Instituto Nacional de Estadística

AN EFFICIENT EDITING AND IMPUTATION STRATEGY WITHIN A CORPORATE-WIDE DATA COLLECTION SYSTEM AT INE SPAIN: A PILOT EXPERIENCE

R. López-Ureña, M. Mancebo, S. Rama and **David Salgado**

david.salgado.fernandez@ine.es

D.G. Methodology, Quality and ICT

Spanish National Statistical Institute

Paris, 24th April 2013

Instituto Nacional de Estadística

Some Preliminaries

- Main goal: **to streamline subprocess 5.3 of GSBPM** (Review, validate & edit, including editing during data collection (subprocesses 4.x)).
- We focus upon the **selection of questionnaires** (detection of errors) under two generic principles:
  - Editing must **minimize the amount of resources** deployed to recontacts, follow-ups and interactive tasks, in general.
  - **Data quality** must be **ensured**.
- **Design of E&I strategies**. Pilot experience with the **ITI and INORI survey**:
  - **Fixed panel** of 11000 (aprox.) industrial establishments selected by **cut-off**.
  - **Monthly** collected data through **CSAQ**, mail, email, fax and telephone at provincial delegations.
  - Laspeyres indices disseminated for **37 publications cells** (NACE Rev. 2). No geographical breakdown. Breakdown into **markets** (national, euro, noneuro, rest of the world).

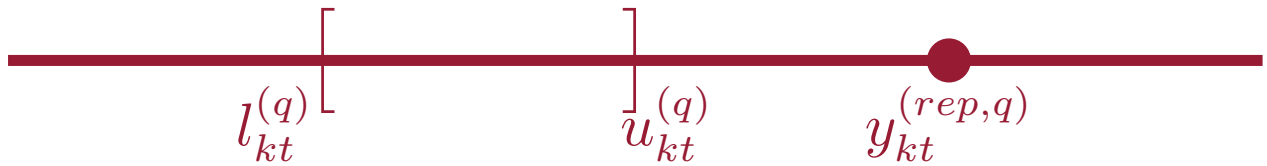
Instituto Nacional de Estadística

Editing Functions

- **Editing function**: type of **task** that has to be performed within a data editing process.
- The interaction between the **statistical methodology** and **information technologies** is fundamental.
- We incorporate this interaction in the design of an E&I strategy by choosing **standardizable editing functions**.
- As a first step in the transition to an industrialized production process, in the editing phase we have focused upon the **selection of questionnaires**.
- We distinguish three types of editing functions:
  - **survey-specific** functions (mainly format and balance edits);
  - **interval-distance** functions;
  - **distribution-angle** functions.

Interval-Distance Editing Function

- General idea: for each **variable of level** $y^{(q)}$ (total turnover and total new orders received in our survey)
  - we construct a **validation interval** for the reference period t for each respondent;
  - we measure the **distance** of the reported value to this interval;
  - we compare this distance with the **threshold** for the reference period t.
- Construction of the **validation interval** $I_{kt}^{(q)} = [l_{kt}^{(q)}, u_{kt}^{(q)}]$
  - $I_{kt}^{(q)} = [\hat{y}_{kt} - s_t \cdot \hat{\sigma}_{kt}, \hat{y}_{kt} + s_t \cdot \hat{\sigma}_{kt}], \quad s_t = \frac{1}{11} s_t^* + \frac{11}{12} s_{t-1},$ where $\hat{y}$ and $\hat{\sigma}$ denote **ARIMA predictions** and $s_t^* = \text{argmax}_s$ **HitRate**.
  - In case of short time series or too many missing/zero values, we use a **ratio edit**.



Instituto Nacional de Estadística

Interval-Distance Editing Function

- Construction of the **distance** $d(y_{kt}^{(rep,q)}, I_{kt}^{(q)})$
  - If the editing function is an **edit**

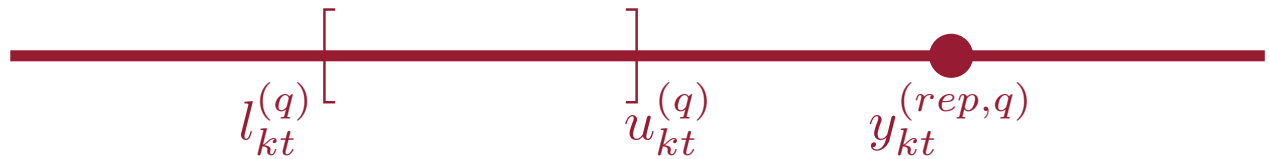
$$d(y_{kt}^{(rep,q)}, I_{kt}^{(q)}) = \begin{cases} 0 & \text{if } y_{kt}^{(rep,q)} \in I_{kt}^{(q)}, \\ \infty & \text{if } y_{kt}^{(rep,q)} \notin I_{kt}^{(q)}. \end{cases}$$

  - If the editing function is a **score function** and $y^{(q)}$ is **discrete**

$$d(y_{kt}^{(rep,q)}, I_{kt}^{(q)}) = \omega_k \begin{cases} 0 & \text{if } y_{kt}^{(rep,q)} \in I_{kt}^{(q)}, \\ y_{kt}^{(rep,q)} - u_{kt}^{(q)} & \text{if } y_{kt}^{(rep,q)} > u_{kt}^{(q)}, \\ l_{kt}^{(q)} - y_{kt}^{(rep,q)} & \text{if } y_{kt}^{(rep,q)} < l_{kt}^{(q)}. \end{cases}$$

  - If the editing function is a **score function** and $y^{(q)}$ is **continuous**

$$d(y_{kt}^{(rep,q)}, I_{kt}^{(q)}) = \omega_k \begin{cases} 0 & \text{if } y_{kt}^{(rep,q)} \in I_{kt}^{(q)}, \\ \frac{y_{kt}^{(rep,q)} - u_{kt}^{(q)}}{u_{kt}^{(q)} - l_{kt}^{(q)}} & \text{if } y_{kt}^{(rep,q)} > u_{kt}^{(q)}, \\ \frac{l_{kt}^{(q)} - y_{kt}^{(rep,q)}}{u_{kt}^{(q)} - l_{kt}^{(q)}} & \text{if } y_{kt}^{(rep,q)} < l_{kt}^{(q)}. \end{cases}$$



Instituto Nacional de Estadística

Interval-Distance Editing Function

- Construction of the **threshold** d_{jt}
  - Compute the distance $d_{k(t-1)} = d(y_{k(t-1)}^{(ed,q)}, I_{k(t-1)}^{(q)})$ between the **final edited values** and their corresponding validation intervals for the **preceding period** $t-1$ for **each unit** k.
  - Divide the sample s into J **minimal publication cells** $s = \bigcup_{j=1}^{J} s_j$.
  - For each domain s_j compute the **quantile** $q_j\left(\{d_{k(t-1)}\}_{k \in s_j}\right)$ over the distribution of distances. The quantile (1st quartile, pth percentile,...) is chosen by a **trade-off between cost and precision**.
  - The **threshold** for unit k is given by $d_{kt} = q_j\left(\{d_{k(t-1)}\}_{k \in s_j}\right)$ if $k \in s_j$.
- An establishment $k \in s_j$ is **flagged** for editing if

$$d(y_{kt}^{(rep,q)}, I_{kt}^{(q)}) > d_{jt}.$$

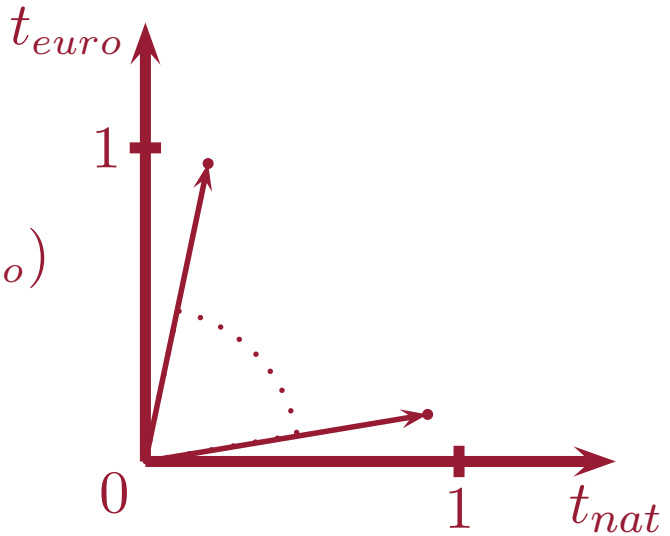
- **Standard input for a data collection application for each variable of level**:

$$l_{kt},\ u_{kt},\ \text{edit}_k\ (0,1),\ \text{continuous}_k\ (0,1),\ d_{kt}.$$

Distribution-Angle Editing Function

- General idea: for each set of **variables of distributions** $\{y^{(q_i)}\}$ (turnover and new orders received by markets in our survey)
  - we define a **vector** $\mathbf{y}_{kt}^{(q)} = \left(y_k^{(q_1)}, \ldots, y_k^{(q_I)}\right) / \sum_i y_k^{(q_i)}$;
  - we determine the **angle** of this vector respect to another ($\mathbf{y}_{k(t-1)}^{(q)}$, $\mathbf{y}_{kt}^{(\tilde{q})}$, etc.);
  - we compare this angle with the **threshold** for the reference period t.
- The angle is trivially computed (**scalar product**).
- The thresholds are determined as **quantiles** over the distribution of angles over each **minimal publication cell**.

$$\mathbf{T} = \frac{(T_{nat}, T_{euro})}{T_{nat}+T_{euro}} = (t_{nat}, t_{euro})$$



Instituto Nacional de Estadística

Macro Editing Phase

- Mathematical translation of
  - Editing must **minimize the amount of resources** deployed to recontacts, follow-ups and interactive tasks, in general.
  - **Data quality** must be **ensured**.
- **Optimization problem**:

  minimize number of questionnaires to edit interactively

  s.t. estimated **mean squared error** of $y^{(q)} \leq \text{bound}^{(q)} \quad p = 1, \ldots, P$

- For editing **field work** considerations, instead of a selection, a **prioritization** of units is determined by concatenating a sequence of optimization problems. This prioritization is carried out for **each publication cell**.
- A fixed number n_{macro} of questionnaires is further edited. These n_{macro} units are **allocated** among the publication cells **proportional to** the estimated **mean squared error**, to the **weights** of the cells within the global index, to the proportion of **questionnaires reporting zero turnover** and to the proportion of **imputed questionnaires** in the preceding time period having reported zero turnover.

Instituto Nacional de Estadística

New E&I Strategy

- **CAWI** mode and editing at **provincial delegations**
  - Editing functions as **edits** (CAWI)/**score functions** (Prov. Del.).
  - **Total turnover** and **total new orders received** controlled by **interval-distance** functions.
  - **Turnover breakdown** controlled by **distribution-angle** with respect to the **preceding** time period.
  - **New orders received breakdown** controlled by **distribution-angle** with respect to **turnover breakdown**.
- Editing at the **central office**.
  - $n_{macro} = 100$.
  - The prediction model is the best among 4 **simple time series models**.
  - The observation model considers the **occurrence of error** as a **Bernoulli variable** whose value in the positive case follows a **normal distribution**.

Instituto Nacional de Estadística

Some conclusions

- **Simulations** have been carried out with **real data** from 13 consecutive months. While maintaining **nearly the same precision**, the **interactive editing rate** has decreased from 55% in the traditional strategy to $15\% - 20\%$ in the proposed strategy.
- This strategy has been applied in **real production conditions** in January 2013 (reference month).
- Preliminary data suggest that simulations were **too optimistic** (interactive editing rate $\approx 30\% - 35\%$). The simulation of the **respondent behaviour during the CAWI is crucial**.
- The distribution-angle editing function can be **reformulated** as an interval-distance editing function.
- The interval construction scheme can be **adapted** to more common sampling designs (rotating panel with stratified random sampling, ...) by (i) **aggregating** units into homogeneous domains and (ii) using **simpler time series models** (random walks, etc.).
- More implementations are **currently under development**.