Master Universitario en “Estadística aplicada y Estadística para el Sector Publico”

Fase de depuración e imputación de datos

previlla@ine.es

“Que errar es humano no sólo significa que hemos de luchar constantemente contra el error, sino también que, aún cuando hayamos puesto el máximo cuidado, no podemos estar totalmente seguros de no haber cometido un error.”

KARL POPPER

Contenido

- Introducción
- Depuración
- Imputación
- Métodos generalizados
- Métodos de depuración selectiva
- Un ejemplo práctico

No hay una definición generalmente aceptada de depuración (I)

“Un procedimiento diseñado y utilizado para detectar datos erróneos y/o sospechosos de error (tanto datos de respuesta como de identificación) con el objetivo de corregir (manualmente y/o automáticamente) tantos datos erróneos como sea posible (no necesariamente todos), normalmente antes de los procesos de imputación y agregación”

(Comité Federal de Metodología Estadística, 1990)

No hay una definición generalmente aceptada de depuración (II)

“Una actividad dirigida a asegurar que los datos cumplen ciertos requerimientos; es decir, que satisfacen condiciones de corrección establecidas”

(Conferencia de Estadísticos Europeos, 1997)

Depuración

- **Conjunto de tareas para el tratamiento de errores**
- **Conjunto de tareas para obtener unas estadísticas de calidad**

Imputación

Estimación ante la falta de respuesta o ante datos considerados erróneos

Actividad Compleja

- **No conocemos la verdad: difícil de contrastar**
- **Métodos sofisticados y caros**

Encuesta para la evaluación del coste de edición, del Federal Committee on Statistical Methodology (EE.UU):

- **20% Estadísticas sociodemográficas**
- **40% Estadísticas económicas**

Depuración

H_O: EL DATO ES BUENO

	VERDADERA	FALSA
ACEPTAR	CORRECTA	ERROR
RECHAZAR	ERROR	CORRECTA

ERROR 1:

RECHAZAR / VERDADERA → **CORRIGE** → **ERROR**

NO SE CORRIGE → **TIEMPO**

ERROR 2:

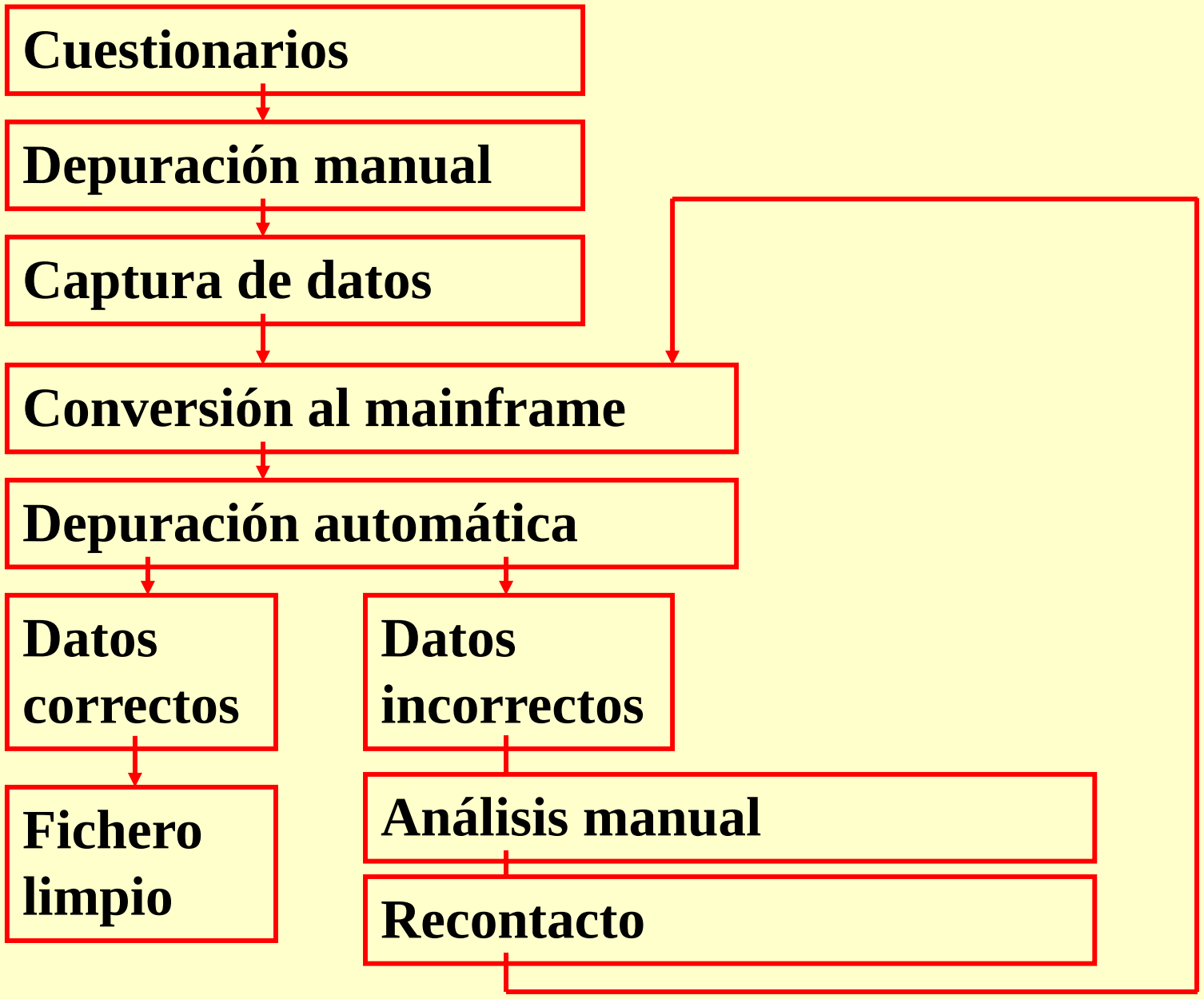
ACEPTAR / FALSA → **ERROR**

Contexto: Producción de datos estadísticos

No se puede aislar la depuración/imputación del resto de las fases de la cadena de producción de datos

Funciones de la depuración

1. **Facilitar el procedimiento automático**
2. **Detectar y corregir errores**
3. **Proporcionar información para medir la calidad de los datos**
4. **Proporcionar la base para futuras mejoras de la encuesta**

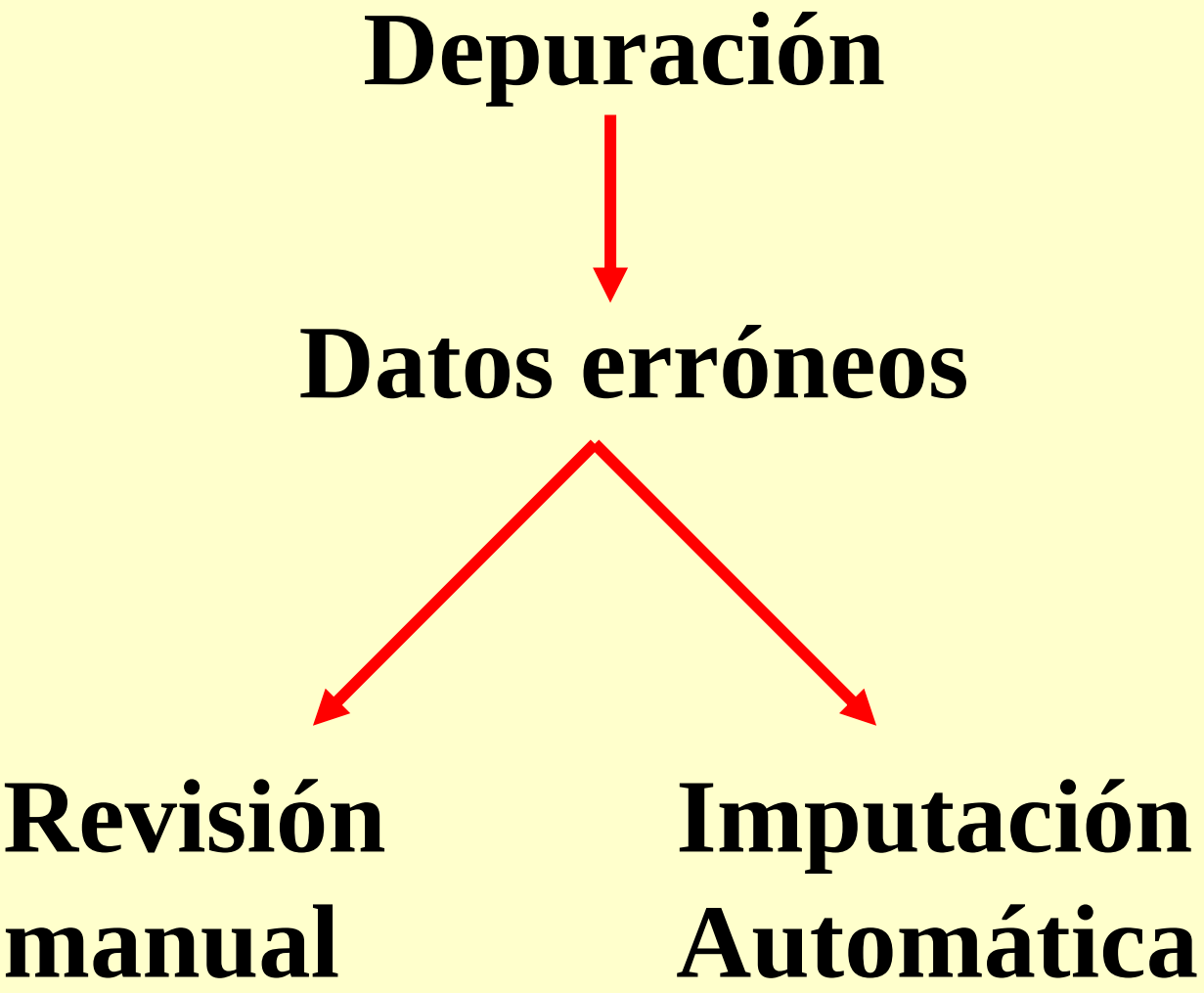

Cuestionarios
Depuración manual
Captura de datos
Conversión al mainframe
Depuración automática
Datos correctos
Datos incorrectos
Fichero limpio
Análisis manual
Recontacto

Enfoques

- **Microdepuración**
- **Macrodepuración**
- **Depuración selectiva**

Errores

- **Errores de muestreo**
- **Errores ajenos al muestreo**


Depuración
Datos erróneos
Revisión manual
Imputación Automática

Detección de errores

Especificación de situaciones erróneas o sospechosas de error.

Herramienta: edits

Especificación de restricciones para los valores de las variables.

- **Individuales: edits de validación**
- **Grupos de variables: edits de consistencia**

Información utilizada

- El mismo cuestionario
- Otros cuestionarios de la misma encuesta
- Datos de la misma encuesta en periodos anteriores
- Otras fuentes externas a la encuesta

Tipos de edits

- **Situaciones imposibles**
- **Situaciones improbables**
- **Restricciones contables**
- **De rutas de respuesta**
- **Outliers estadísticos**

Edits

- **Fatales (“fatal edits”)**
- **Dudosos (“query edits”)**

Fatales → certeza

Dudosos → sospechosos

Edits Fatales

- **Deben ser eliminados**
- **Automatizables**
- **Sistemas generalizados**

Edits dudosos

Difícil de diseñar

¿Cuales?

¿Cuantos?

¿Amplitud?

Edits dudosos

- **Procedimientos manuales**
- **Recontacto con el informante**
- **A veces pocos cambios en los datos**
- **Métodos de depuración selectiva**

Tareas relacionadas con la depuración de datos

- Control de completitud
- Control de grabación
- Control de identificación
- Control de cobertura
- Control del tipo de dato
- Control de formato
- Tablas de análisis

Alternativas ante los datos conflictivos:

1. **Volver al suministrador de la información**
2. **Asignar un valor consistente**
3. **Publicar una categoría “invalido” o “erróneo”**
4. **Ignorar los cuestionarios sospechosos (perdida de información)**
5. **Ajustar las estimaciones finales (no consistencia microdatos-macrodatos)**

Imputación

Estimación ante la falta de respuesta o ante datos considerados erróneos

Imputación

Origen

Falta de información:

- Falta de respuesta
- Respuestas incorrectas

Razones de la falta de información muy diversas y variadas

- **Informantes ausentes a pesar de repetidas llamadas**
- **Informantes rehusan responder a todas o alguna de las cuestiones**
- **Valor inaceptable de acuerdo a los edits**

.....

Falta de respuesta

- **Total (*unit nonresponse*)**
- **Parcial (*item nonresponse*)**

Falta de respuesta

total $\Rightarrow$ reponderación

parcial $\Rightarrow$ imputación

Reponderación **aumentar los pesos de los que responden**

Imputación **los datos que faltan son reemplazados por valores “plausibles”**

Métodos de imputación

Gran divesidad

2 grandes troncos:

- **Registros donantes**
- **Modelo de relación entre variables**

Imputación con registro donante

Asignan a los campos a imputar de un registro el valor que en tales campos tiene otro registro (de la encuesta o de una fuente externa)

Cold-deck

Asigna a los campos a imputar de todos los registros receptores los valores de un “registro tipo” definido por el usuario

Hot-deck

Asigna a los campos a imputar de cada registro receptor los valores de un registro donante diferente

Hot-deck secuencial

El donante es el registro válido inmediatamente anterior

Hot-deck con donante aleatorio

Eligen aleatoriamente uno o varios registros donantes para cada registro candidato

hot-deck con donante aleatorio

1) **Un único registro donante para cada candidato**

2) **Una muestra de registros donantes, y se imputan los valores medios**

Hot-deck métrico (nearest-neighbor)

Se elige como donante el más próximo según una medida de distancia

Imputación lógica o deductiva

El dato a imputar puede deducirse de las respuestas a otras cuestiones

ejemplos

- **edad <10 años ⇒ estado civil = soltero**
- **asalariados = 0 ⇒ sueldos y salarios = 0**
- **falta un solo subcomponente de un total**

Imputación de la media

Reemplaza el dato a imputar por la media de los datos válidos (normalmente de un estrato)

Imputación por regresión

Reemplaza el dato a imputar por la predicción de una regresión (de la variable a imputar sobre otras variables del cuestionario de las que sí se tiene información)

Imputación histórica

Utiliza valores de la misma unidad en periodos anteriores

Métodos generalizados

- **Utilizar una metodología contrastada y eficiente**
- **Coordinar y sistematizar esfuerzos que se repiten de encuesta en encuesta**
- **Ahorrar recursos y tiempo en el estudio y desarrollo de procedimientos específicos**
- **Contribuir a la consistencia entre encuestas similares**

Fases

1) Especificación de los edits

2) Análisis de los edits

3) Detección de errores

4) Identificación variables a imputar

5) Imputación

Metodología de Fellegi y Holt

- **“A Systematic Approach to Automatic Edit and Imputation”, JASA (1976)**
- **Un riguroso modelo matemático sustenta los principios propuestos en su metodología**

Principios

a) Cambio mínimo (los datos deben satisfacer todos los edits, cambiando el menor número posible de campos)

b) Los expertos se limitan a especificar los edits. No es necesario especificar reglas de imputación (se derivan automáticamente de los edits)

c) Mantenimiento de la estructura de frecuencias de los datos sin error

Edits

- **explícitos: originalmente especificados por el experto**
- **implícitos: se deducen lógicamente de los explícitos**

Ejemplo

A< producción / salario < B	**explícito**
C< salario / persona < D	**explícito**
A.C < producción / persona < B.D	**implícito**

Conjunto completo de edits

corazón de la metodología de F&H

conjunto de todos los edits generados implícitamente a partir del conjunto primitivo de edits explícitos

Conjunto completo de edits

- **Proceso iterativo**
- **F&H demuestran que, si el conjunto de edits explícito cumple unas condiciones bastante generales, el proceso es convergente**

Familia de Sistemas Generales de Depuración e Imputación basados en la Metodología de Fellegi&Holt

- **Generalized Edit and Imputation System (GEIS y BANFF)**
- **Structured Programs for Economic Editing and Referrals (SPEER)**
- **CherryPy**

...

Métodos de depuración selectiva

Valores atípicos e influyentes

Atípicos: **“sospechosos” de ser erróneos por algún criterio**

Influyentes: **con “peso” o “influencia” grande en los datos que se difunden**

(Importancia en variables cuantitativas)

Problemas de la depuración tradicional

- **La mayoría de los datos sospechosos detectados no dan lugar a correcciones**

$$\text{Ratio de impacto} = \frac{\text{Nº de correcciones}}{\text{Nº de detectados}}$$

23% Censo Australiano de Comercio
28-47% Lindström (1991)

- **No se detectan errores importantes**

Problemas de la depuración tradicional

Falsa impresión de capacidad de respuesta de los informantes

“Depuración creativa”

Dificultad de aplicar métodos generalizados en datos cuantitativos

Depuración selectiva (I)

Concepto: Detección y corrección selectiva de errores

Pone en relación los microdatos con los macrodatos para determinar los errores de los microdatos que tienen influencia en los macrodatos

Iniciada por L . Granquist (1984)

Depuración selectiva (II)

Objetivo: Agilizar las tareas de depuración - imputación en los datos de una encuesta sin detrimento de la calidad del proceso

Errores (I)

Errores de negligencia (Negligence errors)

- **Es el resultado de la falta de cuidado del informante o del proceso de la encuesta**
- **En una repetición de la encuesta el mismo error probablemente no seria encontrado para la misma variable del mismo cuestionario**

Errores (II)

Errores sistemáticos no anticipados (misunderstanding errors)

- **Aparecen debido a la ignorancia o mala interpretación de las cuestiones, conceptos o definiciones, o son cometidos deliberadamente**
- **En una repetición de la encuesta el mismo error probablemente afectaría a la misma variable del mismo cuestionario**

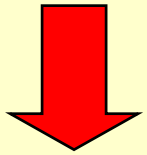
Depuración selectiva

- **Arma contra la sobredepuración**
- **Arma contra los errores sistemáticos no anticipados**

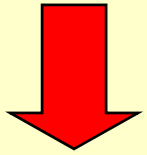
Estrategia de la depuración tradicional

Gran número de edits

Intervalos de aceptación muy cerrados



Principio de “la seguridad lo primero” (cuantos más edits y más cerrados mejor)



Sobredepuración

Crítica al método tradicional

- **No es eficiente con los errores de negligencia**
- **Está lejos de ser aceptable con los errores sistemáticos no anticipados**
- **Ilusoria sensación de confianza**

Contraste de hipótesis

$\alpha = P(E_1) = P(\text{rechazar } H_0 / H_0 \text{ cierta})$

$\beta = P(E_2) = P(\text{aceptar } H_0 / H_0 \text{ falsa})$

$1 - \beta = P(\text{rechazar } H_0 / H_0 \text{ falsa})$

Proceso selectivo de detección de errores (I)

- **No todos los cuestionarios o variables son objeto de investigación**
- **Sólo aquellos que son influyentes**
- **Se ignoran datos cuya magnitud no es significativa o que se cancelan en el proceso de agregación**

Proceso selectivo de detección de errores (II)

- **Filosofía y no un conjunto de procedimientos cerrado**
- **Se concibe generalmente como un proceso interactivo**
- **Preferentemente en datos cuantitativos**

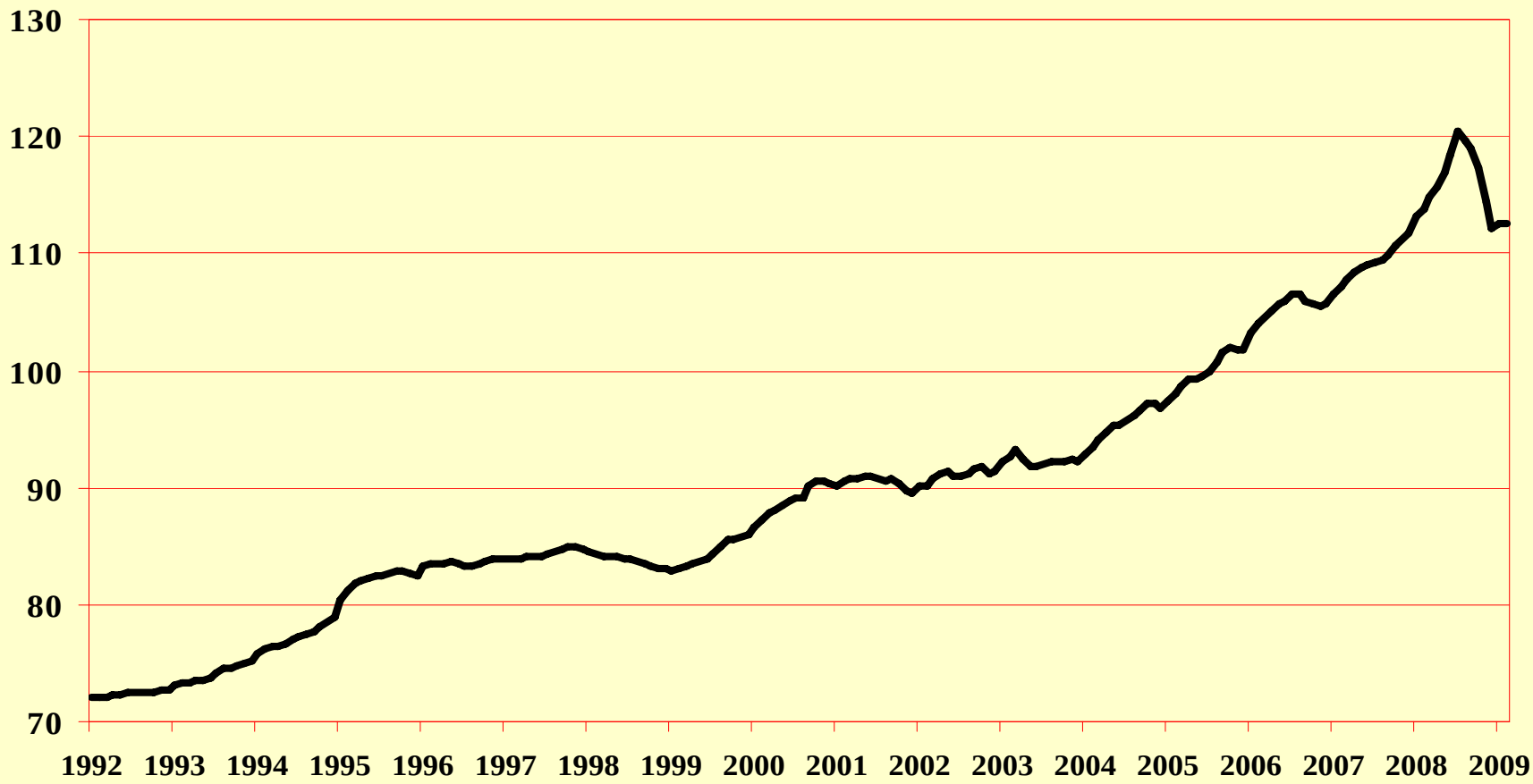
La cuestión no es “¿Quién debe gobernar?” o “¿Quién ha de detentar el poder?” sino más bien “¿Cómo podemos crear las instituciones políticas de forma que incluso los gobernantes incompetentes o poco honestos, que debemos intentar evitar, pero que de todas formas seguramente vamos a tener, no puedan causar mucho daño?”

En busca de un mundo mejor

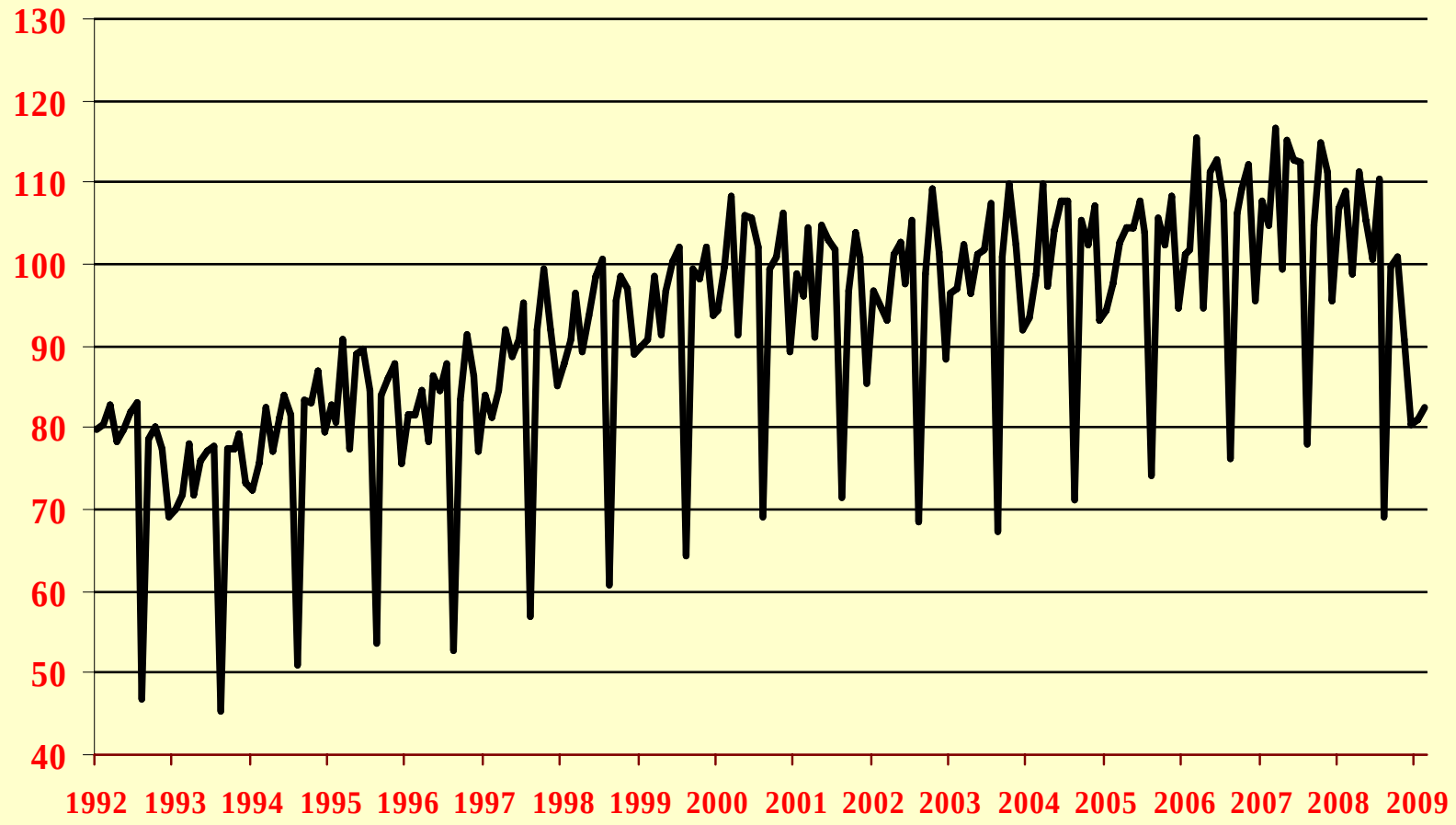
KARL POPPER

Un ejemplo práctico: herramientas de depuración e imputación basadas en modelos de series temporales

IPRI Indice General



IPI Indice General



Información para la depuración e imputación de Indicadores de coyuntura: datos pasados de la misma población

- **Herramientas tradicionales:**
  - **tasa intermensual, tasa interanual**
  - **imputación histórica**

- **Mejora: modelos de series temporales**

Ventajas modelos

1) **Tasas utilizan un único dato, modelos todo el pasado**

2) **Modelos permiten depuración probabilística**

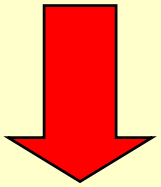
Ventajas modelos

1) **Imputación histórica utiliza pocos datos, modelos todo el pasado**

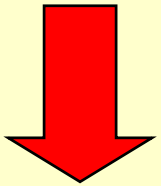
2) **Modelos permiten asertos probabilísticos**

3) **Tratamiento de “ceros”**

Serie Temporal [X_1, X_2,…,X_T]



Modelo teórico



Depuración e imputación

Naturaleza fenómenos económicos y sociales

Probabilística

Dinámica

Modelos de este tipo

Contexto teórico: procesos estocásticos

Modelos ARIMA con análisis de intervención

Índices elementales

$$I_{i,t} = I_{i,t-1} \frac{\sum_j q_{i,j,t}}{\sum_j q_{i,j,t-1}}$$

Índices agregados

$$I_t = \sum_i \omega_i I_{i,t}$$

(índices de Laspeyres)

Imputación histórica

$$\hat{q}_s^t = q_s^{t-1}$$

Imputación histórica

$$q_s^t = q_s^{t-1} \frac{\sum_{j \neq S} q_j^t}{\sum_{j \neq S} q_j^{t-1}}$$

$$\hat{q}_s^t = q_s^{t-1} \frac{\sum_{j \neq S,A} q_j^t}{\sum_{j \neq S,A} q_j^{t-1}}$$

Imputación histórica (series mensuales)

$$\hat{q}_t = q_{t-1} \cdot \frac{q_{t-12}}{q_{t-13}}$$

$$\hat{q}_t = q_{t-2} \cdot \frac{q_{t-12}}{q_{t-14}}$$

autoimputación

$$I_i^t = I_i^{t-1} \times \frac{\sum_{j \neq S} q_{ij}^t}{\sum_{j \neq S} q_{ij}^{t-1}} = I_i^{t-1} \times K$$

$$K = \frac{\sum_{j \neq S} q_{ij}^t}{\sum_{j \neq S} q_{ij}^{t-1}}$$

$$\hat{q}_s^t = q_s^{t-1} \times K$$

models

$$Lnq_{i,j,t} = \frac{\theta_{i,j}(B)\Theta_{i,j}(B^{12})}{\varphi_{i,j}(B)\Phi_{i,j}(B^{12})} a_{i,j,t} + \sum_{h} \frac{\alpha_{i,j,h}(B)}{\delta_{i,j,h}(B)} A_{i,j,h,k}$$

$$LnI_{i,t} = \frac{\theta_{i}(B)\Theta_{i}(B^{12})}{\varphi_{i}(B)\Phi_{i}(B^{12})} a_{i,t} + \sum_{h} \frac{\alpha_{i,h}(B)}{\delta_{i,h}(B)} A_{i,h,k}$$

Variables de regresión

- **Additive outliers**
- **Level shifts**
- **Temporary changes**
- **Calendar effects**
- **Intervention effects**

Outliers

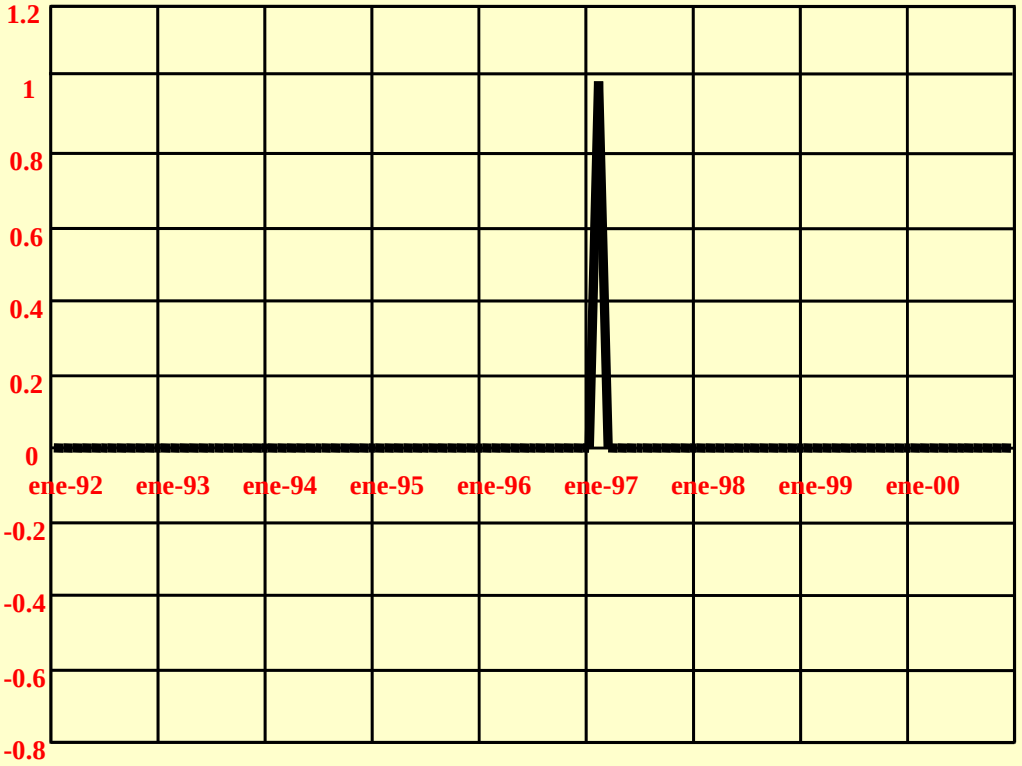
Chang et al. (1988)

Detección de sucesos especiales que pueden afectar la producción

Aditive Outlier (AO)

$$I_t(t_0) = \begin{cases} 0 & \forall_t \neq t_0 \\ 1 & t = t_0 \end{cases}$$

$$Z_t = Y_t + \omega I_t(t_0)$$

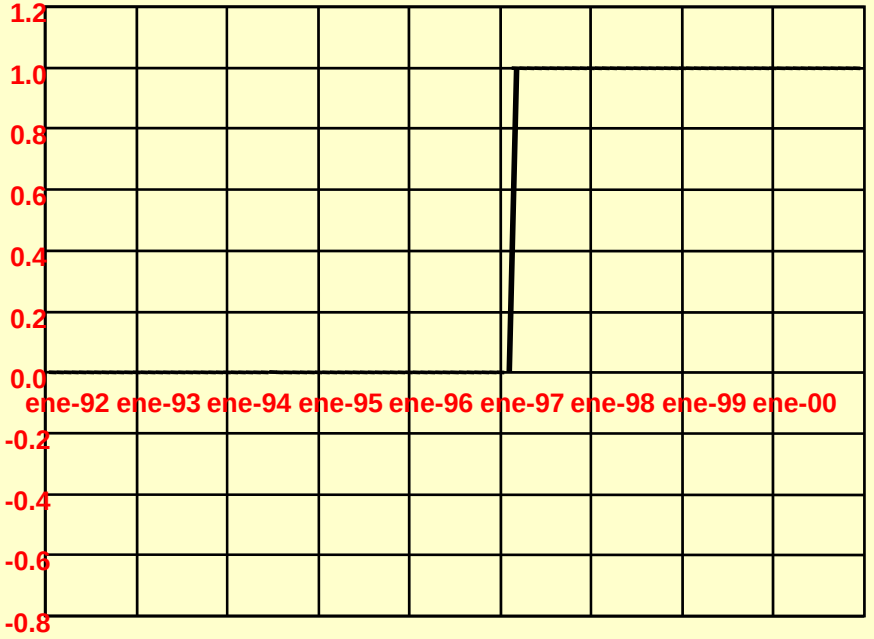


Level Shift (LS)

$$Z_t = Y_t + \frac{1}{1-B}\omega I_t(t_0) = Y_t + \omega E_t(t_0)$$

Where :

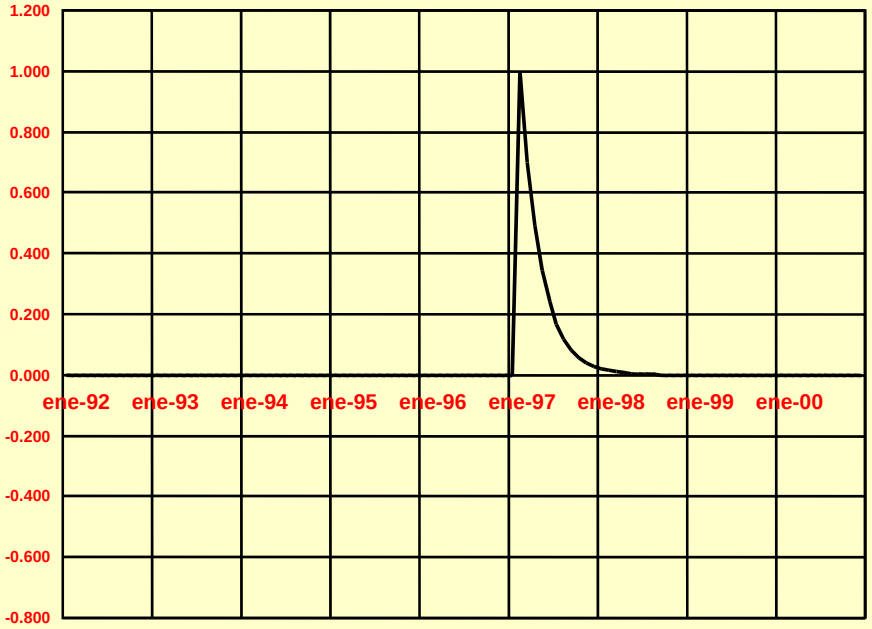
$$E_t(t_0) = \begin{cases} 0 & \forall_t < t_0 \\ 1 & \forall_t \geq t_0 \end{cases}$$



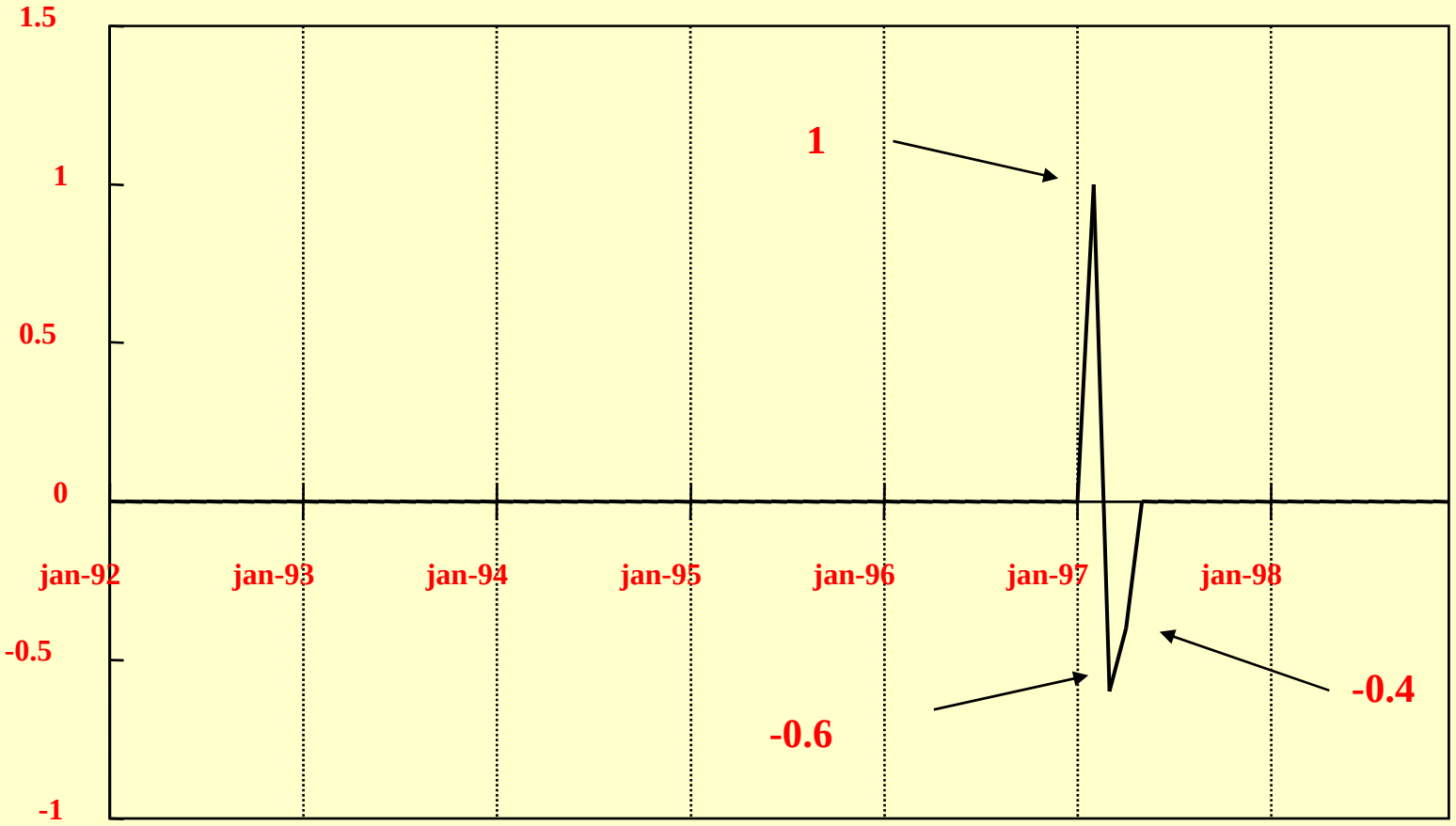
Temporary Change TC

$$Z_t = Y_t + \frac{1}{1 - \delta B}\omega I_t(t_0), 0 < \delta < 1$$

$$\text{When} \begin{cases} \delta = \Rightarrow \text{AO} \\ \delta = \Rightarrow \text{LS} \end{cases}$$



February 1997 Strike Intervention



Modelos

$$\log I_t = \frac{(1 - 0.5478\ B)(1 + 0.1648\ B^{12})}{(1 - B)(1 - B^{12})} a_t$$

$$\sigma = 0.04019$$

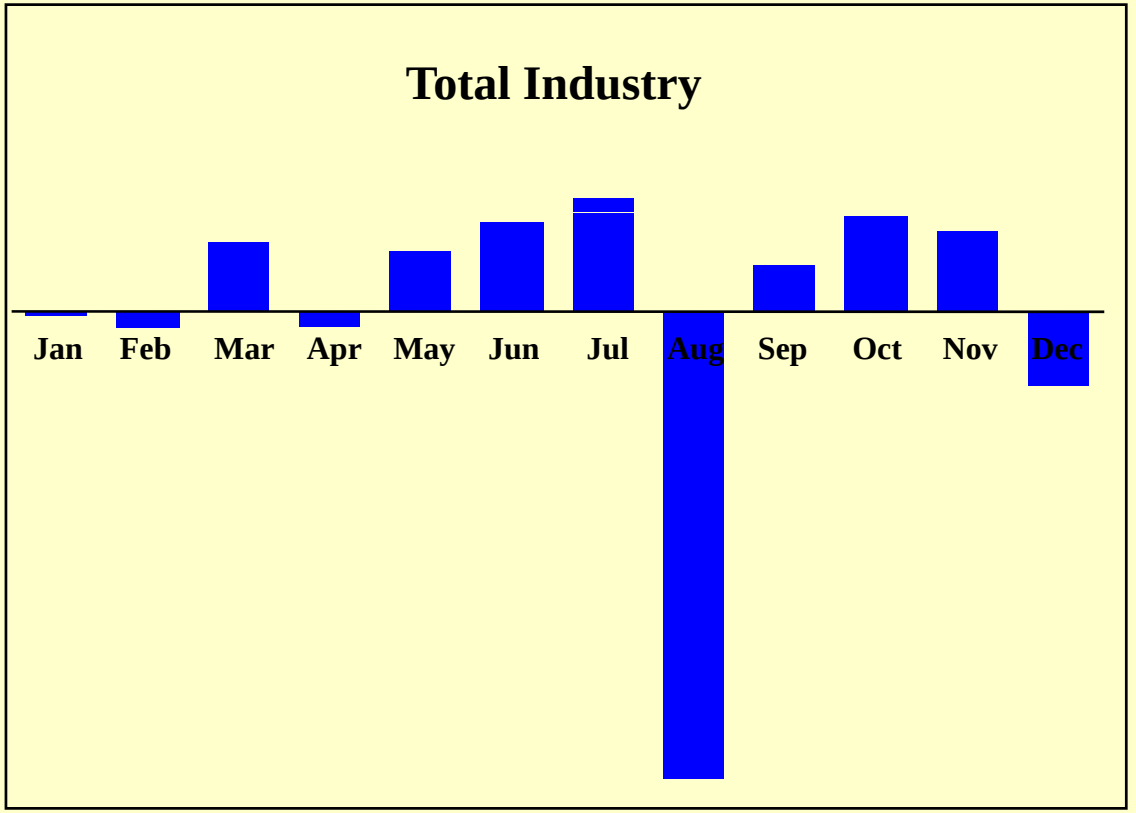
$$\log I_t = \frac{(1 - 0.3060B)(1 + 0.3437B^{12})}{(1 - B)(1 - B^{12})} a_t + 0.0188\,WD_t - 0.0396H_t$$

$$\sigma = 0.02417$$

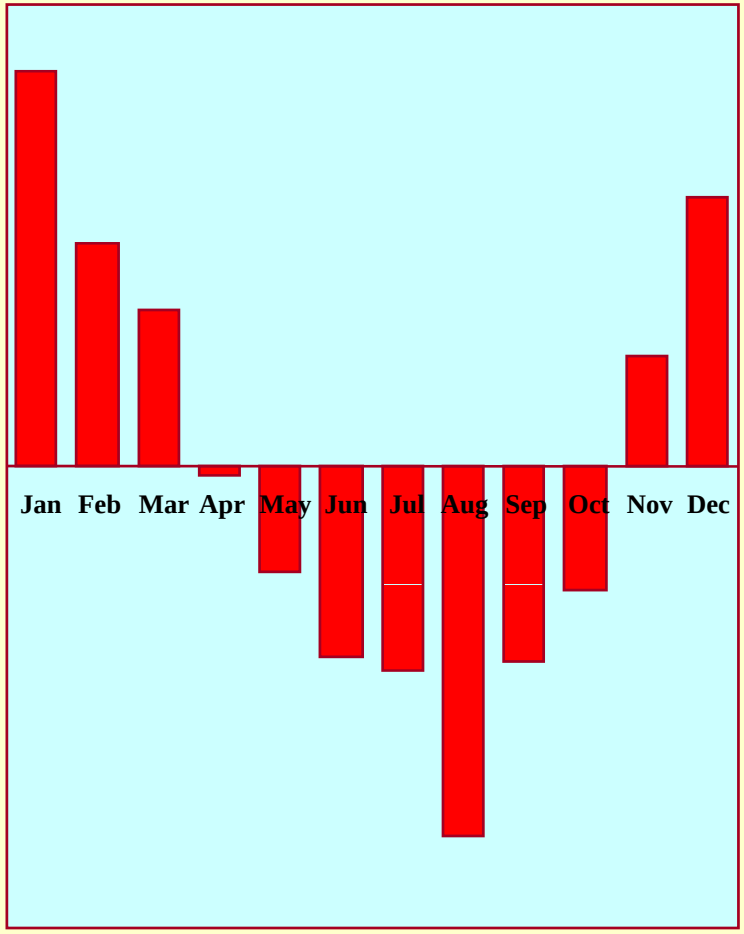
$$\log I_t = \frac{(1 - 0.2645\,B)(1 + 0.3206\,B^{12})}{(1 - B)(1 - B^{12})} a_t + 0.0186\,WD_t - 0.0453\,H_t - 0.0340\,S_t$$

$$\sigma = 0.02310$$

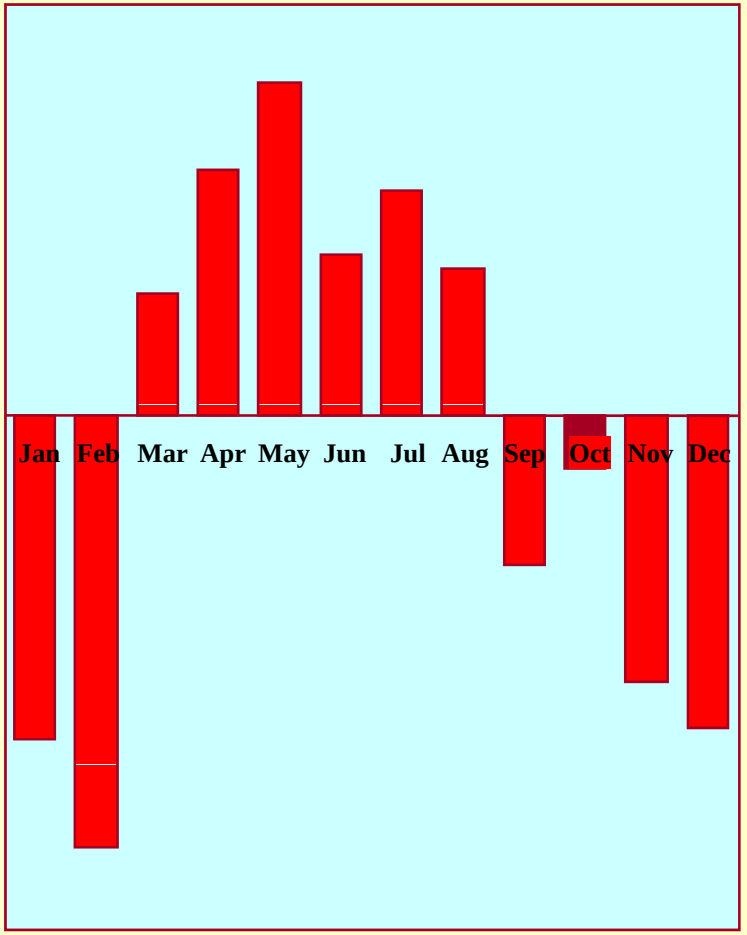
Seasonal behaviour: the models can contain a factor which picks up a seasonal cycle with a period of 12 time units



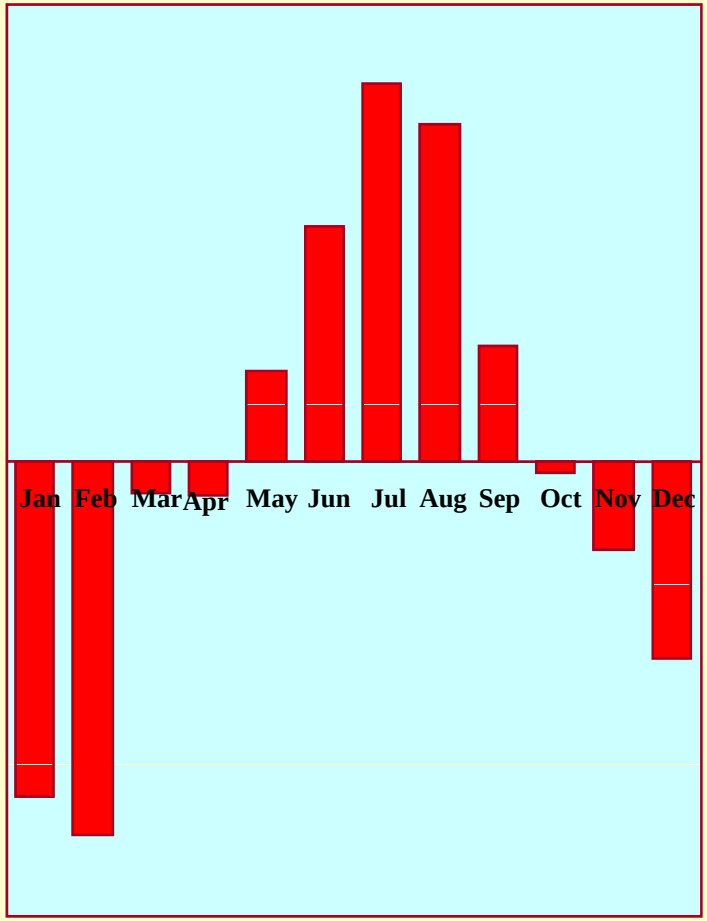
Gas Manufacturing



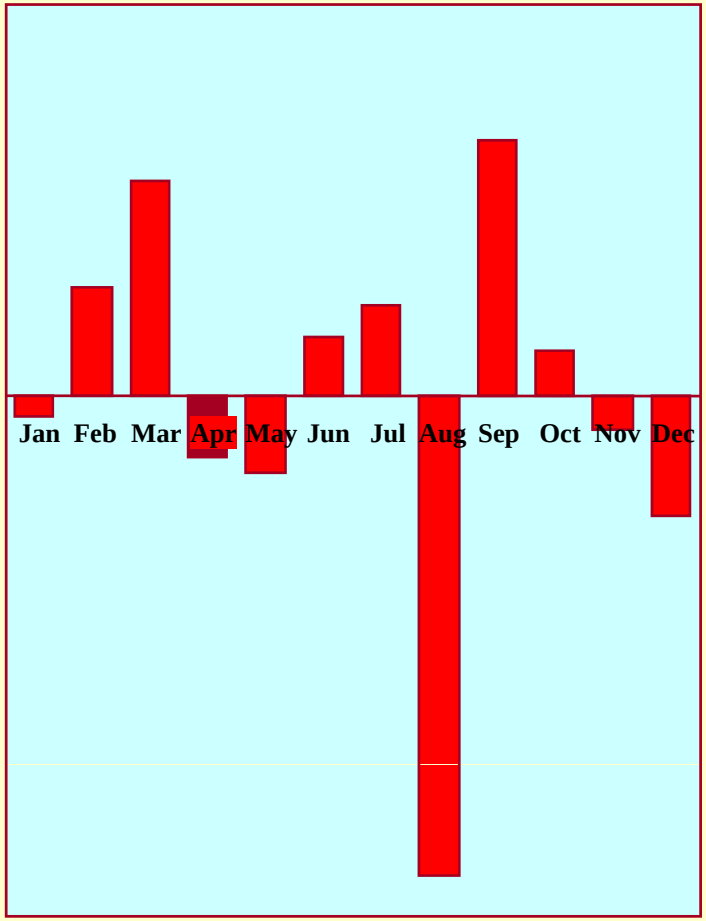
Dairy Industries



Beer Brewing



Clothing Industries



“Sólo una cosa no hay. Es el olvido.

Dios, que salva el metal, salva la escoria

y cifra en su profética memoria

las lunas que serán y las que han sido.”

BORGES

Imputación dinámica

“La profética memoria”

Imputación dinámica: predicción del modelo

Objetivo predecir el valor futuro X_{t+l} estando situados en t

Predicción en el origen t para el horizonte l $X_t(l)$

Imputación dinámica: predicción del modelo

- En qué consiste
- Características estocásticas

La predicción estará en función de los valores presente y pasados de la serie

$$X_t(l) = f(x_t, x_{t-1}, x_{t-2}, \ldots)$$

Dado que $X_t = \Psi(B)\, a_t \qquad \forall_t$

$$\hat{X}_t(l) = f(a_t, a_{t-1}, a_{t-2}, \ldots)$$

Si se considera una predicción de tipo lineal

$$\hat{X}_t(l) = \psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \ldots$$

Donde $\psi_l^*, \psi_{l+1}^*, \psi_{l+2}^*, \ldots$ **son los pesos que deben ser determinados**

Predicción óptima

<u>Idea</u> minimizar el error de predicción

$$e_t(l) = X_{t+l} - \hat{X}_t(l)$$

Entre las diferentes predicciones (es decir, entre los diferentes pesos

$\psi_l^*, \psi_{l+1}^*, \psi_{l+2}^*, \cdots)$ **se busca aquella que minimiza el error medio cuadrático**

$$E\left[e_t(l)\right]^2 = E\left[X_{t+l} - \hat{X}_t(l)\right]^2$$

$$E\left[X_{t+1} - X_{t(l)}\right]^2$$

$$= (1 + \Psi_1^2 + \Psi_2^2 + ... + \Psi_{l-1}^2)\gamma_a^2 +$$
$$+ \left[(\Psi_l - \Psi_l^*)^2 + (\Psi_{l+1} - \Psi_{l+1}^*)^2 + (\Psi_{l+2} - \Psi_{l+2}^*)^2 + ...\right]\gamma_a^2$$

Esta expresión se hara mínima cuando $\Psi_{l+j} = \Psi^*_{l+j}$

<u>Los coeficientes que minimizan el error medio cuadrático son los coeficientes del modelo ARIMA que ha generado la serie</u>

La predicción óptima es

$$\hat{X}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + ...$$

La predicción óptima coincide con la esperanza condicional a los valores presente y pasados de X_t

$$E\left[X_{t+1} / X_t, X_{t-1}, X_{t-2}, \ldots\right] = E\left[a_{t+l} + \Psi_1 a_{t+l-1} + \Psi_2 a_{t+l-2} + \right.$$

$$\left. + \ldots + \Psi_{l-1} a_{t+1} + \Psi_l a_t + \Psi_{l+1} a_{t-1} + \Psi_{l+2} a_{t-2} + \ldots \right] =$$

$$= \Psi_l a_t + \Psi_{l+1} a_{t-1} + \Psi_{l+2} a_{t-2} + \ldots$$

Por tanto:

$$X_t(l) = E\left[X_{t+l} / X_t, X_{t-1}, X_{t-2}, \ldots\right]$$

Error de predicción

$$e_t(l) = X_{t+l} - \hat{X}_t(l) \qquad X_{t+l} = \hat{X}_t(l) + e_t(l)$$

como

$$X_{t+l} = a_{t+l} + \Psi_1 a_{t+l-1} + \Psi_2 a_{t+l-2} + \ldots + \Psi_{l-1} a_{t+1} +$$
$$+ \frac{\Psi_l a_t + \Psi_{l+1} a_{t-1} + \Psi_{l+2} a_{t-2} + \ldots}{X_t(l)}$$

se tiene

$$\boxed{e_t(l) = a_{t+l} + \Psi_1 a_{t+l-1} + \Psi_2 a_{t+l-2} + \ldots + \Psi_{l-1} a_{t+1}}$$

Esperanza matemática y varianza del error de predicción

$$e_t(l) = a_{t+1} + \Psi_1 a_{t+l-1} + \Psi_2 a_{t+l-2} + \ldots + \Psi_{l-1} a_{t+1}$$

$$E\left[e_t(l)\right] = 0$$

$$V\left[e_t(l)\right] = (1 + \Psi_1^2 + \Psi_2^2 + \ldots + \Psi_{l-1}^2)\sigma_a^2$$

Esperanza matemática del error de predicción

$$E\left[e_t(l)\right] = 0$$

La imputación basada en la predicción del modelo es <u>insesgada</u>

Varianza del error de predicción

$$V\left[e_t(l)\right]=(1+\Psi_1^2+\Psi_2^2+\ldots+\Psi_{l-1}^2)\sigma_a^2$$

La varianza crece a medida que se aleja el horizonte de predicción

usaremos como imputación la predicción con el horizonte más cercano

Errores de predicción un período por delante

$$e_t(1) = X_{t+1} - \hat{X}_t(1) = (a_{t+1} + \psi_1 a_t + \psi_2 a_{t-1} + \psi_3 a_{t-2} + \ldots) -$$
$$-(\psi_1 a_t + \psi_2 a_{t-1} + \psi_3 a_{t-2} + \ldots) = a_{t+1}$$

Los errores de predicción un periodo por delante coinciden con las innovaciones o residuos a_t que generan el proceso

$$e_t(1) = a_{t+1}$$

$$V[e_t(1)] = \sigma_a^2$$

- **Los errores de predicción un periodo por delante $e_t(1) = a_{t+1}$ están incorrelacionados**
- **En otro caso, $e_t(1)$ podría predecirse en alguna medida a partir de los errores de predicción anteriores $e_{t-1}(1)$, $e_{t-2}(1)$, …**
- **$\hat{x}_t(1)+\hat{e}_t(1)$ sería mejor predicción de x_{t+1} que $\hat{x}_t(1)$**

la imputación basada en la predicción un periodo por delante garantiza un uso óptimo de la información disponible

Imputación histórica relativamente poco afectada por el mecanismo de no respuesta

Microdepuración

Intervalo de confianza

$$P\left[\hat{q}_{ijt} - 1.96\,\sigma_{ij} < q_{ijt} < \hat{q}_{ijt} + 1.96\sigma_{ij}\right] = 0.95$$

Imputación

$$\hat{q}_{ijt}$$

Macrodepuración

Intervalo de confianza

$$P\left[\hat{I}_{it} - 1.96\sigma_i < I_{it} < \hat{I}_{it} + 1.96\sigma_i\right] = 0.95$$

Imputación

$$\hat{I}_{it}$$

Enfoque Tradicional

1) **Microdepuración**
2) **Cálculo de índices**
3) **Depuración de índices**
4) **Microdepuración**

Desventajas

- **Bajo “hit rate”**
- **Mismos microdatos muchas veces revisados**
- **Esfuerzos idénticos con microdatos con mucho o poco impacto en los macrodatos**
- **Criterios subjetivos**
- **Pérdida de tiempo**

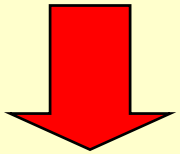
Objetivos

- **Mejorar “Hit Rate”**
- **Integrar las fases de depuración**
- **Priorizar errores con gran impacto**
- **Criterios objetivos**
- **Mejora rapidez**

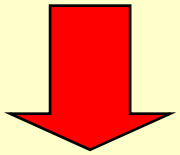
Depuración selectiva

- **Detectar outliers macrodatos**
- **Definir microdatos influyentes**

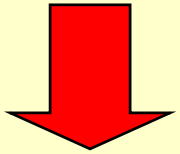
Set of observed data over time



models



Automatic modelling procedures



Surprises

The *Surprise (or simple surprise)* $S_{i,t}$ for the index $I_{i,t}$ is the relative change between the observed and the forecasted data:

$$S_{i,t} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}}$$

If we calculate the one-step ahead forecast Ln $\hat{I}_{i,t}$ for Ln $I_{i,t}$ the one-step ahead forecast error is:

$$e_{i,t} = LnI_{i,t} - Ln\hat{I}_{i,t}$$

Since the one-step ahead forecast error $e_{i,t}$ is a $N(0,\sigma_i)$ white noise process and $\mathrm{Ln}I_{i,t} - \mathrm{Ln}\hat{I}_{i,t} \cong (I_{i,t} - \hat{I}_{i,t}) / \hat{I}_{i,t}$, we have that $S_{i,t}$ is approximately $N(0,\sigma_i)$

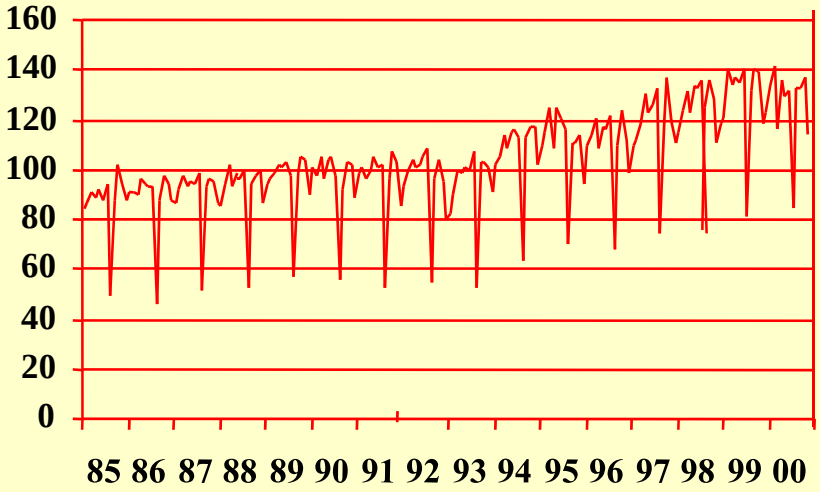
A confidence interval for the surprises can be constructed:

$$P\left[-1.96\sigma_i < S_{i,t} \leq 1.96\ \sigma_i\right] = 0{,}95$$

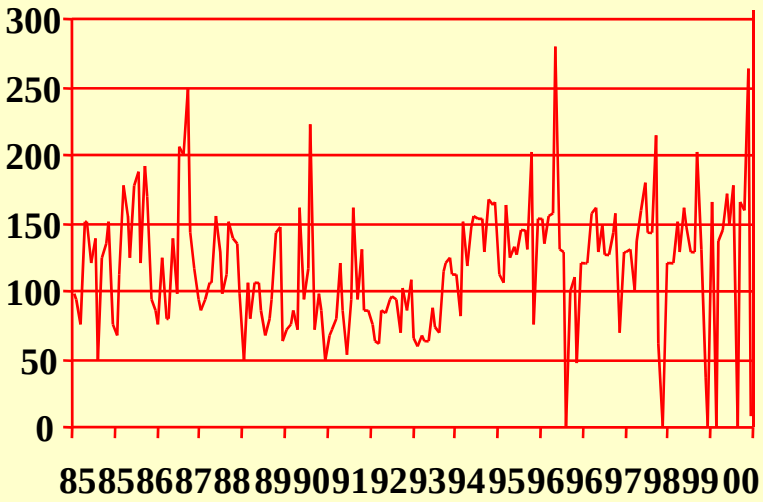
<u>outliers</u> can be defined as the indices with surprise outside the confidence interval

OIL

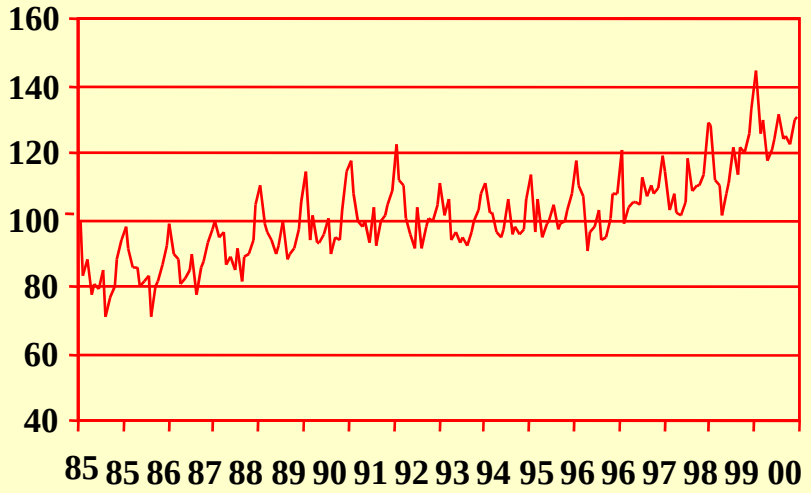
CHEMICAL



RADIOACTIVE MINERAL



ELECTRICAL ENERGY



Influences

- **Once detected and ranked the surprising indices (indices that are not coherent with their past behaviour and can be considered as outliers) we need to measure the impact of each of the microdata on these surprising indices**

- **For this purpose, we use the “influences”.**

The *Influence of an individual datum over an aggregated magnitude* is defined as the difference between the observed aggregated magnitude and the value for this same magnitude when the individual datum is not available.

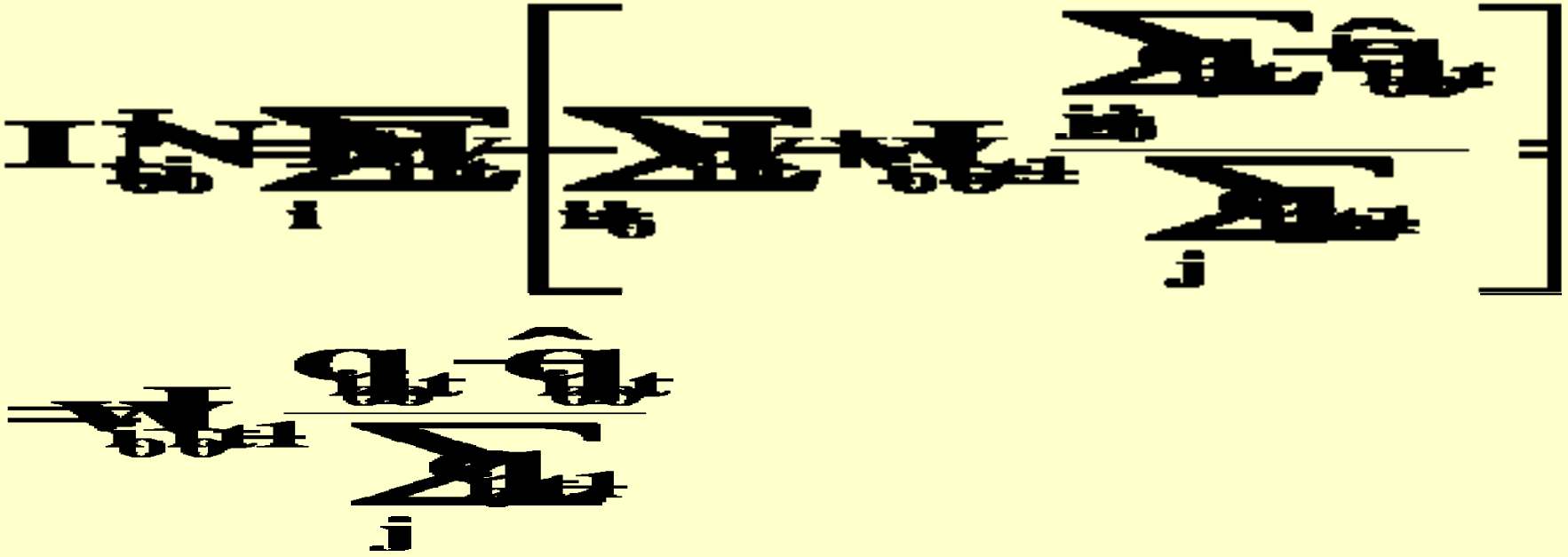
The *Influence of an individual datum* $q_{i_0,j_0,t}$ *over the product index* $I_{i_0,t}$ is:

$$INF_{i_0 j_0} = I_{i_0 t-1} \frac{\sum_j q_{i_0 j t}}{\sum_j q_{i_0 j t-1}} - I_{i_0 t-1} \frac{\sum_{j \neq j_0} q_{i_0 j t} + \hat{q}_{i_0 j_0 t}}{\sum_j q_{i_0 j t-1}} =$$

$$= I_{i_0 t-1} \frac{q_{i_0 j_0 t} - \hat{q}_{i_0 j_0 t}}{\sum_j q_{i_0 j t-1}}$$

Where $\hat{q}_{i_0 j_0 t}$ is an imputed value for the individual datum $q_{i_0 j_0 t}$

and the *Influence over the aggregated index* I_t is:



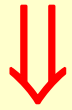
This expression measures the impact of microdata over the index by means of the following factors:

- **The product (or activity) weight** w_{i_0}
- **The index** $I_{t_0,t-1}$ **which "updates" the previous weight.**
- **A measure of the relative discrepancy between the real and the imputed individual datum.**

$$\frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

It can be seen that the microdata more influential over the aggregated index are the more influential over the surprises of that index.

The use of “influences” and “surprises” allows us to prioritise the suspicious values in the microdata in order to verify and recontact less number of enterprises



improvement in timeliness

The most influential suspicious values could immediately be detected

the most important errors can be corrected before the Index is disseminated for the first time

SIMPLE SURPRISE

Serie	Actual Annual Rate	95% confidence Interval for Annual Rate	Forecasted Annual Rate	Si,t	σi
0	-0,79	[-1.90; 7.23]	2,57	-3,27	2,27
1	2,09	[1.02; 14.48]	7,54	-5,07	3,19
11	-16,37	[-12.06; 7.70]	-2,68	-14,07	5,17
12	-37,7	[-40.38; -30.51]	-35,63	-3,21	3,91
13	6,79	[-11.24; 16.10]	1,52	5,20	6,85
14	-6,55	[-11.24; 16.10]	-5,55	-1,05	26,31
15	4,76	[-3.20; 16.25]	6,08	-1,24	4,67
2	-1,39	[-4.19; 4.89]	0,24	-1,63	2,31
21	-62,19	[-65.33; -46.70]	-57,01	-12,05	10,97
22	5,12	[-7.99; 6.66]	-0,94	6,11	3,77

SIMPLE SURPRISE

Serie	Actual Annual Rate	95% confidence Interval for Annual Rate	Forecasted Annual Rate	Si,t	σi
23	13,12	[-4.14; 20.05]	7,28	5,44	5,74
24	3,26	[-1.55; 6.81]	2,55	0,70	2,08
25	-6,45	[-6.10; 9.75]	1,51	-7,84	3,98
3	0,86	[-4.32; 11.35]	3,22	-2,29	3,87
31	-1,75	[-5.72; 11.51]	2,53	-4,18	4,28
32	3,3	[-6.25; 21.58]	6,76	-3,25	6,63
33	-11,75	[-32.70; 73.52]	8,06	-18,34	24,16
34	4,54	[-2.90; 20.41]	8,13	-3,32	5,49
35	-17,74	[-32.37; 15.48]	-11,63	-6,91	13,65
36	3,48	[-.26; 14.63]	6,93	-3,22	3,55

SIMPLE SURPRISE

Serie	Actual Annual Rate	95% confidence Interval for Annual Rate	Forecasted Annual Rate	Si,t	σi
37	-19,3	[-29.38; -10.23]	-20,38	1,35	6,12
38	12,52	[-10.95; 21.85]	4,16	8,03	8,00
39	9,95	[-4.43; 23.15]	8,49	1,35	6,47
4	-2,93	[-2.29; 7.68]	2,57	-5,37	2,48
41	-8,22	[-7.19; 4.10]	-1,71	-6,63	2,93
43	-0,62	[-11.28; 8.82]	-1,74	1,14	5,21
44	-8,59	[-16.14; 7.14]	-5,21	-3,56	6,25
45	0,7	[-5.67; 8.66]	1,24	-0,54	3,61
46	-4,06	[-5.99; 10.97]	2,14	-6,07	4,23
47	3,06	[-3.86; 8.69]	2,23	0,81	3,13
48	2,85	[-1.90; 13.99]	5,74	-2,74	3,83
49	-4,15	[-24.48; 31.57]	-0,32	-3,84	14,16

SIMPLE SURPRISE

Serie	Actual Annual Rate	95% confidence Interval for Annual Rate	Forecasted Annual Rate	Si,t	σ i
4	-2,93	[-2.29; 7.68]	2,57	-5,37	2,48
41	-8,22	[-7.19; 4.10]	-1,71	-6,63	2,93
411	-50,24	[-15.31; 1.52]	-6,89	-70,96	36,95
412	-6,38	[-31.65; 10.22]	-13,21	7,86	12,19
4121	-14,02	[-35.21; 23.47]	-10,56	-3,87	16,45
4123	2,82	[-32.69; 32.72]	-5,48	8,79	17,32
4124	-30,08	[-34.53; -.87]	-19,44	-13,21	10,58
413	1,22	[-39.36; 14.47]	-16,69	21,50	16,21
4131	2,81	[-.95; 11.19]	4,94	-2,03	2,95
4132	-2,58	[-36.56; 28.83]	-9,59	7,75	18,07
4133	16,74	[-15.56; 45.10]	10,69	5,47	13,81

SURPRISES

Serie	Actual Rate	Forecasted Rate	Simple Surprise	Standard Surprise	Weighted Standard Surprise
4243	70,28	3,32	64,73	3,79	17,10
2511	-27,73	-3,29	-25,25	-3,11	-16,93
4110	-50,24	-6,89	-70,96	-6,84	-16,89
2514	-15,92	4,64	-19,62	-3,00	-16,87
2512	39,39	-11,83	58,12	7,22	16,51
4752	-0,74	2,06	-2,75	-1,09	-15,66
3299	-11,97	4,45	-15,70	-2,02	-15,57
4751	22,82	-7,36	32,55	2,34	14,64
3630	-0,28	3,68	-3,82	-0,81	-14,54
3166	15,97	5,83	9,58	1,89	13,92

INFLUENCES

Branch=4243

N	Enterprise	Product	Actual data	Imputed Data	Influence
1	290068	42430120	9.590	3.298	143.41
2	084438	42430110	23.429	4.290	37.60
3	110079	42430110	794	9.896	-22.80
4	084143	42430120	1.416	520	14.38
5	980114	42430110	1.525	510	8.90
6	310302	42430210	2.825	5.708	-7.52
7	110205	42430120	773	315	7.35
8	084169	42430120	424	-	6.81

product 4243.0120 GIN
enterprise 290068.

	t-1							t						
	quantity	A	E	V	value	A	E	V	quantity	A	E	V	value	A	E	V
JAN	1862	1	0	1	9098	1	0	1	4374	1	0	1	16184	1	0	1
FEB	3263	1	0	1	12729	1	0	1	9590	1	0	1	38503	1	0	1
MAR	1871	1	0	1	7600	1	0	1	0	0	0	0	0	0	0	0
APR	2010	1	0	1	7872	1	0	1	0	0	0	0	0	0	0	0
MAY	3769	1	0	1	14529	1	0	1	0	0	0	0	0	0	0	0
JUN	2810	1	0	1	10773	1	0	1	0	0	0	0	0	0	0	0
JUL	1814	1	0	1	7011	1	0	1	0	0	0	0	0	0	0	0
AUG	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0
SEP	2387	1	0	1	9190	1	0	1	0	0	0	0	0	0	0	0
OCT	3796	1	0	1	16803	1	0	1	0	0	0	0	0	0	0	0
NOV	2383	1	0	1	10328	1	0	1	0	0	0	0	0	0	0	0
DIC	4487	1	0	1	17517	1	0	1	0	0	0	0	0	0	0	0

INFLUENCES

Branch=2511

N	Enterprise	Product	Actual data	Imputed Data	Influence
1	993401	25111000	8.123.500	14.085.972	-100.98
2	990842	25112100	28.300.000	15.865.460	47.28
3	990977	25111000	946.060	1.903.815	-16.22
4	084765	25111000	75.590	280.431	-3.47
5	991046	25112100	1.071.000	828.610	1.53

product 2511.1000 hydrocarbon
enterprise 993401

	t-1						t					
	quantity	A	E	V	value	A E V	quantity	A	E	V	value	A E V
JAN	15650800	1	0	1	355020	1 0 1	13370900	1	0	1	156892	1 0 1
FEB	16487800	1	0	1	326753	1 0 1	8123500	1	0	1	99797	1 0 1
MAR	14355746	1	0	1	340572	1 0 1	0	0	0	0	0	0 0 0
APR	11908300	1	0	1	254307	1 0 1	0	0	0	0	0	0 0 0
MAY	14696100	1	0	1	265968	1 0 1	0	0	0	0	0	0 0 0
JUN	17997400	1	0	1	305320	1 0 1	0	0	0	0	0	0 0 0
JUL	12225200	1	0	1	203319	1 0 1	0	0	0	0	0	0 0 0
AUO	12884600	1	0	1	210218	1 0 1	0	0	0	0	0	0 0 0
SEP	13151200	1	0	1	200491	1 0 1	0	0	0	0	0	0 0 0
OCT	15299900	1	0	1	227920	1 0 1	0	0	0	0	0	0 0 0
NOV	12970000	1	0	1	196903	1 0 1	0	0	0	0	0	0 0 0
DIC	12437200	1	0	1	134224	1 0 1	0	0	0	0	0	0 0 0

INFLUENCES

Branch=4110

N	Enterprise	Product	Actual data	Imputed Data	Influence
1	230158	41101000	729.250	2.645.309	-78.66
2	230177	41101000	887.188	2.271.395	-56.83
3	230118	41101000	422.963	1.641.680	-50.03
4	230130	41101000	371.852	1.508.894	-46.68
5	230171	41101000	398.865	1.447.508	-43.05
6	230151	41101000	674.838	1.682.347	-41.36
7	230117	41101000	381.500	1.268.393	-36.41
8	230068	41101000	252.875	1.112.257	-35.28

INFLUENCES

Branch=4110

N	Enterprise	Product	Actual data	Imputed Data	Influence
9	230021	41101000	108.710	740.571	-25.94
10	230053	41101000	330.416	956.224	-25.69
11	230211	41101000	130.000	696.276	-23.25
12	230023	41101000	263.240	829.336	-23.24
13	140180	41101000	166.228	730.612	-23.17
14	450080	41101000	47.395	609.610	-23.08
15	230199	41101000	304.200	856.656	-22.68
16	230201	41101000	1.955.000	2.463.949	-20.89

Imputaciones

1) **Tradicional basada en los datos de otras empresas**

2) **Imputación basada en modelos para q_{ijt}**

3) **Imputación basada en modelos para I_{it}**

Trabajos en curso

- **Enfoque más sistemático**
- **Intensificar depuración selectiva**
- **Modelizar microdatos**

¡Cuantas estupideces cometemos con aire de riguroso razonamiento! Claro, razonamos bien, razonamos magníficamente sobre las premisas A, B y C. Sólo que no habíamos tenido en cuenta la premisa D.

Sobre héroes y tumbas

(ERNESTO SABATO)

“Basta, al que empieza, aborrecer el vicio, y el ánimo enseñar a ser modesto; después le será el cielo más propicio.”

Epístola moral a Fabio

(ANDRÉS FERNÁNDEZ DE ANDRADA (?))

Master Universitario en "Estadística Aplicada y Estadística para el Sector Público"

Indicadores Estadísticos y Socio-económicos

Tema 3. Índices Económicos

Apuntes preparados por Pedro Revilla Novella (previlla@ine.es)
Instituto Nacional de Estadística

Índices económicos

En estos apuntes se describen algunos de los principales índices económicos disponibles en España para el análisis de la coyuntura del sector industrial, comentando también sus perspectivas futuras.

I. Introducción

Tradicionalmente, para el seguimiento de la coyuntura del sector industrial se dispone de dos indicadores cuantitativos fundamentales: el Índice de Producción Industrial (IPI) y el Índice de Precios Industriales (IPRI). Estos índices tienen una amplia difusión, tanto para el estudio de las macromagnitudes a nivel agregado (índice general, bienes de equipo, bienes de consumo, etc.) como para el estudio detallado de los sectores y subsectores industriales.

En los últimos años, la información disponible sobre el sector industrial ha aumentado de forma considerable, principalmente a través de la Encuesta Industrial de Empresas y de la Encuesta Industrial de Productos. Estas encuestas ofrecen un amplio conjunto de datos, cubriendo diversos aspectos de la actividad de las empresas, como son la producción, las ventas, las exportaciones, los consumos intermedios y las inversiones. Adicionalmente, la integración de variables permite analizar la productividad, los costes laborales unitarios, la competitividad, la concentración, la rentabilidad, etc. Estos análisis, dados los diseños y tamaños muestrales, pueden hacerse a un elevado detalle de desagregación territorial y sectorial.

Sin embargo, esta información tiene únicamente carácter anual y, aunque se dispone de ella cada vez con mayor rapidez (actualmente se difunde al máximo detalle dentro del año

siguiente al de referencia), no ofrece la inmediatez ni la frecuencia que exige el análisis de la coyuntura.

La trascendencia de los indicadores coyunturales ha crecido en los últimos años por el desarrollo del Mercado Único y la implementación de la Unión Monetaria Europea. Al mismo tiempo, ha crecido el interés por indicadores más desagregados territorialmente, por ejemplo al nivel de las Comunidades Autónomas.

Para hacer frente a esta demanda de información, el IPI y el IPRI resultan insuficientes, siendo necesario que se complementen con otros indicadores que ofrezcan una perspectiva más amplia de la coyuntura del sector industrial. A medio plazo, las perspectivas deben orientarse a una cobertura completa de indicadores de oferta, demanda y precios, incluyendo índices de producción, de ventas, de exportaciones e importaciones, de inversiones, de stocks, de cartera de pedidos, de empleo y de precios en los diferentes mercados.

En el ámbito de la UE, el aumento de la demanda de indicadores coyunturales por parte de las instituciones y usuarios europeos, junto a la necesidad de crear un marco común para la producción de estadísticas coyunturales, han dado lugar a la adopción, en mayo de 1998, del Reglamento de Estadísticas Coyunturales, al que siguen todo un conjunto de disposiciones y normativas. En este sentido, existe ya una reciente modificación del citado Reglamento, que ha entrado en vigor, tras su aprobación por el Parlamento y Consejo europeos, el 6 de Julio de 2005. La modificación del Reglamento supone nuevas exigencias de información a los Estados Miembros, como la inclusión de nuevas variables (por ejemplo, los precios de importación), el acortamiento de los plazos de difusión, etc.

En este contexto, el INE se plantea como objetivo estratégico la elaboración de un conjunto de indicadores coyunturales, de manera coordinada con el resto de países de la UE. En el año 2003 se presentaron cuatro proyectos fundamentales: los cambios de base de los Índices de Producción y Precios Industriales y la publicación de dos nuevos indicadores, los Índices de Cifras de Negocios (ICN) y los Índices de Entradas de Pedidos (IEP). A finales de 2006 se difundieron por primera vez otros dos nuevos indicadores: los Índices de Precios de Exportación y los Índices de Precios de Importación. A continuación se describen los principales rasgos de estos proyectos.

II. Los Índices de Producción y de Precios Industriales

En el año 2003 se difundieron los primeros Índices de Producción y de Precios Industriales elaborados en base 2000. Los nuevos índices, además de la necesaria puesta al día de las ponderaciones, presentan una completa actualización de su arquitectura analítica, adaptándola a la estructura industrial española de la nueva década. Al mismo tiempo, se introducen un conjunto de novedades metodológicas que permitirán una mayor rapidez en la difusión de los resultados, mayor representatividad y precisión, disminución de las revisiones de las primeras estimaciones y mayor comparabilidad con los índices de los países de la Unión Europea. Una de las principales novedades es el cálculo de índices por Comunidades Autónomas.

Objetivos y principales utilizaciones

Los Índices de Producción Industrial tienen como objetivo medir la evolución mensual del volumen del valor añadido bruto generado por las distintas ramas industriales y por el total de la industria. Por tanto, miden la evolución conjunta de la cantidad y la calidad, excluyendo la influencia de los precios. Cuentan con gran tradición en el análisis coyuntural de la mayor

parte de los países, siendo demandados por diferentes tipos de usuarios. Constituyen uno de los elementos básicos para la elaboración de la Contabilidad Nacional Trimestral. También, calculados por destino económico, son utilizados como indicador de la demanda agregada, teniendo especial relevancia para el seguimiento de la inversión en bienes de equipo. A nivel desagregado, son utilizados por los empresarios industriales para comparar la evolución de su producción con el conjunto de empresas de su mismo sector, o para seguir la evolución de su sector dentro del conjunto de la industria.

Por su parte, los Índices de Precios Industriales tienen como objetivo medir la evolución de los precios de los productos fabricados por el sector industrial en la primera etapa de su comercialización. Tienen diversos usos, tanto analíticos como administrativos. En primer lugar, tienen una utilización directa para el análisis de los precios, detectando, en sus inicios, las presiones inflacionistas. Su estudio conjunto con el IPC es de interés para conocer el mecanismo de formación de los precios. Por otra parte, tienen una utilización indirecta como deflactores, por ejemplo para deflactar las series en valor del IPI. Por último, también se emplean para llevar a cabo las revisiones en los precios de determinados contratos.

La necesidad de un cambio de base

Los índices de base fija pierden paulatinamente su representatividad, a medida que la estructura industrial del periodo que se analiza se va distanciando de la establecida según el año base. Por este motivo, cada cierto tiempo resulta necesario actualizar dicha estructura analítica, llevando a cabo lo que se conoce como un cambio de base. En el caso del IPI y del IPRI españoles, se hacía conveniente una renovación, dadas la antigüedad de sus bases (1990) y las profundas transformaciones que se han producido en el tejido industrial desde entonces.

El año elegido como nueva base es 2000, año común de referencia en los países de la Unión Europea.

Para llevar a cabo el cambio de base se realiza un estudio de la estructura y de la producción en la industria por medio de la información estadística existente, fundamentalmente la que proporcionan la Encuesta Industrial de Empresas y la Encuesta Industrial de Productos. El adelanto que se ha producido en la elaboración de estas encuestas ha permitido que el desfase entre el año base y el de publicación de los primeros índices, uno de los principales problemas prácticos a los que se enfrentan los indicadores de base fija, no sea excesivo. Los índices en la nueva base estuvieron disponibles a comienzos de 2003, primera fecha prevista en las normativas europeas para que los países miembros los transmitan a Eurostat.

El cambio de base supone una profunda modificación en la arquitectura de los índices. La renovación de las ponderaciones, de la cesta de productos representativos y del panel de unidades informantes va acompañada de la inclusión de nuevas ramas de actividad y del aumento en el tamaño de la muestra.

La puesta al día de las ponderaciones

Las nuevas ponderaciones se calculan de acuerdo a la importancia relativa de las ramas de actividad y de los productos en el año 2000, según la información que suministra la Encuesta Industrial de Empresas y la Encuesta Industrial de Productos. En el caso del IPI, las ponderaciones se establecen según el valor añadido bruto, a nivel de ramas de actividad, y según el valor de producción, a nivel de productos. En el caso del IPRI, según el valor de la cifra de negocios en ambos niveles.

Las nuevas ponderaciones, comparadas con las antiguas, reflejan las transformaciones que se han producido en la estructura industrial española en los años 90. Así, pierden peso sectores tradicionales como el textil, la confección, el cuero, el calzado, la metalurgia básica, y los alimentos, y, por el contrario, ganan participación el sector de refino, la fabricación de productos metálicos, la maquinaria y la edición.

La renovación de la cesta de productos representativos

La cesta de productos representativos de cada rama de actividad se actualiza para introducir nuevos productos no tenidos en cuenta en la base precedente y para descartar productos cuyo peso económico ha perdido importancia. La nueva cesta se amplia para hacer frente al proceso de diversificación de la producción que ha caracterizado al sector industrial español en los últimos años y para poder calcular índices por Comunidades Autónomas que sean suficientemente representativos. La selección de productos se efectúa, para cada rama de actividad, atendiendo a la importancia de su valor de producción.

Un nuevo panel de establecimientos informantes

El nuevo panel de informantes se selecciona teniendo en cuenta la aportación de los establecimientos en la producción de los artículos contenidos en la cesta. Por tanto, para cada producto, se selecciona un conjunto de establecimientos que representen la mayor parte de la producción del mismo. En la mayoría de los productos, los establecimientos incluidos en el panel superan el 90% de la producción o de la cifra de negocios.

Ampliación de la cobertura y del tamaño de la muestra

Los nuevos índices investigan un conjunto de ramas de actividad que no eran objeto de seguimiento en los antiguos. Con la inclusión de estas nuevas actividades la cobertura de los índices se extiende a prácticamente la totalidad de las ramas de actividad del sector industrial. Únicamente se dejan de investigar aquellas que tienen una producción poco significativa, que representan actividades próximas a la artesanía, o que son poco relevantes para el análisis coyuntural.

Los nuevos índices incorporan un conjunto de ampliaciones cuantitativas en los distintos elementos que integran su estructura analítica. De esta forma, además de incluir nuevas actividades, se aumenta el número de productos de la cesta y el tamaño de la muestra. Estos aumentos no son consecuencia únicamente de la inclusión de nuevas actividades, sino que van encaminados a incrementar la precisión y robustez de los índices y a poder realizar las estimaciones a nivel de Comunidades Autónomas.

Armonización con la metodología de la Unión Europea

La metodología de los nuevos índices se ha adaptado a las normativas y acuerdos comunitarios en materia de índices industriales, expresados en el Reglamento de Estadísticas Coyunturales y en el correspondiente manual metodológico.

Estas directrices, que han sido el resultado de frecuentes reuniones entre los expertos estadísticos de los distintos países comunitarios, coordinados por Eurostat, se extienden no sólo al período que se toma como referencia, el año 2000, sino a todo el conjunto de aspectos

metodológicos: cobertura, clasificaciones de actividad y de destino económico, definición de las variables a investigar, método de cálculo de las ponderaciones, etc. Esta armonización supone una importante mejora de cara al seguimiento coyuntural de la producción y de los precios españoles en relación con los del resto de países europeos, al poder utilizarse índices directamente comparables, incluso a un elevado detalle de desagregación sectorial y territorial.

Nueva base 2005=100

En el año 2008 se procederá a efectuar el cambio a la nueva base, 2005=100, en el IPI y en el IPRI, de acuerdo a las normativas de la UE.

III. Los Índices de Cifras de Negocios y de Entradas de Pedidos

En el mes de septiembre del año 2003 el INE publicó por primera vez los Índices de Cifras de Negocios y de Entradas de Pedidos. Estos dos nuevos índices, junto con el Índice de Producción Industrial, se orientan a medir la evolución mensual de la actividad. Mientras que el IPI tiene por objeto medir la actividad por el lado de la oferta que genera el sector industrial, los otros dos indicadores lo hacen por el lado de la demanda que se dirige a este sector, tanto presente (Índices de Cifras de Negocios) como futura (Índices de Entradas de Pedidos).

Las cifras de negocios son las cantidades en términos monetarios facturadas por la venta de bienes y servicios prestados a terceros. Por su parte, las entradas de pedidos son las cantidades facturadas o por facturar definidas de igual forma que la cifra de negocios y referidas a los

pedidos recibidos y definitivamente aceptados. Puede suponerse que los pedidos se reciben en la empresa antes que la producción y esta se realiza antes que las ventas. Por tanto el IEP será un indicador adelantado del IPI y este a su vez del ICN. De hecho, el principal uso del IEP es como indicador adelantado del ciclo industrial.

Tanto las cifras de negocios como los pedidos se desglosan en los diferentes mercados geográficos. De este modo, se distingue entre el mercado interior y el mercado exterior. A su vez, el mercado exterior se divide entre la Unión Europea y el resto del mundo. Por último, la Unión Europea se subdivide entre la Zona Euro y la Zona no Euro.

El interés del seguimiento de la variable cifra de negocios deriva de su propia importancia dentro de la actividad empresarial. La cifra de negocios constituye normalmente la parte fundamental de los ingresos de explotación de las empresas, es decir, de los ingresos que proceden del ejercicio de su actividad habitual. Estos, a su vez, constituyen la parte fundamental de los ingresos totales.

El objetivo de un índice de cifras de negocios es mostrar la evolución de los mercados de los sectores industriales. Esta información tiene interés para los inversores y para la toma de decisiones de política económica y monetaria. Del mismo modo, las propias empresas industriales pueden evaluar la evolución de sus cuotas de mercado y así controlar el éxito de sus estrategias productivas y comerciales.

Si bien en algunos casos se considera que el ICN tiene grandes similitudes con el IPI, en realidad las diferencias entre los dos indicadores son notables. En primer lugar, el ICN es un indicador que mide la evolución de la actividad en términos corrientes, mientras que el IPI lo hace en términos constantes.

Por otra parte, el ICN es un indicador de ventas en tanto que el IPI es un indicador de producción, de forma que las cantidades producidas que no son vendidas en el mes e incrementan los stocks entrarían a formar parte del IPI pero no del ICN, y, alternativamente, las cantidades vendidas procedentes de los stocks que han sido producidas en meses anteriores entrarían a formar parte del ICN pero no del IPI.

Otra diferencia entre los dos indicadores se encuentra en las mercaderías, es decir, en la reventa de productos en el mismo estado que se adquirieron, que formarían parte del ICN pero no del IPI.

Por último, el ICN es un indicador que se basa en el concepto de sector como suma de la actividad de las empresas pertenecientes al mismo. Por tanto, contempla las ventas totales, incluyendo las actividades secundarias; en tanto que el IPI, construido a partir de una cesta de productos representativos, se basa en el concepto más homogéneo de rama de actividad, incluyendo únicamente la actividad principal y no las secundarias.

La cifra de negocios es un concepto de uso común dentro de la contabilidad de las empresas. Su definición puede establecerse en términos de las partidas contables del Plan General de Contabilidad, lo que facilita notablemente la recogida de la información. Por otra parte, su definición está armonizada con la información anual que recoge la Encuesta Industrial de Empresas, lo que permitirá la coherencia entre la información estructural y coyuntural.

La variable pedidos tiene mayor dificultad de definición dentro de los esquemas contables, presentando además una notable heterogeneidad dentro de las distintas prácticas empresariales, variando de sector a sector, e incluso de empresa a empresa. En principio, se

puede considerar como pedido el valor del acuerdo, cualquiera que sea la forma que éste adopte (verbal, escrito, etc.), por el cual el productor se obliga a suministrar unos bienes o a prestar unos servicios a un tercero, tanto si son realizados por él como si proceden de la subcontratación. Al implicar el pedido una venta futura de bienes y servicios, se toman como rúbricas a considerar en su definición las mismas que las de las cifras de negocios. Se consideran las Entradas de Pedidos como el valor de los pedidos recibidos y aceptados en firme por la empresa durante el mes de referencia.

La cobertura de los índices se extiende a todos los sectores industriales, excluida la construcción. En términos de la Clasificación Nacional de Actividades Económicas (CNAE-93) cubre las secciones C, D y E. Por tanto, investiga la minería, las manufacturas y las utilidades públicas (gas y electricidad).

Dado que no todas las empresas organizan su producción sobre la base de los pedidos y aquellas que lo hacen suelen estar concentradas en determinadas actividades (por ejemplo, las de largo periodo de fabricación), la cobertura de un índice de pedidos no se extiende necesariamente a todos los subsectores industriales. Dentro del ámbito europeo se considera (y se establece de forma obligatoria su estudio en el Reglamento) como subsectores que habitualmente trabajan a pedido los siguientes: textil, confección, papel, química, metalurgia, fabricación de productos metálicos, maquinaria y equipo mecánico, máquinas de oficina y equipos informáticos, maquinaria y material eléctrico, material electrónico, instrumentos de precisión y equipo médico, automoción y otro material de transporte. No obstante, el INE ha optado por tener una cobertura más amplia, y estudiar todos los sectores industriales.

Para obtener la información básica necesaria para el cálculo de los índices, el INE ha iniciado en el año 2002 una encuesta mensual dirigida a más de 13.000 establecimientos industriales.

La encuesta recoge conjuntamente los dos tipos de variables. El panel de unidades informantes está coordinado con el del IPI, con el objeto de hacer comparables los tres indicadores de actividad del sector industrial.

En el año 2008 se procederá a efectuar el cambio a la nueva base, 2005=100, en el ICN y en el IEP, de acuerdo a las normativas de la UE.

IV. Los Índices de Precios de Exportación y de Importación

Los Índices de Precios de Exportación (IPRIX) y los Índices de Precios de Importación (IPRIM) constituyen dos nuevos indicadores dentro de ámbito de los indicadores industriales coyunturales. Con carácter provisional, los primeros resultados comenzaron a difundirse a partir de noviembre de 2006. Los índices se elaboran con base 2005=100.

Estos indicadores responden a la necesidad de hacer frente a la demanda de información sobre la evolución de los precios de comercio exterior, en una economía crecientemente internacionalizada como es la economía española.

Tradicionalmente, para el seguimiento de los precios de los flujos de exportación e importación, se contaba únicamente en España, como variable “proxy”, con los Índices de Valor Unitario de comercio exterior (IVU), que elabora mensualmente el Ministerio de Economía y Hacienda, a partir de los datos de comercio exterior del departamento de Aduanas e Impuestos Especiales de la Agencia Estatal Tributaria.

Por este motivo, se hace necesario elaborar índices específicos de precios, calculados a partir de la observación directa de los precios de determinadas subvariedades de productos, que reflejen variaciones "puras" de precios y no variaciones ocasionadas por la calidad de los productos o por el efecto composición, como es característico en los valores unitarios.

La necesidad de la elaboración de estos índices de precios de comercio exterior responde a la satisfacción de la demanda, tanto de usuarios en el ámbito español como en el de la UE. De hecho, la implantación de estos dos indicadores es de obligado cumplimiento para los países miembros de la Unión Europea, tras la adopción, en mayo de 1998, del Reglamento (CE) 1165/98 del Consejo sobre las Estadísticas Coyunturales, y su posterior modificación en Julio de 2005.

En este contexto, como se ha comentado en la introducción, el INE se plantea la implantación de un conjunto de indicadores coyunturales, de manera coordinada con el resto de países de la UE. El plan estratégico del INE se orienta a la obtención de una cobertura completa de indicadores, plenamente integrados, que permita un análisis coherente de la coyuntura industrial desde diversas perspectivas. Fruto de este plan estratégico fue la implantación, en el mes de septiembre del año 2003, de dos nuevos indicadores: los Índices de Cifras de Negocios y de Entradas de Pedidos. Estos dos nuevos índices, junto con el Índice de Producción Industrial, se orientan a medir la evolución mensual de la actividad.

Para el análisis de los precios, se hace necesario completar el IPRI, que mide los precios de los bienes industriales vendidos en el mercado interior, con indicadores que midan los precios de los flujos de comercio exterior, tanto de los bienes industriales exportados (Índices de Precios de Exportación de Productos Industriales) como importados (Índices de Precios de Importación de Productos Industriales).

Bibliografía

- Dollt, A. (2001). The industrial production index. Methodology of industrial short-term statistics. Rules and recomendations. Eurostat. www.forum.europa.eu.int.

- Pereira, H. (2001). Turnover index in industry. Methodology of industrial short-term statistics. Rules and recomendations. Eurostat. www.forum.europa.eu.int.

- Revilla, P. (2001). Spanish methods to improve timeliness in the industrial production indices. International Seminar on Short-Term Statistics. Eurostat. Office for Official Publications of the European Communities.

- Revilla, P. (1994). La modernización de las estadísticas industriales. Hacia un sistema integrado de encuestas. Economía Industrial. Nº 299.