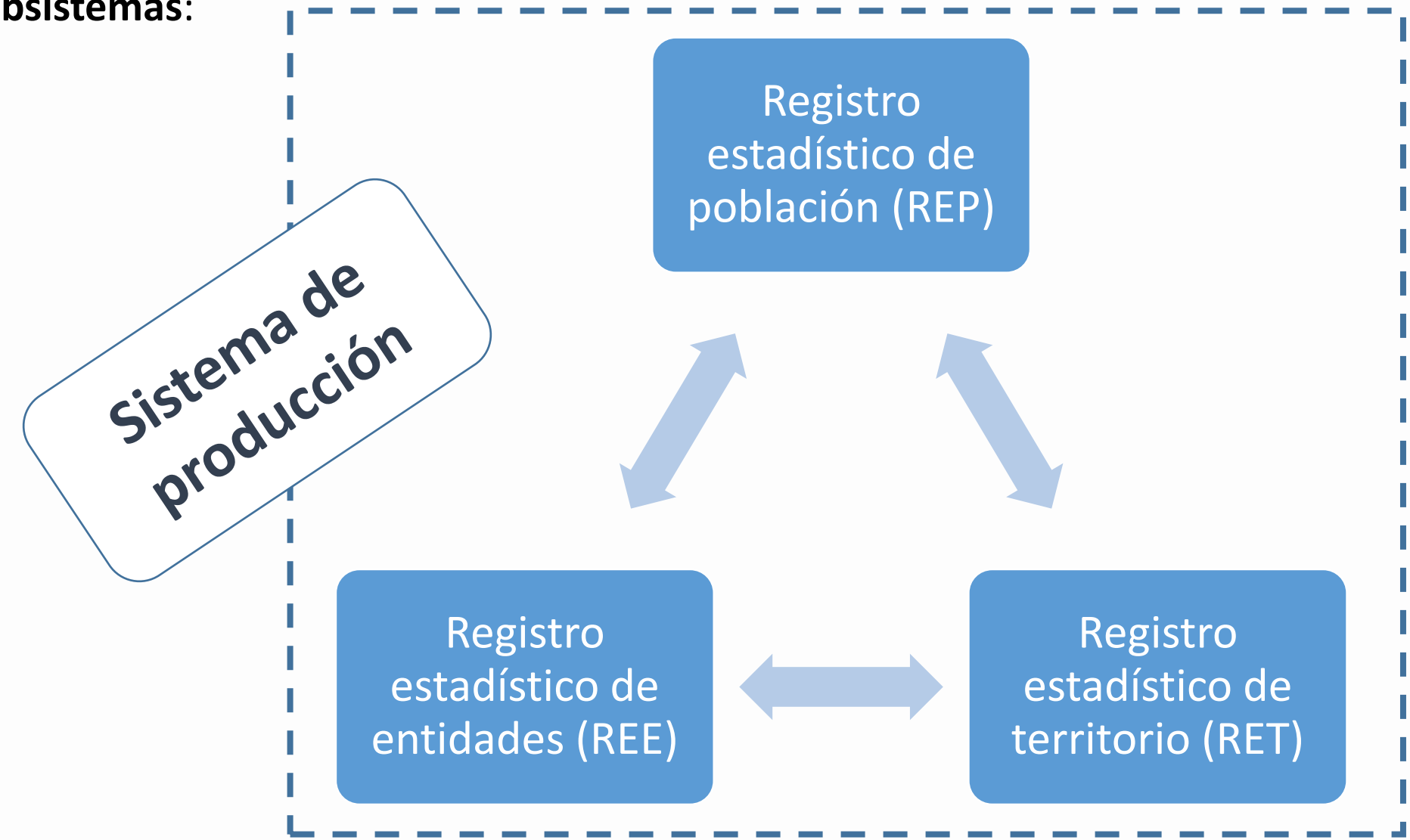
Generalitat de Catalunya
Institut d'Estadística de Catalunya

Análisis y consideraciones sobre la implementación del Registro estadístico de territorio.

D. Ibáñez Vidal dibanez@Idescat.cat
E. Suñé Luís esl@Idescat.cat

Sistema de producción

Construcción de un **sistema de producción** basado en **registros administrativos** formado por **3 subsistemas**:



Registro estadístico de territorio

Registro estadístico de territorio (RET)

Objetivos

- Geocodificar microdatos
- Validar direcciones postales

Fuentes

- Direcciones postales del REP y REE. Idescat, INE…
- Portales. Base de datos municipal de direcciones de Cataluña. Instituto Cartográfico y Geológico de Cataluña (ICGC)
- Bienes inmuebles. Dirección General del Catastro (DGC)

geocodificar

validar

REP,REE

TVIA	NVIA	RPOB	PRO	MUN	DIS	SEC	CVIA	TINUM	NUMER	CNUMER	NUMERS	CNUMERS	KMT	HMT	BLOQ	PORT	ESCA	PLAN	PUER
PG	GRAN PASSEIG RONDA	20141	25	120	04	015	02980		0001	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	P01	0001
PG	GRAN PASSEIG RONDA	20141	25	120	04	015	02980		0001	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	P01	0003
PG	GRAN PASSEIG RONDA	20141	25	120	04	015	02980		0001	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	P02	0001
PG	GRAN PASSEIG RONDA	20141	25	120	04	015	02980		0001	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	P02	0002

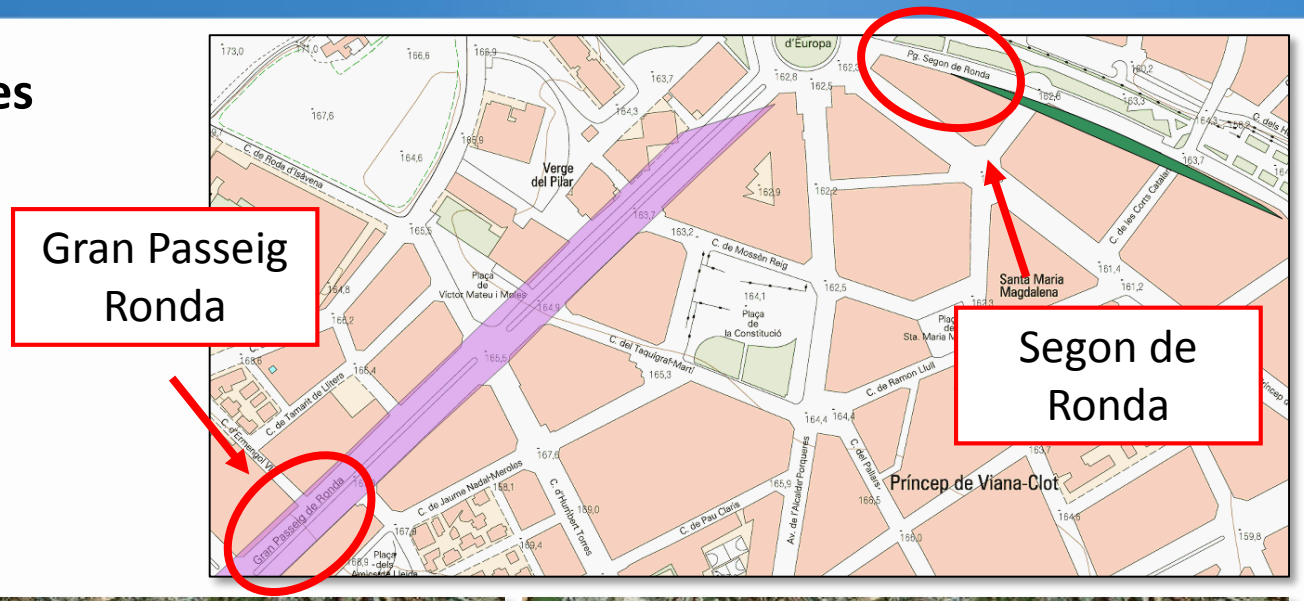
DGC

tipus_via text	nom_via text	pro_ine text	mun_ine text	codi_via_dgc text	num text	lletra text	num_segon text	lletra_segon text	km text	bloc text	escala text	planta text	porta text	parcela_catastral text	num_ordre_bi integer
PS	RONDA GRAN	25	120	00119	0001		0000		00000		1	01	01	1394909CG0019C	117
PS	RONDA GRAN	25	120	00119	0001		0000		00000		1	01	02	1394909CG0019C	118
PS	RONDA GRAN	25	120	00119	0001		0000		00000		1	01	03	1394909CG0019C	119
PS	RONDA GRAN	25	120	00119	0001		0000		00000		1	02	01	1394909CG0019C	120

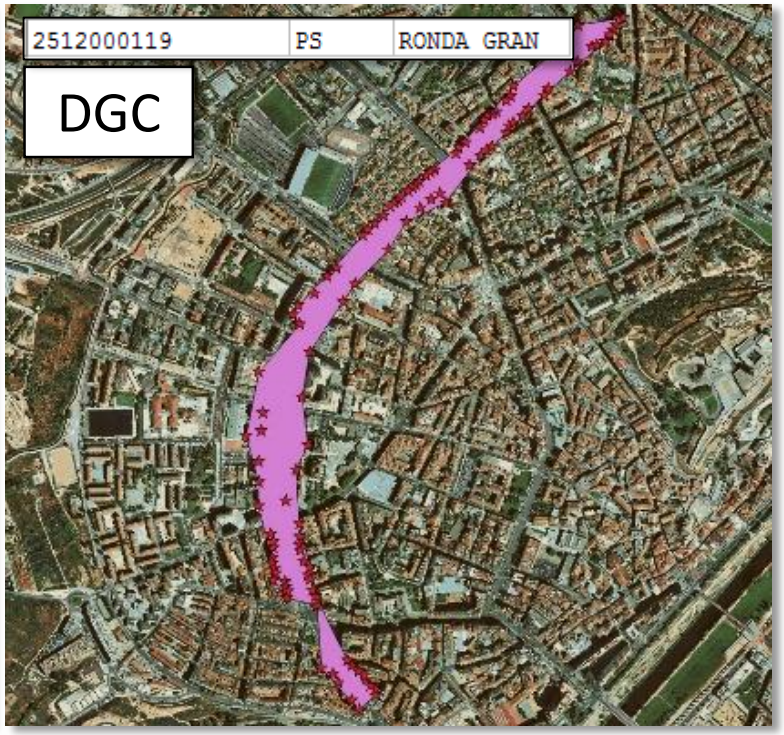
El problema (I)

- La **búsqueda mediante métricas entre literales** puede conducir a **falsos positivos**

'GRANPASSEIGRONDA'	TRIM(NOM_VIA)	JARO_S	EDIT_S
GRAN PASSEIG RONDA	SEGON DE RONDA	64	45
GRAN PASSEIG RONDA	RONDA GRAN	62	34
GRAN PASSEIG RONDA	DE LA TORRE PATRIOT	52	11
GRAN PASSEIG RONDA	RONDA	47	28



- Una vez se han geocodificado las direcciones a validar, podemos construir el '*concave hull*' (prioritariamente) o el '*convex hull*' de todos los puntos de una vía, para las dos fuentes (INE y DGC)
- Así, si dos polígonos de diferentes fuentes (DGC e INE) representan la **misma realidad física** la **intersección** entre ellos tiene que ser muy **elevada**, así como muy parecido su **azimut**

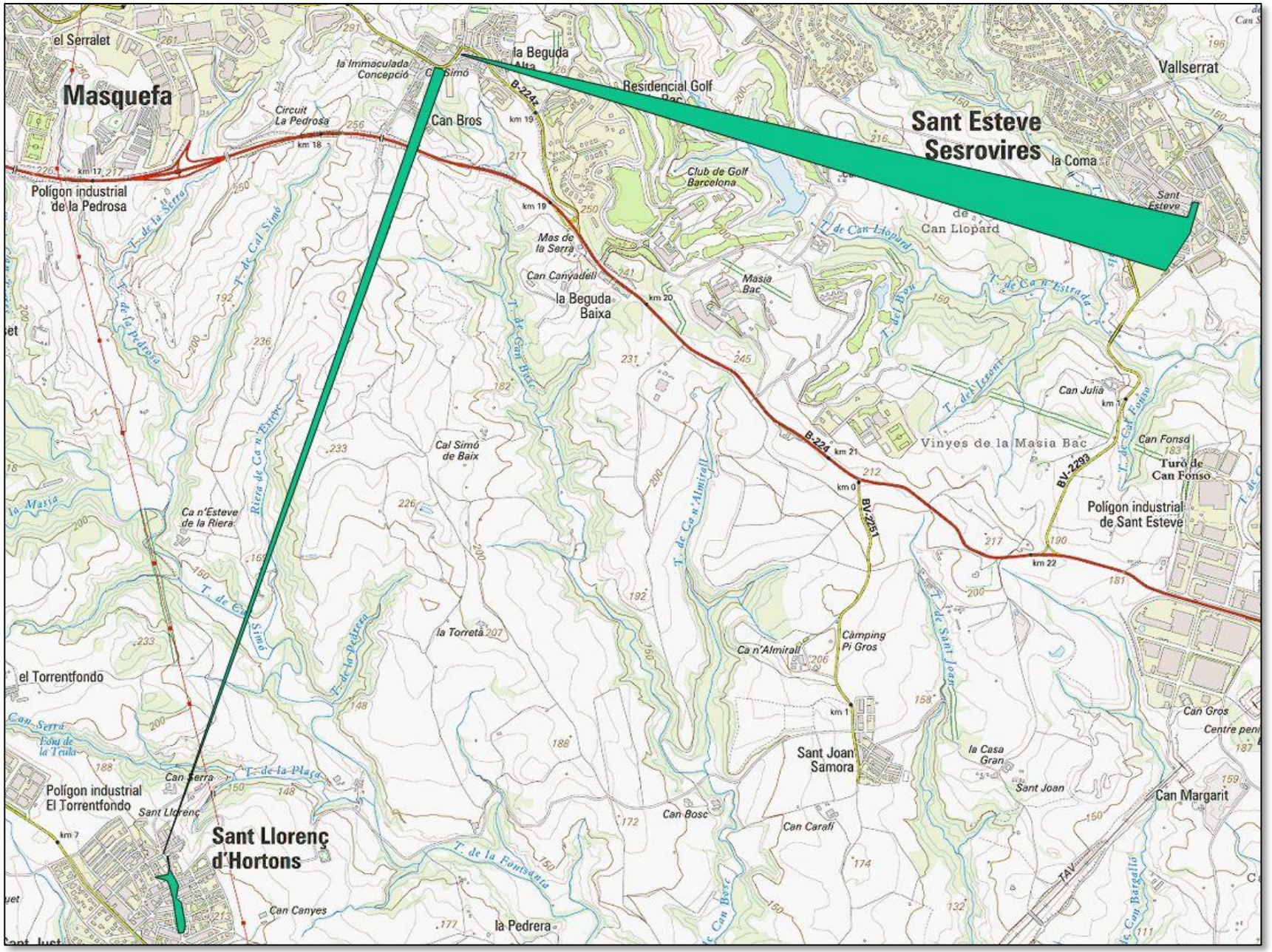


El problema (II)

Pero existen polígonos con algún **punto anómalo ***

OBJETIVO:

Detectar el máximo de polígonos erróneos



* Análisis sobre la Fuente DGC

Creación de la base de datos

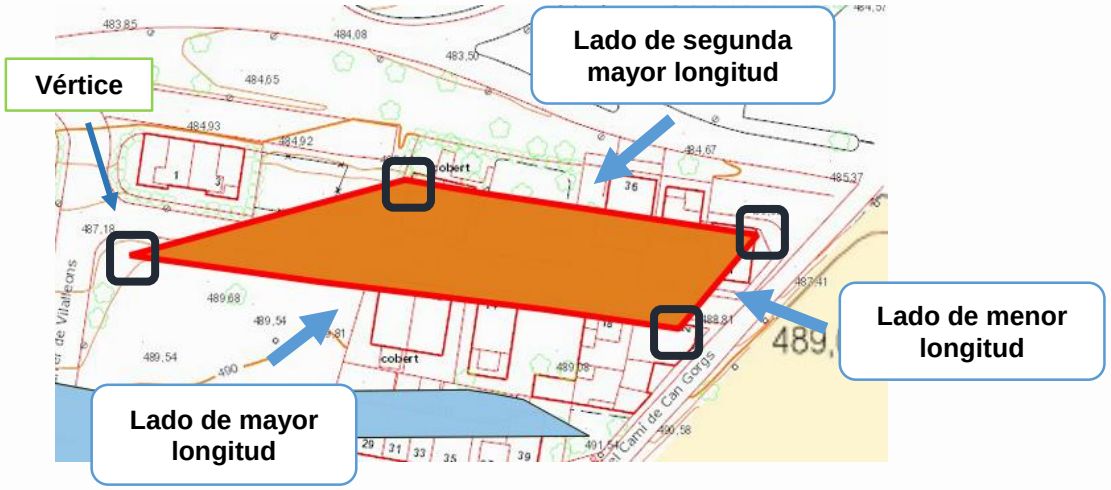
Base de datos de polígonos: 81.240 polígonos

Estudio previo: Análisis del tipos de vía
Decisión: Excluir los polígonos que representen diseminados, despoblados o afueras

Base de datos de polígonos final: 79.344 polígonos

Contiene información sobre diferentes variables agregadas sobre sus longitudes, número de vértices, diferencias de longitudes…

Ejemplo de polígono de 4 vértices



Proceso de análisis

Selección de una muestra etiquetada (aprendizaje y test)
1.150 polígonos

↓

Estudio univariante
Método de comparación: Área bajo la curva ROC

↓

Uso de modelos multivariantes

↓

Desarrollo de indicadores

↓

Propuesta de indicador de calidad

Proceso de análisis: selección de una muestra

- Para validar el proceso, se necesitan polígonos etiquetados con la información real sobre si son erróneos o no. Además, para predecir bien el comportamiento, se necesitan bastantes polígonos erróneos en ésta muestra.

- **Muestra discrecional de 1.150 polígonos** (inspección y etiquetaje de todos los polígonos de la muestra) distribuidos aleatoriamente entre:
  - Muestra de **aprendizaje (900 casos)**
  - Muestra de **test (250 casos)**

- Polígonos erróneos presentes en la muestra: **30,3%**

Proceso de análisis: estudio univariante

- **Método de comparación**: área bajo la curva ROC

- **Primer estudio**: Comportamiento no deseado de las variables relacionadas con el **lado de mayor longitud** del polígono

*Ejemplo de polígono **erróneo** con **dos lados largos** asociados*

*Polígono **correcto** con sólo **un lado largo***

- Creación de base de datos auxiliar con variables relacionadas con el polígono **sin tener en cuenta el lado de mayor longitud**

Proceso de análisis: estudio univariante

- **Variables a destacar**: las relacionadas con el polígono sin tener en cuenta el lado de mayor longitud
- **Método de comparación**: área bajo la curva ROC

$$Sensibilidad = \frac{VP}{VP + FN}$$

$$Especificidad = \frac{VN}{VN + FP}$$

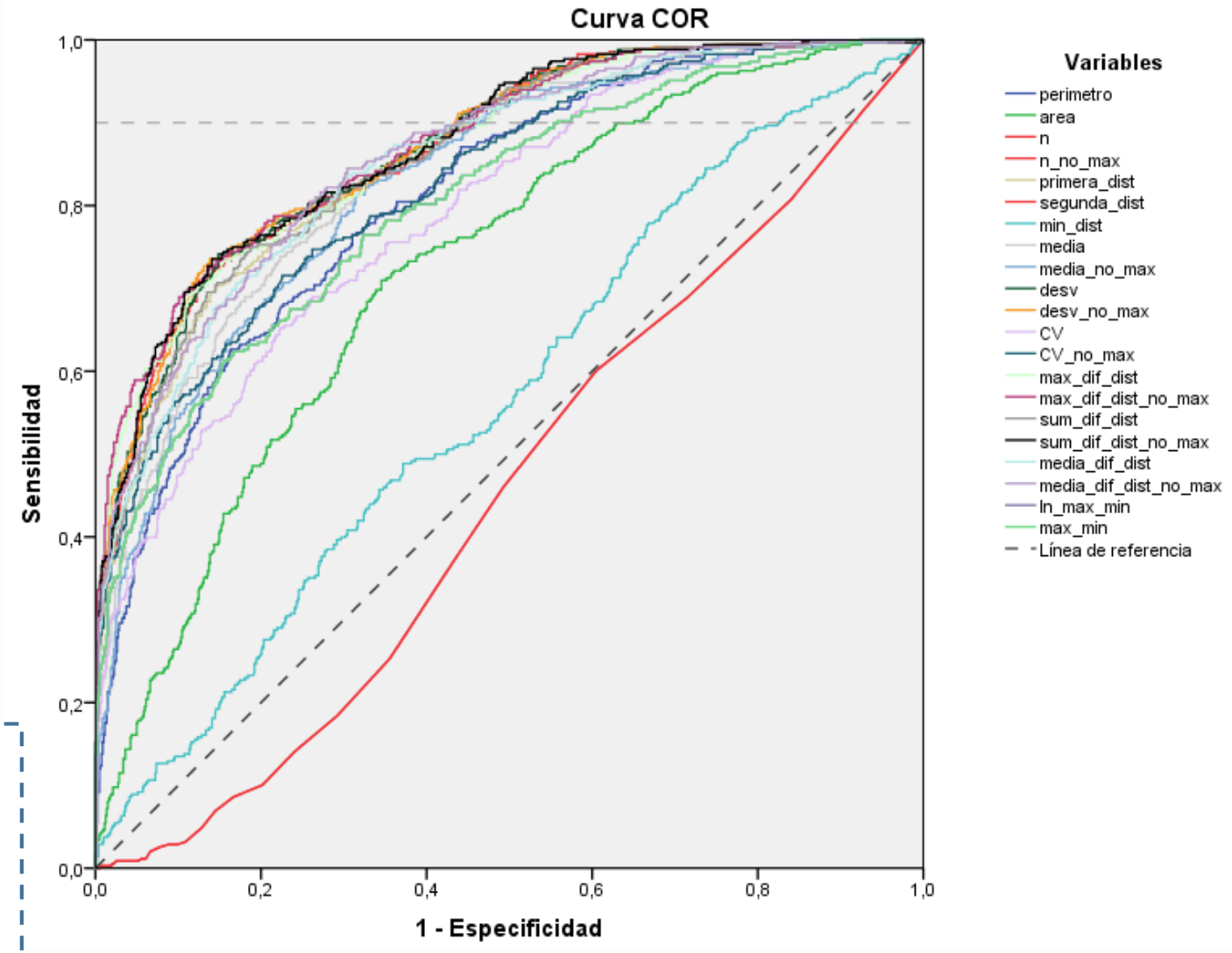
$VP = Verdaderos\ positivos$

$VN = Verdaderos\ negativos$

$FP = Falsos\ positivos$

$FN = Falsos\ negativos$

*El estado **positivo** es el asociado a un polígono **erróneo**, mientras que el estado **negativo** es el de un polígono **correcto***



Proceso de análisis: uso de modelos multivariantes

- Comparación con diferentes métodos multivariantes de clasificación supervisada. Se muestran algunos de los mejores resultados obtenidos:

Regresión logística

		REALIDAD	
		Polígono erróneo	Polígono correcto
PREDICCIÓN	Polígono erróneo	47	6
	Polígono correcto	42	155
		89	161
Sensibilidad		52,8%	

Árbol de inferencia condicional

		REALIDAD	
		Polígono erróneo	Polígono correcto
PREDICCIÓN	Polígono erróneo	51	15
	Polígono correcto	38	146
		89	161
Sensibilidad		57,3%	

Análisis discriminante lineal

		REALIDAD	
		Polígono erróneo	Polígono correcto
PREDICCIÓN	Polígono erróneo	20	1
	Polígono correcto	69	160
		89	161
Sensibilidad		22,5%	

Análisis discriminante cuadrático

		REALIDAD	
		Polígono erróneo	Polígono correcto
PREDICCIÓN	Polígono erróneo	41	7
	Polígono correcto	48	154
		89	161
Sensibilidad		46,1%	

Proceso de análisis: desarrollo de indicadores

- Creación de indicadores mediante la **combinaciones no lineales** de variables
- Uso de un umbral que proporcione aproximadamente un **90% de sensibilidad** en la muestra aprendizaje y parecido en la muestra de test (objetivo inicial cumplido), además de una alta especificidad.

Objetivo adicional:

Diferenciar el subconjunto de polígonos con **probabilidad más alta de contener puntos anómalos** entre los polígonos clasificados como erróneos

En otras palabras, **identificar polígonos prioritarios para su revisión**.

Proceso de análisis: propuesta de indicadores

Indicador seleccionado (ICP)

$$\mathrm{ICP}_p = CV_p{}^2 \cdot max\{longitud_i - longitud_{i+1}\}_{i=1\ldots v_p-1}$$

Donde "i" recorre los lados del polígono según su longitud decreciente, "v_p" es su número de vértices y "CV_p" es el coeficiente de variación de las longitudes de los lados.

Las variables se refieren al polígono **sin tener en cuenta el lado de mayor longitud**.

Este indicador, además de clasificar y ordenar los polígonos según su calidad estimada, permite calcular:

- porcentaje de polígonos susceptibles de ser erróneos (grupo 1) → **sensibilidad 90%**
- porcentaje de polígonos que necesitan una revisión prioritaria (grupo 2) → **especificidad 95%**

Resultados (I)

Tabla. Prevalencia de polígonos erróneos

	Prevalencia de polígonos erróneos	
	Muestra aprendizaje (900 casos)	Muestra test (250 casos)
Grupo de calidad 1	48,2	45,1
Grupo de calidad 2 (Revisión prioritaria)	81,0	90,9

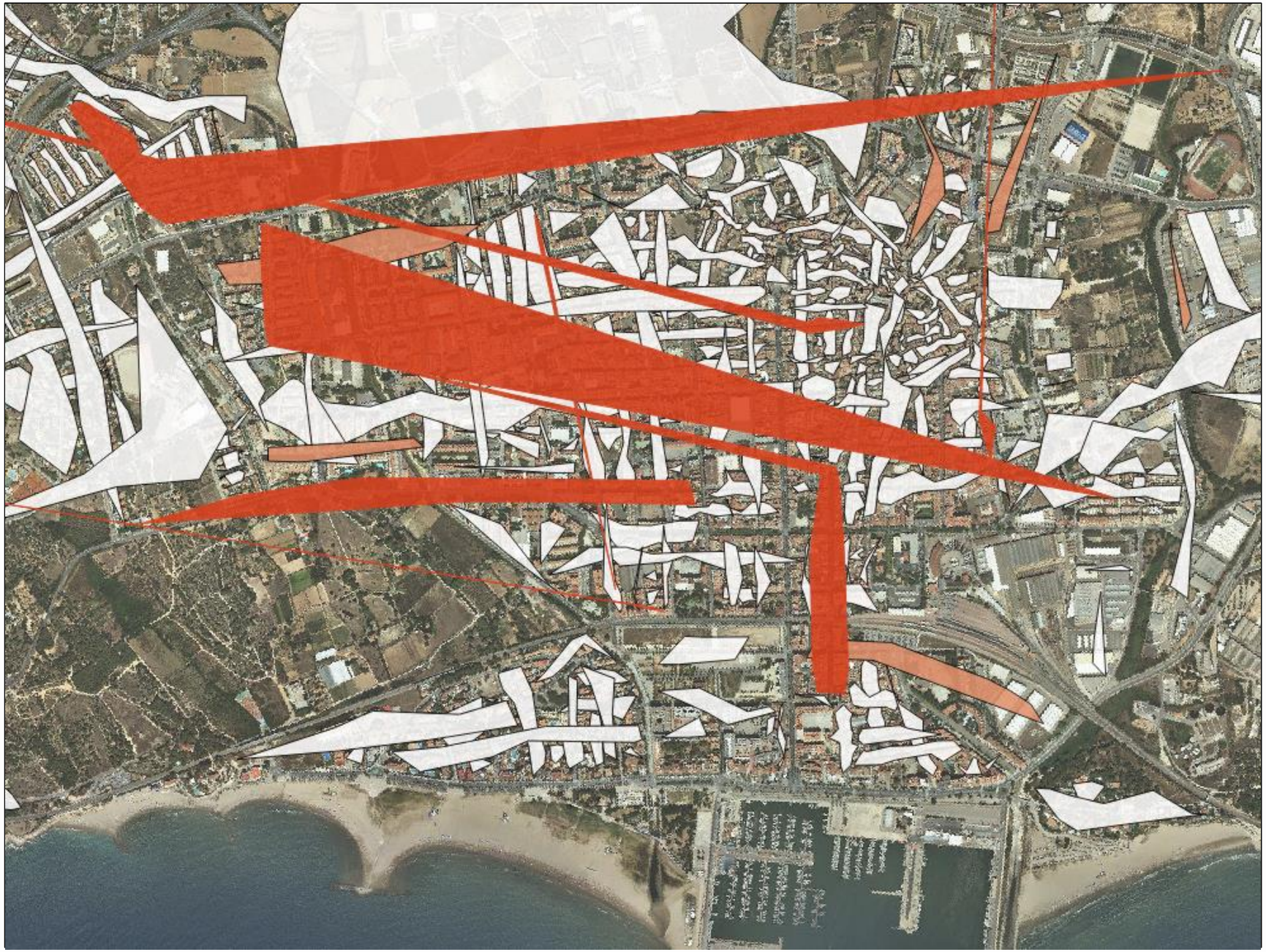
Tabla. Distribución en la base poblacional

	% polígonos en la base de datos
Grupo de calidad 1	8,9
Grupo de calidad 2 (Revisión prioritaria)	2,9

El grupo de polígonos con **más probabilidad de ser erróneos** (grupo 2) es de un **2,9%** del total la base de datos de Cataluña, lo que conlleva que el número de puntos anómalos sea aún menor.

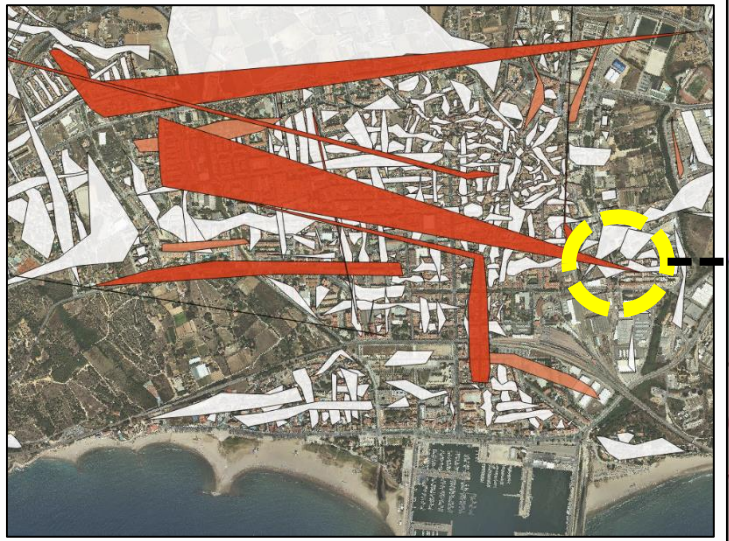
Resultados (II)

ICP del municipio de Vilanova i la Geltrú.



Resultados (III): causas

Causa: vía mal asignada



Calle Joan Ricart ≠ Calle Joan Fuster

CARRER DE JOAN RICART

Información de parcelas e inmuebles

PARCELA CATASTRAL 4146004CF9644N

Croquis

Fotografía fachada

Parcela construida sin división horizontal
CL JOAN FUSTER 40
VILANOVA I LA GELTRÚ (BARCELONA)
131 m²

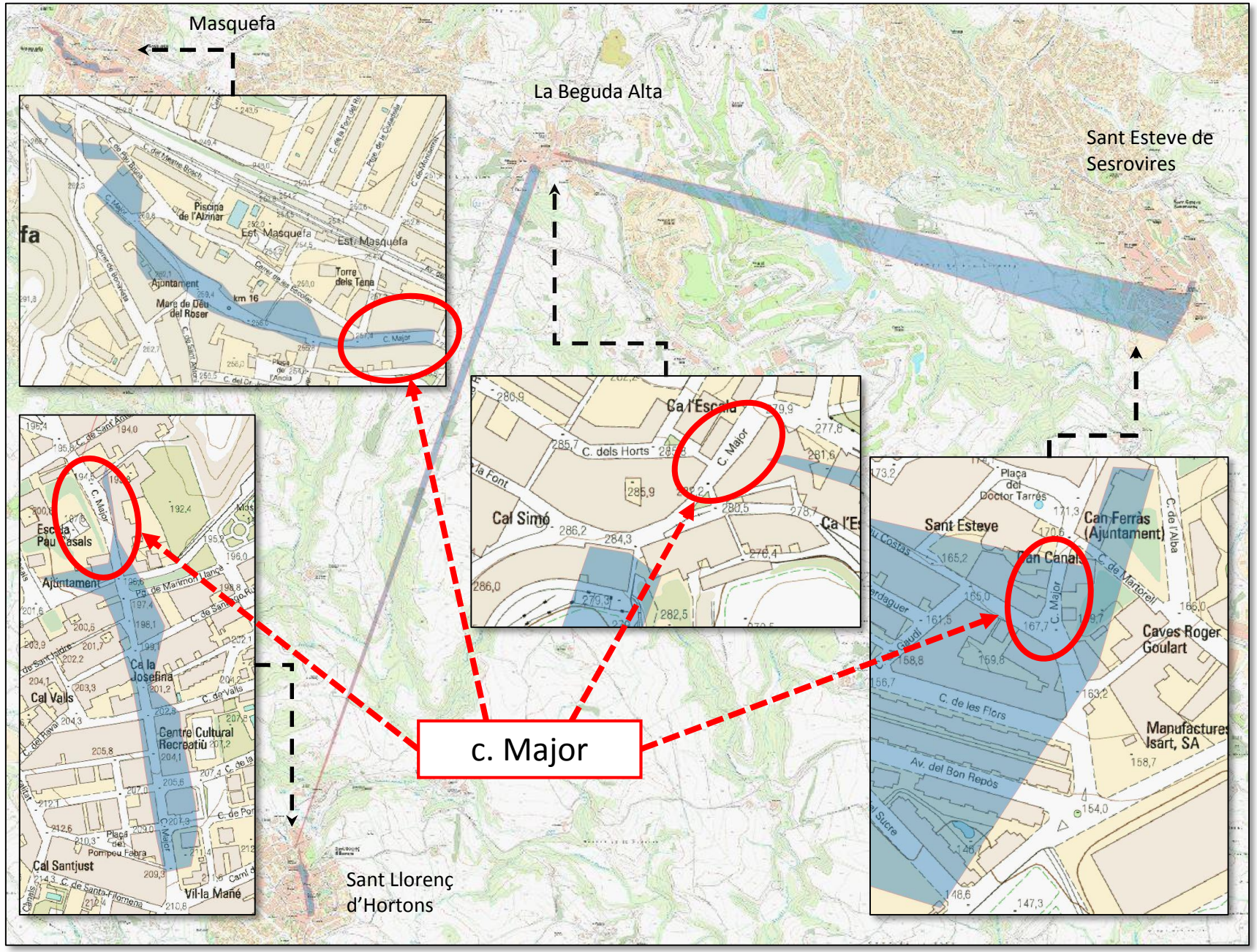
INFORMACIÓN DE LOS INMUEBLES

Excel

4146004CF9644N0001ZK CL JOAN FUSTER 40
Residencial | 293 m² | 100,00% | 1999

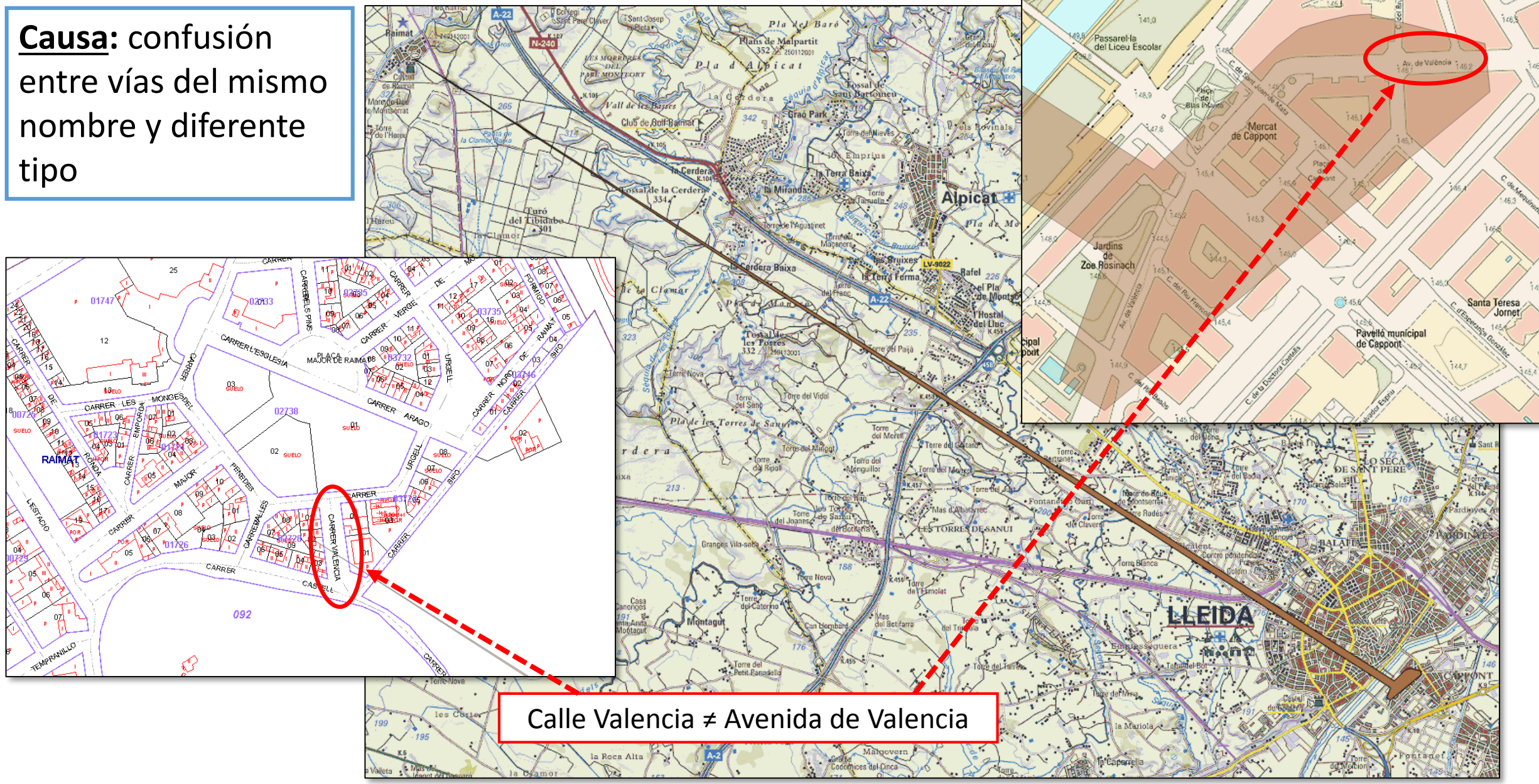
Resultados (IV): causas

Causa: confusión entre vías del mismo nombre y tipo (c. Major)



Resultados (V): causas

Causa: confusión entre vías del mismo nombre y diferente tipo



Conclusiones

- Se ha propuesto un indicador (ICP) que **ordena y clasifica** polígonos según su **calidad estimada** (separa polígonos presuntamente correctos de los **erróneos**).

- Adicionalmente, se han podido diferenciar los polígonos con una probabilidad más alta de contener puntos anómalos entre los polígonos etiquetados como erróneos (**revisión prioritaria**).