

# **Introducción a la Depuración**

## **Curso Selectivo 2016**

Miguel Á. Martínez Vidal y David Salgado

Dpto. Metodología y Desarrollo de la Producción Estadística

27 Junio, 2016



- **Cuestiones generales**

- Definiciones
- Estándares
- Métodos tradicionales

- **Diseño** de estrategias de depuración

- Función de producción
- Comunicación de la estrategia

- Fase **longitudinal**

- Construcción de controles If-Then
- Construcción de controles Interval-Distance

- Fase **transversal**

- Principios generales
- Modelo de observación-predicción para variables continuas
- Ordenación y afijación de unidades

- Trabajos en curso y a la espera

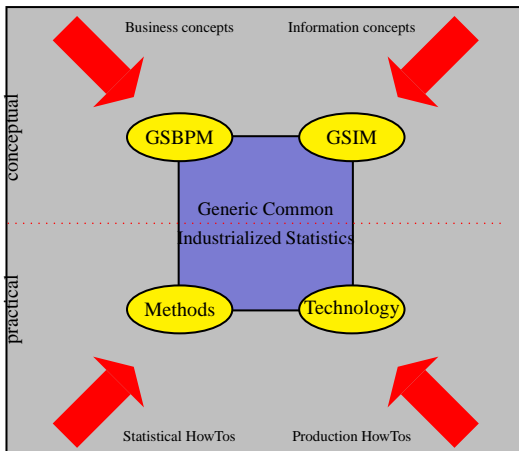
# Visión estratégica del HLG

Cuestiones  
generales

Diseño

Longitudinal

Transversal



# Definiciones de Depuración de Datos

- **U.S. Federal Committee on Statistical Methodology (1990)**

Procedure(s) designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much of the erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures.

- **United Nations Economic Commission for Europe (1994)**

An activity aimed at the acquirement of data which meet certain requirements.

- **United Nations Economic Commission for Europe (1997)**

Procedure for detecting, by means of edit rules, and for adjusting, manually or automatically, errors resulting from data collection or data capture.

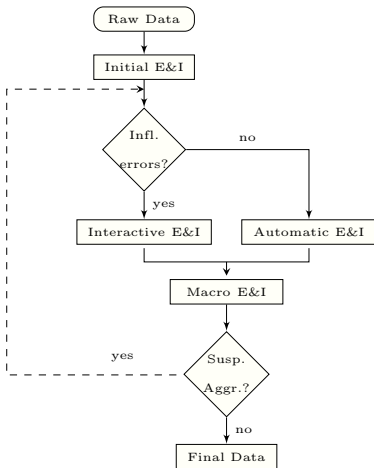
Cuestiones  
generales

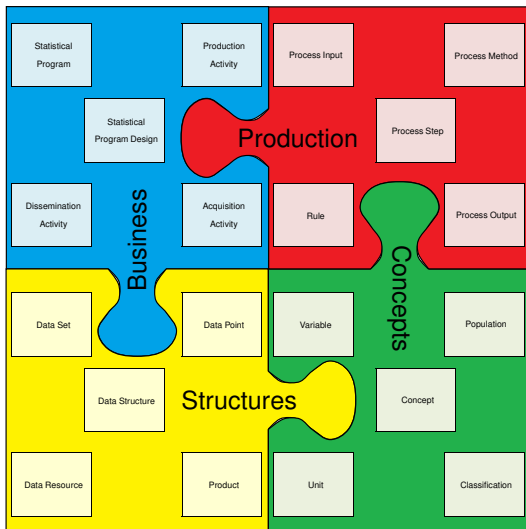
Diseño

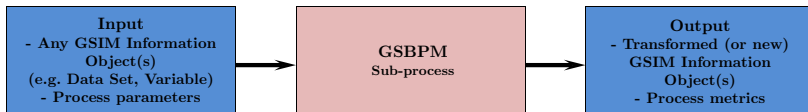
Longitudinal

Transversal

Quality Management / Metadata Management							
Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objective	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review& validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.1 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			









Conjunto de **tareas**, generalmente basadas en **métodos estadísticos** y **metodología** de encuestas, necesarias para la **ejecución** de un proceso o subproceso de producción.

- Conjunto(s) de **datos de entrada**;
- **Parámetro(s)** de entrada;
- Descripción de las **tareas** a ejecutar;
- Conjunto(s) de **datos de salida**;
- **Métrica(s)** del proceso.

- Depuración **interactiva**

- Requiere la intervención de un experto en la materia;
- Asistida por software específico de depuración;
- A menudo conlleva el recontacto con el informante;
- Alcanza grandes cotas de calidad en los datos depurados.

- Depuración **automática**

- Efectuada por ordenador sin la intervención humana;
- Más económica que la depuración interactiva;
- Más rígida que la depuración interactiva.

- Depuración **selectiva**

- Identifica errores influyentes: aquellos con mayor impacto en los agregados resultantes (índices, estimaciones. . . );
- Optimiza los recursos financieros y humanos, redireccionándolos hacia una mejora de la calidad;
- Se emplea para datos cuantitativos.

- **Macro**-depuración

- Examina los agregados (índices, estimaciones. . . ) para identificar microdatos sospechosos;
- Emplea sobre todo técnicas de análisis exploratorio de datos;
- Necesita que al menos una fracción significativa de los datos estén recogidos.

Esta técnica otorga una puntuación a cada cuestionario, obtenida en cuatro etapas:

- (i) Cálculo de valores **anticipados**  $\hat{y}_k$ .
- (ii) Cálculo de las funciones **score locales**  $s_k$ .
- (iii) Cálculo de la función **score global**  $S_k$ .
- (iv) Determinación del valor **umbral**  $t_k$ .

Los valores anticipados  $\hat{y}_k$  son, en general, **predicciones** de **baja calidad** (en comparación con p.ej. la imputación).

En encuestas periódicas suele emplearse el valor depurado final del último período

$$\hat{y}_k^{(t)} = y_k^{(*,t-1)},$$

o en presencia de estacionalidad, el valor depurado del último período estacional

$$\hat{y}_k^{(t)} = y_k^{(*,t-s)}.$$

Cuestiones  
generales

Diseño

Longitudinal

Transversal

La **estructura genérica** de una función *score* para la unidad  $k$  y una variable cuantitativa  $y$  es

$$s_k(y_k, \hat{y}_k) = F_k(y_k, \hat{y}_k) \times R_k(y_k, \hat{y}_k),$$

donde  $F_k$  es la componente de **influencia** y  $R_k$  es la componente de **riesgo**.

- La componente de influencia  $F_k$  mide la **influencia o contribución de la unidad  $k$  al agregado estimado**.
- La componente de riesgo  $R_k$  mide la **dimensión y probabilidad (*likelihood*) del error potencial cometido por la unidad  $k$** .

La forma más general de construir una función *score* global es emplear **funciones (ponderadas) de Minkowski**:

$$S_k^{(\alpha)} = \left( \sum_{p=1}^P w_p [s_k^{(p)}]^\alpha \right)^{1/\alpha},$$

$$S_k^{(\infty)} = \max \left\{ w_1 s_k^{(1)}, \dots, w_P s_k^{(P)} \right\}, \quad w_p \geq 0.$$

Cuestiones  
generales

Diseño

Longitudinal

Transversal

- Dada la función *score* global  $S_k^{(\alpha)}$ , debemos determinar el **umbral**  $C_0$ :

La unidad  $k$  se depura si, y sólo si,  $S_k^{(\alpha)} \geq C_0$ .

- Se realiza un **estudio de simulación** sobre datos depurados (interactivamente) y no depurados en un período anterior:
  - Se determina el valor de  $S_k^{(\alpha)}$  **para cada unidad**  $k$ .
  - Se fija un primer valor alto a priori de  $C_0^{(1)}$  y en todas las unidades  $k$  con  $S_k^{(\alpha)} \geq C_0^{(1)}$  se sustituyen los valores no depurados de las variables por los correspondientes depurados interactivamente. Se obtiene así un **conjunto**  $E(C_0^{(1)})$  **de datos depurados selectivamente**.



- – Se calculan las **estimaciones** empleando este conjunto de datos.
- Se escoge un nuevo valor  $C_0^{(2)}$ , se obtienen un nuevo conjunto de datos  $E(C_0^{(2)})$  y nuevas estimaciones correspondientes.
- Se repite este procedimiento varias veces  $s = 1, 2, \dots, S$ .
- Se tabulan o representan los resultados y a la vista se escoge el valor de  $C_0$  más conveniente (**juicio del experto**) para el período en curso.

## Función **RATIO**

$$1. r_k^{(t)} = \frac{y_k^{(t)}}{y_k^{(*,t-1)}} \rightsquigarrow \text{Me}(r_k^{(t-1)}) : \text{Mediana de } r_k^{t-1}.$$

$$2. \bar{r}_k^{(t)} = \begin{cases} \left| \frac{r_k^{(t)}}{\text{Me}(r_k^{(t-1)})} - 1 \right| & \text{si } r_k^{(t)} > \text{Me}(r_k^{(t-1)}), \\ \left| 1 - \frac{\text{Me}(r_k^{(t-1)})}{r_k^{(t)}} \right| & \text{si } r_k^{(t)} \leq \text{Me}(r_k^{(t-1)}). \end{cases}$$

$$3. g_k^{(t)} = d_k^{(t)} \times \bar{r}_k^{(t)} \times \left( \max\{y_k^{(t-1)}, y_k^{(*,t)}\} \right)^U, \quad U \in [0, 1].$$

$$4. s_k^{(t)} = \frac{|g_k^{(t)} - \text{Me}(g_k^{(t-1)})|}{\text{IR}(g_k^{(t-1)})}, \text{ donde } \text{IR}(g_k^{(t-1)}) : \text{recorrido intercuartílico de } g_k^{(t-1)} \rightsquigarrow \text{Para cada variable } y^{(p)}: s_k^{(p,t)}.$$

$$5. \text{RATIO}_k^{(t)} = S_k^{(1,t)} = \sum_{p=1}^P w_p^{(t)} s_k^{(p,t)}, \text{ donde } w_p^{(t)} \geq 0.$$

Cuestiones  
generales

Diseño

Longitudinal

Transversal

## Función **FLAG**

1. Escoger a juicio del experto la variable  $p^*$  más importante.
2.  $\text{FLAG}_k^{(t)} = d_k^{(t)} \left( \max\{y_k^{(p^*, t)}, y_k^{(*, p^*, t-1)}\} \right)^U w_{p^*}^{(t)}$ ,  $U \in [0, 1]$ ,  $w_{p^*}^{(t)} \geq 0$ .

Cuestiones  
generales

Diseño

Longitudinal

Transversal

## Función **DIFF**

1.  $\hat{Y}^{(p, t-1)} = \sum_{k \in S} \omega_{ks}^{(t-1)} y_k^{(*, p, t-1)}$ .
2.  $\text{DIFF}_k^{(t)} = \sum_{p=1}^P w_p^{(t)} \times \frac{d_k^{(t)} |y_k^{(p, t)} - y_k^{(*, p, t-1)}|}{\hat{Y}^{(p, t-1)}}$ .

Examinar el **impacto potencial** sobre agregados para identificar datos sospechosos en unidades.

Puede verse como una **forma diferente de depuración selectiva**.

- La gran diferencia radica en el **momento** del proceso de depuración en el que se aplica.
- Es necesario disponer de la **gran mayoría de datos** ya recogidos para aplicar estas técnicas.
- Trata el **conjunto de datos como un todo**, frente a otros enfoques que trabajan registro a registro.

Conduce en general a una **notable reducción del trabajo** de depuración sin disminución de la calidad de las estimaciones.

## Métodos de **agregación**.

- Método de agregación (*aggregate method*).
- Método *top-down*.

## Métodos de **distribución**.

- Método *box-plot*.
- Técnicas de análisis exploratorio de datos.

“[...] it seems desirable, to the extent feasible, to avoid estimates or inferences that need to be defended as judgments of the analysts conducting the survey.”

(Hansen, Madow and Tepping, 1983).

Cuestiones  
generales

Diseño

Longitudinal

Transversal

**Información disponible** para la ejecución de tareas:

- **Longitudinal** ✓
- **Multivariante** ✓
- **Transversal**

$$\left\{ \begin{array}{l} n_{col} : \text{Número de unidades recogidas} \\ n_{cyc}^{(k)} : \text{Número de ciclos de depuración } k \\ n_{cyc} : \text{Número de ciclos de la muestra} \end{array} \right.$$

Cada función se especificará como

$$\text{FunctionName}(n_{col}, n_{cyc}^{(k)}, n_{cyc})$$

$$\text{FunctionName}(n_{col}, n_{cyc})$$

1. Depuración durante la **recogida** de datos

$$\text{EditColl}(n_{col} \leq n, n_{cyc}^{(k)} = 0, n_{cyc} = 0)$$

2. **Selección** de unidad

$$\text{UnitSelection}(n_{col} \leq n, n_{cyc}^{(k)}, n_{cyc})$$

3. Depuración e imputación **interactivas**

$$\text{InterEI}(n_{col} \leq n, n_{cyc}^{(k)}, n_{cyc})$$

4. Depuración e imputación **automáticas**

$$\text{AutoEI}(n_{col} \leq n, n_{cyc}^{(k)}, n_{cyc})$$



## 5. Validación de la muestra

$$\text{Validation}(n_{col} = n, n_{cyc})$$

Cuanto menor sea el valor de  $n_{cyc}^{(k)}$  y  $n_{cyc}$ , más eficiente serán las funciones.

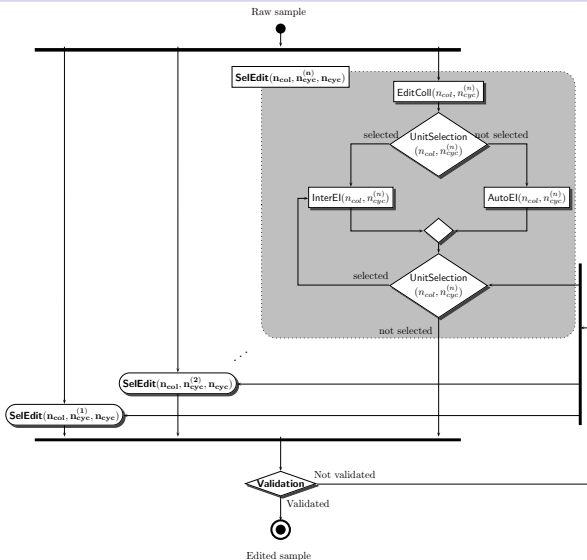
La tecnología tiene consecuencias directas en el diseño de las funciones  $\rightsquigarrow n_{col}$ .

Cuestiones  
generales

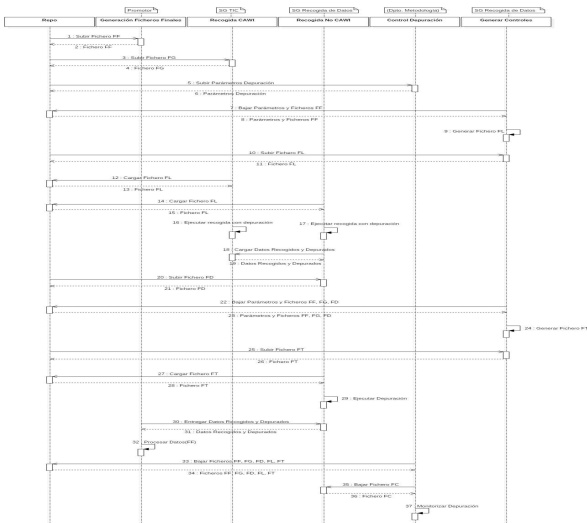
Diseño

Longitudinal

Transversal



# Implementación: el repositorio



El principio **safety-first** no es válido: **sobredepuración**

Eliminar *edits* que producen **avisos innecesarios**

Basarse en estadísticos a partir de los datos, no subjetivos:  
**depuración creativa**

Procedimiento genérico:

1. *Edits* que garantizan la **consistencia lógica** del cuestionario  $\rightsquigarrow$  **Duros**: Formato, Rango, Balance . . .
2. *Edits* que garantizan la “**consistencia estadística**” del cuestionario  $\rightsquigarrow$  **Blandos**
3. *Edits* **multivariantes** mejor que univariantes

Básicamente, la estrategia es un **conjunto** de *edits*

Edits: Forma **If-Then generalizada**

$$\text{IF } C(\mathbf{y}_{[t]}^{(k)}, \mathbf{x}_{[t]}^{(k)}) \text{ THEN } d_t^{(k)} \left( f(\mathbf{y}_t^{(k)}, \mathbf{z}_t^{(k)}), [l_t^{(k)}, u_t^{(k)}] \right) \leq u_t^{(k)}$$

Control	Variable	Cond	LimInf	LimSup	TipoD	Umbral
Nombre <i>edit</i>	$f(\mathbf{y}, \mathbf{z})$	$C(\mathbf{y}_{[t]}^{(k)}, \mathbf{x}_{[t]}^{(k)})$	$l_t^{(k)}$	$u_t^{(k)}$	$d_t^{(k)}$	$u_t^{(k)}$
Mensaje:	<i>Mensaje de texto para informante o agente.</i>					
	Modos de recogida		TipoE	Plataformas		
	CAWI, Resto		Duro, Blando	IRIA, GRECO		
Observaciones:	<i>Texto con observaciones.</i>					

En encuestas **económicas coyunturales**, se ha escogido la función *score* **global** para las fases de recogida y longitudinal

$$S^{(\infty)} = \max .$$

Además, cada *edit* **activo**, ya sea duro o blando, conlleva la activación del correspondiente **mensaje**.

Cuestiones  
generales

Diseño

Longitudinal

Transversal

Cuestiones  
generales

**Diseño**

Longitudinal

Transversal



IF  $C_1(\mathbf{y}_k, \mathbf{x}_k)$  THEN  $C_2(\mathbf{y}_k, \mathbf{z}_k)$

Son muy versátiles: modos CA y post-captura.

A priori, las condiciones  $C_1$  y  $C_2$  son arbitrarias, pero su uso puede estar limitado por la disponibilidad de las variables auxiliares  $\mathbf{x}$  y  $\mathbf{z}$ .

Es mejor comprobar la condición  $C_1$  cuanto antes, si es posible.



$$\text{IF } C_1(\mathbf{y}_k, \mathbf{x}_k) \text{ THEN } d_k(z_k, [l_k, u_k]) \leq \xi_k$$

Puede entenderse como un caso **particular** del control IF-THEN.

Los parámetros se determinan para **cada unidad**  $k$  y **cada período de recogida**.

Cuestiones  
generales

Diseño

Longitudinal

Transversal

# Construcción de intervalos

Los detalles metodológicos según **disponibilidad** de datos.

Objetivo: control tanto de la **precisión** como del **trabajo de campo**.

Opción 1: Predicción del **centro** y determinación del **radio**.

- Se **predice** un valor  $\hat{c}_k^{(q, T+\tau)}$  para el **centro** del intervalo mediante un modelo de series temporales, regresión, etc.
- Se **determina** un valor  $\hat{r}_k^{(q, T+\tau)}$  para el **radio** del intervalo.

$$I_k^{(q, T+\tau)} = [\hat{c}_k^{(q, T+\tau)} - \hat{r}_k^{(q, T+\tau)}, \hat{c}_k^{(q, T+\tau)} + \hat{r}_k^{(q, T+\tau)}]$$

Opción 2: **Predicción de ambos extremos**.

- Se **predicen** sendos valores  $\hat{l}_k^{(q, T+\tau)}, \hat{u}_k^{(q, T+\tau)}$  para los **extremos**.

$$I_k^{(q, T+\tau)} = [\hat{l}_k^{(q, T+\tau)}, \hat{u}_k^{(q, T+\tau)}]$$

En caso de paneles **fijos**:

- ajustar sendos **modelos de series temporales** (ARIMA, random walk. . . ) sobre las variables  $z_k^{(q,t)}$  para cada informante  $k$ ;
- el centro de cada intervalo viene dado por la **predicción**

$$\hat{c}_k^{(q)} = \hat{z}_k^{(q, T+\tau)}.$$

Cuestiones  
generales

Diseño

**Longitudinal**

Transversal

# Opción 1: Predicción del Centro

En caso de paneles **rotantes** y diseños muestrales con **baja permanencia**:

- agregar informantes en **dominios**  $U_i$ ;
- calcular **ratios**  $R_k^{(q,t)} = \frac{z_k^{(q,t)}}{z_k^{(q,t-\tau)}}$  cuando sea posible;
- calcular una **medida de posición**  $\mu_i^{(q,t)}$  (media, mediana, media recortada. . . ) de los ratios  $R_k^{(q,t)}$  en cada dominio  $U_i$ ;
- ajustar un **modelo de series temporales** (ARIMA, random walk. . . ) sobre la variable  $\mu_i^{(q,\cdot)}$  para cada dominio  $U_i$ ;
- el centro de cada intervalo viene dado por la **predicción**

$$\hat{c}_k^{(q,T+\tau)} = \hat{\mu}_i^{(q,T+\tau)} \times z_k^{(q,T)}.$$

En caso de paneles **fijos**:

- ajustar sendos **modelos de series temporales** (ARIMA, random walk. . . ) sobre las variables  $z_k^{(q,\cdot)}$  para cada informante  $k$ ;
- estimar **medidas de dispersión** (desviaciones típicas. . . )  $\hat{s}_k^{(q,T+\tau)}$  de la variable  $z_k^{(q,\cdot)}$ ;
- el **radio**  $\hat{r}_k^{(q,T+\tau)}$  viene dado por  $\hat{r}_k^{(q,T+\tau)} = \beta^{(q,T)} \hat{s}_k^{(q,T+\tau)}$ , donde  $\beta^{(q,T)}$  optimiza algún **indicador global de la calidad** de la depuración en el período anterior (*Hit Rate*, . . . ).

# Opción 1: Determinación del Radio

En caso de paneles **rotantes** y diseños muestrales con **baja permanencia**:

- agregar informantes en **dominios**  $U_i$  (pref. celdas de publicación minimales);
- calcular **ratios**  $R_k^{(q,t)} = \frac{z_k^{(q,t)}}{z_k^{(q,t-\tau)}}$  cuando sea posible;
- calcular una **medida de dispersión**  $s_i^{(q,t)}$  (desviación típica, recorrido intercuartílico. . . ) de los ratios  $R_k^{(q,t)}$  en cada dominio  $U_i$ ;
- ajustar un **modelo de series temporales** (ARIMA, random walk. . . ) sobre la variable  $s_i^{(q,\cdot)}$  para cada dominio  $U_i$ ;
- el **radio** viene dado por  $\hat{r}_i^{(q,T+\tau)} = \beta^{(q,T)} \hat{\sigma}_k^{(q,T+\tau)} z_k^{(q,T)}$ , donde  $\beta^{(q,T)}$  optimiza algún **indicador global de la calidad** de la depuración en el período anterior (*Hit Rate*, *CED*. . . ).

## Opción 2: Construcción del Intervalo

Ajustar un **modelo de series temporales** sobre  $z_k^{(q,t)}$  para cada unidad  $k$ .

Calcular sendas **predicciones**  $\hat{z}_k^{(q,T+\tau)}$ .

Agregar informantes en **dominios**  $U_i$ .

Calcular **cuantiles**  $c_{j_1,i}^{(q,t)}$  y  $c_{j_2,i}^{(q,t)}$  ( $j_1 < j_2$ ) de las variables en cada dominio  $U_i$ .

Ajustar un **modelo de series temporales** (ARIMA, random walk. . . ) sobre  $c_{j_h,i}^{(q,\cdot)}$  para cada dominio  $U_i$ .

Calcular sendas **predicciones**  $\hat{c}_{j_h,i}^{(q,T+\tau)}$ .

Los extremos vienen dados por

$$\begin{aligned}\hat{l}_k^{(q,T+\tau)} &= (1 - \lambda) \cdot \hat{c}_{j_1,i}^{(q,T+\tau)} + \lambda \cdot \hat{z}_k^{(q,T+\tau)} & k \in U_i, \\ \hat{u}_k^{(q,T+\tau)} &= (1 - \lambda) \cdot \hat{c}_{j_2,i}^{(q,T+\tau)} + \lambda \cdot \hat{z}_k^{(q,T+\tau)} & k \in U_i.\end{aligned}$$

## Opción 2: Construcción del Intervalo

Dos modelos de predicción para cualesquiera series temporales  $\{y_k^{(t)}\}_{k \in s}^{t=1, \dots, T}$ :

1. Calcular **ratios**  $R_k^{(t)} = \frac{y_k^{(t)}}{y_k^{(t-\tau)}}$  cuando sea posible.

Agregar informantes en **dominios**  $U_i$ .

Calcular **cuantiles**  $c_{j,i}^{(\text{rat}, t)}$  de los ratios en cada dominio  $U_i$ .

Ajustar un **modelo de series temporales** (ARIMA, random walk. . .) sobre  $c_{j,i}^{(\text{rat}, \cdot)}$  para cada dominio  $U_i$ .

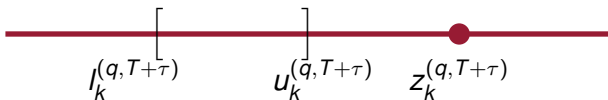
Calcular sendas **predicciones**  $\hat{c}_{j,i}^{(\text{rat}, T+\tau)}$ .

Las predicciones vienen dadas por

$$\hat{y}_k^{(T+\tau)} = \hat{c}_{j,i}^{(\text{rat}, T+\tau)} \times y_k^T.$$

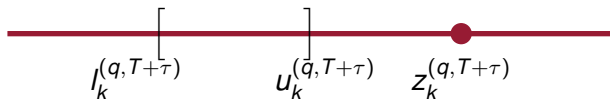
2. Ajustar un **modelo de series temporales** (ARIMA, random walk. . .) directamente sobre cada  $y_k^{(t)}$ .





Distancia Tipo 1 (**edit** tradicional)

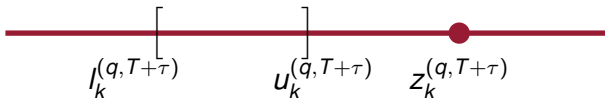
$$d(z_k^{(q,t)}, l_k^{(q,t)}) = \begin{cases} 0 & \text{if } z_k^{(q,t)} \in l_k^{(q,t)}, \\ \infty & \text{if } z_k^{(q,t)} \notin l_k^{(q,t)}. \end{cases}$$



Distancia Tipo 2 (distancia geométrica ponderada: función **score**)

$$d(z_k^{(q,t)}, l_k^{(q,t)}) = \omega_k \times \begin{cases} 0 & \text{if } z_k^{(q,t)} \in l_k^{(q,t)}, \\ z_k^{(q,t)} - u_{kt}^{(q)} & \text{if } z_k^{(q,t)} > u_{kt}^{(q)}, \\ l_{kt}^{(q)} - z_k^{(q,t)} & \text{if } z_k^{(q,t)} < l_{kt}^{(q)}. \end{cases}$$

# Construcción de la distancia



Distancia Tipo 3 (función **score**, especialmente útil para variables **continuas**)

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

Calcular las **series temporales de intervalos**  $I_k^{(q,t)}$ .

Calcular las **distancias**  $d_k^{(q,t)}$  entre los valores depurados  $z_k^{(q*,t)}$  y sus correspondientes intervalos para cada cuestionario  $k$ .

Dividir las muestras  $s_t$  en **dominios**  $s_t = \bigcup_{i=1}^I s_{ti}$ .

Para cada celda  $s_{ti}$  calcular la serie de **cuantiles**  $c_j^{(q,t)}$  sobre la distribución de distancias. Los cuantiles  $c_j$  se escogen buscando un equilibrio entre (i) la **precisión** y (ii) el **coste** y la **carga al informante**.

Calcular las **predicciones**  $\hat{c}_i^{(q,T+\tau)}$  para cada dominio  $s_{ti}$ .

Calcular las **predicciones**  $\hat{d}_k^{(q,T+\tau)}$  para cada unidad  $k$ .

El umbral para cada cuestionario  $k$  viene dado por

$$\xi_k^{(q,T+\tau)} = (1 - \lambda_k) \cdot \hat{c}_i^{(q,T+\tau)} + \lambda_k \cdot \hat{d}_k^{(q,T+\tau)}, \quad k \in s_i^{(T+\tau)}.$$

Si  $d = d_1$ ,

$$\bar{l}_k = [l_k, u_k].$$

Si  $d = d_2$ ,

$$\bar{l}_k = \left[ l_k - \frac{\xi_k}{\omega_k}, u_k + \frac{\xi_k}{\omega_k} \right].$$

Si  $d = d_3$ ,

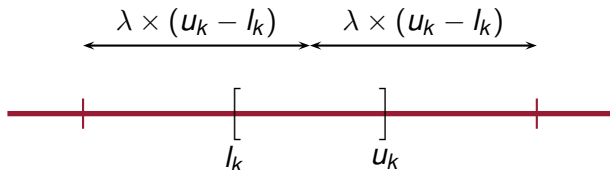
$$\bar{l}_k = \left[ l_k - \frac{\xi_k}{\omega_k} (u_k - l_k), u_k + \frac{\xi_k}{\omega_k} (u_k - l_k) \right].$$

Cuestiones  
generales

Diseño

Longitudinal

Transversal



Para distancia 2 :  $\xi_k \rightarrow \omega_k \times \min \left( \frac{\xi_k}{\omega_k}, (u_k - l_k) \cdot \left( \lambda - \frac{1}{2} \right) \right)$

Para distancia 3 :  $\xi_k \rightarrow \omega_k \times \min \left( \frac{\xi_k}{\omega_k}, \lambda - \frac{1}{2} \right)$

Cuando no hay disponibilidad de datos (series demasiado cortas, falta de respuesta. . . ), dependemos de las **circunstancias**.

Como regla global, para aquellos cuestionarios  $k$  en que no puedan encontrarse  $\hat{\gamma}_k^{(q, T+\tau)}$ ,  $\hat{u}_k^{(q, T+\tau)}$  puede

- **imputarse** por la media o algún cuantil;
- encontrarse un **donante**;
- emplearse otro procedimiento *ad hoc*.

De la variable **benchmark**  $X \rightsquigarrow F_{X,s_i}^*$

De la variable a imputar  $Z \rightsquigarrow F_{Z,r_i}^*$ ,  $s_i = r_i \cup \bar{r}_i$

Los valores **ausentes**  $z_k$  se imputan mediante:

$$z_k = \left( F_{Z,r_i}^{*-1} \circ F_{X,s_i}^* \right) (x_k)$$

Como variable *benchmark* puede emplearse una **predicción**.



Modelo aditivo de error:

$$y_k^*(\mathbf{r}) = y_k^0 + r_k \cdot (y_k - y_k^0).$$

Dos principios genéricos:

- editing must **minimize** the amount of resources deployed to **interactive tasks**:

$$\min \sum_{k \in s} (1 - r_k) \Rightarrow \max \sum_{k \in s} r_k$$

- data **quality** must be **ensured**:

$$\text{MSE} \left( \hat{Y}^*(\mathbf{r}) \right), |\hat{Y}^*(\mathbf{r}) - Y^0|, L \left( \hat{Y}^*(\mathbf{r}), Y^0 \right)$$

Con  $L(a, b) = |a - b|$  y explotando información **transversal** de la muestra:

$$\begin{aligned} & [P_{co}(\tilde{\eta}, \Omega_0)] \quad \max \mathbf{1}^T \mathbf{r} \\ \text{s.t.} \quad & \mathbf{r}^T M^{(q)} \mathbf{r} \leq \tilde{\eta}_q, \quad q = 1, 2, \dots, Q, \\ & \mathbf{r} \in \Omega_0 \subset [0, 1]^{\times n}, \end{aligned}$$

donde

$$M_{kk}^{(q)} = \mathbb{E}_m \left[ \omega_{ks} \cdot |y_k^{(q)} - y_k^{(0,q)}| | \mathbf{z}_k^{cross} \right].$$

$M_{kk}^{(q)} \rightsquigarrow$  una función *score* local de la variable  $y^{(q)}$ .

Modelo de **observación**  $y_k^{obs} = y_k^0 + \epsilon_k^{obs}$

Modelo de **predicción**  $y_k^0 = \hat{y}_k + \epsilon_k^{pred}$ ,  $\xi \rightsquigarrow \hat{y}_k$

**Especificaciones:**

1.  $\epsilon_k^{obs} = e_k \cdot \delta_k^{obs}$ .
2.  $e_k \simeq Be(p_k)$ , where  $p_k \in (0, 1)$ .
3.  $\left( \epsilon_k^{pred}, \delta_k^{obs} \right) \simeq N \left( \mathbf{0}, \begin{pmatrix} \nu_k^2 & 0 \\ 0 & \sigma_k^2 \end{pmatrix} \right)$ .
4.  $\epsilon_k^{pred}$ ,  $\delta_k^{obs}$  and  $e_k$  are jointly independent of  $\mathbf{Z}_k^{cross}$ .
5.  $e_k$  is independent of  $\epsilon_k^{pred}$  and  $\delta_k^{obs}$ .

$$M_{kk}(y) = \sqrt{\frac{2}{\pi}} \cdot \omega_{ks} \cdot \nu_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; -\frac{(y_k - \hat{y}_k)^2}{2\nu_k^2}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}, p_k, \sigma_k\right),$$

$$\text{donde } \zeta_k(x, p_k, \sigma_k) = \frac{1}{1 + \frac{1-p_k}{p_k} \left(\frac{\nu_k^2}{\sigma_k^2 + \nu_k^2}\right)^{-1/2} \exp\left(-\frac{1}{2} \frac{\sigma_k^2}{\sigma_k^2 + \nu_k^2} x^2\right)}.$$

Cuestiones  
generales

Diseño

Longitudinal

Transversal

Se estiman los parámetros mediante la serie histórica del doble conjunto de datos brutos y depurados:

$$\hat{M}_{kk}(y) = \sqrt{\frac{2}{\pi}} \cdot \omega_{ks} \cdot \hat{\nu}_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; -\frac{(y_k - \hat{y}_k)^2}{2\hat{\nu}_k^2}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\hat{\nu}_k}, \hat{p}_k, \hat{\sigma}_k\right).$$

- Se ordena por el valor de la función *score* global

$$S_k = S \left( \hat{M}_{kk}^{(1)}, \dots, \hat{M}_{kk}^{(Q)} \right).$$

- Escoger  $S$  es equivalente a (i) escoger una **sucesión de cotas**  $\eta_i$ , (ii) **fijar**  $i$  e (iii) **iterar**  $i \leftarrow i + 1$ .
- En la práctica hemos escogido

$$\tilde{S}_k^{(\alpha)} = \left( S^{(\alpha)} \circ \left( F_{\text{diag}(\hat{M}^{(1)})}^*, F_{\text{diag}(\hat{M}^{(2)})}^* \right) \right) \left( \hat{M}_{kk}^{(1)}, \hat{M}_{kk}^{(2)} \right),$$

con

- $S = S^{(\alpha)}$  funciones *score* minkowskianas;
- $F^*$  funciones de distribución empíricas.

Cuestiones  
generales

Diseño

Longitudinal

Transversal

# Algoritmo de afijación

Problema **multivariante**: repartir  $n_{cross}$  unidades en dominios  $U_i$ .

Sea  $n_i$  el número de unidades de la celda  $i$ :

- $n_i^{(0)} \leftarrow n_{i0}$ , con  $n_{im} \leq n_{i0} \leq n_{iM}$ .
- $n_i^{(1)} \leftarrow \min \left( n_{iM}, \lfloor \Lambda_i \cdot (n_{cross} - \sum_i n_i^{(0)}) \rfloor \right)$ , con  $\Lambda_i = \sum_f \lambda_f E_{if}$  siendo

$E_{if} \geq 0$  factores de **relevancia** o medidas de **error** de cada celda  $i$ ;

$\lambda_f \geq 0$  **pesos relativos**.

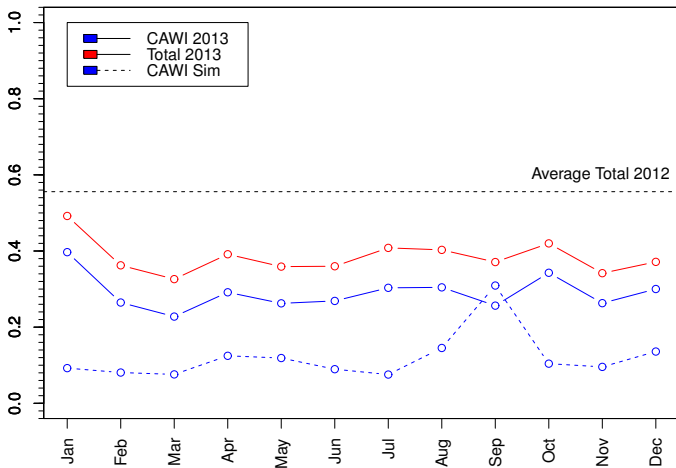
- Si  $n_{cross} - \sum_i (n_i^{(0)} + n_i^{(1)}) > 0$ , afijar una unidad por turno en orden **decreciente** de los valores  $\Lambda_i$  en cada celda no saturada. Se obtiene  $n_i^{(2)}$ .

$$n_i = n_i^{(0)} + n_i^{(1)} + n_i^{(2)}$$

Se seleccionan las  $n_i$  **primeras unidades de cada celda**.

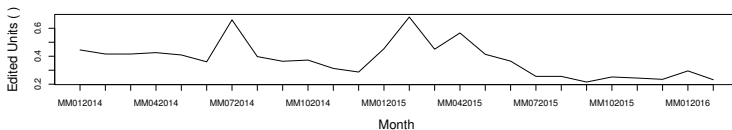
# Resultados en producción

Sampling fraction of selected units.  
Implementation in production

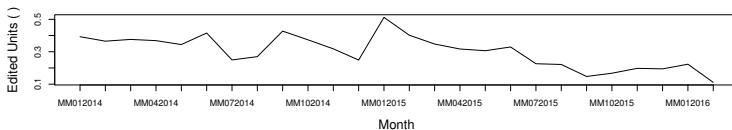


# Resultados en producción

**Retail Trade Survey**



**Sector Service Indicators Survey**





- **Soporte** en la extensión a más operaciones estadísticas: **simulaciones, formación, ...**
- Mejora de funciones: **predicción, infactibilidad, afijación, opt. estocástica, ...**
- Construcción de **indicadores de monitorización** de las estrategias de depuración.
- Extensión a encuestas económicas **estructurales**.
- Depuración de variables **cualitativas**: nuevos modelos de observación-predicción.
- **Estandarización** de las estrategias de depuración a encuestas de **hogares/personas y económicas**.
- Construcción de funciones para la depuración **macro**.

- A.E. Anderson, S. Cohen, E. Murphy, E. Nichols, R. Sigman, and D.K. Willimack. *Changes to Editing Strategies when Establishment Survey Data Collection Moves to the Web*, 1–36 (2003).
- I. Arbués, M. González, and P. Revilla. *La depuración selectiva como un problema de optimización estocástica*. Boletín de Estadística e Investigación Operativa, **25**, 32–41 (2009).
- I. Arbués, M. González, and P. Revilla. *A class of stochastic optimization problems with application to selective data editing*. Optimization **61**, 265–286 (2012); published online on Feb 4, 2010.
- I. Arbués and P. Revilla (2014). *Score functions under the optimization approach*. UNECE Work Session on Statistical Data Editing WP 1, pp. 1 –9.
- I. Arbués, P. Revilla, and D. Salgado. *Optimization as a theoretical framework to selective editing*. UNECE Work Session on Statistical Data Editing, WP1, 2012.
- T. de Waal. *An overview of statistical data editing*. Discussion paper (08018), Statistics Netherlands (2008).
- T. de Waal. *Statistical data editing*, in D. Pfefferman and C.F. Rao, eds. (2009), *Sample Surveys: Design, Methods and Applications*, North Holland, Amsterdam.
- T. de Waal. *Selective editing: a quest for efficiency and data quality*. Journal of Official Statistics **29**, 473–488 (2013).
- T. de Waal and W. Coutinho. *Automatic editing for business surveys: an assessment of selected algorithms*. International Statistical Review **73**, 73–102 (2005).

- EDIMBUS (O. Luzzi *et al.*). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Eurostat, 1997.  
Available at [http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM\\_EDIMBUS.pdf](http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf).
- K. Farwell. *The general application of significance editing to economic collections*. Technical report, Australian Bureau of Statistics (2004).
- I.P. Fellegi and D. Holt. *A systematic approach to automatic edit and imputation*. Journal of the American Statistical Association **71**, 17–35 (1976).
- L. Granquist. *Macro-editing: methods for rationalizing the editing of quantitative data*. Eustat (1991).
- L. Granquist. *On the current best methods document: edit efficiently*. UNECE Work Session on Statistical Data Editing, W.P. No. 30 (1997).
- L. Granquist and J.G. Kovar (1997). *Editing of survey data: how much is enough?*, in L.E. Lyberg *et al.*, eds. (1997), *Survey Measurement and Process Quality*. Wiley, New York.
- R.M. Groves. *Survey errors and survey costs*. Wiley, New York (1989).
- D. Hedlin. *Score functions to reduce business survey editing at the U.K. Office for National Statistics*. Journal of Official Statistics, **19**, 177–199 (2003).
- D. Hedlin. *Local and global score functions in selective editing*. UNECE Work Session on Statistical Data Editing, W.P. 31, 1–8 (2008).

- M.A. Hidirolou and J.M. Berthelot. *Statistical editing and imputation for periodic business surveys*. Survey Methodology, **12**, 73–84 (1986).
- J. Hoogland. *Selective editing by means of plausibility indicators*. UNECE Work Session on Statistical Data Editing, W.P. 33 (2002).
- M. Latouche and J.M. Berthelot. *Use of a score function to prioritize and limit recontacts in editing business surveys*. Journal of Official Statistics, **8**, 389–400 (1992).
- D. Lawrence and C. McDavitt. *Significance editing in the Australian survey of average weekly earnings*. Journal of Official Statistics, **10**, 437–447 (1994).
- D. Lawrence and R. McKenzie. *The general application of significance editing*. Journal of Official Statistics, **16**, 243–253 (2000).
- R. López-Ureña, M. Mancebo, S. Rama, and D. Salgado. *An efficient editing and imputation strategy within a corporate-wide data collection system at INE Spain: a pilot experience*. 2013 Meeting on the Management of Statistical Information Systems, pp. 1–9. UNECE, Paris (2013).
- R. López-Ureña, M. Mancebo, S. Rama, and D. Salgado. *Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey*. To appear as INE Spain Working Paper. INE, Madrid (2014).
- E.M. Nichols, E.D. Murphy, A.E. Anderson, D.K. Willimack, and R.S. Sigman. *Designing Interactive Edits for U.S. Electronic Economic Surveys and Censuses: Issues and Guidelines*. U.S. Census Bureau Research Report Series, Survey Methodology #2005-03 (2005).

- J. Pannekoek, S. Scholtus, and M. van der Loo. *Automated and manual data editing: a view on process design and methodology*. J. Off. Stat. **29**, 511–537 (2013).
- M. Pierzchala. *A review of the state of the art in automated data editing and imputation*. Journal of Official Statistics **6**, 355–377 (1990).
- S. Rama and D. Salgado. *Standardising the editing phase at Statistics Spain: a little step beyond EDIMBUS*. To appear as INE Spain Working Paper. INE, Madrid (2014).
- D. Salgado, I. Arbués, and M.E. Esteban. *Two greedy algorithms for a binary quadratically constrained linear program in survey data editing*. INE Working Papers Doc. 03/2012.
- A. Salvador, and D. Salgado. *A generalization of the stochastic optimization approach to selective editing*. In preparation. INE, Madrid (2014).

# Estadístico oficial = Estadístico + *Computer Scientist*

