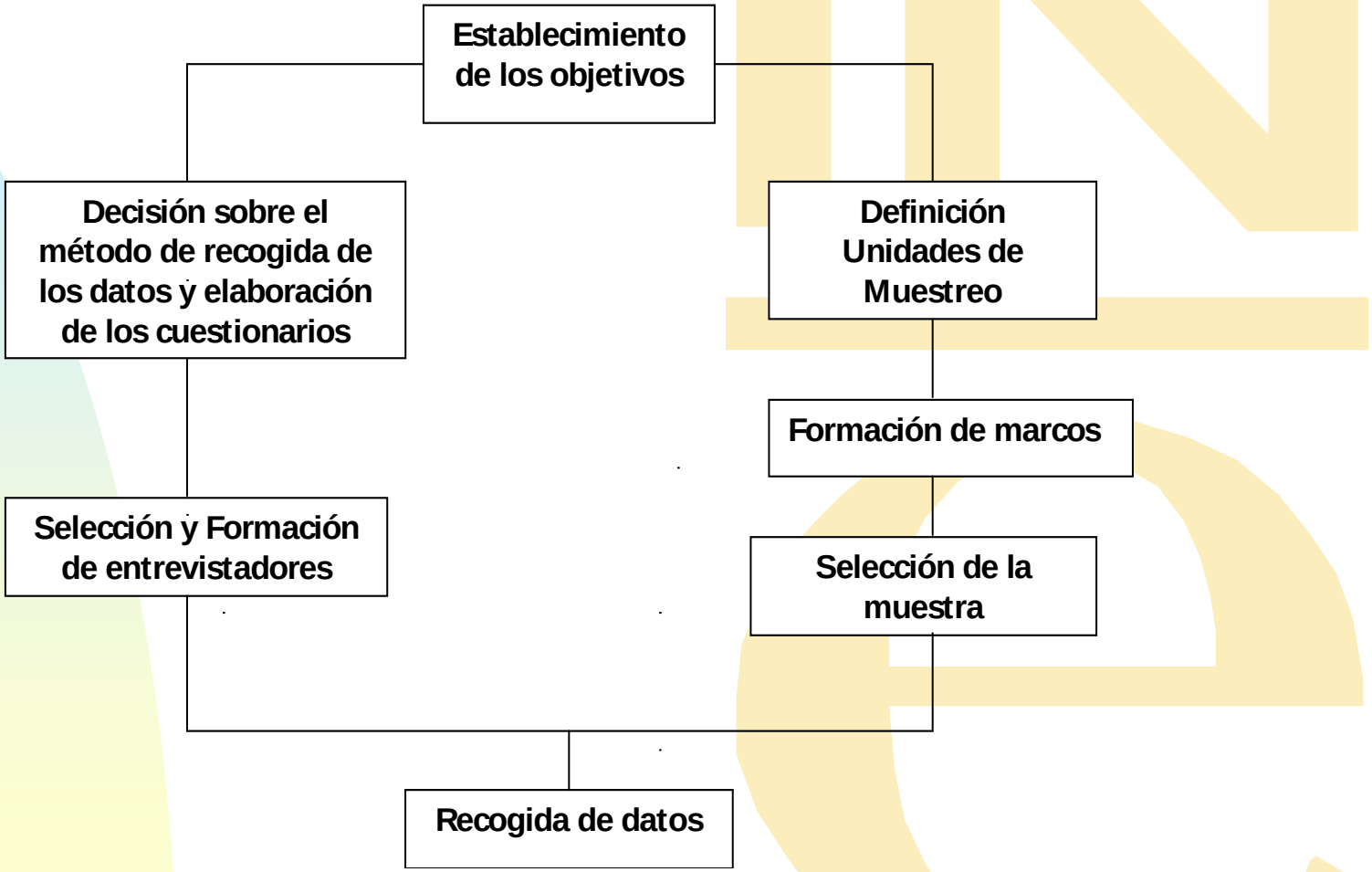
CIFF
Centro Internacional
de Formación Financiera

Juana Porras Puga

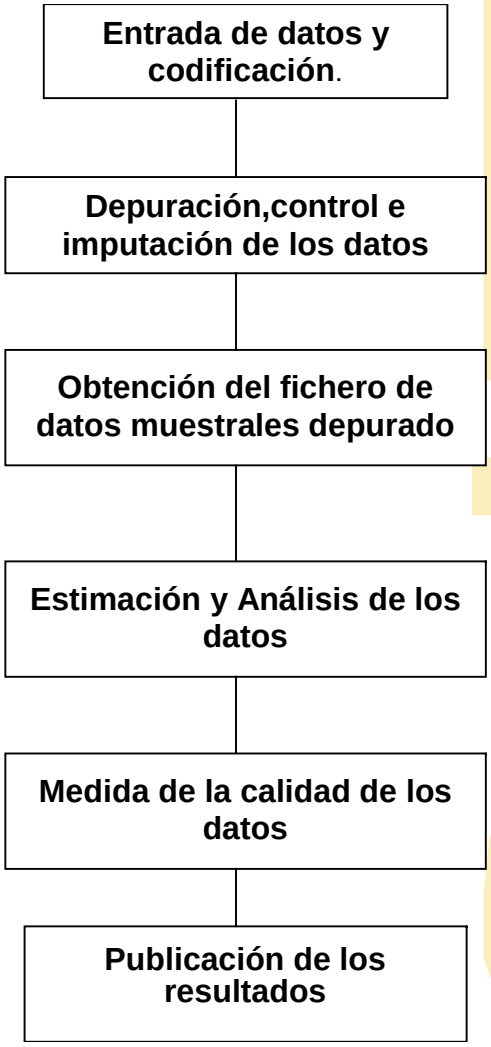
Diseño muestral de las principales encuestas realizadas en el INE.

Juana Porras Puga
Subdirección General de Metodología y Técnicas Estadísticas.
2 de Abril de 2009

ESQUEMA DE LAS FASES DE UN PROYECTO ESTADÍSTICO(I)



ESQUEMA DE LAS FASES DE UN PROYECTO ESTADÍSTICO(II)



Estas fases pueden agrupar en las siguientes:

1. SELECCIÓN DE LA MUESTRA: Resultado de la aplicación de todas las etapas del diseño muestral**.**

2. RECOGIDA DE DATOS. Los datos se recogen con los instrumentos de medida planificados.

3. PROCESO DE LOS DATOS. Preparación de los datos para la estimación y análisis.

4. ESTIMACIÓN Y ANÁLISIS. Cálculo de las estimaciones de acuerdo con el diseño muestral, uso de información auxiliar y ajustes de no respuesta.

5. PUBLICACIÓN DE LOS RESULTADOS. Publicación de los resultados , incluyendo la metodología general del proyecto.

TIPOS DE ENCUESTAS REALIZADOS EN EL INE:

Según el *ámbito poblacional*:

- Encuestas de **Población**:Dirigidas a los hogares.
- Encuestas **Económicas**:Dirigidas a las empresas.

Según la *periodicidad:*

- Encuestas **Estructurales**: (mayor de un año)
- Encuestas **Coyunturales**: (Inferior a un año)

- **Encuestas continuas**
  - Encuesta de Población Activa
  - Encuesta Continua de Presupuestos Familiares
  - Encuesta de Condiciones de Vida
  - Encuesta de uso del TIC y comunicación en los Hogares.
- **Encuestas esporádicas**
  - Encuesta de Empleo del Tiempo
  - Encuesta de Discapacidades, Autonomía Personal y Situaciones de Dependencia
  - Encuesta Nacional de Salud
  - Encuesta de Hábitos Sexuales

- **Encuestas económicas estructurales:**
  - Encuesta Anual de Servicios
  - Encuesta Industrial Anual de Empresas
  - Encuesta de Innovación Tecnológica
  - Encuesta de uso del TIC y Comercio Electrónico
  - Encuestas medioambientales
  - Encuesta Industrial de Productos
  - Encuesta sobre la Estructura de las Explotaciones Agrícolas
- **Encuestas económicas coyunturales**
  - Índices coyunturales (ICM,IASS,ETCL)
  - Encuesta mensual de transportes de viajeros
  - Encuestas de Turismo

Encuesta	Periodicidad	U. elementales	Tamaño	Ambito Geog.	Rotación(%)
(Hogares)					
Encuesta de Población Activa(EPA)	Trimestral	Viviendas	70.000	Provincia	16,66
(Encuesta Continua de Presupuestos Familiares(ECPF)	Anual	Hogares	22.000	Com. Autonoma	50
Encuesta de Condiciones de Vida(EU-SILC)	Anual	Hogares	16.000	Com. Autonoma	25
Tecnologías de la Información y la Comunicación(TIC-H)	Semestral	Hogares	21.000	Com. Autonoma	25
Encuesta Nacional de Salud(ENS)	Esporádica	Hogares	31.000	Com. Autonoma	
Transición E_F e Inserción Laboral(ETEFIL)	Esporádica	Alumnos	45.000	Com. Autonoma	
Encuesta Nacional de Inmigrantes(ENI)	Esporádica	Inmigrantes	15.000	Com. Autonoma	
Encuesta de Personas sin Hogar(EPSH)	Esporádica	Peronas sin hogar	4.000	Com. Autonoma	
(Empresas)					
Encuesta Industrial Anual(EIA)	Anual	Empresa industrial	47.000	Com. Autonoma	
Encuesta Anual de Servivcios(EAS)	Anual	Empresas	147.000	Com. Autonoma	
Innovación tecnológica	Anual	Empresas	32.000	Com. Autonoma	
I+D(exhaustiva)	Anual	Empresas	21.000	Com. Autonoma	
Generación de Residuos(GR)	Anual	Empresas y estab.	15.000	Nacional	
Estructura de las Explotaciones Agrícolas(EEAA)	Bienal	Explotaciones	55.000	Com. Autonoma	
Coste Laboral(ETCL)	Trimestral	Centros	28.000	Com. Autonoma	20
Transporte de viajeros por carretera(TV)	Mensual	Empresas	700	Nacional	25
Indicadores de Actividad del Sector Servicios(IASS)	Mensual	Empresas	27.000	Com. Autonoma	25(anual)
Indice de Comercio al por menor(ICM)	Mensual	Empresas	13.000	Com. Autonoma	25(anual)
Ocupación Turística	Mensual	Establecimientos	15.000	Provincial	25(anual)

ESQUEMA GENERAL DEL DISEÑO MUESTRAL.

Las siguientes características son **comunes** a todas las encuestas tanto de hogares como económicas.

- **Ámbito**: Poblacional, Geográfico y Temporal

- **Marco:** Relación de unidades de muestreo junto con la información auxiliar sobre las mismas.

  Marco de Áreas: Muestreo de conglomerados
  Marco de Listas: Muestreo de unidades elementales

- **Estratificación**: Criterios y variables utilizadas en la estratificación: Variables contenidas en el marco.

- **Tamaño de la muestra:** Se establece en función de:

1- La desagregación requerida para las estimaciones.

2- La dispersión de la(s) variables(s) objetivo.

3- Límites establecidos por el Servicio Promotor.

Resulta muy útil la experiencia de otras encuestas similares

- **Tipo de muestreo**: Proceso mediante el cual se selecciona la muestra.

- **Estimadores**: El estimador se obtiene siguiendo los siguientes pasos:

1. **Peso de diseño**: Diferentes probabilidades de selección. (Estimador de Horvitz-Thompson)
2. **Corrección de falta de respuesta**: Corrección del sesgo en las estimaciones.
3. **Aplicación de Técnicas de calibrado**:Mejoran la precisión de las estimaciones, con la información proporcionada por fuentes externas.

- **Calidad de los datos.**

  - **Errores de muestreo**: Debidos al hecho de estimar las características de la población a partir del estudio de una muestra. En diseños complejos se calculan mediante métodos indirectos. Semimuestras reiteradas, Conglomerados últimos, Jacknife etc..

  - **Errores ajenos al muestreo**:Se presentan en todas las etapas del desarrollo de una encuesta.Son difíciles de medir.

Encuestas a hogares

- Problemas de marco: viviendas vacías,...
- Muestreo Multietápico
- Marcos de areas y de lista
- Recogida de datos por entrevista personal o telefónica
- Carga de trabajo del informante.
- Coste elevado
- Variables cualitativas
- Afijación de compromiso
- Distribución temporal de la muestra

Encuestas a empresas

- Problemas de marco: empresas mal clasificadas...
- Muestreo monoetápico
- Marco de lista
- Recogida de datos por correo y apoyo telefónico
- Mayor carga de trabajo
- Menor coste
- Variables cuantitativas
- Afijación óptima
- En general no es necesaria la distribución temporal

ENCUESTA DE POBLACIÓN ACTIVA

Diseño de la muestra

INTRODUCCIÓN

- **La encuesta de Población Activa (E.P.A) es una encuesta de tipo continuo, con periodicidad trimestral que se viene realizando ininterrumpidamente desde 1964.**

- **Su objetivo es el *conocimiento de la actividad económica del país en lo relativo al componente humano*.**

¿Por qué no utilizar el Censo para conseguir este objetivo?

-Larga periodicidad

-Datos insuficientes

-Autoenumeración

Presentación de resultados:

-Detallados a escala nacional

-Principales características al nivel de comunidad autónoma y provincia.

ESQUEMA

1. ÁMBITO
2. MARCO DE LA ENCUESTA
3. TIPO DE MUESTREO
4. CRITERIOS DE ESTRATIFICACIÓN
5. TAMAÑO Y AFIJACIÓN DE LA MUESTRA
6. SELECCIÓN
7. DISTRIBUCIÓN DE LA MUESTRA EN EL TIEMPO
8. RENOVACIÓN PARCIAL DE LA MUESTRA
9. ESTIMADORES
10. ERRORES DE MUESTREO
11. ACTUALIZACIÓN DE LAS UNIDADES DE MUESTREO

1. ÁMBITO

Poblacional:
Población que reside en viviendas familiares principales, Se excluyen los hogares colectivos.

Geográfico:
Territorio nacional.

Temporal:
Resultados de la encuesta: Trimestre.
Información recogida: Semana anterior a la de la entrevista (semana de referencia).

2. MARCO DE LA ENCUESTA

Relación de unidades que van a ser muestreadas junto con toda la información complementaria que se puede utilizar en el diseño de la encuesta

Se utilizan dos marcos:

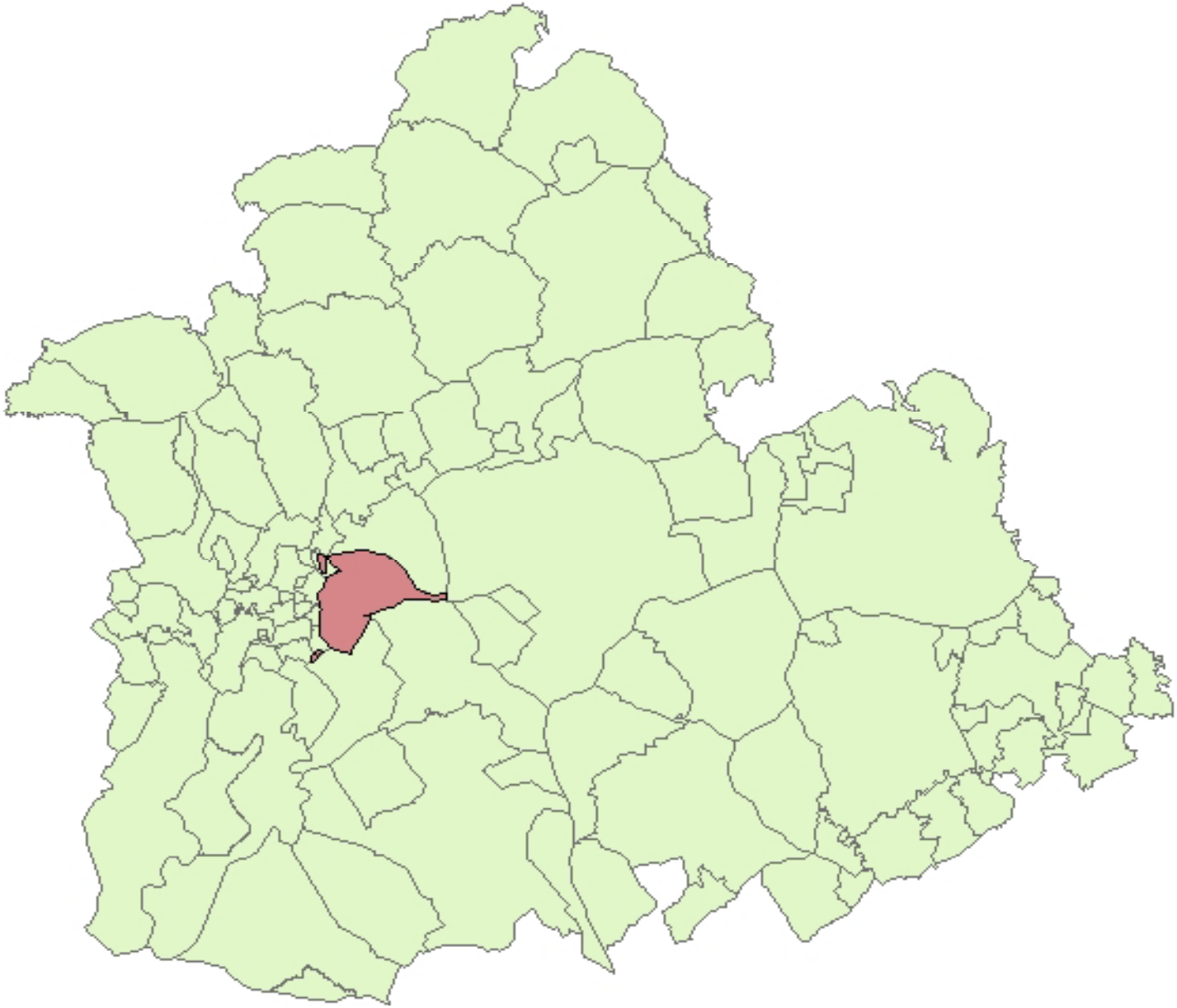
- **Marco de áreas geográficas**:
  - Comunidades Autónomas
  - Provincias
  - Municipios - Actualmente 8.200
  - Distritos municipales
  - Secciones censales. Aproximadamente 34.000

2.1 Sección censal

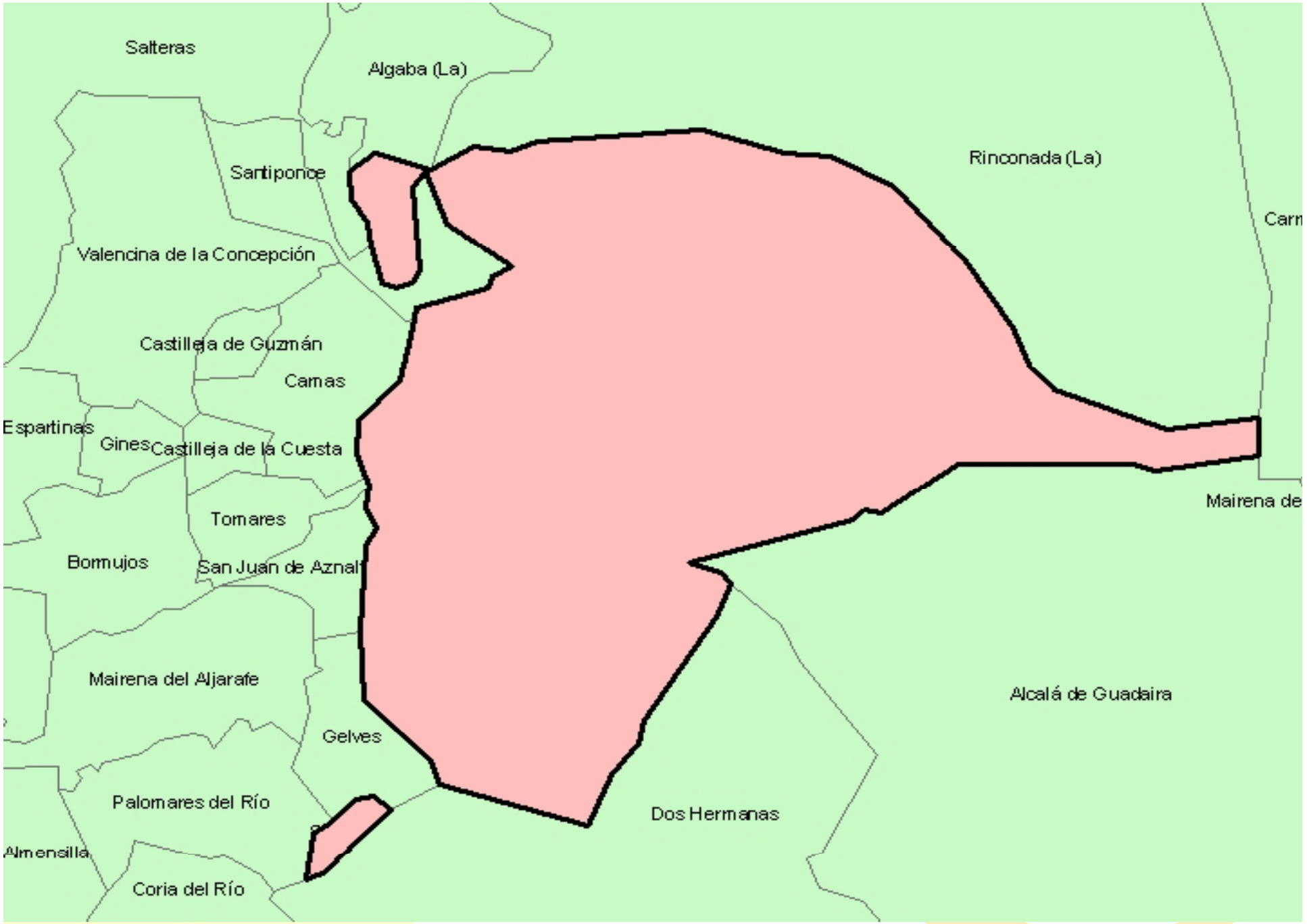
Área geográfica en que se divide el territorio nacional, utilizada con fines estadísticos y electorales.

Características:

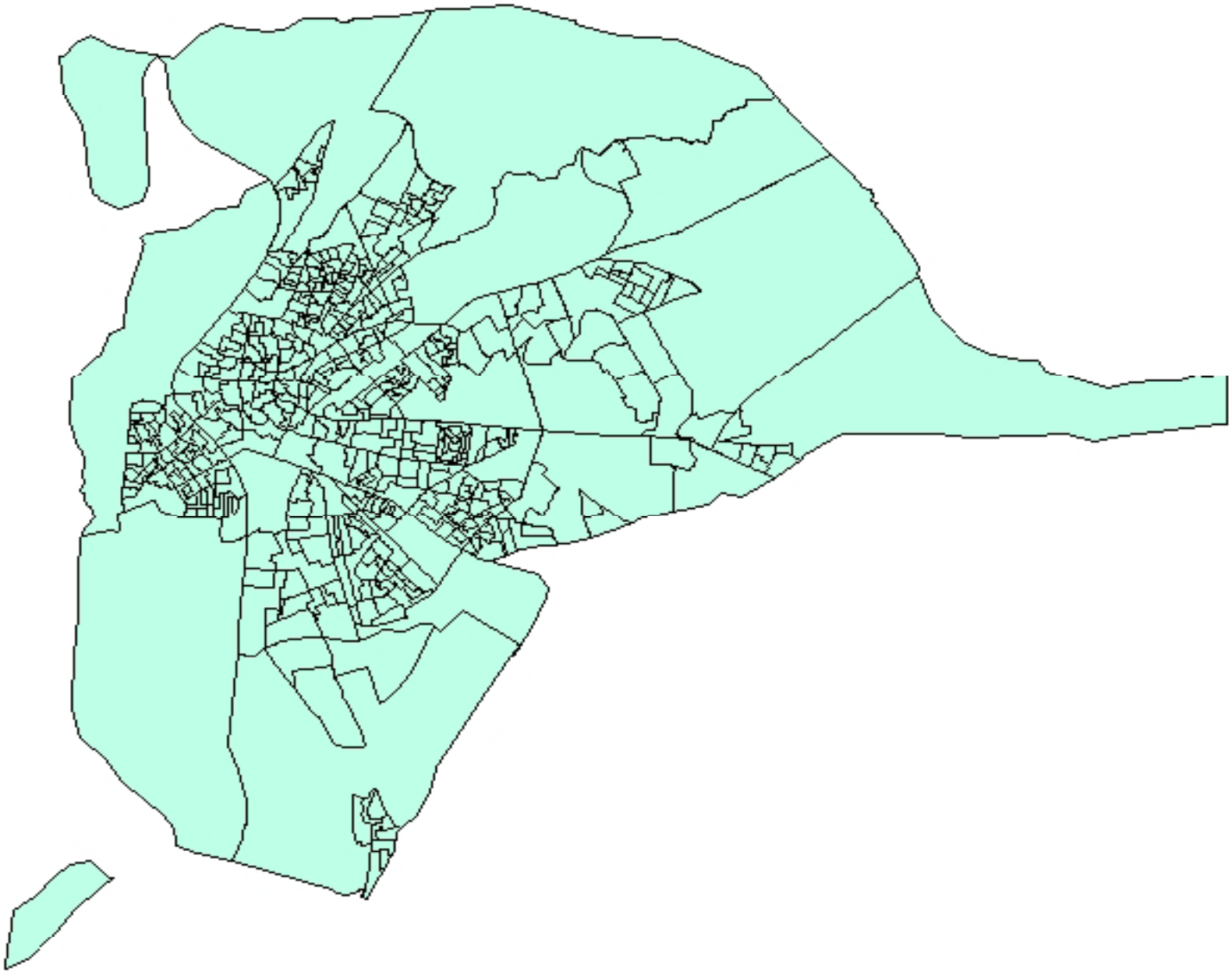
- **Perfectamente definida con límites fácilmente identificables**
- **El tamaño debe variar de acuerdo con la Ley General Electoral entre 500 y 2.000 electores**
- **Cualquier parte del territorio nacional debe pertenecer a una sola sección**



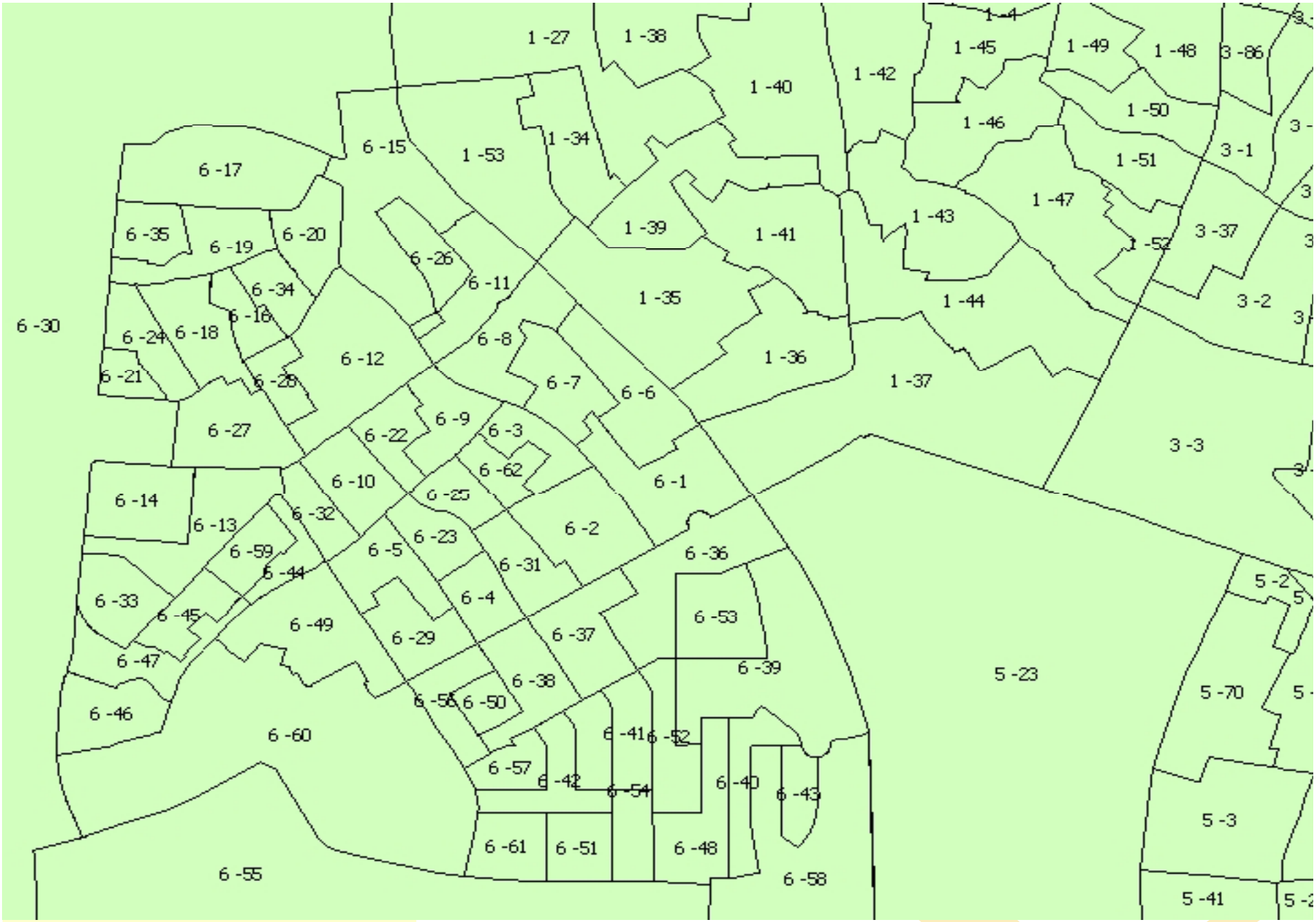
SEVILLA


Salteras
Algaba (La)
Rinconada (La)
Santiponce
Valencina de la Concepción
Castilleja de Guzmán
Camas
Espartinas
Gines
Castilleja de la Cuesta
Tomares
Bormujos
San Juan de Aznalf
Mairena del Aljarafe
Gelves
Palomares del Río
Almensilla
Coria del Río
Dos Hermanas
Alcalá de Guadaira
Mairena de

SEVILLA



SEVILLA-Municipio



SEVILLA-Municipio-Secciones

2.2 Marco de viviendas

- Relación de todas las viviendas familiares con sus direcciones postales, en cada una de las secciones censales seleccionadas para la encuesta

- En el caso de la EPA se obtiene en cada censo a partir de los cuadernos de recorrido utilizados en los trabajos censales y se elabora una vez finalizada la fase de recogida de los cuestionarios censales. Se actualiza periódicamente

- En las encuestas esporádicas se obtiene de la explotación del Padrón Continuo

2.3 Utilización del Censo en la formación del marco

- Fuente de información desagregada a nivel de unidades primarias de muestreo: Estratificación y subestratificación

- Instrumento para la formación del marco de viviendas, unidades de segunda etapa

- Actualización de la cartografía

Instituto Nacional de Estadística

CALLEJERO DE LA SECCIÓN

Provincia 41
Municipio 91
T. Secc **De Núcleo** **Distrito** 01 **Sección** 25

CCSSNN	Nombre de la Vía	Manzana	Tp	Numeración de	Numeración hasta	Numeración Clave	Observaciones
00 02 01	ANICETO SAENZ,CALLE	1		001	009Z	21	1
00 02 01	ANICETO SAENZ,CALLE	2		001	0023	25	1
00 02 01	ANICETO SAENZ,CALLE	3		002	004	24	2
00 02 01	ANICETO SAENZ,CALLE	4		002	0030	34	2
00 02 01	ANICETO SAENZ,CALLE	14		001	001	1	1
00 02 01	ANTONIA SAENZ,CALLE	2		002	0012	12	2
00 02 01	ANTONIA SAENZ,CALLE	3		001	003	7	1
00 02 01	ANTONIA SAENZ,CALLE	4		002	002	6	2
00 02 01	CETINA,CALLE	16		002	002	16	2
00 02 01	DUQUE DE MONTEMAR,CALLE	8		001	001	27	1
00 02 01	EUSTAQUIO BARRON,CALLE	14		002	004	8	2
00 02 01	EUSTAQUIO BARRON,CALLE	15		001	001	5	1
00 02 01	FLECHA,CLLON	12		001	001	1	1

Instituto Nacional de Estadística
Índice de viviendas

ENCUESTA DE POBLACIÓN ACTIVA .....

CÓDIGO ENCUESTA		
41	02106	2
PROV	SECCIÓN	D.C

Provincia: SEVILLA | 41 |

Municipio SEVILLA | 091 |

Distrito: | 01 | **Sección:** | 025 |

Fecha		Total inscritos	Viviendas				Colectivos	Locales
Mes	Año		Total	Habitadas	Vacías	Secundarias		
NOVIEMBRE	2001	1438	1246	817	338	91	2	190
JULIO	2005	1438	1247	966	238	43	2	189

Ent. Colectiva 0

Ent. Singular 02 SEVILLA

Núcleo/Dic 01 SEVILLA

Encuesta de Población Activa

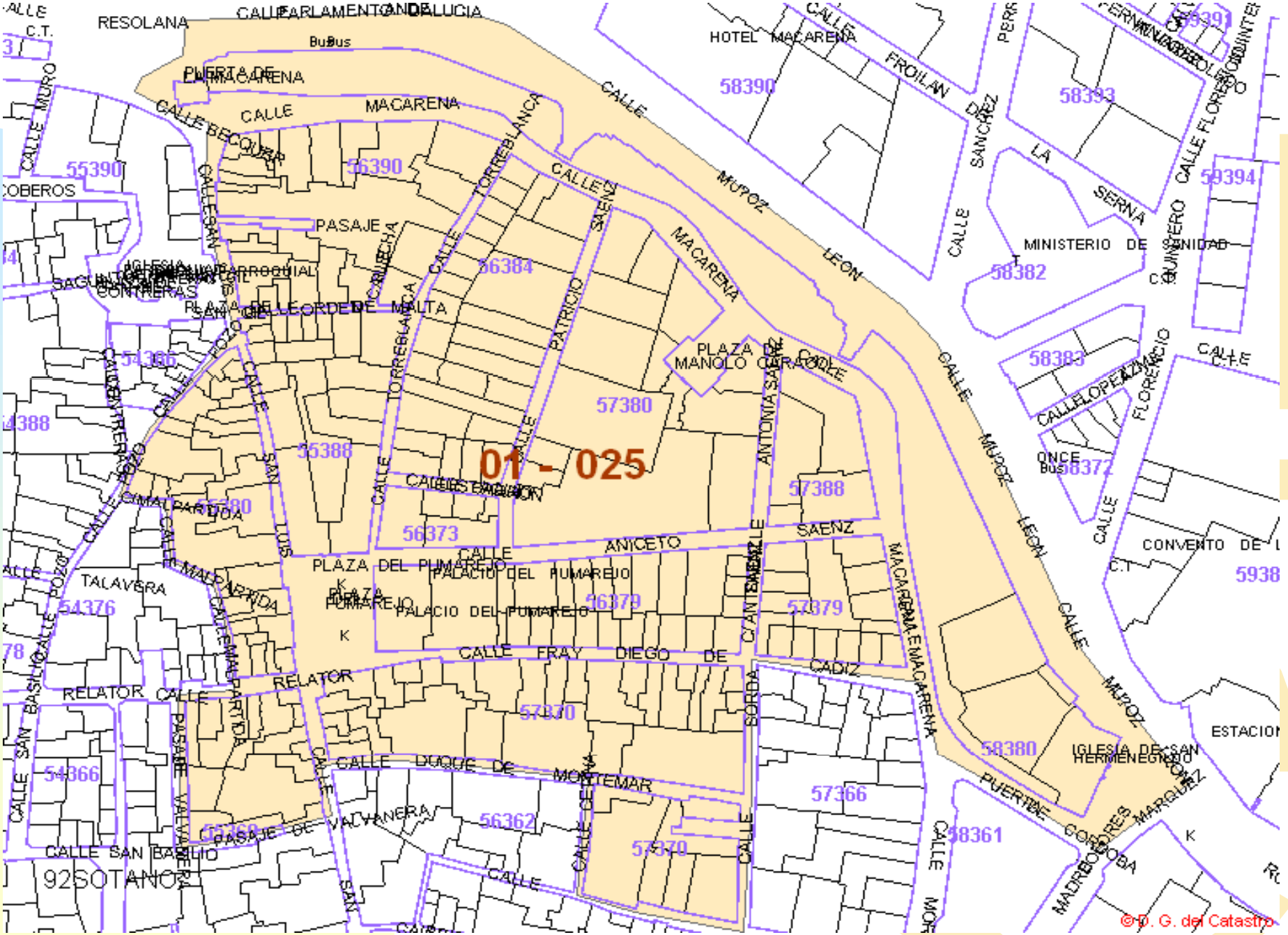
Código de encuesta 4102106 2

Provincia 41 SEVILLA

Municipio 091 SEVILLA

Sección 01 025

Num De Manz	Dirección postal TVIA NVIA	TP	NUM	BQ	PR	ES	PLN	PUER	Núm. de orden	Cod de Hueco	Viv ien das	Col ect ivos	Loc ales	Ocupante Principal	Núm de Inc	Núm de Sel	Encuestas Anteriores	
1	CALLE ANICETO GARNS	N	9	2			PBJ		0006	00045		C		ARRIAZA BARRIO, REGLA	1	0		0
1	CALLE ANICETO GARNS	N	21			1	P01	F	0013	00100	H			GARCIA CUBILLO, FRANCISCA	2	1		0
1	CALLE ANICETO GARNS	N	21			1	P01	G	0013	00101	H			ABSALOM , KEITH	3	2		0
1	CALLE ANICETO GARNS	N	21			1	P02	G	0013	00102	S			,	4	0		0
1	CALLE ANICETO GARNS	N	21			1	P02	IZ	0013	00103	H			RIAO VILLALOBOS, RAFAEL	5	3		0
1	CALLE ANICETO GARNS	N	21			1	P02	JI	0013	00104	H			NIETO MARTINEZ, FRANCISCO	6	4		0
1	CALLE ANICETO GARNS	N	21			1	PBJ	0001	0013	00099			L	,	7	0		0
1	CALLE ANICETO GARNS	N	21			1	PBJ	A1	0013	00097	H			GALLARDO GODOY, EMILIO JOSE	8	5		0
1	CALLE ANICETO GARNS	N	21			1	PBJ	B1	0013	00098	V			,	9	0		0
1	CALLE ANICETO GARNS	N	21			2	P01	DR	0013	00109	H			HIDALGO BOTELLO, FRANCISCO	10	6		0
1	CALLE ANICETO GARNS	N	21			2	P02	D	0013	00110	H			LOPEZ ARNESTO, ISABEL	11	7		0
1	CALLE ANICETO GARNS	N	21			2	P02	E	0013	00111	H			RODRIGUEZ VIVERO, DAVID	12	8		0
1	CALLE ANICETO GARNS	N	21			2	P02	IZ	0013	00112	H			BARRERA MARQUEZ, MANUEL	13	9		0
1	CALLE ANICETO GARNS	N	21			2	PBJ	A	0013	00106	H			GARCIA PAREDES FRAILE, PATRICIA	14	10		0
1	CALLE ANICETO GARNS	N	21			2	PBJ	B2	0013	00105	H			ZAMORA ANGULO, JUAN ANTONIO	15	11		0
1	CALLE ANICETO GARNS	N	21			2	PBJ	CN	0013	00107	H			GUERRA FALCON, CARLOS	16	12		0
1	CALLE ANICETO GARNS	N	21			2	PBJ	DR	0013	00108	H			CASO ARMERO, ANA	17	13		0
1	CALLE MACARENA	N	30				P01	E	0089	01623	H			GOMEZ PIAMBA, ROSALBA PATRICIA	18	818		0
1	CALLE MACARENA	N	30				P01	F	0089	00539	H			PEREZ GOMEZ, JUAN CARLOS	19	819		0
1	CALLE MACARENA	N	30				P01	G	0089	00540	H			VIDAL LEMUS, ELVIRA	20	14		0
1	CALLE MACARENA	N	30				P02	E	0089	00541	H			JIMENEZ PEREZ, ESPERANZA	21	15		0
1	CALLE MACARENA	N	30				P02	F	0089	00542	H			RAGEL BONILLA, MARIA ROSARIO	22	16		0
1	CALLE MACARENA	N	30				P02	G	0089	00543	H			JALON RODRIGUEZ, MANUEL	23	17		0
1	CALLE MACARENA	N	30				PBJ	1	0089	01622	H			,	24	820		0
1	CALLE MACARENA	N	32				P01	1	0090	00544	H			ARREDONDO PEREZ, FRANCISCO	25	18		0


01 - 025
RESOLANA
PUERTA DE LA MACARENA
CALLE MACARENA
CALLE PARLAMENTO DE ANDALUCIA
CALLE MUÑOZ LEON
HOTEL MACARENA
CALLE FROILAN
CALLE SANCHEZ DIEZ
CALLE LA SERNA
MINISTERIO DE SANIDAD
CALLE QUINTERO
CALLE FLORENCIO
CALLE LOPEZ AZCONA
ONCE
CONVENTO DE
ESTACION
IGLESIA DE SAN HERMENEGILDO
CALLE CORDOBA
CALLE MADRE DOLORES MARQUEZ
CALLE PUERTA
CALLE BECQUER
CALLE MURO
CALLE SAN BASILIO
CALLE TORREBLANCA
CALLE PATRICIO SAENZ
CALLE ANTONIA SAEZ
CALLE SAN LUIS
CALLE FUENCARRAL
CALLE POZO
CALLE CONTRERAS
CALLE MALPARTIDA
CALLE TALAVERA
CALLE RELATOR
CALLE FRAY DIEGO DE CADIZ
CALLE DUQUE DE MONTEMAR
CALLE PASAJE DE VALVANERA
CALLE SAN BASILIO
PASAJE
PLAZA DEL PUMAREJO
PALACIO DEL PUMAREJO
PLAZA MANOLO GARRIDO
IGLESIA PARROQUIAL SAGRADO CORAZON DE JESUS
PLAZA DE LA ORDEN DE MALTA
CALLE ANICETO SAENZ
CALLE CANTABRANA
CALLE BORDA
55390
56390
56384
56388
56373
56379
56362
54386
54388
54380
54376
54366
57380
57388
57379
57370
57366
57369
58390
58393
59394
58382
58383
58372
58380
58361
59391
59380
92SOTANOS
© D. G. del Catastro

3.TIPO DE MUESTREO. Muestreo bietápico con estratificación de unidades de primera etapa

- **Unidades de primera etapa: Secciones censales.**

La muestra de secciones permanece fija indefinidamente salvo:

- Resultados censales que aconsejan otra afijación
- Agotamiento de los hogares consultables
- Actualización de probabilidades de selección

- **Unidades de segunda etapa: Viviendas familiares principales y alojamientos fijos**

Dentro de las unidades de segunda etapa no se realiza submuestreo alguno.

4. CRITERIOS DE ESTRATIFICACIÓN

- **Geográfico(Estratos)**: Según la importancia demográfica del municipio al que pertenecen las unidades primarias

Municipios *Autorrepresentados*: Estratos 1-2-3

Municipios *Correpresentados* :

Estratos	Población		
4	50.000	-	100.000
5	20.000	-	50.000
6	10.000	-	10.000
7	5.000	-	10.000
8	2.000	-	5.000
9			< 2.000

- **Socioeconómico (Subestratos)**: Dentro de cada estrato las secciones se clasifican según la categoría socioeconómica de la población activa de la sección

4.1 Criterios de Subestratificación

- Las secciones censales se han agrupado, dentro de cada estrato, en subestratos.

- Se han tenido en cuenta aquellas características que se consideran más correlacionadas con las variables de interés de la encuesta.

- La información sobre las variables de subestratificación al nivel de sección censal procede del Censo 2001 y de la Agencia Tributaria

4.1 Criterios de Subestratificación(2).

- En el proceso de subestratificación se han considerado dos grupos de secciones:

  1, Las de los estratos 7, 8 y 9 a las que se les asigna como subestrato la comarca (NUTS4) del municipio al que pertenecen

  2, Las del resto de estratos a las que se agrupan, dentro de sus estratos, aplicando técnicas de análisis de conglomerados (cluster).

Variables de subestratificación(1)

En la subestratificación de las secciones de los estratos 1 a 6, se han utilizado las siguientes variables:

- **Porcentaje de parados en la sección.**
- **Porcentaje de inactivos.**
- **Porcentaje de ocupados.**
- **Porcentaje de extranjeros.**
- **Porcentaje de personas entre 0 y 14 años.**

Variables de subestratificación(2)

- Porcentaje de personas entre 15 y 24 años.
- Porcentaje de personas de 65 o más años.
- Porcentaje de personas con nivel de estudios realizado de analfabetos, sin estudios o nivel de estudios de primer grado.
- Porcentaje de personas con nivel de estudios realizado de ESO, EGB, Bachillerato, FP.
- Porcentaje de personas con nivel de estudios realizado de diplomatura, licenciatura o doctorado.

Variables de subestratificación (3)

Se ha considerado también como variable las 18 modalidades de la variable **condición socioeconómica.**

Las **variables fiscales** que se han utilizado son:

Renta total por vivienda con perceptores.

Renta Capital mobiliario e inmobiliario sobre renta total.

Renta agraria sobre renta total.

- El **algoritmo** usado para obtener los subestratos (conglomerados) ha sido el de Ward (JASA 1963), Este es un algoritmo multivariante de análisis de conglomerados jerarquizado basado en la minimización de las distancias dentro de los conglomerados, Este método está disponible en el procedimiento CLUSTER, del módulo SAS/STAT de SAS.

CPRO	CMUN	DIST	NSECC	Población	% de jóvenes (0-19)	% de jóvenes (15-24)	% de Mayores	% de parados en la sección	% de inactivos	% de ocupados	% de extranjeros
41	091	01	022	1.146,0	9,34	21,29	20,24	10,38	53,66	35,95	3,14
41	091	01	023	1.487,0	9,75	21,52	16,75	11,97	49,83	37,26	2,69
41	091	01	024	1.261,0	10,55	17,76	20,38	10,47	54,48	34,66	2,38
41	091	01	025	2.036,0	11,25	19,40	17,58	9,48	49,85	37,28	2,65
41	091	01	027	1.391,0	9,99	22,00	21,21	5,97	54,57	39,47	1,01
41	091	01	028	773,0	12,55	20,83	17,21	11,25	52,65	34,67	2,85
41	091	01	029	1.915,0	9,92	23,86	13,68	11,96	47,42	35,67	1,04
41	091	01	030	762,0	8,27	23,23	22,18	6,96	53,67	37,53	0,79
41	091	01	031	758,0	8,84	17,81	26,65	10,16	56,20	33,64	1,72

				% de personas con nivel de estudios							
CPRO	CMUN	DIST	Nº SECC	inferiores	medios	superiores	Renta total por vivienda con percentores	Renta por desempleo entre renta total	Renta Capital mobiliario e inmobiliario sobre renta total	Renta agraria sobre renta total	Subestrato
41	091	01	022	39,70	37,87	22,43	19160,6	2,0	4,6	0,1	4
41	091	01	023	36,85	42,57	19,64	17464,7	2,2	3,1	0,0	4
41	091	01	024	44,96	33,23	21,41	19662,2	1,6	5,2	0,3	4
41	091	01	025	43,71	34,33	18,57	18711,8	1,8	3,7	0,3	4
41	091	01	027	26,82	37,60	35,59	44987,0	0,5	23,4	1,2	6
41	091	01	028	46,18	33,12	19,28	19579,7	1,5	4,6	0,4	4
41	091	01	029	37,08	39,95	18,02	19480,2	1,7	4,6	0,3	4
41	091	01	030	29,27	43,96	24,93	33633,7	1,2	6,6	0,0	4
41	091	01	031	41,29	39,58	19,13	17857,5	2,7	4,1	0,1	4

5. TAMAÑO DE LA MUESTRA

Se establece en función de:

1- La desagregación requerida para las estimaciones

2- La dispersión de la(s) variables(s) objetivo

3- Límites establecidos por el Servicio Promotor

De acuerdo con lo anterior se determina:

- **El número de secciones muestrales por estrato**
- **Un número fijo de viviendas por sección**

Resulta muy útil la experiencia de otras encuestas similares

5.1 TAMAÑO DE LA MUESTRA en la EPA

Tamaño: En función del *coste*(Q) y del *coeficiente de variación*(C):

$$Q = nQ_s + nmQ_v$$

δ = *coeficiente de correlación intraclásica*, Para la población activa se estimó δ=0,05,

$$C^2(\hat{P}) = \frac{V(\hat{P})}{\hat{P}^2} = \frac{1-\hat{P}}{\hat{P}^2} \cdot \frac{1+\delta(m-1)}{nm}$$

El mínimo para un coste dado se obtuvo para:

n = 3.000 secciones.

m = 20 viviendas por sección.

5.2. AFIJACIÓN.

Objetivos : Estimaciones **provinciales** fiables.

Estimaciones **nacionales** fiables.

Número exacto de bloques en cada provincia.

(13 secciones por trimestre)

Entre **provincias**: De compromiso entre uniforme y proporcional

Entre **estratos**: Estrictamente proporcional

6. SELECCIÓN DE LA MUESTRA

Secciones: Probabilidad proporcional al tamaño medido por el número de viviendas familiares principales

Viviendas: Probabilidad igual(muestreo sistemático)

De esta forma en cada estrato, las viviendas familiares tienen la misma probabilidad de pertenecer a la muestra (**muestra autoponderada**)

$$P\left(V_{ijh}\right) = P\left(S_{jh}\right).P\left(\frac{V_{ijh}}{S_{jh}}\right) = K_h \cdot \frac{V_{jh}}{V_h} \cdot \frac{m}{V_{jh}} = \frac{K_h \cdot m}{V_h}$$

7. DISTRIBUCIÓN DE LA MUESTRA EN EL TIEMPO

La muestra se distribuye de forma ***uniforme*** a lo largo del ámbito temporal en el que se desarrolla.

Para ello las variables que, normalmente, se toman en consideración son:

- Semana
- Provincia o Comunidad autónoma
- Estrato
- Turno de rotación

7.1 DISTRIBUCIÓN DE LA MUESTRA EN la EPA

- Cada período de la encuesta es de un **trimestre** siendo cada una de las secciones de la muestra visitada en una de las 13 semanas del mismo.
- La distribución de la muestra es **uniforme** en el tiempo, Para ello se han considerado las variables provincia, estrato, turno de rotación y semana.
- La totalidad de la muestra está dividida en **tres submuestras independientes** representativas, cada una de ellas, de toda la población.
- Las submuestras correspondientes a cada turno de rotación son representativas, aunque su reducido tamaño impide las estimaciones en dominios medios o pequeños.

EPA- 2T/06. Muestra de secciones de Málaga

		SEMANA												
		01	02	03	04	05	06	07	08	09	10	11	12	13
CPRO	ESTRATO													
29	1	3	3	3	2	3	3	2	2	3	3	3	3	3
	4	1	1	1	.	1	1	.	1	.	1	1	1	1
	5	2	1	1	2	1	1	2	2	2	1	1	1	1
	6	.	.	.	1	1	.	1	.	1	.	.	.	1
	7	.	1	1	1	.	1	1	1	.	1	1	1	.
	All	6	6	6	6	6	6	6	6	6	6	6	6	6

EPA- 2T/06. Muestra de secciones de Málaga

		SEMANA												
		01	02	03	04	05	06	07	08	09	10	11	12	13
CPRO	TR													
29	1	1	1	1	1	.	2	1	1	1	1	1	1	1
	2	1	.	1	2	1	1	1	1	.	2	1	1	1
	3	1	1	1	.	2	1	1	1	1	1	1	1	1
	4	1	2	1	1	1	.	.	1	2	1	.	2	1
	5	1	1	1	1	1	.	2	1	1	.	2	1	1
	6	1	1	1	1	1	2	1	1	1	1	1	.	1
	All	6	6	6	6	6	6	6	6	6	6	6	6	6

8. RENOVACIÓN PARCIAL DE LA MUESTRA

Unidades primarias: Las secciones censales permanecen fijas indefinidamente en la muestra (salvo las excepciones señaladas).

Unidades secundarias: Las viviendas familiares de la muestra son renovadas parcialmente cada trimestre, Esta renovación afecta a una sexta parte de las secciones (5/6 permanecen de un trimestre a otro).

Turnos de rotación: El conjunto de las secciones de la muestra está repartido en 6 grupos llamados turnos de rotación.

Cada trimestre, se renueva la muestra de viviendas correspondientes a las secciones de un determinado turno de rotación.

9. ESTIMADORES

El proceso habitual para la obtención de estimadores en encuestas demográficas en general, y en la EPA en particular es:

1. ***Estimador insesgado de expansión*** *(Horvitz-Thompson)*: Compensa las desiguales probabilidades de selección.
2. ***Corrección de la falta de respuesta***: Corrige el sesgo producido en las estimaciones por la falta de respuesta total de algunos elementos.
3. ***Calibrado con fuentes externas***: Reduce la varianza de las estimaciones mediante la utilización de fuentes auxiliares externas y puede actualizar la estimación en el tiempo.

Como resultado de este proceso se obtiene finalmente un factor de elevación para cada elemento de la muestra efectiva.

9.1. Estimador insesgado de expansión(H-T)

Recordamos que la probabilidad de *pertenecer a la muestra* de una vivienda ‘i’ de la sección ‘j’ del estrato ‘h’ viene dada por:

$$P\left(V_{ijh}\right) = P\left(Sec_{jh}\right).P(V_{ijh}/Sec_{jh}) = K_h \cdot \frac{V_{jh}}{V_h} \cdot \frac{m}{V_{jh}} = \frac{K_h \cdot m}{V_h}$$

Donde K_h son las secciones de la muestra en el estrato “h”, y “m” es el número de viviendas muestrales por sección.

Según lo anterior, la probabilidad de pertenecer a la muestra se puede expresar por:

$$P\left(V_{ijh}\right)=\frac{v_h^t}{V_h}$$

Siendo v^t_h el número teórico de viviendas de la muestra en el estrato “h”.

Por tanto el estimador H-T tendrá la expresión:

$$\hat{Y}_{H-T}=\sum_h \frac{V_h}{v_h^t}\cdot\sum_{i\in h} y_i$$

9.2. Corrección de la falta de respuesta

La probabilidad de respuesta por estrato la podemos estimar por:

$$P_{Rh} = \frac{v_h}{v_h^t}$$

Donde v_h representa la muestra efectiva de viviendas en el estrato h.

Por tanto el estimador corregido será:

$$\hat{Y}_{H-TCorr} = \sum_h \frac{V_h}{v_h^t} \cdot \frac{v_h^t}{v_h} \sum_{i \in h} y_i = \sum_h \frac{V_h}{v_h} \sum_{i \in h} y_i = \sum_h \hat{Y}_{H-TCorr(h)}$$

9.3. Calibrado con fuentes externas.

Se utiliza un **estimador de razón** y se aplican **Técnicas de reponderación** con objeto de ajustar las estimaciones de la encuesta a la información procedente de fuentes externas.

Estimador de razón:

$$\hat{Y} = \sum_h \frac{\hat{Y}_h}{\hat{P}_h} \cdot P_h$$

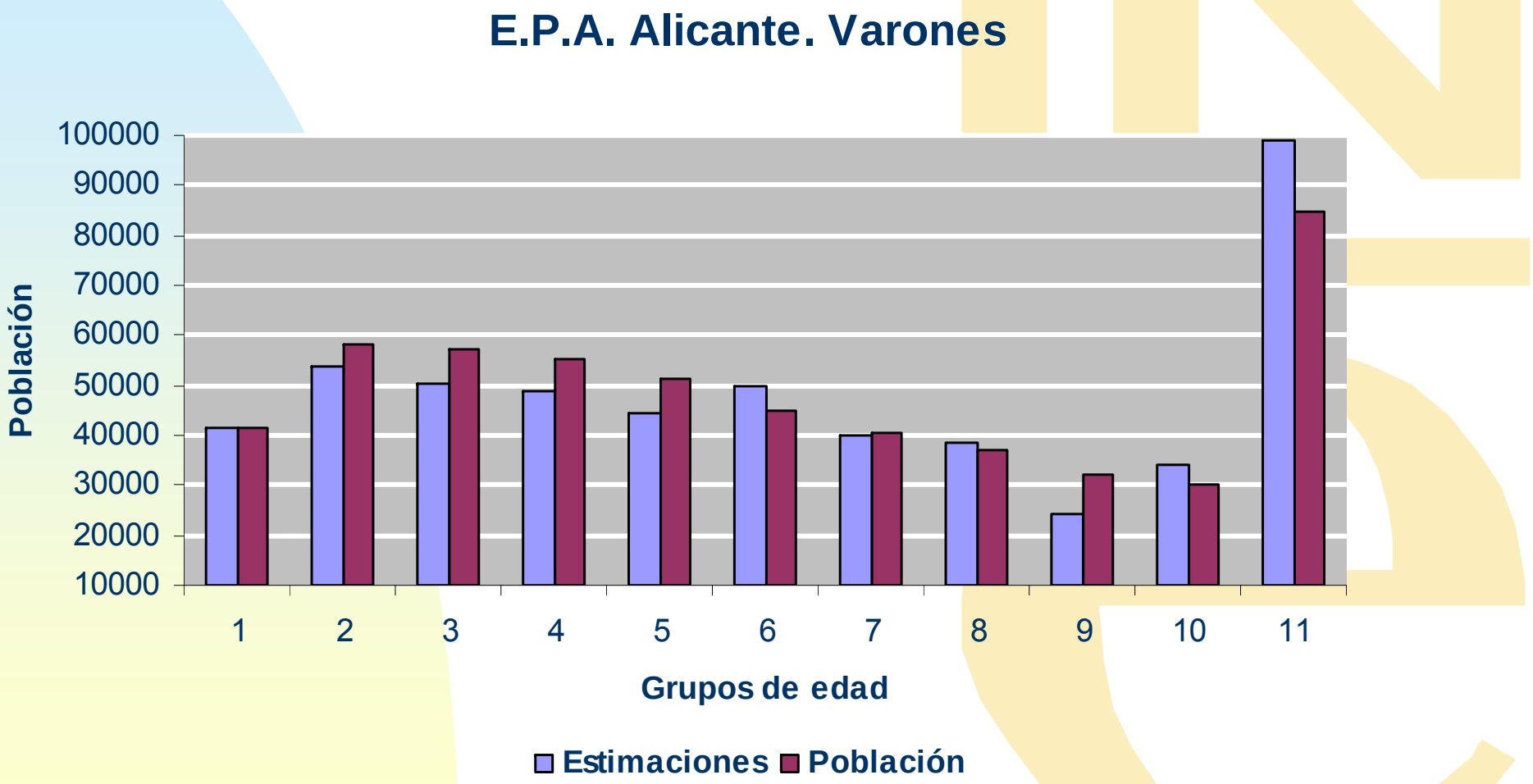
Siendo:

$$\hat{Y}_h = \sum_{i \in h} \frac{1}{\frac{v_h}{V_h}} \cdot y_{hi} \qquad \hat{P}_h = \sum_{i \in h} \frac{1}{\frac{v_h}{V_h}} \cdot p_{hi}$$

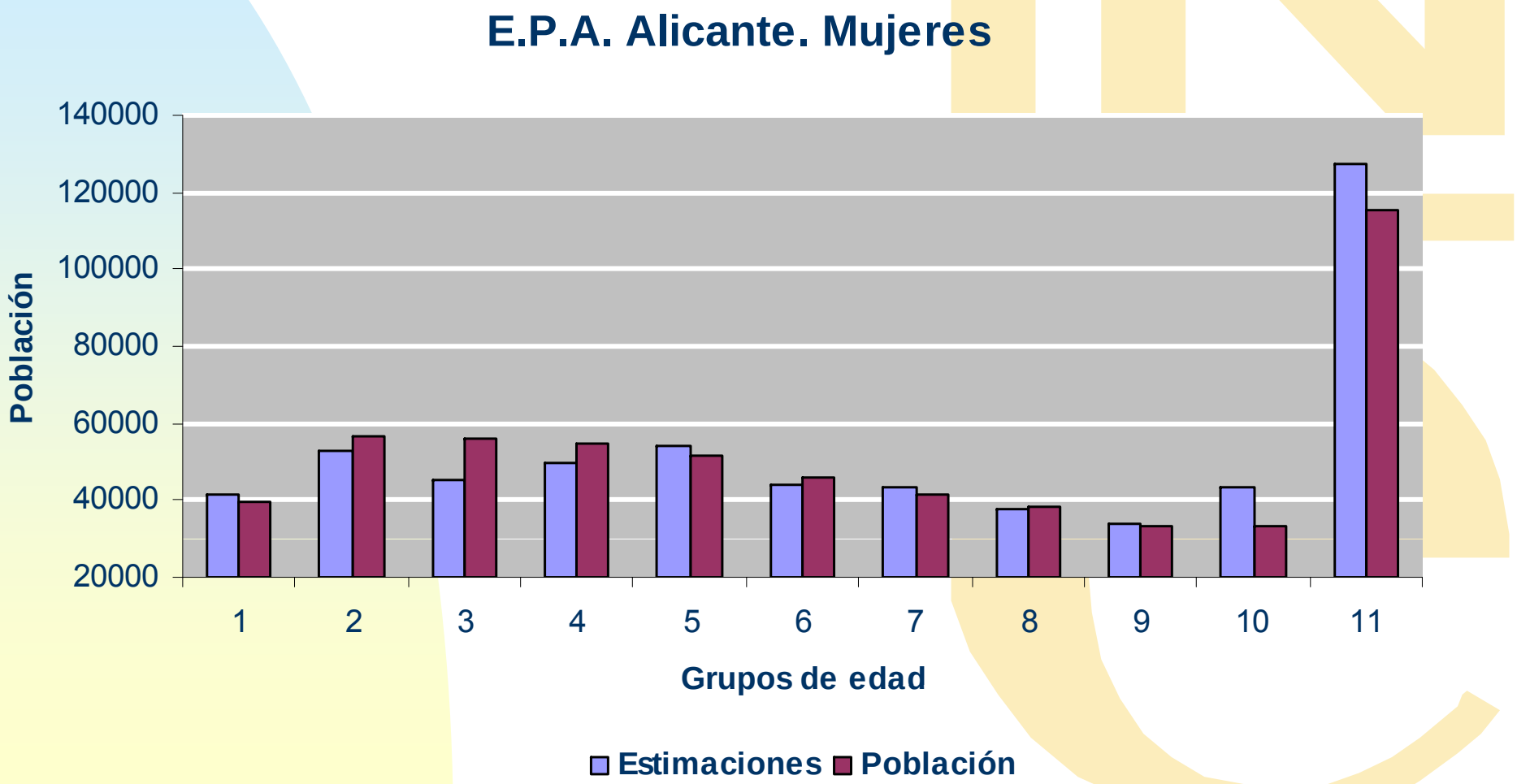
Por tanto:

$$\hat{Y} = \sum_h \frac{\sum_{i \in h} \frac{1}{\frac{v_h}{V_h}} \cdot y_{hi}}{\sum_{i \in h} \frac{1}{\frac{v_h}{V_h}} \cdot p_{hi}} \cdot P_h = \sum_h \frac{P_h}{p_h} \sum_{i \in h} y_{hi}$$

Los datos muestrales elevados no se ajustan a los datos de la población



Los datos muestrales elevados no se ajustan a los datos de la población



9.4. Aplicación de Técnicas de calibrado(1)

La expresión del estimador de razón es:

$$\hat{Y} = \sum_h \frac{P_h}{p_h} \sum_{i \in h} y_{hi}$$

Esta expresión puede escribirse como:

$$\hat{Y} = \sum_{k \in s} d_k \, y_k$$

Se dispone de J variables auxiliares cuyos valores son conocidos para la muestra y cuyos totales son conocidos para la población

$$X_j = \sum_{k \in U} x_{jk}$$

Generalmente la muestra no es equilibrada:

$$X_j \neq \hat{X}_j = \sum_{k \in s} d_k \, x_{jk}$$

9.4. Aplicación de Técnicas de calibrado(2)

Objetivo de la reponderación: Obtener unos nuevos pesos w_k ,lo mas parecido posible a los pesos d_k , que equilibren la muestra, es decir:

$$\hat{X} = \sum_{k \in s} w_k \, x_k$$

Solución matemática del problema : Encontrar unos valores que hagan mínima la expresión:

$$\sum_{k \in s} d_k \, G\left(\frac{w_k}{d_k}\right)$$

con la condición

$$\sum_{k \in s} w_k \, X_k = X$$

siendo:

G = Función de distancia.

X= Vector de dimensión (J,1) con los totales de las variables auxiliares.

X_k= Vector de dimensión (J,1) con los valores de las variables auxiliares en la unidad muestral k.

La solución del problema depende de la función de distancia G que se utilice.

9.4. Aplicación de Técnicas de calibrado(3)

En la EPA se ha optado por utilizar la **función de distancia lineal** con objeto de aprovechar las propiedades del estimador de regresión, de pequeña varianza y mínimo sesgo

Además se ha utilizado un algoritmo que permite acotar las variaciones de los factores finales respecto de los iniciales con el fin de evitar factores finales negativos.

Para la resolución práctica de este problema se ha utilizado el software CALMAR (**CAL**age sur **MAR**ges) programado por el INSEE (Institut National de la Statistique et des Études Économiques) de Francia

9.4. Aplicación de Técnicas de calibrado(4)

Las variables auxiliares que se han empleado son:

1- Población de 16 y más años por grupos de edad y sexo a nivel de Comunidad Autónoma

2- Población de 16 y más años por provincia

3- País de nacionalidad

10. ACTUALIZACIÓN

En la E.P.A. hay que considerar tres tipos de actualizaciones:

Actualización en el marco de unidades primarias (secciones censales) es la que se produce en los periodos intercensales como consecuencia de modificaciones en las unidades primarias seleccionadas para la muestra.

Actualización en el marco de viviendas, restringida a las secciones de la muestra.

Actualización con carácter general, relativa a todas las secciones y viviendas, que se realiza cada tres años y se actualizan las probabilidades de selección.

11. ERRORES DE MUESTREO

Se aplica el método de las semimuestras reiteradas

Consiste en:

- Obtención de sucesivas semimuestras de la muestra total.
- Estimación de la característica con cada semimuestra

El estimador de la varianza es:

$$\hat{v}\left(\hat{x}\right) = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{x}_i - \hat{x}\right)^2$$

donde:

r es el número de semimuestras

$\hat{x}_i$ es la estimación con la i-ésima reiteración.

$\hat{x}$ es la estimación obtenida con la muestra completa.

En la E.P.A. el número de reiteraciones es de 40.

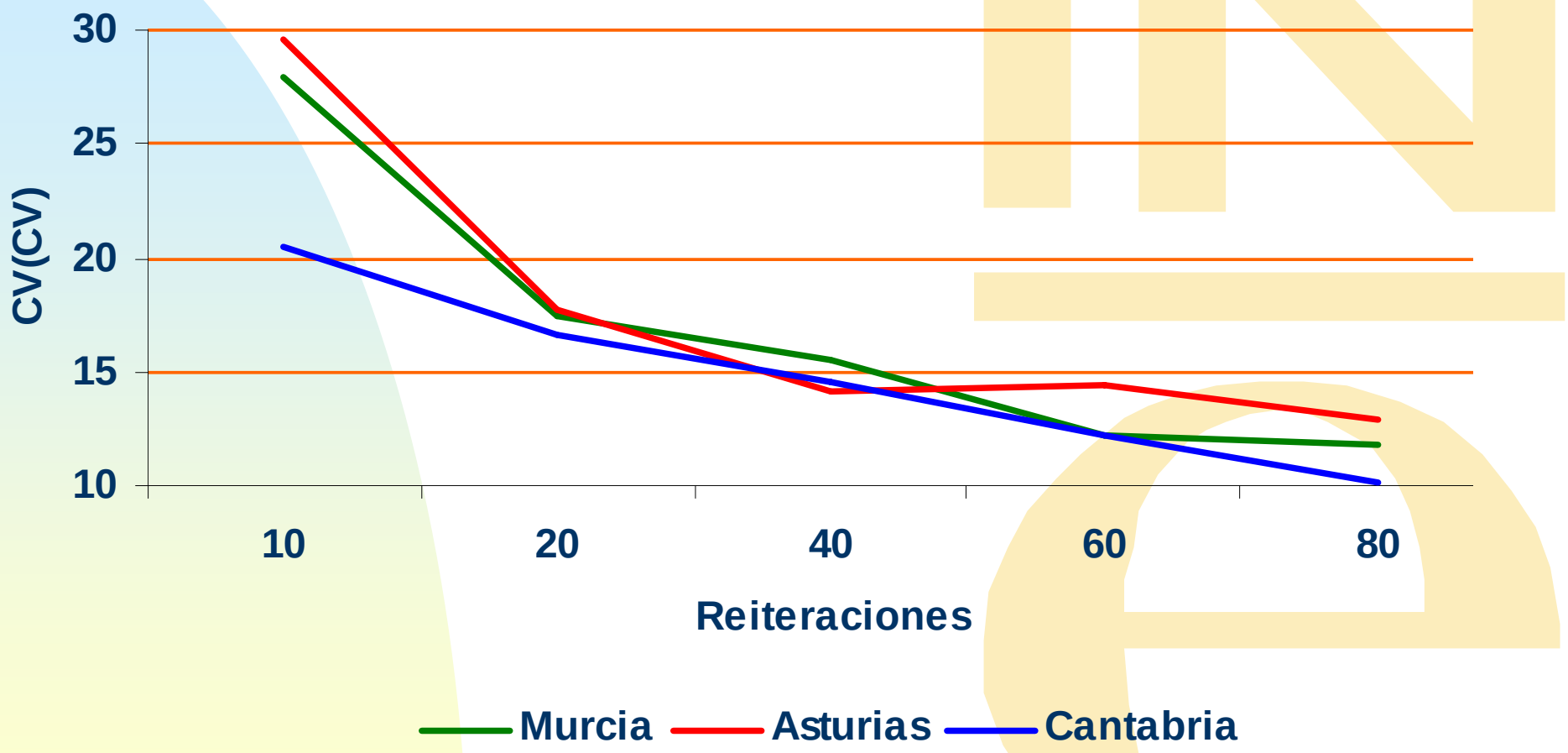
Formación de las reiteraciones:

- Se agrupan las secciones de cada estrato por pares.
- Se asigna aleatoriamente la primera sección a 20 reiteraciones y la otra sección a las otras 20.

De esta forma:

- Cada reiteración queda constituida por un número de secciones equivalente al 50% de la muestra.
- Cada sección aparece en la mitad de las reiteraciones

C.V. DE LA VARIABLE CV (PARO)



11.1. UTILIZACIÓN DEL ERROR DE MUESTREO

La teoría del muestreo determina que:

$$\text{Prob}\left\{\hat{X} - 1{,}96\sigma(\hat{X}) < \hat{X} < \hat{X} + 1{,}96\sigma(\hat{X})\right\} = 0{,}95$$

es decir, en el intervalo comprendido entre la estimación menos 1,96 veces el error de muestreo y la estimación más 1,96 veces el error de muestreo existe una confianza del 95 por ciento de que se encuentre el valor verdadero o parámetro que se pretende estimar.

A este intervalo se le denomina **Intervalo de confianza del 95%**

- **INTERPRETACIÓN**: En promedio, de cada 100 muestras obtenidas bajo el mismo diseño y condiciones generales, los intervalos de confianza obtenidos a partir de cada una de ellas contendrían el valor verdadero en 95 casos de los 100.

- **Ejemplo**: la estimación del total de ocupados en un determinado trimestre es 14.706.600 con un error de muestreo relativo del 0,44 por ciento. Esto significa que existe una gran confianza, en términos de probabilidad una confianza del 95 por ciento, de que el valor verdadero del total de ocupados se encuentre en el intervalo comprendido entre 14.579.770 y 14.833.430 (esto es, 14.706.600x1,96 x 64.709).

EPA. Errores de muestreo relativos,en porcentaje, de la población de 16 y más años según su relación con la actividad económica, por comunidades autónomas. Tercer trimestre 2008

Comunidades autónomas	Activos	Ocupados	Parados			Inactivos
			Total	Buscan primer empleo	Han trabajado antes	
Total	0,24	0,31	1,45	4,81	1,58	0,37
Andalucía	0,60	0,94	3,08	7,47	3,30	0,81
Aragón	1,15	1,43	10,65	20,42	12,36	1,68
Asturias (Principado de)	1,45	1,60	10,60	22,59	12,46	1,67
Balears (Illes)	0,87	1,38	9,59	23,36	9,60	1,76
Canarias	0,93	1,28	3,90	17,63	4,00	1,45
Cantabria	1,33	1,41	10,65	30,44	10,90	1,78
Castilla y León	0,67	0,77	3,43	17,89	4,01	0,82
Castilla - La Mancha	0,78	1,03	6,45	20,74	6,97	1,04
Cataluña	0,60	0,76	4,78	11,29	5,36	1,04
Comunitat Valenciana	0,83	0,97	5,41	14,49	5,76	1,28
Extremadura	0,95	1,31	5,92	13,52	5,78	1,11
Galicia	0,74	0,98	5,78	15,43	6,41	0,93
Madrid (Comunidad de)	0,62	0,84	6,10	20,10	6,66	1,14
Murcia (Región de)	0,96	1,91	6,59	23,68	7,69	1,55
Navarra (Comunidad Foral de)	1,15	1,46	12,10	39,52	11,52	1,84
País Vasco	0,93	1,07	7,04	19,03	8,08	1,32
Rioja (La)	1,24	1,52	9,38	25,67	11,09	1,85
Ceuta	6,06	8,10	17,08	35,76	18,89	7,41
Melilla	6,25	5,36	21,71	41,19	16,69	7,72

ENCUESTA INDUSTRIAL ANUAL DE EMPRESAS

Diseño de la muestra

Esquema

1. OBJETIVOS
2. ÁMBITO
3. MARCO DE LA ENCUESTA
4. TIPO DE MUESTREO
5. CRITERIOS DE ESTRATIFICACIÓN
6. TAMAÑO Y AFIJACIÓN DE LA MUESTRA
7. SELECCIÓN. COORDINACIÓN DE MUESTRAS
8. ESTIMADORES
9. ERRORES DE MUESTREO

1. OBJETIVOS

- La Encuesta Industrial de Empresas permite disponer de una información básica para el conocimiento de la realidad industrial y el análisis de las principales características estructurales.

- Implantada desde 1993 proporciona **anualmente** una visión general de la estructura industrial.

- Su metodología se atiene a las recomendaciones de Eurostat, especificadas en sus reglamentos y directivas.

2. ÁMBITO DE LA ENCUESTA

• **Poblacional**: Conjunto de empresas con una o más personas ocupadas remuneradas y cuya actividad principal está incluida en las secciones C a E de la CNAE-93

La encuesta cubre las industrias *extractivas, manufactureras* y la *producción y distribución de energía eléctrica,gas y agua.*

ACTIVIDAD PRINCIPAL de la empresa es aquélla que genera el **mayor valor añadido**. Si no se dispone de esta información, se considera aquélla que proporcione el **mayor valor de producción**, o en su defecto, la que emplee un **mayor número de personas ocupadas**

•**Geográfico**: **Todo el territorio nacional** excepto Ceuta y Melilla.

La encuesta está diseñada para obtener resultados a nivel de Comunidades Autónomas.

•**Temporal**: La encuesta se lleva a cabo con carácter **anual**.

Lo datos solicitados se refieren al **año natural** objeto de la encuesta.

Excepcionalmente, las empresas que funcionan por temporadas o campañas que comprenden dos años distintos, refieren la información a la temporada o campaña que terminó en dicho año.

3. MARCO :DIRCE

•Constituye el marco de referencia para la mayoría de las encuestas económicas.

•Es un marco de listas.

•Los variables estadísticas del marco son *código de actividad y el tamaño de la empresa*.También figura la cifra de negocios.

•La población objeto de estudio se ha dividido **a efectos del diseño** de la muestra en una serie de SECTORES INDUSTRIALES(128).

•Cada sector constituye una población independiente a efectos de muestreo.

• Los sectores se corresponden, en su mayoría, con el nivel de 3 dígitos (grupo) de la CNAE-93. En algunos casos se ha optado por una mayor desagregación a nivel de clase (4 dígitos), con el fin de recoger adecuadamente los datos de ciertas actividades industriales que a 3 dígitos resultan muy agregadas.

•Dentro de cada sector: Las empresas **con 20 o más** trabajadores son investigadas de forma **exhaustiva**, las de **menos de 20** trabajadores son investigadas **por muestreo**.

•Algunos sectores o estratos cuyo pequeño tamaño poblacional impide seleccionar una muestra representativa, han sido considerados exhaustivos.

4. TIPO DE MUESTREO

•Muestreo aleatorio estratificado.

5. VARIABLES DE ESTRATIFICACIÓN

En la formación de los estratos se utilizan las siguientes variables:

- Comunidad Autónoma
- Intervalo de tamaño (asalariados)
- Código de actividad(CNAE)

En consecuencia,cada estrato dentro de una población a muestrear (Sector de actividad) viene determinado por el cruce de las variables comunidad autónoma y tamaño.

A efectos del muestreo y del proceso posterior de estimación se han considerado los siguientes **Intervalos de Tamaño**:

TAMAÑO	PERSONAS OCUPADAS
1	1-3
2	4-9
3	10-19
4	20-49
5	50-99
6	100-199
7	200-499
8	500-999
9	1000 Y más

6. TAMAÑO DE LA MUESTRA

- **Tamaño final: Aprox. 45000 empresas.**
- Una parte de los estratos se investiga de forma exhaustiva, y el resto son muestreados.

  -**Estratos exhaustivos:** Tamaños 4, 5, 6, 7, 8, 9

  -**Estratos no exhaustivos:** Tamaños 1, 2, 3

- En los estratos no exhaustivos se utiliza una **Afijación de Neyman** prefijando los siguientes errores relativos de la variable personas ocupadas.

Errores prefijados: 1% sector

5% sector y comunidad

20% sector, comunidad y tamaño

- Cálculo del tamaño:

  Dada la expresión de la varianza de un total en un muestreo estratificado:

$$V\left(\hat{X}\right) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right).\frac{S_h^2}{n_h}$$

  donde,

$$S_h^2 = \frac{\sum_{i=1}^{N_h} \left(X_{hi} - \overline{X}_h\right)^2}{N_h - 1}$$

  La afijación óptima o de mínima varianza obtiene unos tamaños:

$$n_h = n\frac{N_h S_h}{\sum_h N_h S_h}$$

donde n viene dada por la precisión prefijada:

$$n = \frac{\left(\sum_h N_h S_h\right)^2}{V + \sum_h N_h S_h^2}$$

Con:

$$\begin{cases} V_1 = (0{,}01 X_1)^2 \\ V_2 = (0{,}05 X_2)^2 \\ V_3 = (0{,}20 X_3)^2 \end{cases}$$

X = nº de trabajadores

X_1 = nº de trabajadores en cada sector de actividad

X_2 = nº de trabajadores en cada sector de actividad y Comunidad Autónoma

X_3 = nº de trabajadores en cada sector de actividad , Comunidad Autónoma y tamaño

7. SELECCIÓN DE LA MUESTRA

•Dentro de cada estrato, la muestra se selecciona mediante la asignación de un numero aleatorio,que también permite la coordinación con otras encuestas.

• El proceso de selección es independiente de un año a otro, es decir, para un determinado estrato, la probabilidad de que una empresa sea seleccionada en el año t es independiente de que haya o no sido seleccionada en el año t-1.

8. ESTIMADORES.

• ESTIMADOR BASICO: Estimador insesgado de expansión en un muestreo estratificado

$$\hat{X} = \sum_h \frac{N_h}{n_h} \sum_i x_i$$

Siendo:

N_h: Número total de empresas en el directorio en el estrato h

n_h: Número de empresas seleccionadas para la muestra en el estrato h

$X_{i:}$ Valor de la variable observada X en la empresa i del estrato h

•**ESTIMADOR CORREGIDO**: Estimador obtenido por la corrección introducida en el factor de elevación debido a la existencia de diversos tipos de incidencias.

- **Tipos de incidencias**
  1. Bajas, cerrados, duplicados,…
  2. No respuesta (negativas, ilocalizables).
  3. Cambios de estrato.

- **Efecto sobre las estimaciones**

  Incidencia 1: Corrección del marco y disminución del tamaño muestral con el consiguiente incremento del error de muestreo.

  Incidencia 2: Disminución de la muestra efectiva y aparición de sesgos.

  Incidencia 3: Redistribución de las unidades del marco y aumento de la varianza.

- **Factor de elevación final**

  - Si hay cambio de estrato: $\frac{N_h}{n_h}$

  - Si no hay cambio de estrato: $\frac{\hat{N}_h^*}{n_h^*}$

siendo:

n_h^* : Número de empresas de la muestra efectiva que no ha cambiado de estrato

$\hat{N}_h^*$: Número de empresas en el directorio en el estrato h obtenido al deflactar en función de las bajas y cambios de estrato

$$\hat{N}_h^* = N_h \left(1 - \frac{b_h}{n_h}\right) - \sum_{h \neq k} \frac{N_h}{n_h} n_h^k$$

b_h : Número de empresas que son bajas en la muestra (incidencia de la muestra del grupo 1)

n_h^k : Número de empresas seleccionadas en el estrato h y que realmente pertenecen al estrato k

- **Expresión final del estimador**

$$\hat{X} = \sum_h \left\{ \sum_{i=1}^{n_h^*} \frac{\hat{N}_h^*}{n_h^*} x_i + \sum_{k \neq h} \frac{N_k}{n_k} \sum_{i=1}^{n_k^h} x_i \right\}$$

El primer sumando representa la aportación de las empresas que no han cambiado de estrato.

El segundo sumando representa la aportación de las empresas seleccionas en el estrato k y que realmente pertenecen al h.

9. ERRORES DE MUESTREO

Se expresan en términos relativos

$$CV\left(\hat{X}\right) = \sqrt{\frac{\hat{V}\left(\hat{X}\right)}{\hat{X}}}.100$$

donde

$$\hat{V}\left(\hat{X}\right) = \sum_{h} \hat{V}\left(\hat{X}_h\right)$$

El valor de $\hat{V}\left(\hat{X}_h\right)$ tiene tres componentes:

- $\hat{N}_h^* \left(\hat{N}_h^* - n_h^*\right) \dfrac{\sum_{i=1}^{n_h^*} \left(x_i - \overline{X}_h^*\right)^2}{n_h^* \left(n_h^* - 1\right)}$

  debida a la variación de la variable

- $\overline{X}_h^{*2} . \hat{N}_h^* \left(N_h - \hat{N}_h^*\right) \dfrac{N_n - n_h}{N_h \left(n_h - 1\right)}$

  debida a la variación de $\hat{N}_h^*$

- $\sum_{k \neq h} N_k \left(N_k - n_k \right) \frac{S_k^{h\,2}}{n_k}$ debida a los cambios de estrato

Siendo:

$$\overline{X}_h^* = \frac{\sum_{i=1}^{n_h^*} x_i}{n_h^*}$$

y

$$S_k^h = \frac{\sum_{i=1}^{n_k^h} x_i^2}{n_k - 1} - \frac{\left(\sum_{i=1}^{n_k^h} x_i \right)^2}{n_k \left(n_k - 1 \right)}$$

la cuasivarianza muestral de las empresas que pasan de un estrato k cualquiera, al estrato h.