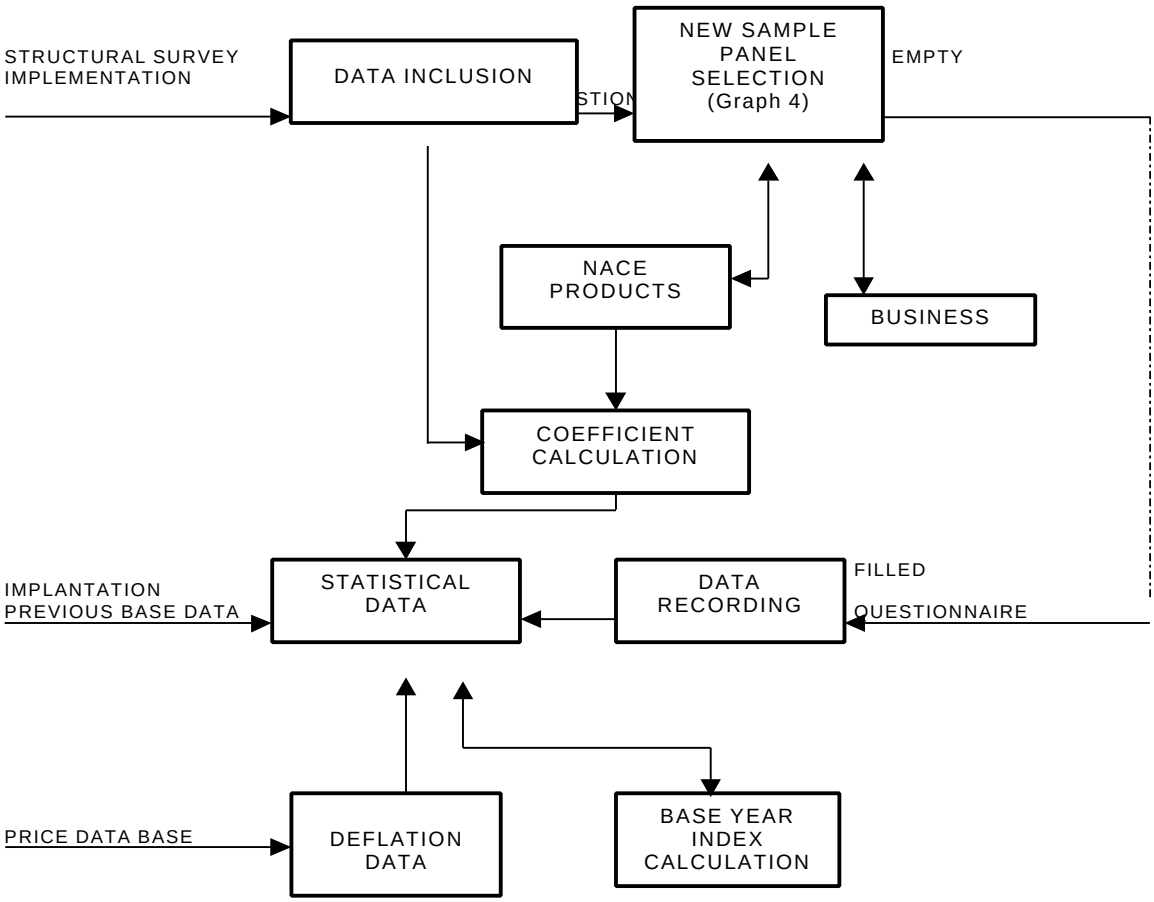
In what it follows we will describe these three steps in more detail.

Chapter 3

Sample panel renovation and index rebase

All trend statistics based on index numbers terms do require determining what it is known as the Base Year. Such year is taken as reference to select firms and establishments and, in some cases, also the products that constitute the sample panel. In this year the variables are made equal to one hundred so as to examine their index evolution in the following years (see graph 3).

Graph 3 Index number operations rebase processes



.

A rebase over some period of time is used, as well, to introduce all improvements and methodological modifications considered appropriated. This is important since once these elements are defined stay constant until the next rebase.

EUSTAT, following the 1995 European System of Accounts, makes rebases once every five years, this is the case for instance of the Industrial Production index number or

Industrial price index. However in some other sectors of great mobility, such as Retail, the base renovation is made more frequently, i.e., Domestic Trade Index changes its base every three years.

This base renovation is, therefore, of great importance for the index to work properly and to reflect the reality that intends to measure. However, the rebase implies an extra load of work over the normal tasks done (panel selection work of establishments, products and units, weight changes, series links, methodological improvements, and so on). This work must be done together with the publication of monthly and quarterly series in the previous base.

Those tasks are a very heavy load of work when a high degree of automation is not available and ad-hoc procedures not included in the general operation system are used.

The renovation of EUSTAT trend statistics system shows, among their main novelties, the integration into the system of the tasks related to the rebase. Such tasks can be comprise in the following points:

1. The automated generation of the new database and the new base year index processes.

2. The sample panel analysis and its sector representation.

3. Weight calculations.

The automated generation of the new database and the new base year index processes

This system process guarantees the generation of a new database to use together with the index new base. Such database contains the current panel ready to admit the required increases and modifications as well as all the calculation, imputation processes adapted to the index new base.

Sample panel analysis and its sector representation

Before describing the analysis facilities provided by the system it is convenient to comment on the type of sample selection used in the operations based on index numbers.

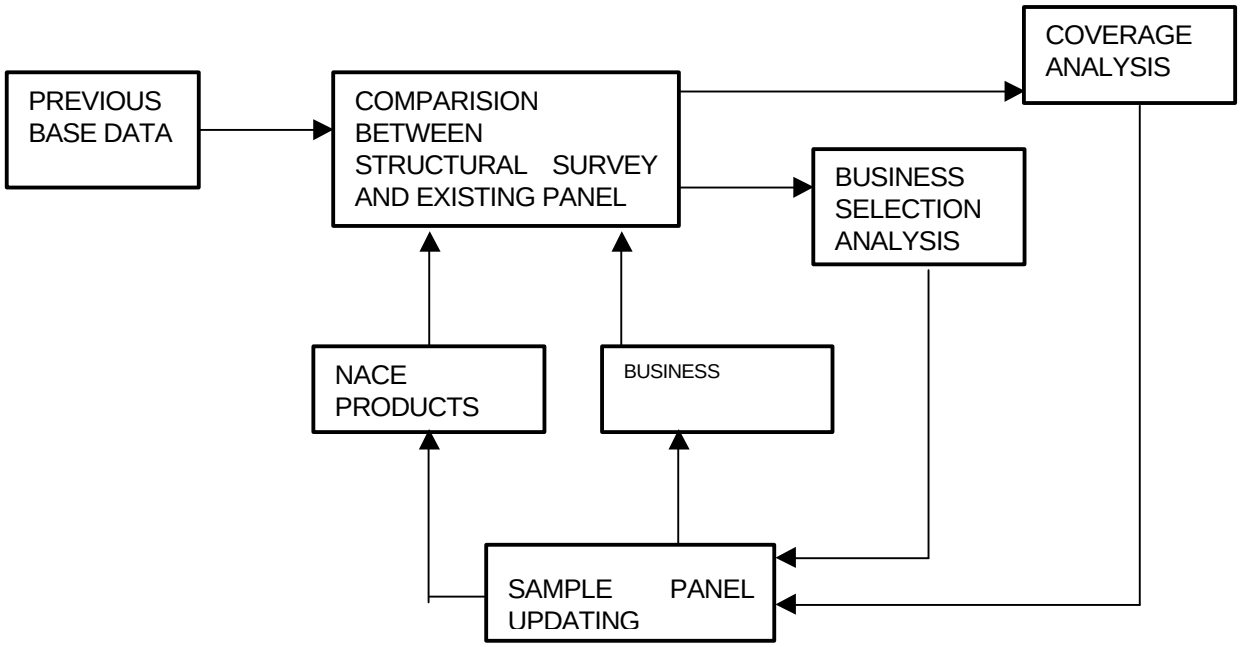
The building of the panel is made through a selection named “by sub-populations” and, this selection is applied to the different aggregation levels. At every level the most important parts within the sub-populations are chosen in such a way to guarantee a sufficient coverage. The procedure continues by obtaining the sample of establishments of the panel. Thus, in the Industrial Production Index (IPI) A31 and A60 classification orderings within each sector are considered in each NACE93 Class[2]and, eventually, within products until reaching the establishments.

The first facility provided by the system helps both to define or renovate the sample panel in the rebase periods and, also, to measure the current panel representation whenever more data is available from the corresponding structural survey (Industrial

[2] A31, A60 are classifications proposed by the 1995 European System of Accounts, which divides the Economy in 31 and 60 activities respectively. Class is named the 5 digits codification of 1993 NACE.

Business Survey, Domestic Trade Survey, and so on). This would make possible to take steps to improve the index, either introducing or substituting firms into the panel.

Graph 4: New sample panel selection process.



The system makes this task easier automating some processes and specially:

1. The sector coverage analysis.

2. The analysis of the panel establishment selection

Sector coverage analysis

This process provides, through some sector orderings of the different NACE93 aggregations, the weight of each of them with respect the superior level. This makes the activities selection work (up toNACE93 5 digits) that should be represented in the panel to reach the coverage required.

At the same time this analysis allows observing the coverage of the current sample in each of the aggregation levels.

Analysis for the panel establishment selection

This process, as the previous one based on structural data, is used to help in the sector selection and allows for a selection within each establishment more preponderant activities.

In the some cases such as IPI (Industrial Production Index) where a product panel is required, those processes are very easily included.

In other operations such as Domestic Trade Index with a great atomisation level in some activities (butchers, food providers, chemists) other type of analysis has been included to conduct the panel selection work. This consists of providing, for different error levels –10%, 15% or 20%- the optimum level of employees and sales to be considered in the panel. Such calculation is based in the following formulation.

$$\text{Optimum sample size (employees, sales)} = \frac{\frac{(1.96)^2 * S_p^2}{(error * \bar{X}_p)^2}}{1 + \left(\frac{\frac{(1.96)^2 * S_p^2}{(error * \bar{X}_p)^2}}{N_p} \right)}$$ ⊕

S_p = Employees or sales standard deviation

X_p= Employees or sales mean

N_p = Population employees or sales total

error = 0,10 ó 0,15 or 0,20

⊕ Francisco Azorín y José Luis Sánchez Crespo.

The inclusion of these error levels both guides the number of establishments to include in the rebase phase and, also, informs about current panel error margins, provided there are new structural data of the corresponding economic sector.

Weight calculation

The index numbers realised by EUSTAT are Laspeyres type and, thus, are calculated as an arithmetic and aggregated average of the panel indexes weighted by their significance in the base year.

Such formulation is:

$$I_i = \sum_{j \in i} I_j * w_j$$

$$w_j = \frac{V_j}{V_i}$$

Where

I_i = Aggregation level index i

I_j = Aggregation level index j, inferior to i

W_j = Weight of aggregation level j

V_j = Added value, production value or sales value depending on the index being calculated for aggregation level j inferior to aggregation level i

V_i =Added value, production value or sales value depending on the index being calculated for aggregation level j

Given the index definition the weight calculation is required when the Index base is renovated. Such calculation was not included in the process in previous versions and was done through an ad-hoc procedure. Currently the system allows the calculation and provides a double possibility:

- The weight calculation for the new base year coming from the corresponding structural data and taking into account the current panel
- The weight calculation for the new base year incorporating new activities into the current panel.

We can conclude that the automation and incorporation of new analysis that provides facilities to select the panel and to measure its representation allows for a new perception of the rebase. That is to say, from being considered as a phase with an extra load of work, generated basically by the sample panel selection to being seen as a phase of weight change and of normalisation of new base to 100. This is possible because the panel updating and its representation can be annually analysed or as frequently as structural information is available.

This new approach of the rebase makes much easier to adopt new base years and reduces the period of maintaining simultaneously two base years. This allows for a cost reduction in the associated tasks.

Chapter 4

Data Validation Process

The data validation is one of the key phases in any statistical operation and, obviously, in the economical statistics validation. The validation should be done quickly and exhaustively to be successful. (Graph 5)

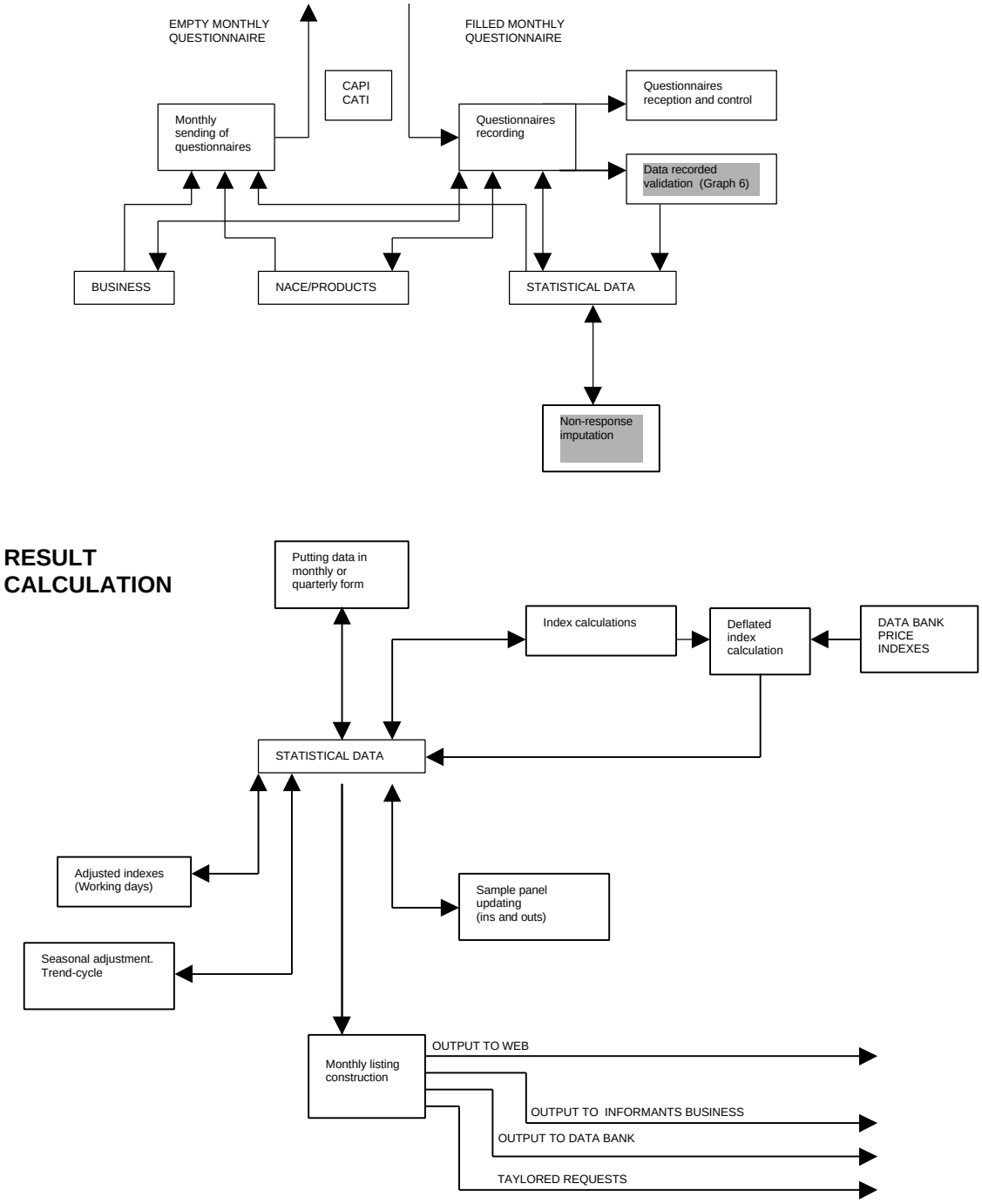
The updating of the trend statistics computerised processes being carried out by EUSTAT has provoked the improvement and renovation of the different statistical phases of such operations. Among those phases there is the data validation. This has been orientated basically to adjust mainly the economical information that comes from the questionnaires.

In general terms, index number trend operations (IPI, Industrial price index, Domestic Trade Index,…) use very simple questionnaires for collecting the information required. Such questionnaires contain a reduced number of economical variables.

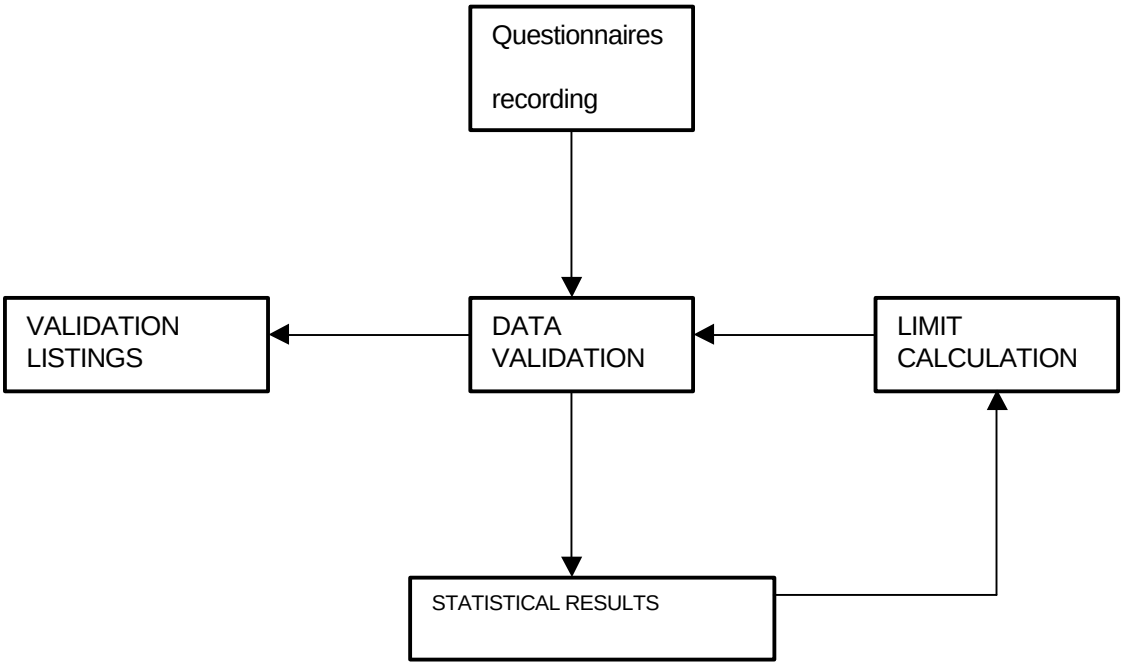
The validation processes of these variables, besides verifying the identification data and the questionnaire internal consistency, check the data variation value with respect previous periods (previous month, previous quarter, last year same month, and so on) and study whether it stays within adequate limits. Not such limits are fulfilled the questionnaire is refused and must be revised.

Therefore, the improvement incorporated into the validation process has been focused to the adjustment of the limits to accept or to refuse the collected information values.

Graph 5: Monthly and quarterly processes of an index number statistical operation.



Graph 6: Data validation process.



In previous validation processes, the mentioned limits were unique for all type of economical activities[3] and were fixed by the analysts, based on their knowledge of the corresponding economical sector. This working procedure was the only possible one due to the lack of historical series.

The problem of adopting general limits for all activities is, logically, the refusal of a much greater number of questionnaires that in the case of using more specific limits, making the revision task much harder. This is the main point where it is based the greater efficiency in the information validation process.

The current procedure tries to correct and optimise this feature and allows for determining such limits in terms of previous data. To this end the Tunkey formulation of extreme values for asymmetrical distribution has been utilised. Before adopting this formulation a very detailed analysis of these economical series was made to find out that such series follow an asymmetrical distribution

Those limits were tailored for each of the economical activities in order to get a greater precision. (5 digits of NACE93)

The Tunkey extreme value formula are defined as:

$$IL = \text{Max}\{\text{Min}\{a\}, q_1 - (q_3 - q_1) * 1{,}5\}$$

$$SL = \text{Min}\{\text{Max}\{a\}, q_3 + (q_3 - q_1) * 1{,}5\}$$

[3] The five digits of 1993 National Classification of Economical activities (NACE) define the economical activity.

IL = Inferior limit of a

LS = superior limit of a

a = Variable variation value (Variation over the same previous period, over the previous month,…)

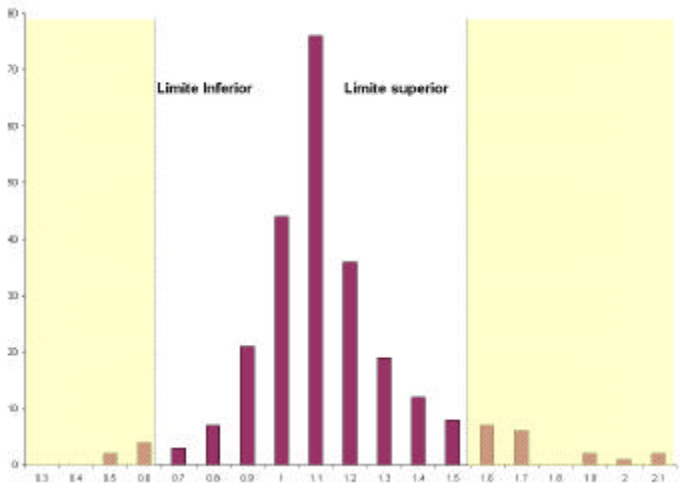
q_1 = First quartile, that is the value that leaves 25% of observations underneath

q_3 = First quartile, that is the value that leaves 75% of observations underneath

Max = Maximum value of *a*

Min = Minimum value of *a*

Graph 7: Distribution of limits and frequencies example over the previous period in the Domestic Trade Index



The new system integrates this process into two totally automated options:

- The data recorded validation in terms of the calculated limits for its activity
- The calculation of the limits for all the activities using all the data accumulated up to that moment; such limits are incorporated into the data validation when there is a great change of the limits currently used.

The integration within the same environment of both calculations makes very easy to maintain those limits updated and, therefore, a greater adjustment of the validation process to real value series.

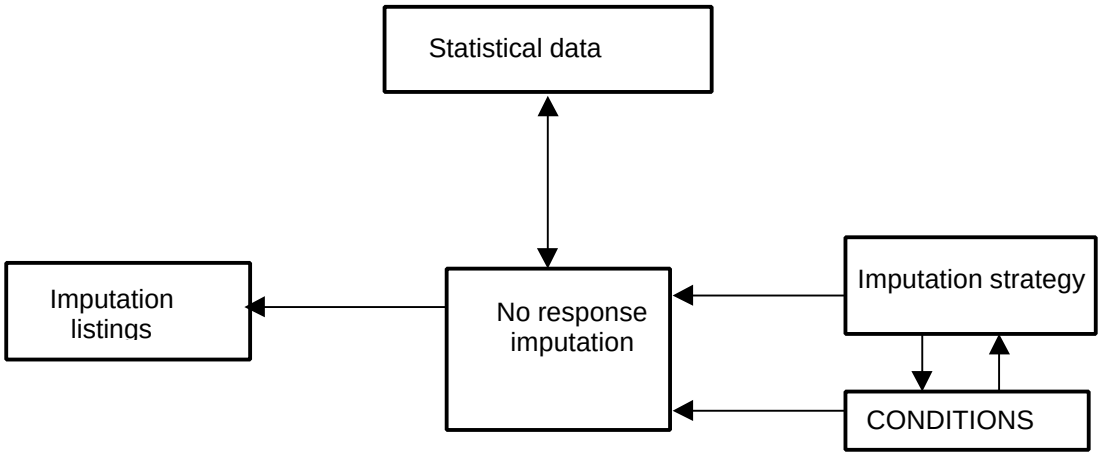
Chapter 5

No Response Imputation

A no response efficient treatment is basic for trend statistics since, usually, they are subject to deadlines in their publication. The imputation of the data not available is fundamental to obtain valid indexes in the required dates.

The know-how and innovation incorporation together with tests of new techniques have been the core of the imputation process definition strategy.

Graph 8: No response process imputation.



The definition process consisted of the application of several imputation methods to the different variables that build the indexes. Time series methods and traditional techniques based on the every sector growth coefficient calculation were applied.

The methods finally employed have been the ones that gave the best estimation according to the evaluation previously carried out.

The evaluation process designed was deleting the last observation available from every series. The values of this last observation were imputed using several techniques: linear adjustment, moving average method and the coefficient method.

Once the imputation was defined for every variable, the imputation strategy was to introduce it into the integrated system production of index numbers. (Graph 8)

The linear adjustment method estimates the following equation and uses it to extrapolate.

$$x_t = a + bt$$

The following step would be to impute the model in the period t+1. The adjustment is made by ordinary least squares.

In the moving average technique the OLS adjustment is made over the moving average series calculated from the original series. That is, for quarterly series those are calculated as:

$$y_t = \left(y_{t-2}/2 + y_{t-1} + y_t + y_{t+1} + y_{t+2}/2 \right) / 4$$

For monthly series the following model is used:

$$y_t = \left(y_{t-6}/2 + y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3} + y_{t+4} + y_{t+5} + y_{t+6}/2 \right) / 12$$

This method implies a shorter series of, at least, four observations less for quarterly models and twelve less for the monthly ones. This makes that to extrapolate further into the future more information is lost. Once the extrapolation and the estimation of y_{t+1} are made the same corresponding seasonal coefficient is applied to it. The seasonal coefficients of each period (month, quarter and so on) are estimated in the following way. For every observation of the series x_t having its corresponding moving average y_t, x_t / y_t coefficient is calculated. This is named seasonal relationship. We will obtain several seasonal relationships depending on the series length. For every period we will obtain several seasonal relationships, the more the better. An average is made to obtain the period seasonal coefficient.

The coefficient method consists of estimating a coefficient from all the changes that the sector, where the enterprise belongs to, has gone through.

Once the variable to be imputed is fixed, for every firm the following formulae is applied in t ,: $x_{tj}^k = x_{t-1j}^k * coefficient$, that is, the variable value of the previous period is multiplied by an incremental or decremental coefficient. This coefficient measures the increment variability within the firms in t.

The evaluation carried out helped to fix the criteria to determine the imputation implementation strategy for every index. The defined method is applied for each basic variable in each group of enterprises. The rest of the variables are estimated either calculating totals or distributing totals in their various components.

The distribution method takes into account the enterprise history using to distribute the total into several elements. To this end various previous distributions are considered and an average distribution is calculated. This is used in the period to be imputed and the distributed variables are calculated from the total variable value.

Let $vtot$ the variable to be distributed and $vdis(d) \quad d = 1, \ldots, D$, the variables to be distributed.

Let us assume that the firm is k and the total variable index is j, then we have

$$x_{tj}^k = vtot_t$$

Now if j is $vdis(d)$ index then

$$x_{tj}^{k} = vdis(d)_t$$

If num_per is the number of periods to consider the distribution of each of these previous periods is calculated as:

$$\forall m = t-1,\ldots,t-num\_per \;\; vector_m(d) = vdis(d)_m / vtot_m \;\; \forall d$$

And the average distribution among all of them is:

$$vector\_med(d) = \sum_m vector_m(d) \Big/ num\_per \;\; \forall d$$

Finally, the imputed data of distributed variables from the total variable value in the period to be imputed and the average distribution vector is

$$vdis(d)_t = vector\_med(d) * vtot_t \;\; \forall d$$

Chapter 5

Conclusions and Future Developments

As we have seen the introduction of automated processes has meant in the generation of index numbers and to consider further improvements we will have to turn our attention to data collection. Having in mind this goal the following step would be to use new technologies to collect data.

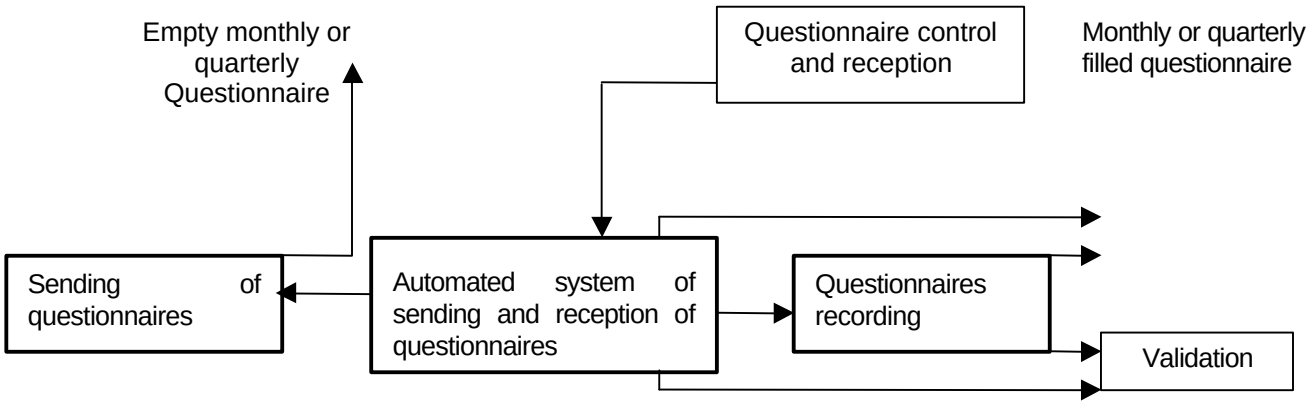
The introduction of Internet and electronic means in the enterprises and business has been gradual. In recent times many firms have sent the questionnaire via e-mail. This has provided their electronic address and, therefore, an automated system of sending and reception of questionnaires has been designed. The system itself generates the sending, reception and recording of the questionnaires. Several configuration options will help to determine the required parameters in all the questionnaires in the sending and reception management process.

There will be a continued updating of the firms' e-mail, the queries to enterprises and the control of the questionnaires sent and received.

The system includes the questionnaire generation in html format making and an encryption process. It also provides a way to create and send messages generating the data required (address, sender, subject, contents, and so on). Finally the system processes the messages received, verifies the information, makes the validation and records them as valid or erroneous and, also, manages the claming processes and the reminding messages in the case of received or no received questionnaires.

Other methods to improve the quality of the information through the implementation of new technologies are being investigated, among them the computerised or phone assisted interviews. Another alternative to collection via e-mail could be the auto filling via Internet. This can be done either on line or downloading the questionnaire and sending it once it is fulfilled.

Graph 9 Information collection phase diagram



There is a great deal of work to be done to have an integrated information system, but we are learning a lot about it through the implementation of automated processes in all trend statistics.

Chapter 6

References

[1] F. Azorín y J. L. Sánchez.

Métodos y aplicaciones del muestreo. Alianza Universidad textos.

[2] EUSTAT.

Índices de Producción Industrial base 1990. Working paper.

[3] EUSTAT

Índices de Precios Industriales base 1990. Working paper.

[4] EUSTAT.

Documento de Análisis de Sistemas del Índice de Producción e Índice de Precios. Abril 2000. Working paper.

[5] EUSTAT.

Documento de Análisis de Sistemas del Índice de Comercio Interior. Diciembre de 2000. Working paper

[6] J. Fourastie.

Análisis de series cronológicas: los índices estadísticos. Statistics International Seminar. 1985.

[7] Robert S. Pindyck and Daniel L. Rubienfeld.

Econometric Models

[8] R. Platek.

Métodos de Imputación. Statistics International Seminar. 1986

[9] C. Prado.

La Elaboración de Índices de Producción e Índices de Precios. Economiaz nº11. 1998

[10] Statistics Canada.

Monográficos especiales sobre la imputación. Tecniques d'Enquete. June-December. 1986.

[11] I.Villán and S. Bravo.

La imputación automática. Sistemas generales de depuración de datos. Statistics International Seminar 1990