Generalitat de Catalunya
Institut d'Estadística de Catalunya

Consideraciones sobre la utilización de Quad Trees en la difusión de datos geocodificados y la preservación del secreto estadístico

Introducción.

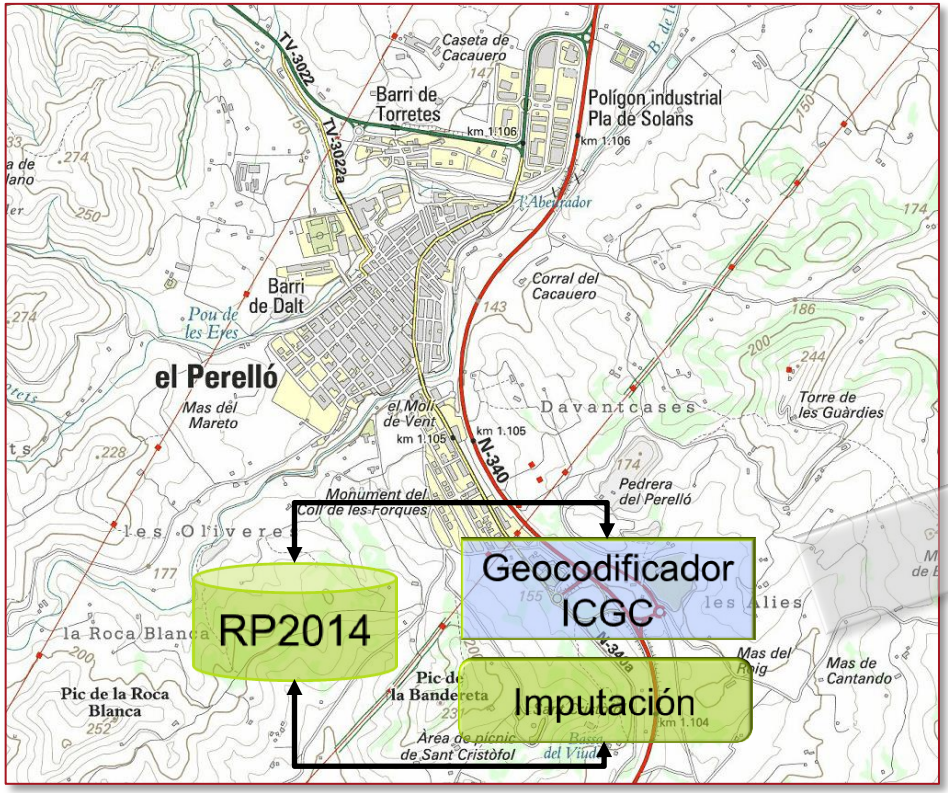
Quadtree. Construcción.

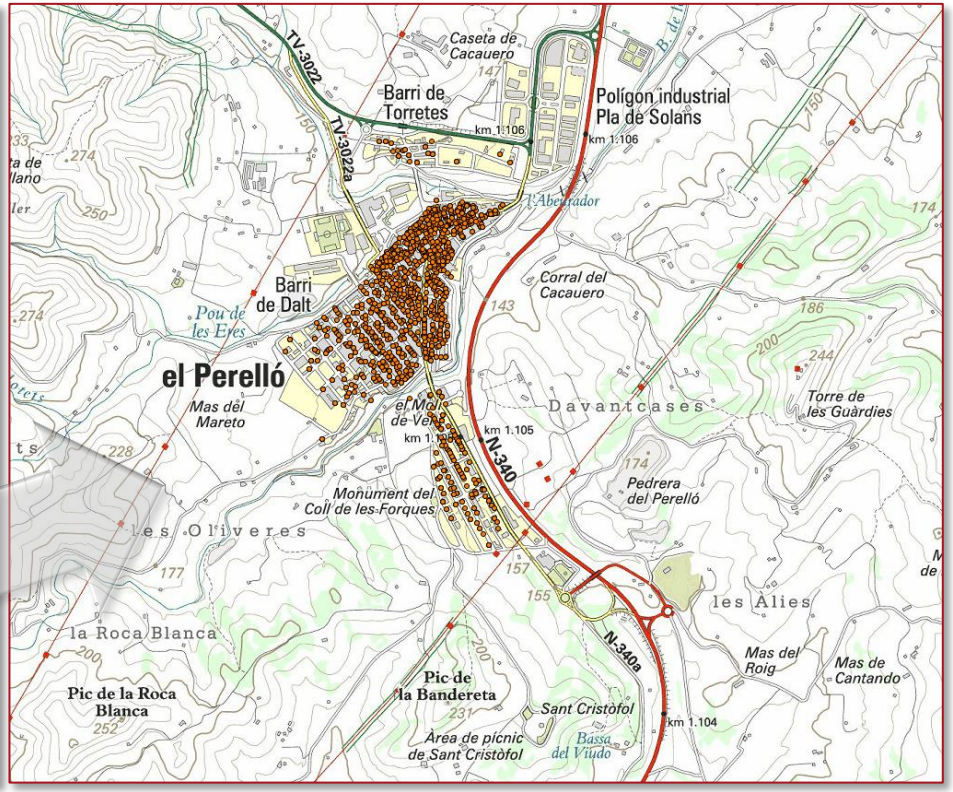
Quadtree. Efecto frontera.

Quadtree. Estimación de errores.

Conclusiones.

Introducción.





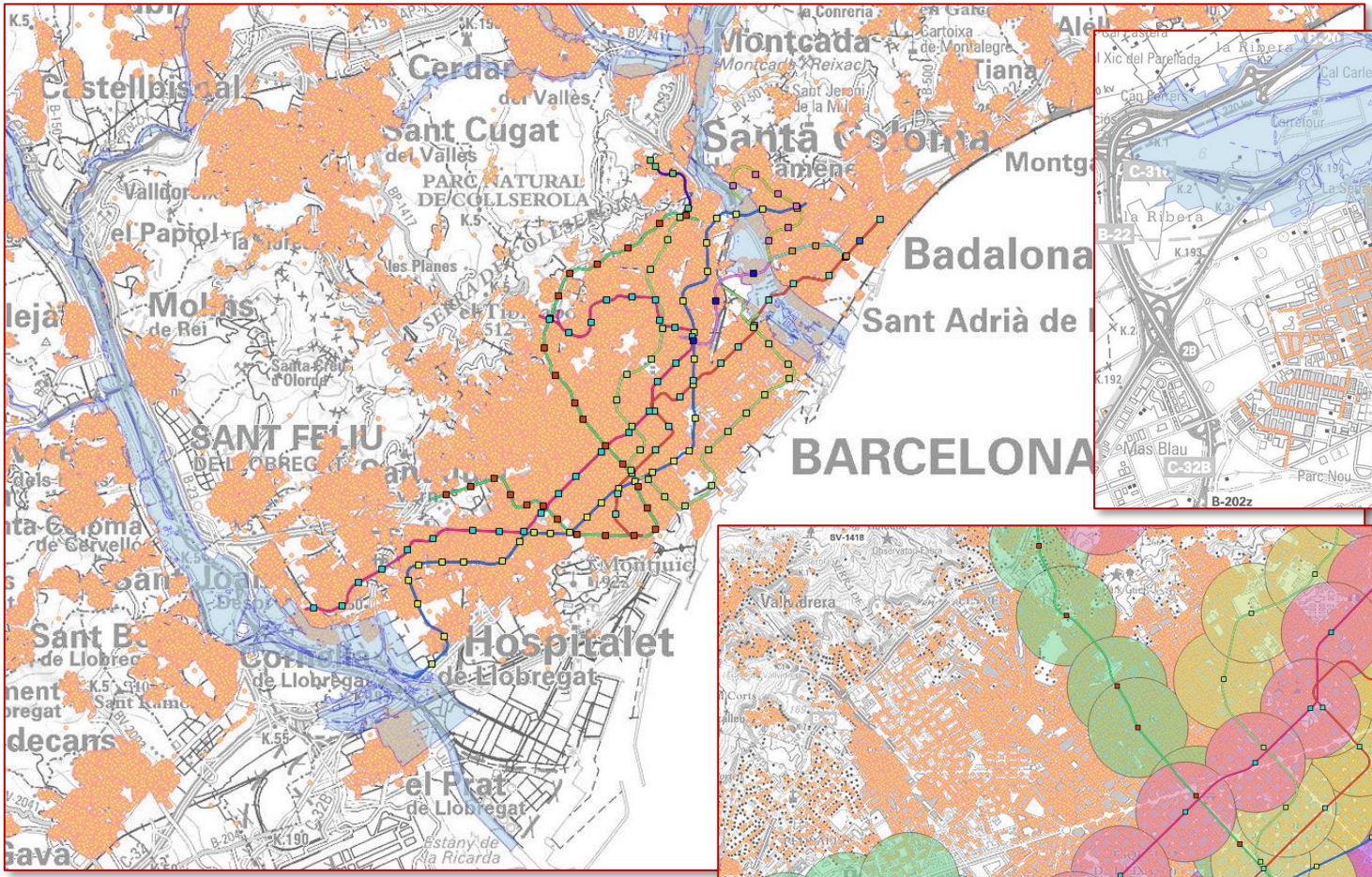
Exactitud (RP2014)	N	% imputados
Portal	2581826	
Portal interpolado	3204766	
Portal interpolado más cercano	1351764	
Rectángulo rodeando la calle	93852	
Portal asignado a finca vía tabla ine-dgc numeración coincidente	127047	1.68
Portal asignado aleat. a finca vía tabla ine-dgc numeración más cercana dentro del convex	8164	0.11
Portal asignado aleat. a finca vía tabla ine-dgc numeración más cercana fuera del convex	69161	0.91
Portal asignado aleat. a finca según tabla hogares-bien inmuebles dentro del convex	113861	1.50
Portal asignado aleat. a finca dentro del convex	16023	0.21
Total	**7566464**	**4.42**

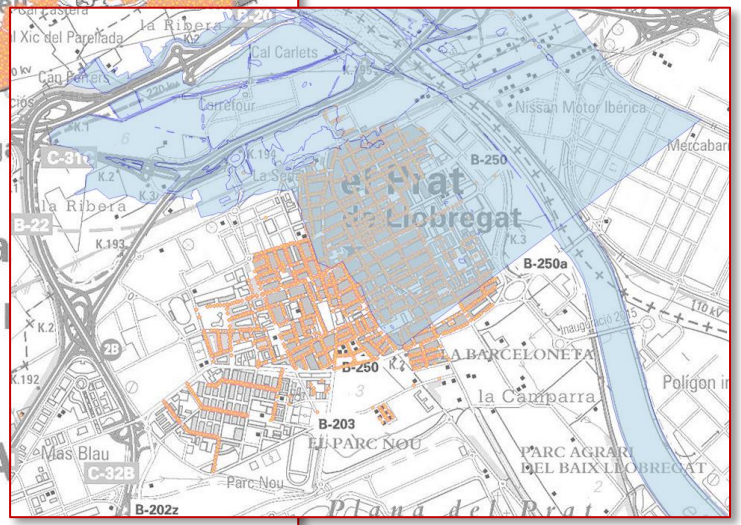
Usos:

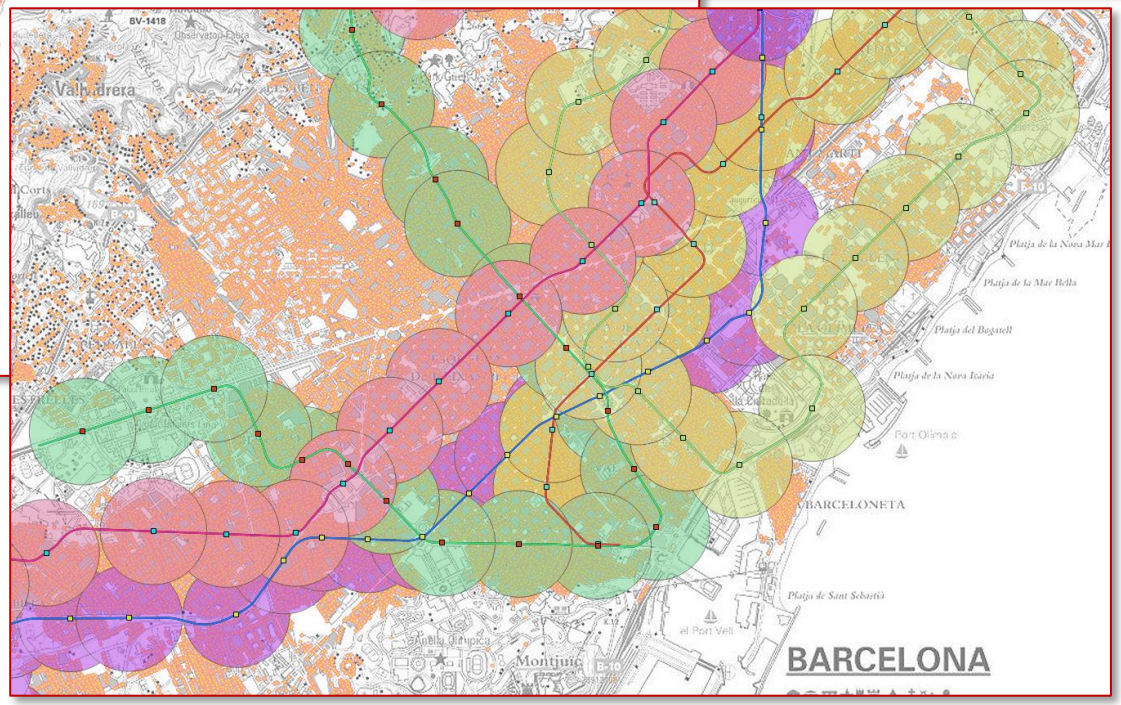
- Planificación del transporte
- Protección civil
- Localización de infraestructuras
- Calidad del aire

  etc...

Introducción



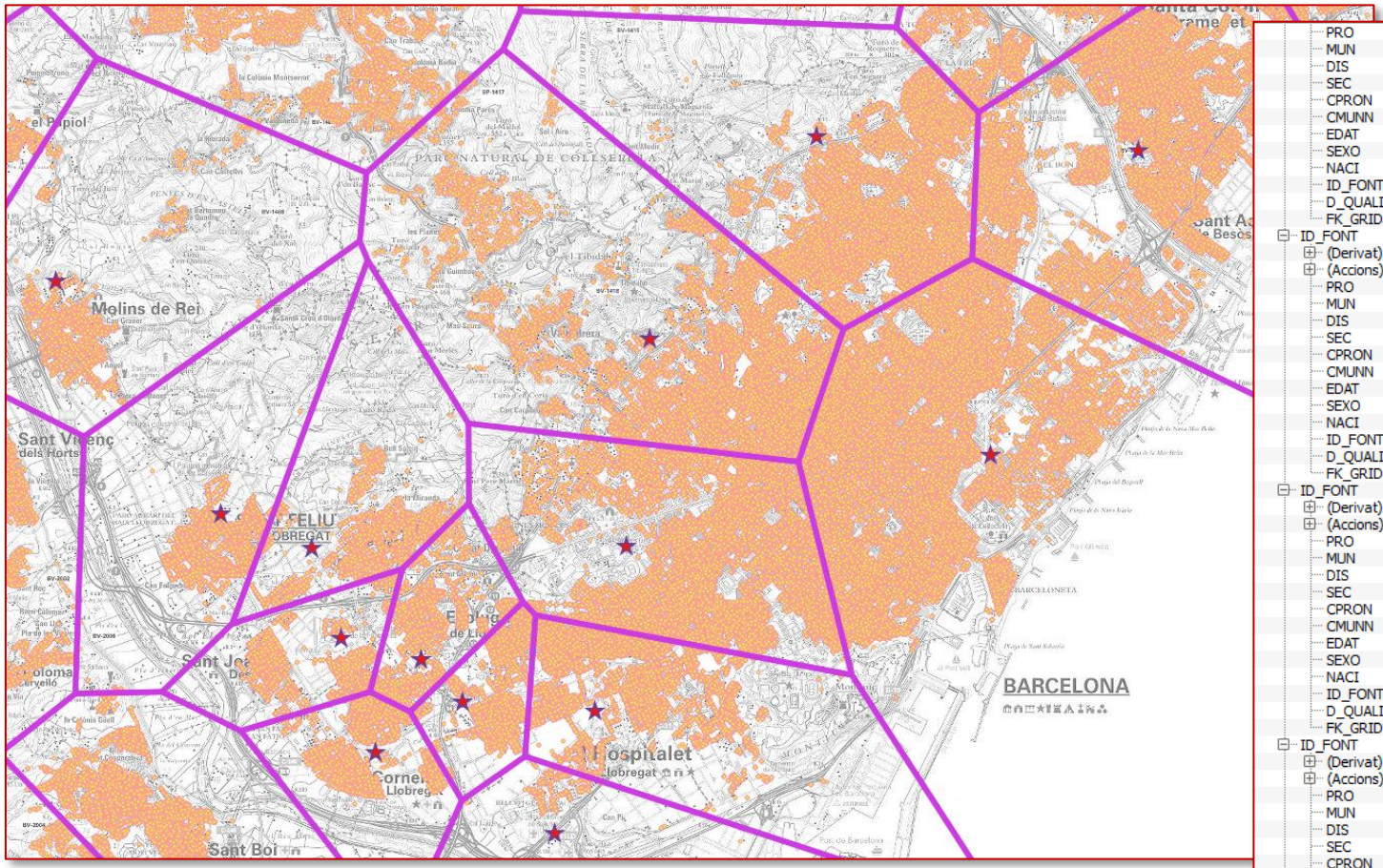




Usos:

<u>cálculo de población dentro de S</u>

Introducción.



- PRO: 08
- MUN: 205
- DIS: 01
- SEC: 012
- CPRON: 08
- CMUNN: 205
- EDAT: 81.0000000000
- SEXO: 6
- NACI: 108
- ID_FONT: 1.0000000000
- D_QUALITAT: 1.0000000000
- FK_GRID250: 251620

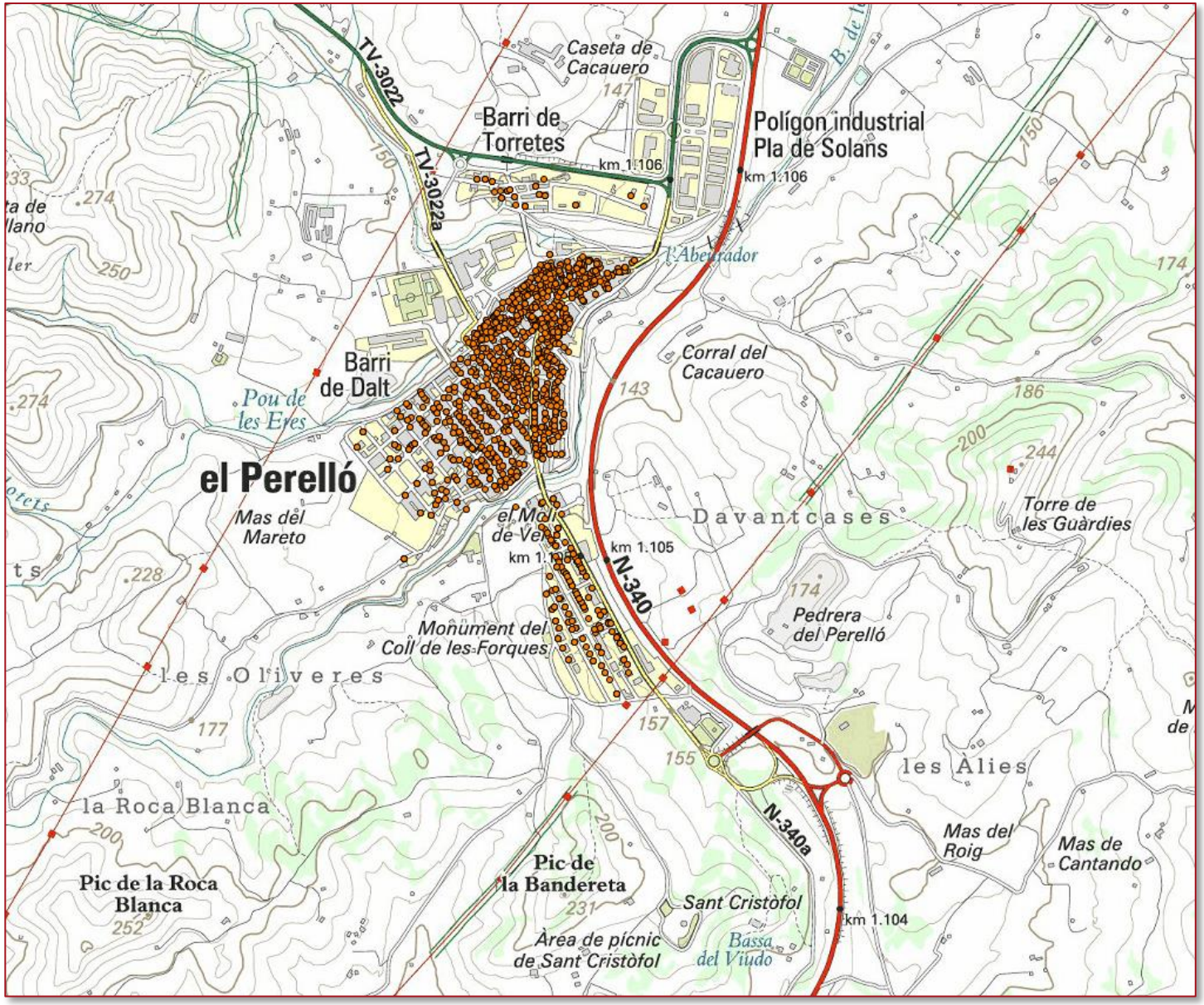
ID_FONT 1.0000000000
- (Derivat)
- (Accions)
- PRO: 08
- MUN: 205
- DIS: 01
- SEC: 012
- CPRON: 08
- CMUNN: 019
- EDAT: 56.0000000000
- SEXO: 6
- NACI: 108
- ID_FONT: 1.0000000000
- D_QUALITAT: 1.0000000000
- FK_GRID250: 251620

ID_FONT 1.0000000000
- (Derivat)
- (Accions)
- PRO: 08
- MUN: 205
- DIS: 01
- SEC: 012
- CPRON: 08
- CMUNN: 205
- EDAT: 59.0000000000
- SEXO: 1
- NACI: 108
- ID_FONT: 1.0000000000
- D_QUALITAT: 1.0000000000
- FK_GRID250: 251620

ID_FONT 1.0000000000
- (Derivat)
- (Accions)
- PRO: 08
- MUN: 205
- DIS: 01
- SEC: 012
- CPRON: 08
- CMUNN: 205
- EDAT: 30.0000000000
- SEXO: 1
- NACI: 108
- ID_FONT: 1.0000000000
- D_QUALITAT: 1.0000000000
- FK_GRID250: 251620

ID_FONT 1.0000000000
- (Derivat)
- (Accions)
- PRO: 08
- MUN: 205
- DIS: 01
- SEC: 012
- CPRON: 08
- CMUNN: 205
- EDAT: 31.0000000000
- SEXO: 6
- NACI: 108

Introducción.

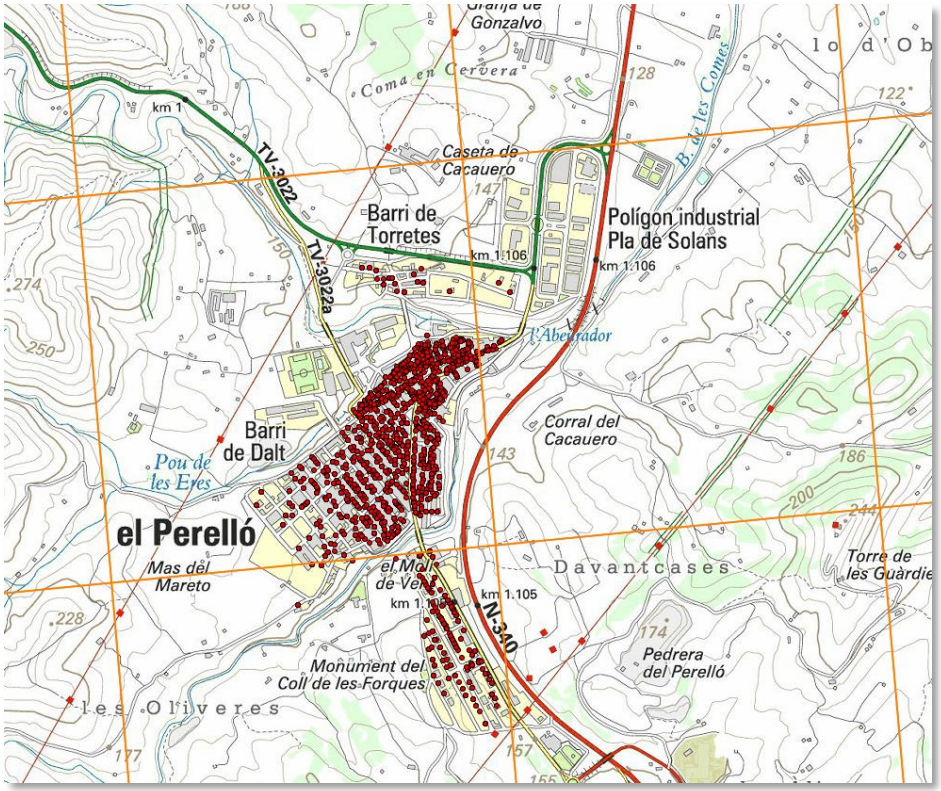


En el momento de la difusión, como preservar el secreto estadístico?

Perturbando posiciones (puntos → puntos)

Agregando espacialmente (puntos → polígonos)

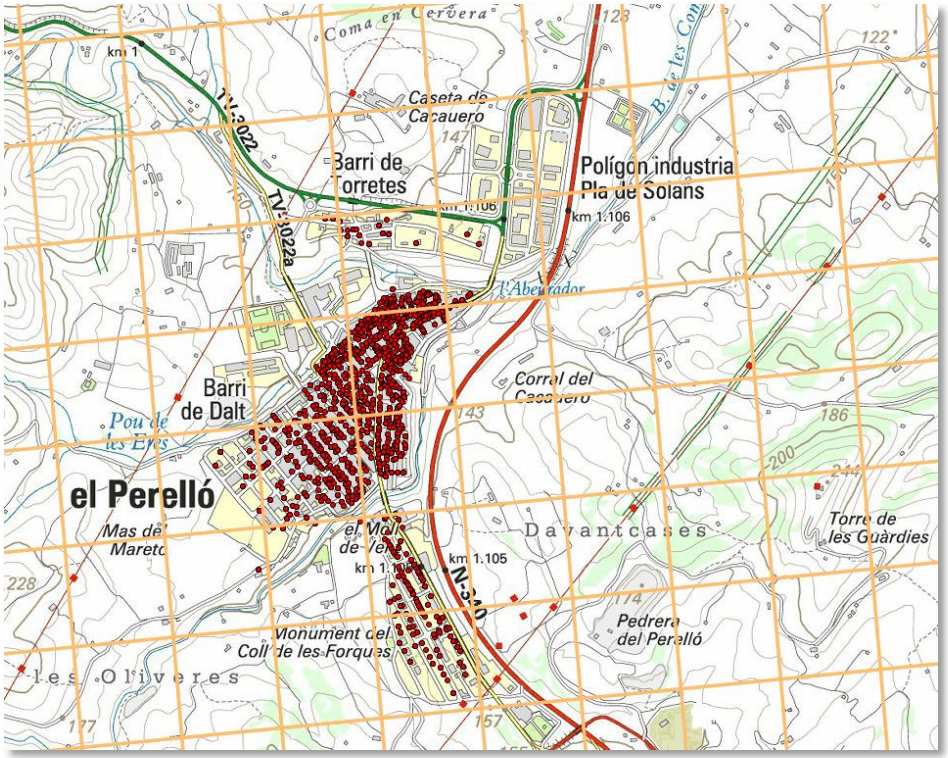
Introducción.



Mayor resolución → mayor riesgo de revelación
Menor resolución → error al evaluar poblaciones aumenta

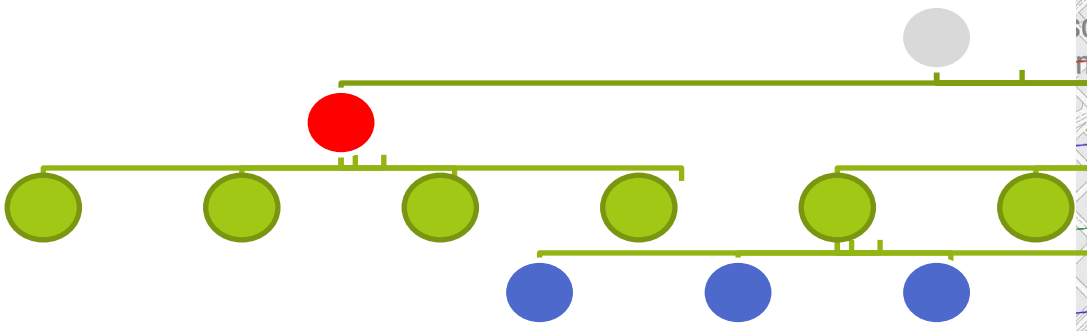
Agregando espacialmente
(puntos → polígonos)

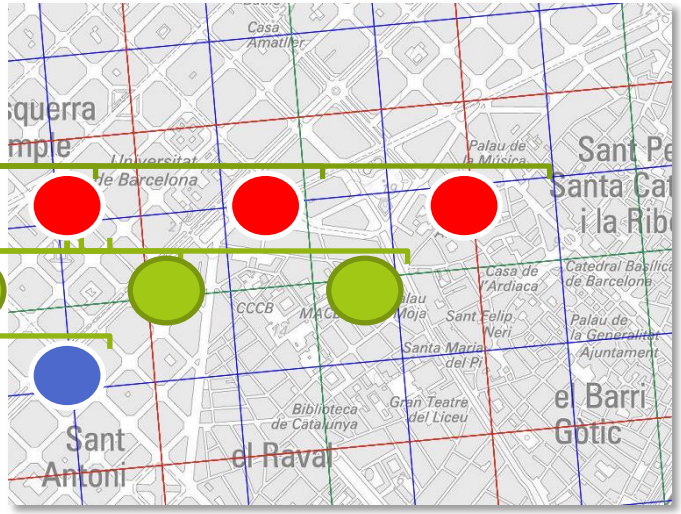
¿Qué tipo de polígonos?
¿Con qué resolución?



Quadtree.

Estructura jerárquica en que los padres o tienen cuatro hijos o ninguno.



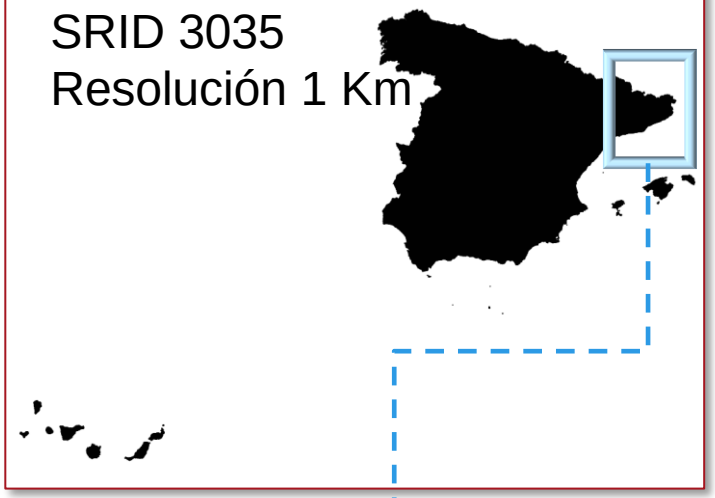


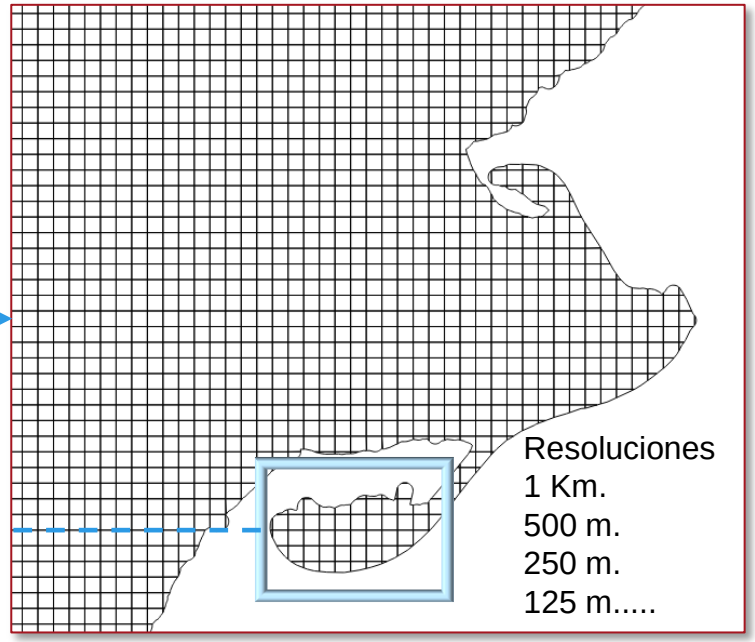
la resolución se adapta localmente para preservar el secreto estadístico:

si hay <u>suficiente población</u> se divide el área... y así recursivamente..

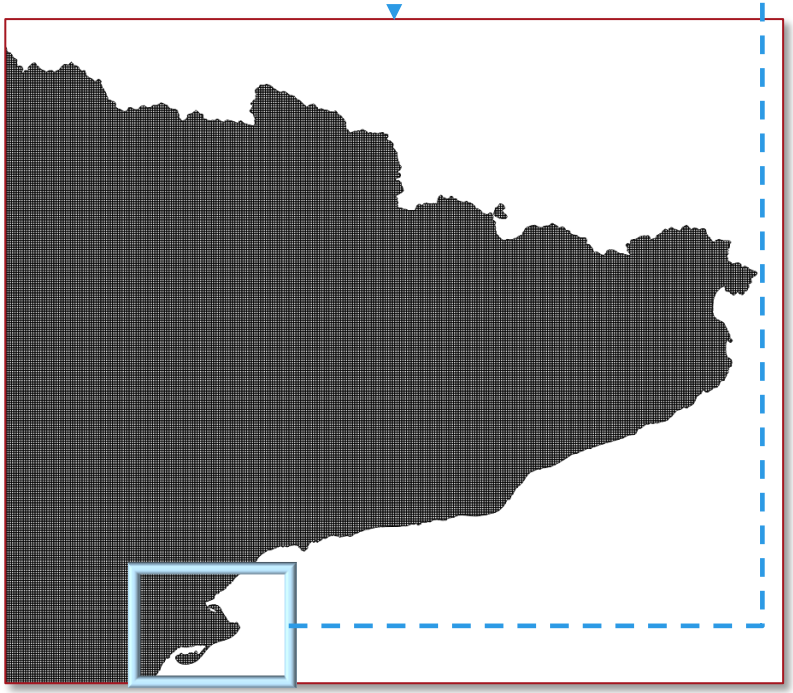
Grid estándar europeo.

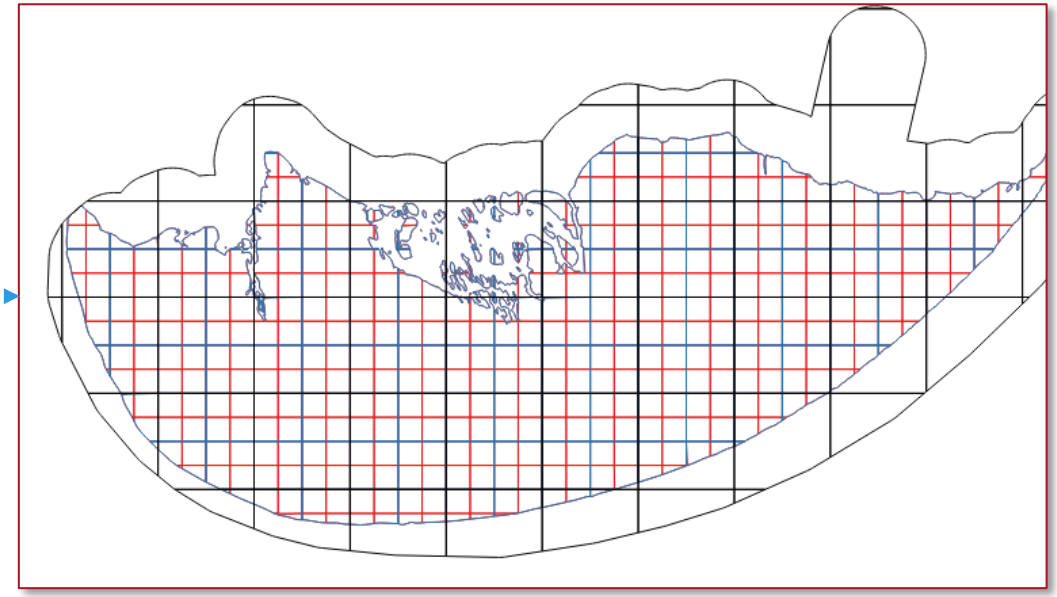
SRID 3035
Resolución 1 Km



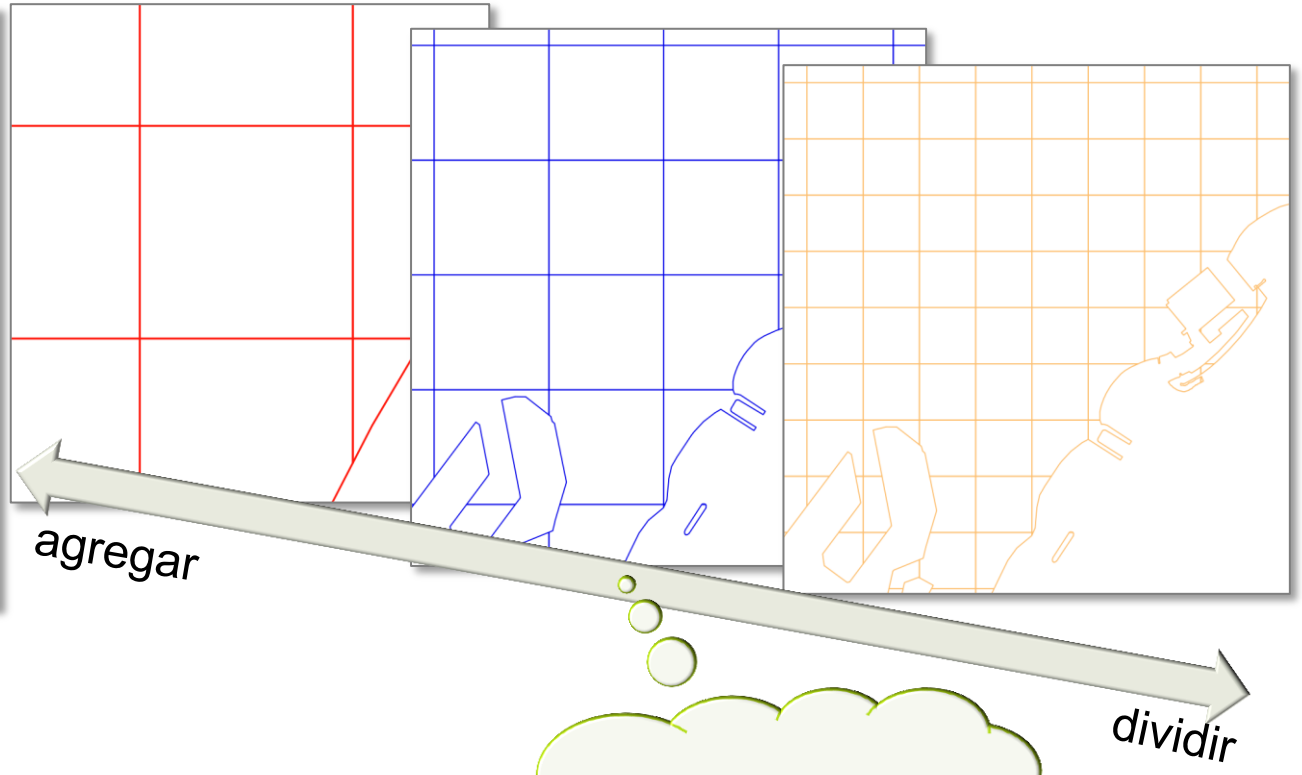


Resoluciones
1 Km.
500 m.
250 m.
125 m.....

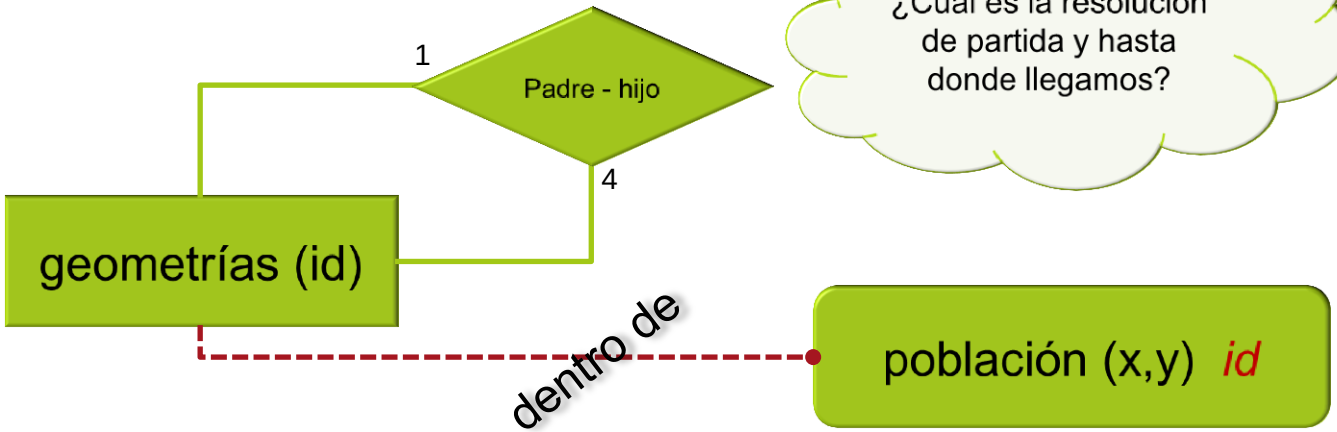




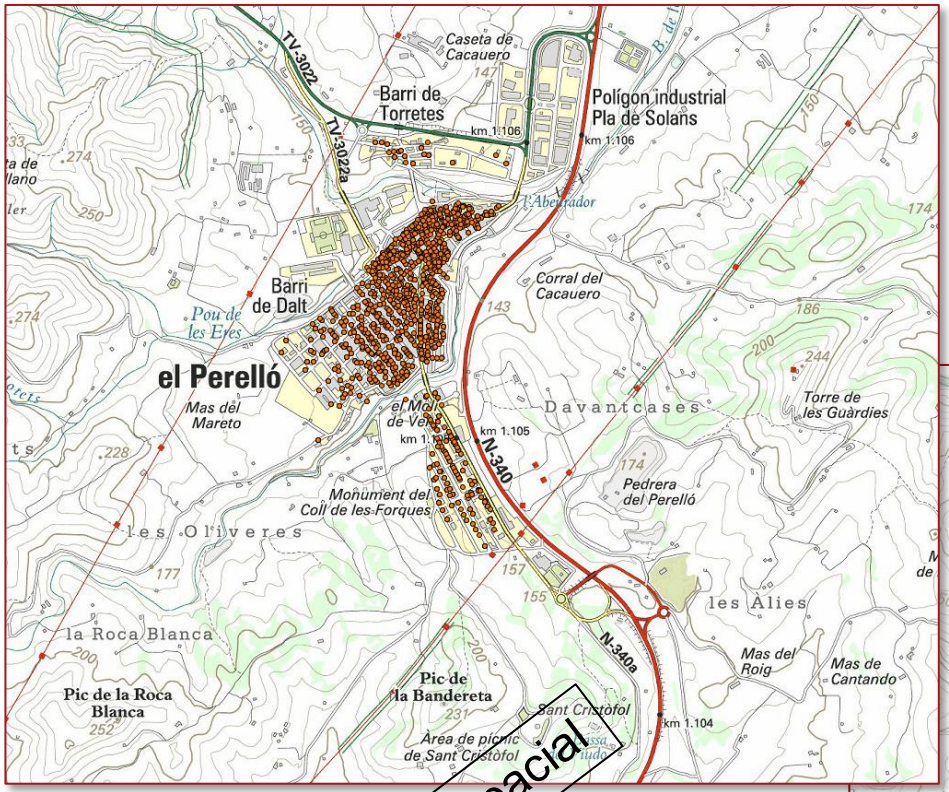
Quadtree. Construcción.

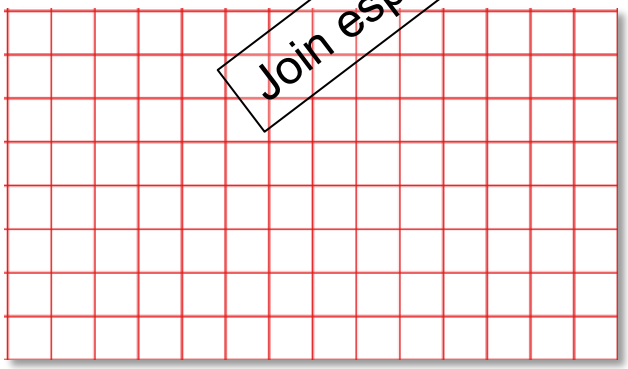


¿Cual es la resolución de partida y hasta donde llegamos?

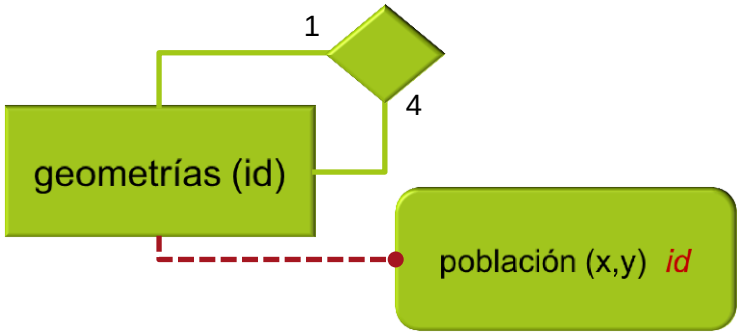


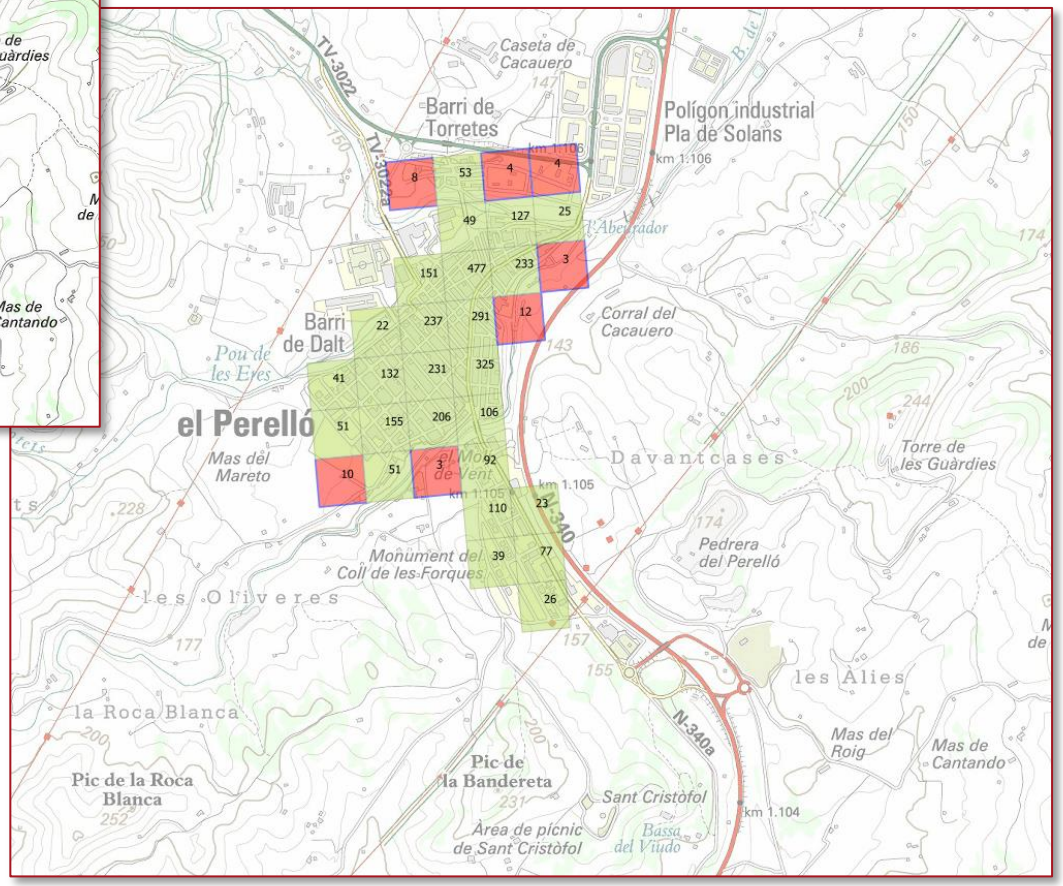
Quadtree. Construcción.





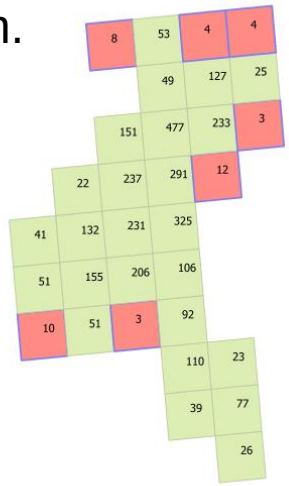
Join espacial

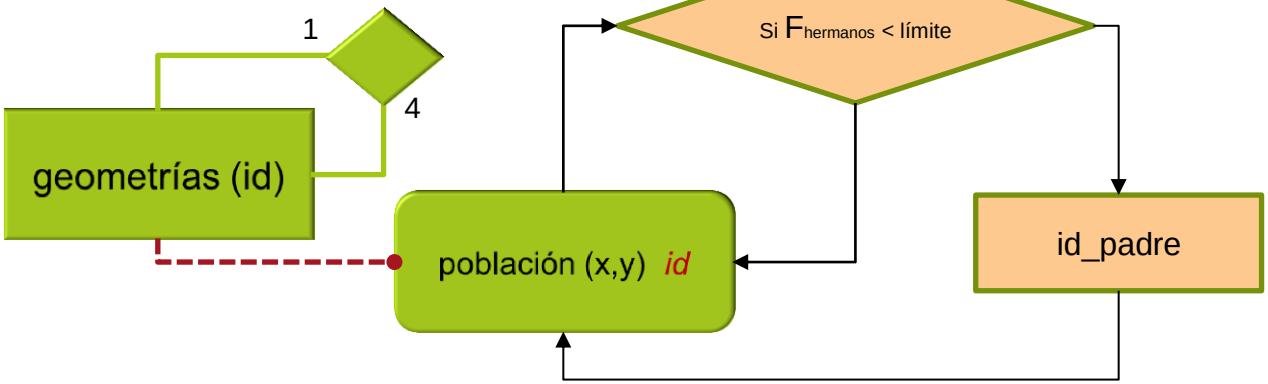




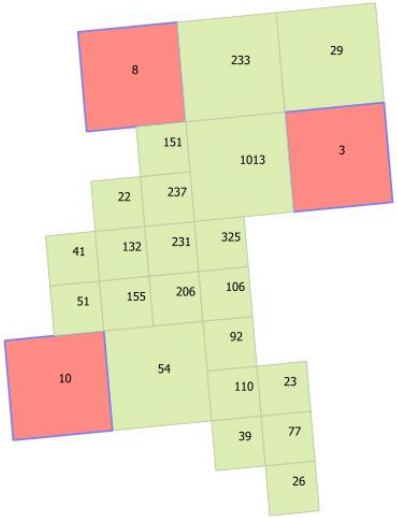
Quadtree. Construcción.

125 m.

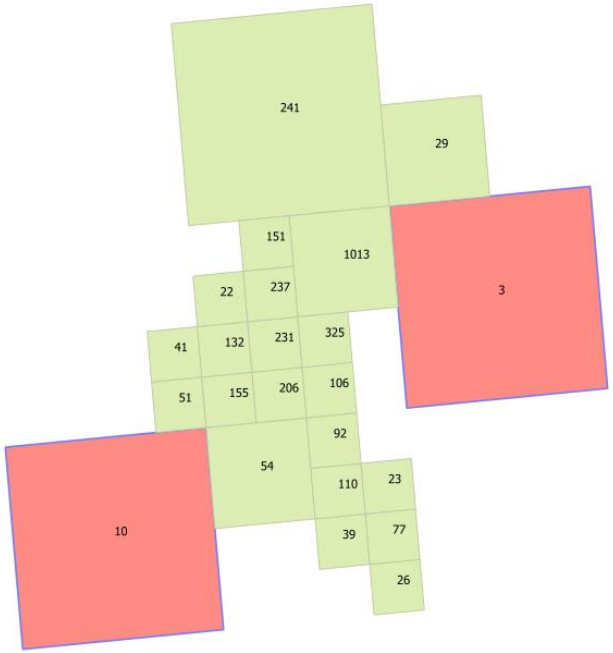




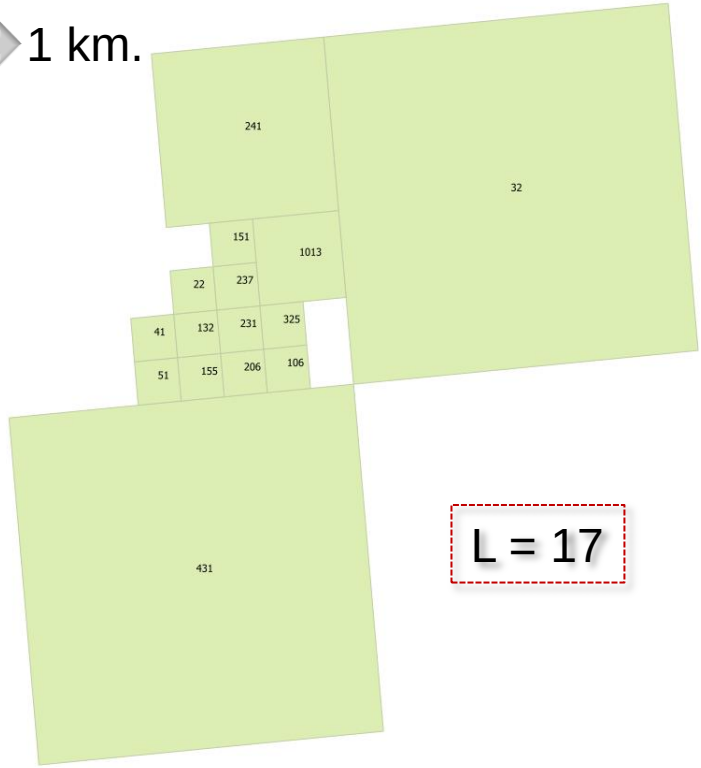
250 m.



500 m.

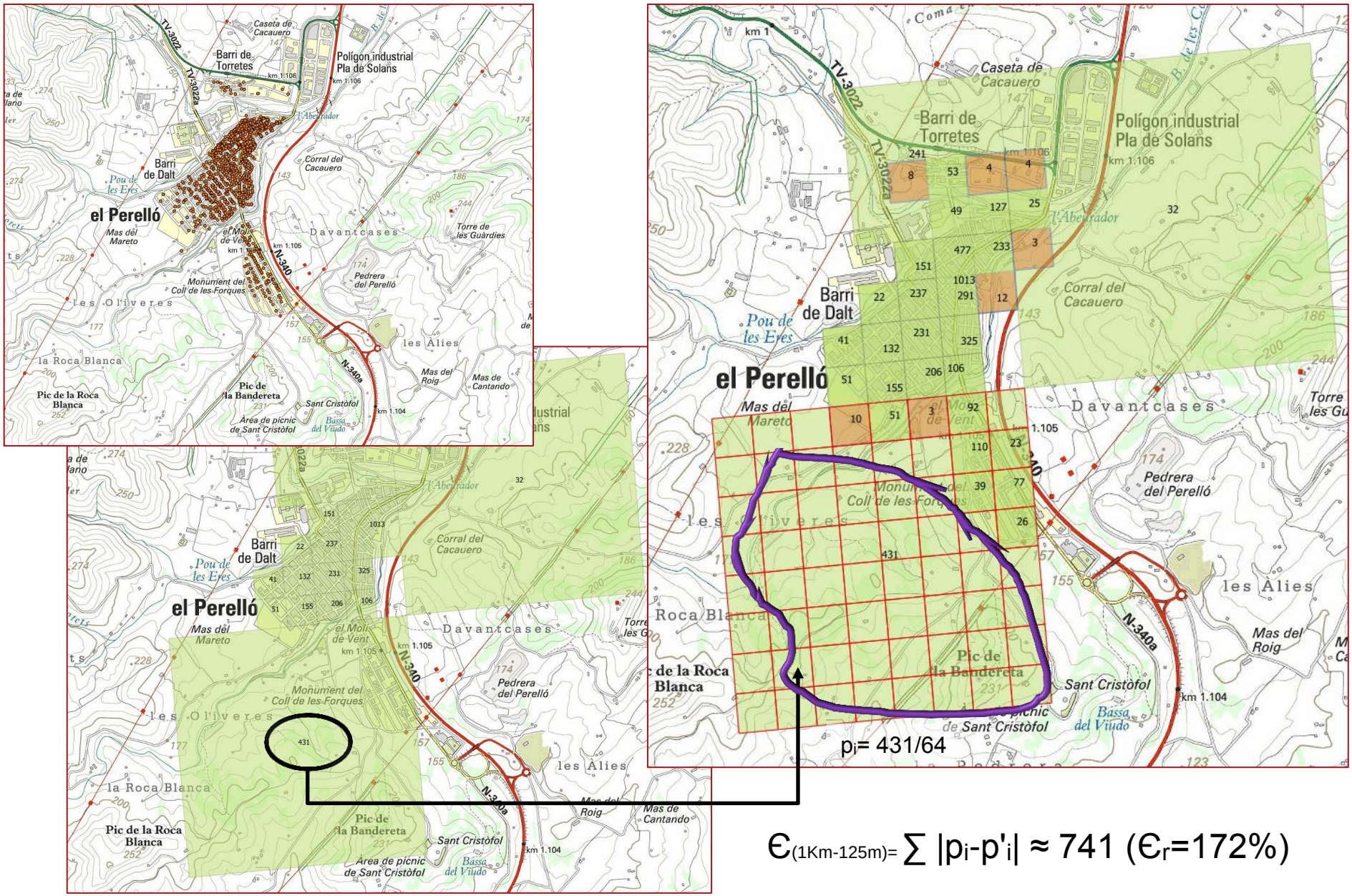


1 km.



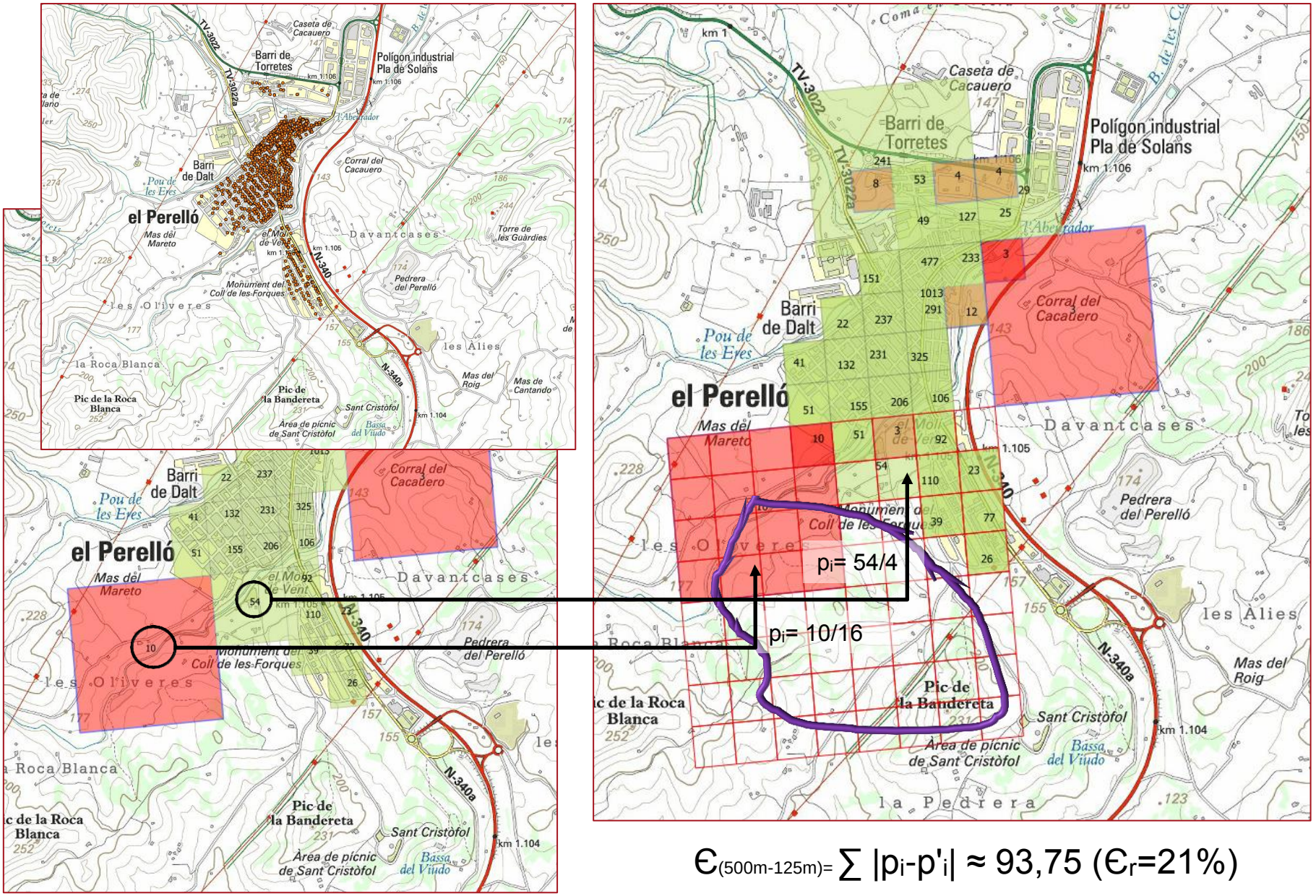
L = 17

Quadtree. Resolución.



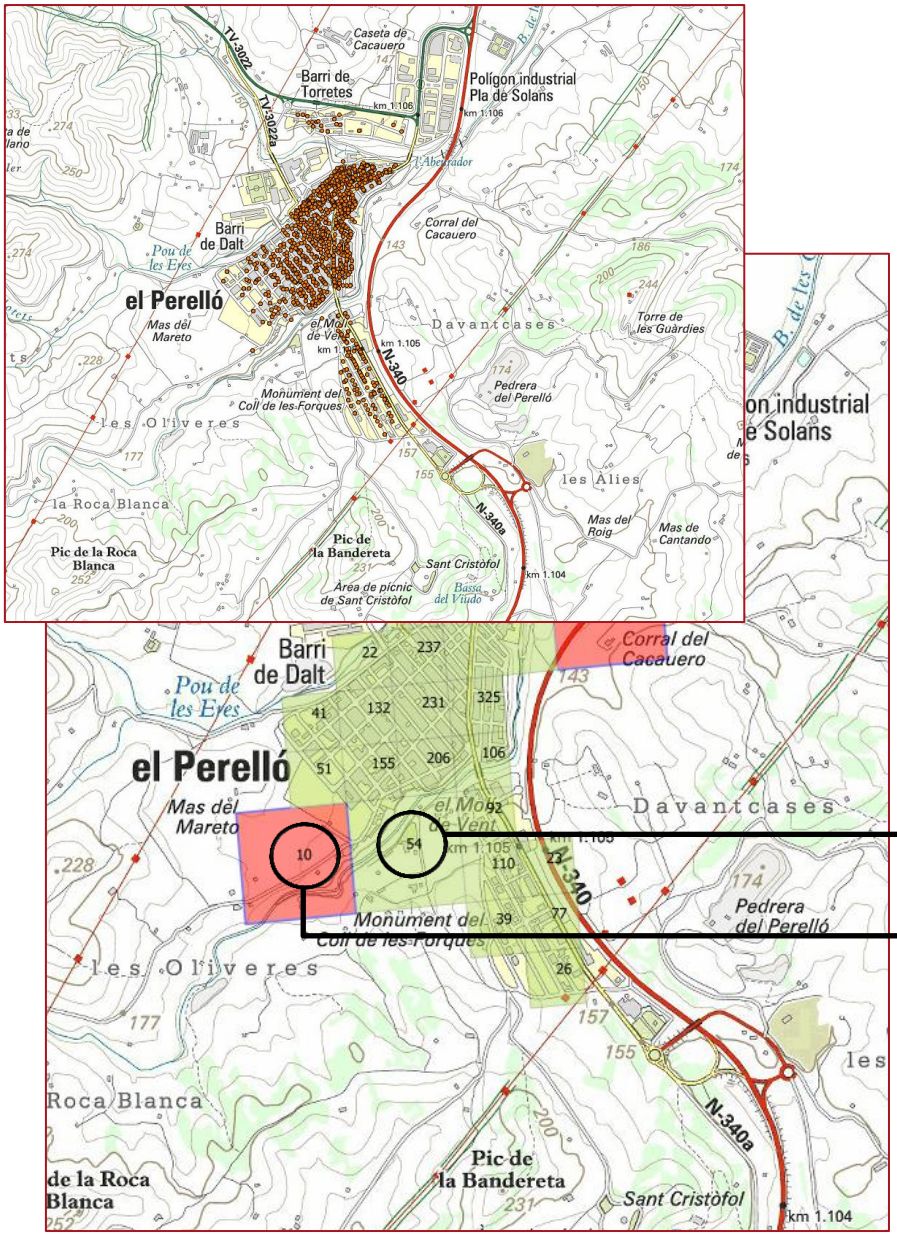
$$\epsilon_{(1Km-125m)} = \sum |p_i - p'_i| \approx 741 \; (\epsilon_r = 172\%)$$

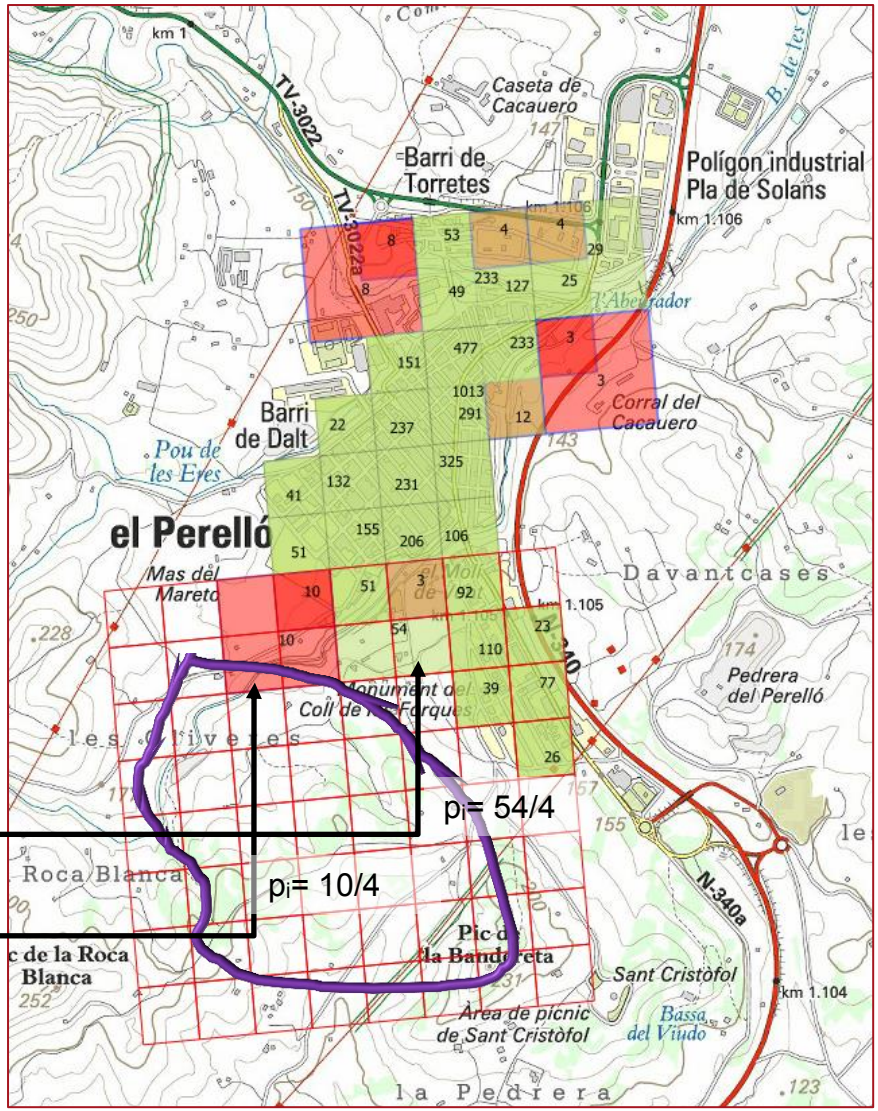
Quadtree. Resolución.



$$\epsilon_{(500m-125m)} = \sum |p_i - p'_i| \approx 93{,}75 \ (\epsilon_r = 21\%)$$

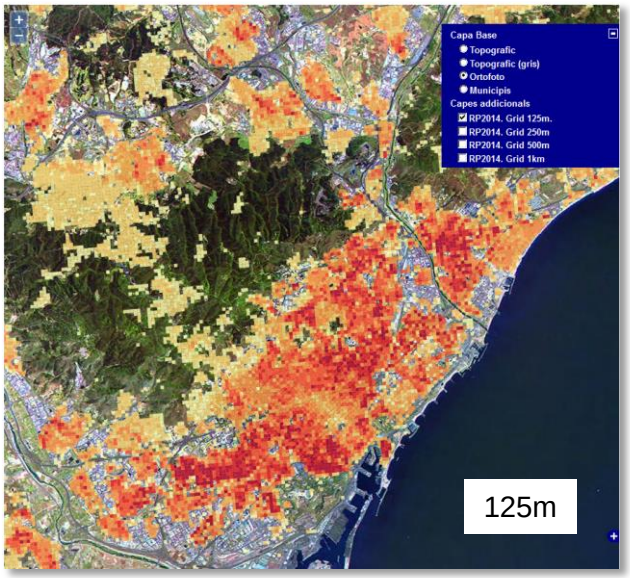
Quadtree. Resolución.

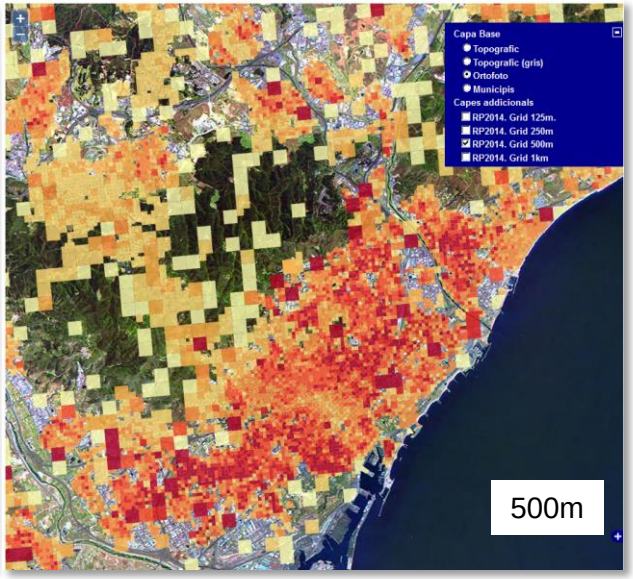


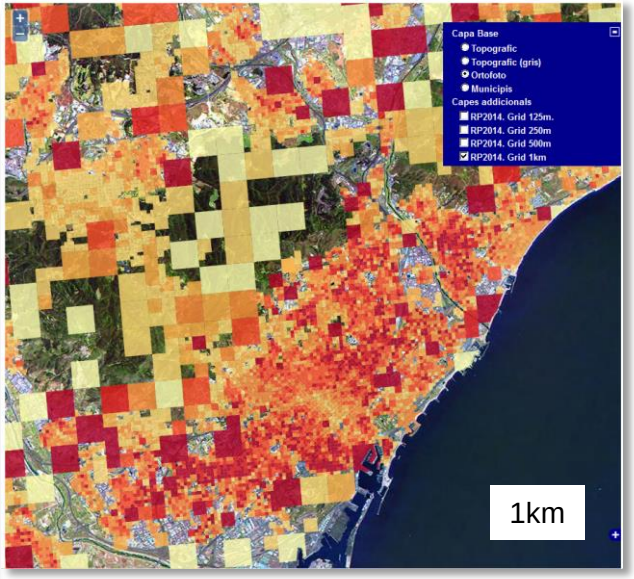


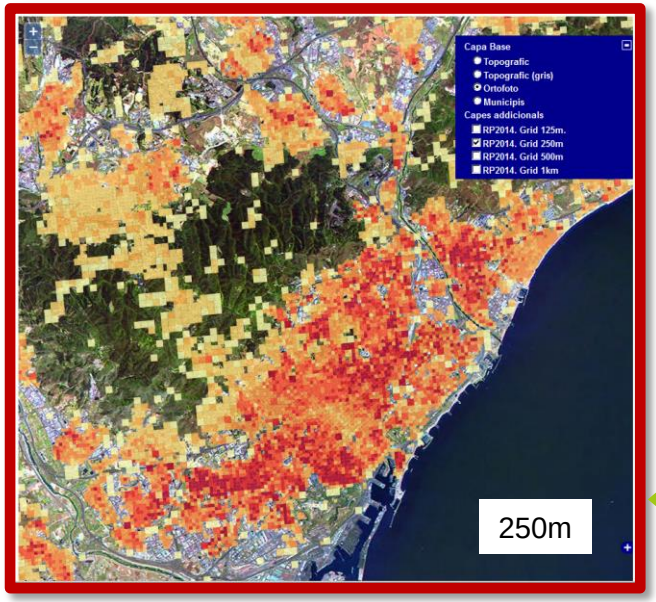
$$\epsilon_{(250m-125m)} = \sum |p_i - p'_i| \approx 90 \ (\epsilon_r = 20{,}8\%)$$

Quadtree. Parámetros de definición.









Quadtree	total	< 17	17 -100	100 - 250	250 - 500	500- 750	750 - 1000	> 999	% < 17	Pob. < 17	% pob (q < 17)
125 m.	85994	38272	29181	9762	5297	2047	893	542	44,50	249043	3,29
250 m.	58189	15071	23807	10155	5519	2123	932	582	25,90	87667	1,15
500 m.	45715	7431	19191	9724	5517	2171	983	698	16,25	41917	0,55
1 Km.	36807	3561	15429	8540	5248	2161	1002	866	9,67	20926	0,27

Quadtree {l:17,Max 125;Min. 250m; RP2014}

Cuadrados de	F	Población
250 m	26.420	1.208.272
125m	31.769	6.358.192
Total	58.189	7.566.464

Qué hacen en otros países.

2013 ERDKUNDE Vol. 67 · No. 2 · 151–166

USING QUADTREE REPRESENTATIONS IN BUILDING STOCK VISUALIZATION AND ANALYSIS

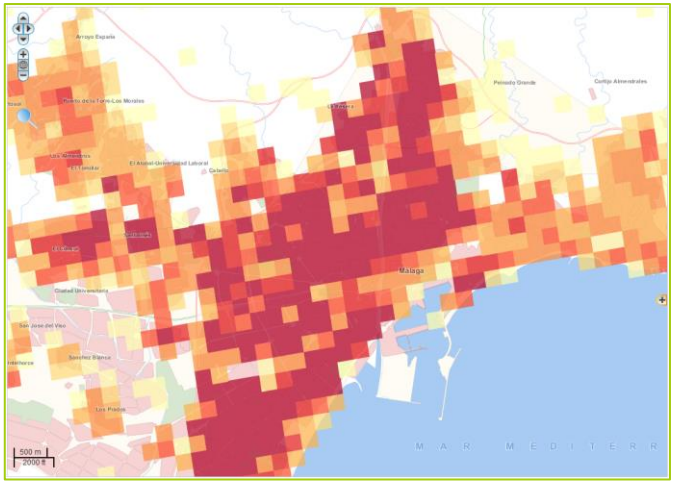
MARTIN BEHNISCH, GOTTHARD MEINEL, SEBASTIAN TRAMSEN and MARKUS DIESSELMANN

With 7 figures and 1 table
Received 14. September 2012 · Accepted 25. February 2013

https://www.erdkunde.uni-bonn.de/

Instituto de Estadística y Cartografía de Andalucía
JUNTA DE ANDALUCIA CONSEJERÍA DE ECONOMÍA Y CONOCIMIENTO

Grid 250 m
Llindar: ? (< 8)

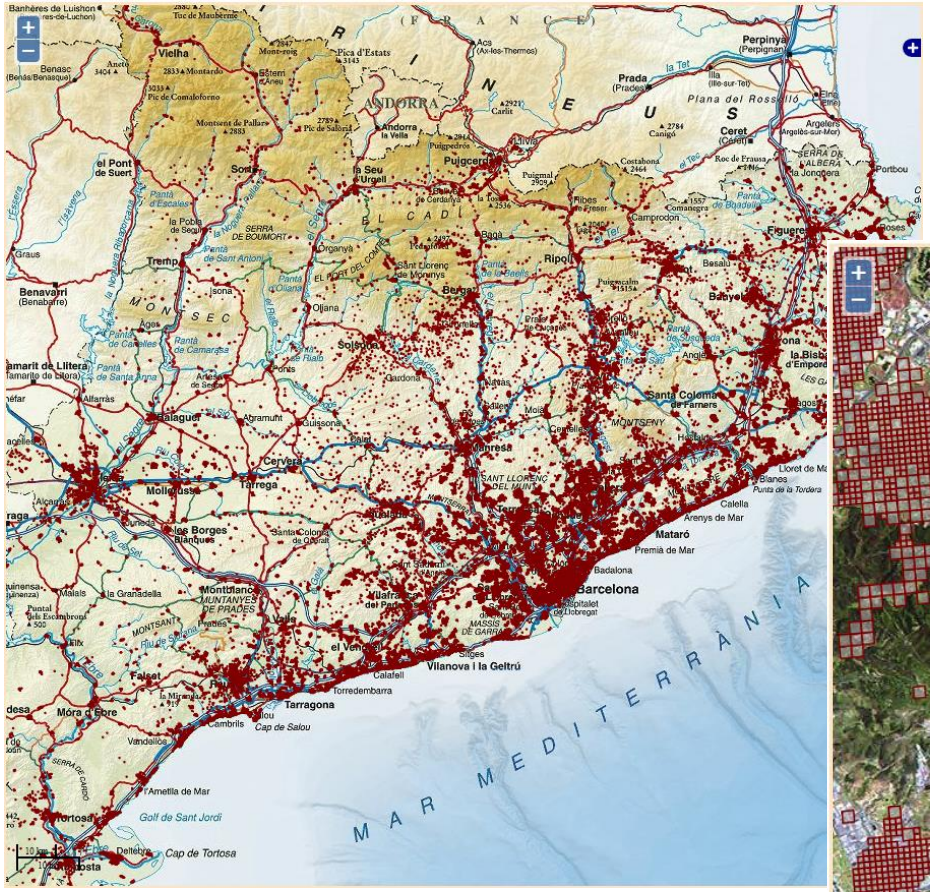


Tab. 1: Spatial grids and threshold values for buildings (B) and population (P)

Country	Census Data	Thresholds B	Thresholds P	Business data
AT	For grid maps, statistical sample sizes (e.g. number of persons with principal residence, number of households) are published without any restriction when using a grid ≥ 125 metres. A classification depending on the attributes of the sample sizes (e.g. civil status of persons at the principal residence) is also published when using grids ≥ 250 metres.	>3	>30	For grid maps, statistical sample sizes are published without any restriction when using a grid ≥ 125 metres. Classifications of attributes of the sample sizes are published without any restriction when using grids ≥ 250 metres.
EST	Grid maps with a minimal grid spacing of 500 metres. Grid cell values, which account for less than three statistical units are blocked. A special code is specified for blocked cells.	≥ 3 cell	>30	no details
FIN	Grid maps with a minimal grid spacing of 250 metres. Attributes for which confidentiality is not necessary (e.g. population, age, gender) and those that must be kept confidential (e.g. education, employment, consumer structures, income, specifications to residential buildings, apartments and households) are strictly differentiated. Grid cell values that account for less than ten statistical units, or grid cells, that represent only one building, are protected by specifying a value of -1.	>1	≥ 10	Grid maps without any lower bound. For company locations geographic coordinates are available. Data of company locations are published in tables or maps for each building. Grid maps of company locations were not published until now, as a precise specification of coordinates is not allowed. Compliance with statistical confidentiality is only necessary for the attributes number of employees and turnover. This is achieved by classification of data.
N	Grid maps with a minimal grid spacing of 250 metres. Special methods are achieved for compliance with statistical confidentiality, when data of less than 20 persons account to a grid cell value.		>20	Grid maps with a minimal grid of 250 metres. Special methods are achieved for compliance with statistical confidentiality, when data of less than 50 employees account to a grid cell value.
CH	Grid maps with a minimal grid spacing of 100 metres. A pre-selection of attributes which are published in grid maps is performed. Sensitive attributes, such as religious affiliation or nationality, are not published. There is no indication of grid cell values, which account for less than four statistical units.		≥ 3 cell	Grid maps with a minimal grid spacing of 100 metres. A pre-selection of attributes which are published in grid maps is performed. Sensitive attributes are not published. There is no indication of grid cell values, which account for less than four statistical units.

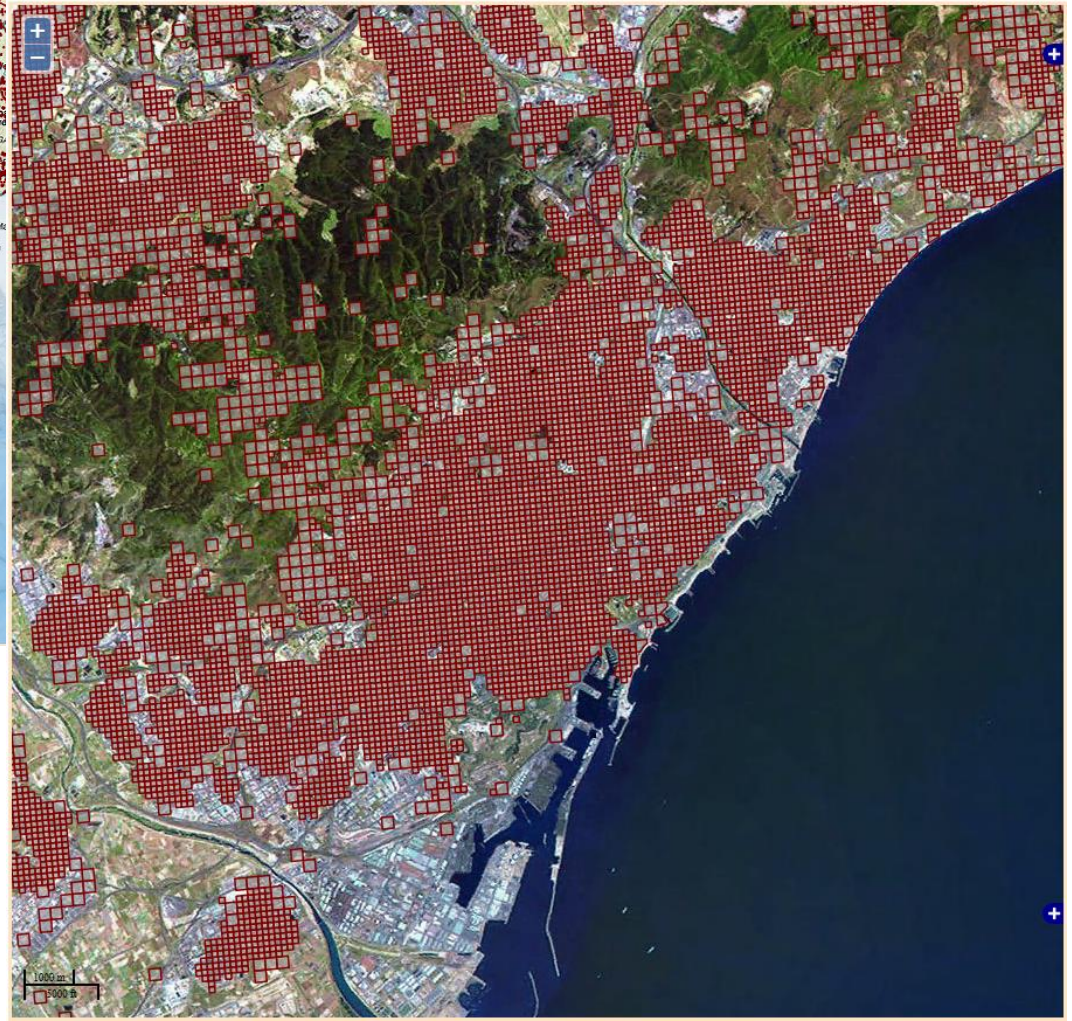
Source: Own contribution referring to SZIBALSKI (2007)

Quadtree. Parámetros de definición.



Quadtree queda definido por

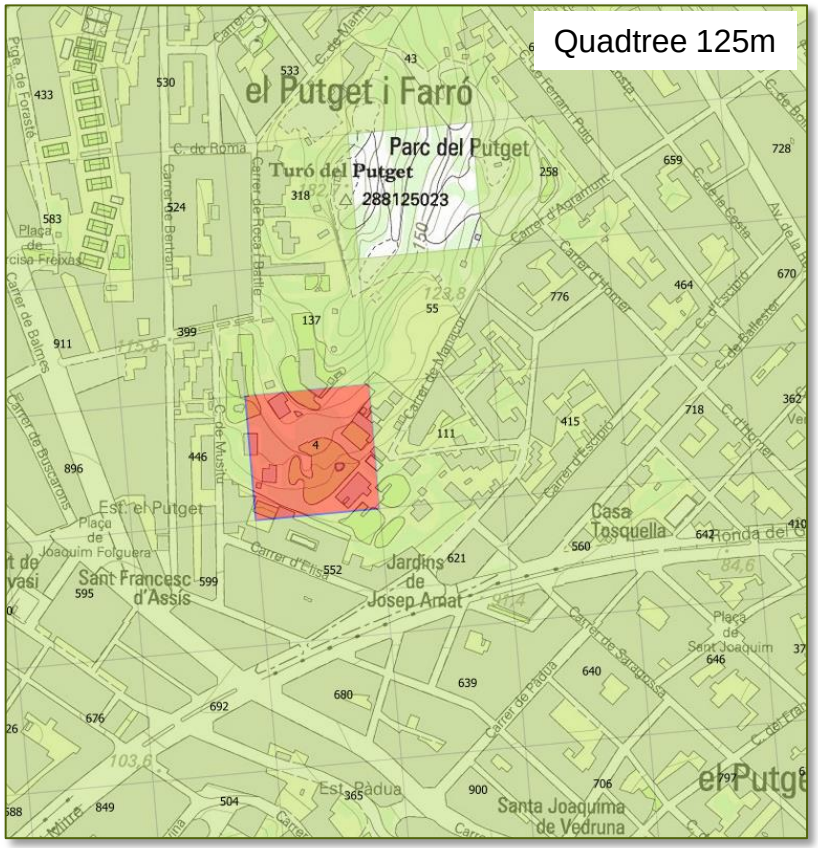
- Límite
- Resolución máxima
- Resolución mínima
- Datos a difundir

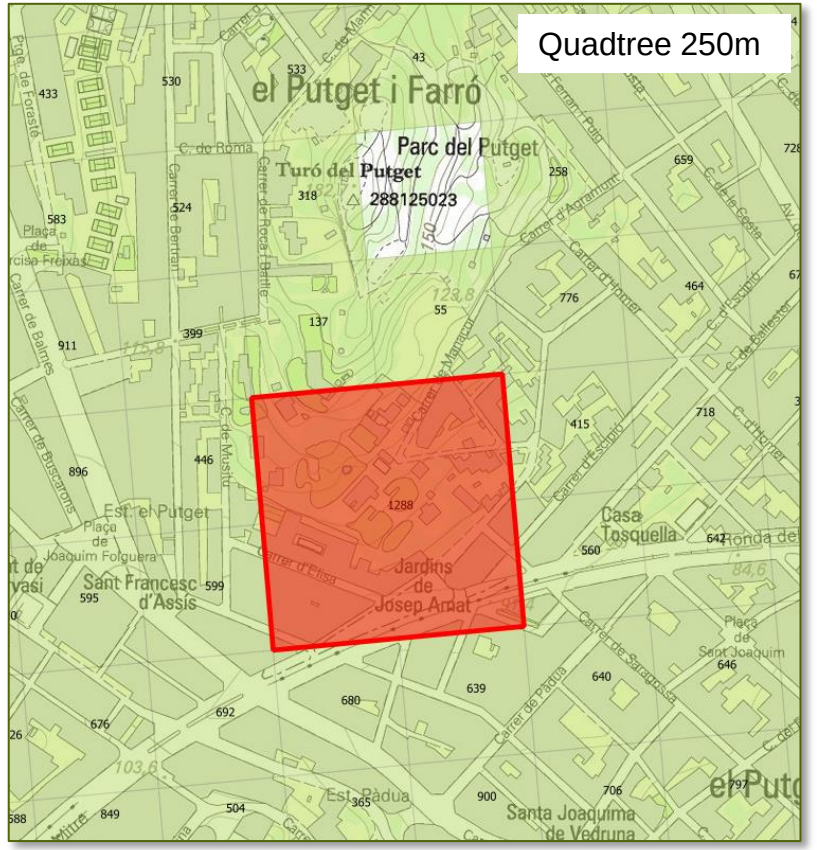


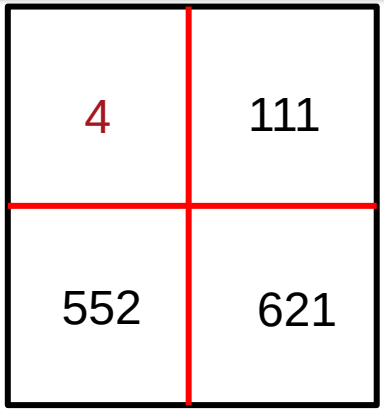
Límite: 17 habitantes
Resolución máxima: 125 m.
Resolución mínima: 250 m.

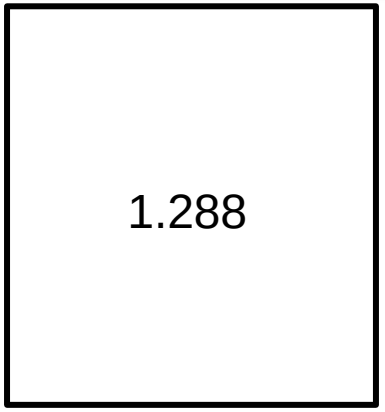
Para celdas a mínima resolución con población por debajo del límite no se da información.

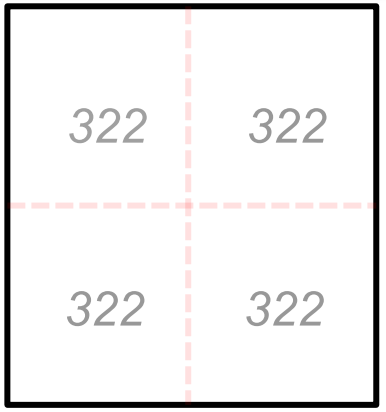
Quadtree. Efecto frontera.





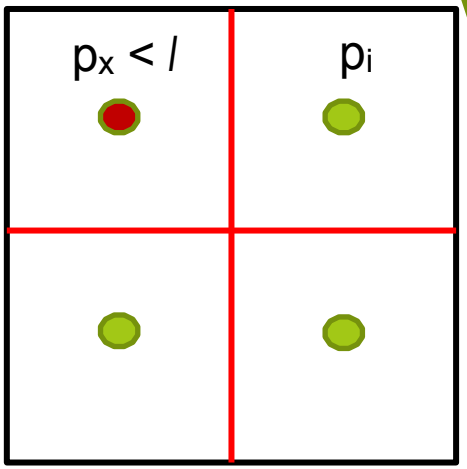


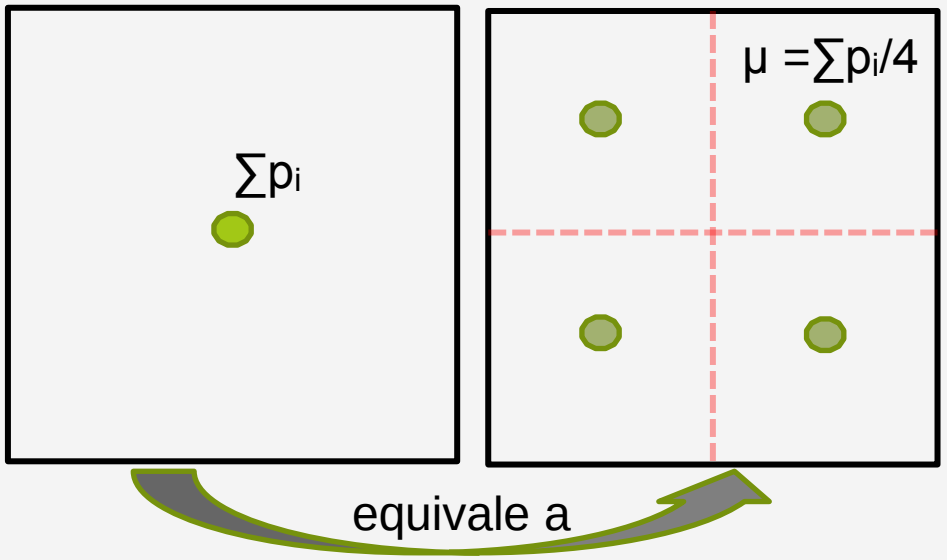




Quadtree. Solución al efecto frontera.

¿Qué ve el usuario con el quadtree ?





Error absoluto

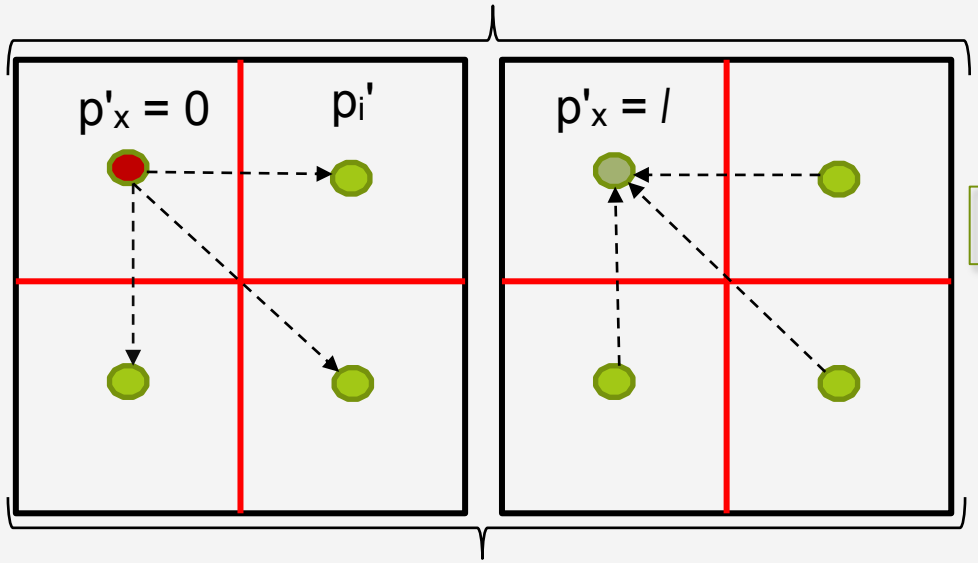
$$\epsilon = \sum | p_i - \mu |$$

Si

$$\epsilon' < \epsilon$$

es mejor desplazar

¿Qué vería el usuario si desplazáramos algunos puntos?

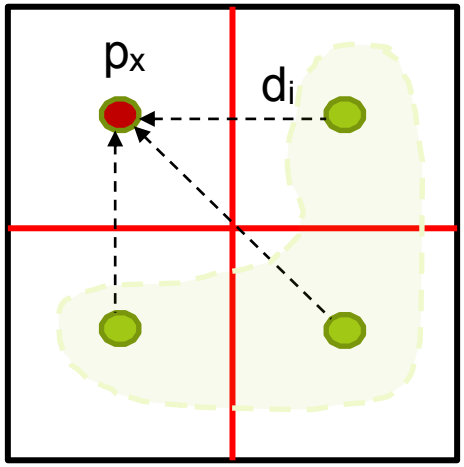


$$\epsilon' = \sum | p_i - p_i' |$$

Error absoluto

Quadtree. Criterios de aplicación.

- Aplicar cuando Є > 2*N (N : número de elementos trasladados)
- Aplicar siempre añadiendo hasta llegar al límite l
- Repartir los movimientos según las frecuencias relativas del subconjunto de donantes

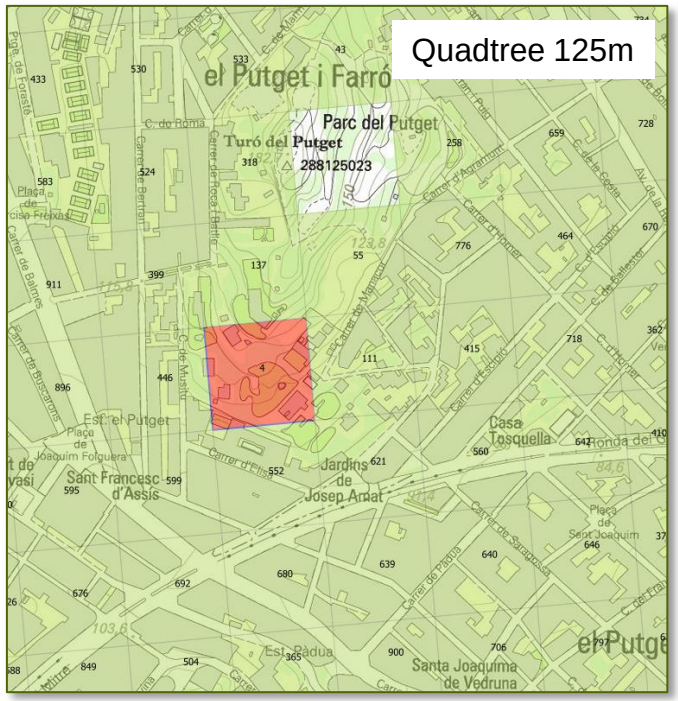


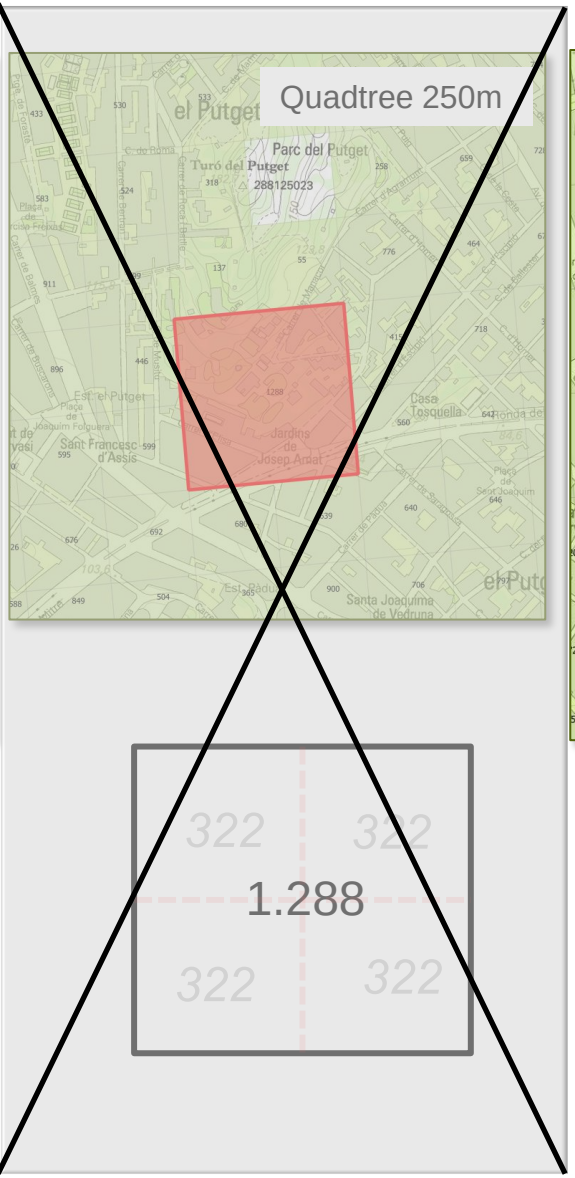
$$\sum d_i = \sum l - p_x$$

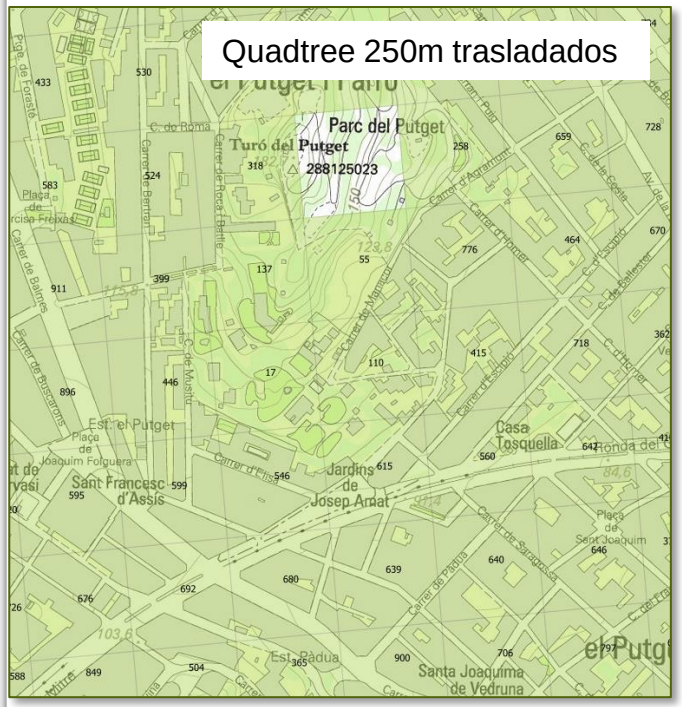
$$d_i = |l - p_x| * (p_i / \sum p_i) \quad \text{para todo i tal que } p_i - d_i \geq l$$

- Escoger aleatoriamente los elementos a mover del subconjunto de donantes

Quadtrees. Ejemplo de solución al efecto frontera.







4	111
552	621

17	110
546	615

Aplicabilidad de la solución. RP2014.

Casos **group by id_pare** paso de 125m a 250m según numero de cuadrados per debajo o por encima del límite

	Casos	n quad. <>0	n < 17	n >= 17	Є > 2N	% amb Є > 2N
	12017	1	1	0		
	2401	2	1	1	1425	59,35
	3313	2	2	0		
	1497	3	1	2	1444	96,45
	1262	3	2	1	433	34,31
	1217	3	3	0		
	1908	4	1	3	1872	98,11
	1336	4	2	2	936	70,06
	900	4	3	1	122	13,55
	569	4	4	0		
Suma posibles	9304				6232	66,98
Total	26420					23,59

Construcción del quadtree con translaciones.

pk, sexo, edad,……p (POINT SRID 25831) id,level (3)

count(*) …. group by id

f(sexe… edat..) id,level (3)

Calcular per a cada id a nivell superior

n_q bigint	f_inf bigint	f_sup bigint	id_pare integer	level integer	allinfoquadtree double precision[]	matriu_translacio integer[]
4	1	3	416022	3	{1058,264.5,13,552,621,4,111,5,6,13,14,2210419,2210420,2210427,2210428}	{-6,-6,13,-1,0}

pk…… id’

left outer join

Trasladar casos aleatoriamente

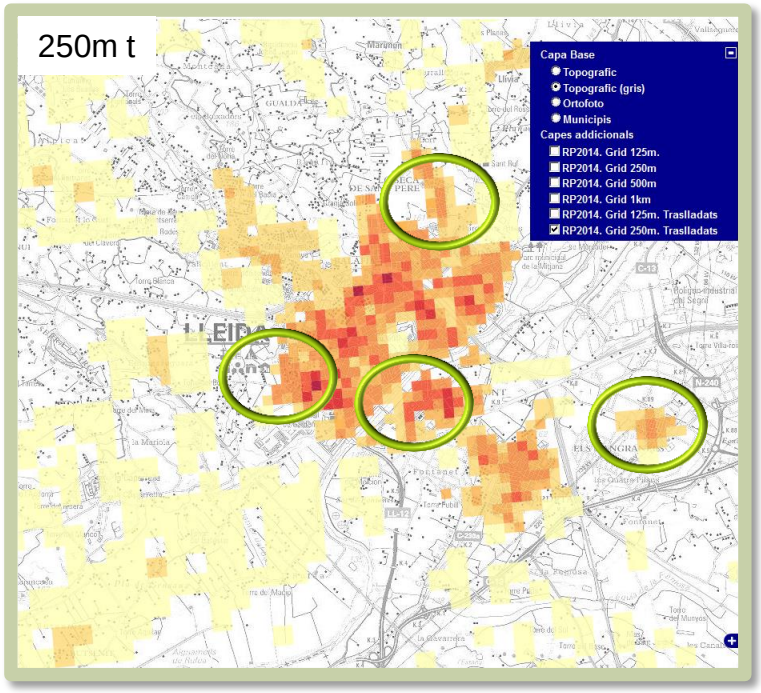
count(*) …. group by id’

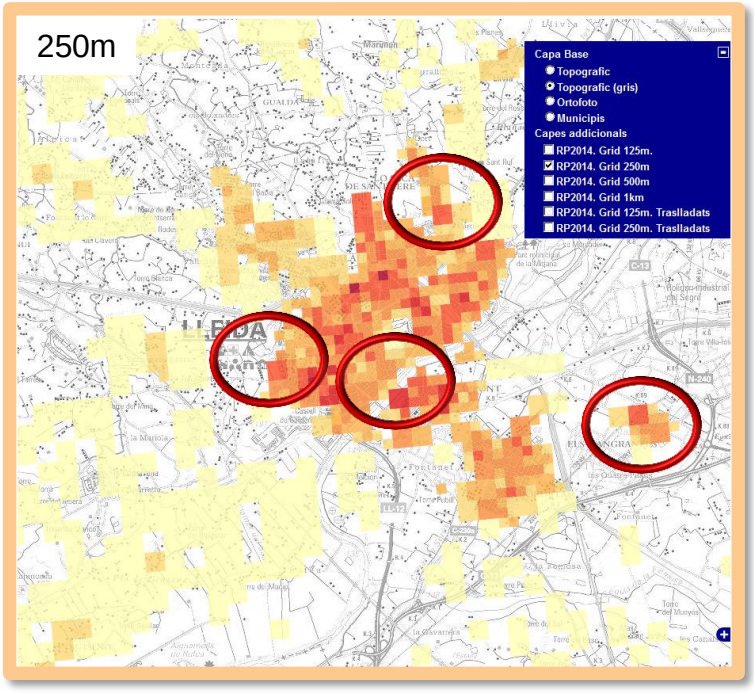
Construir el quadtree

f(sexo… edad..) id’,level 3

pk…… new_id

Resultados. RP2014.





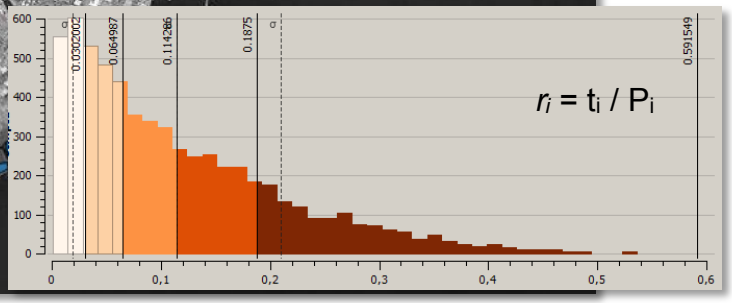
Resolución 250m	Trasladados		Sin translación		% (población)		con población < 17				con población ≥ 17			
							Trasladados		Sin traslación		Trasladados		Sin traslación	
Cuadrados de	F	Población	F	Población	Trasladados	Sin Translación	F	Población	F	Pob.	F	Población	F	Población
250 m	20.263	266.639	26.420	1.208.272	3,52	15,97	15.071	87.667	15.071	87.667	5.192	178.972	11.349	1.120.605
125m	51.744	7.299.825	31.769	6.358.192	96,48	84,03	0	0	0	0	51.744	7.299.825	31.769	6.358.192
Total	72.007	7.566.464	58.189	7.566.464	100	100	15.071	87.667	15.071	87.667	56.936	7.478.797	43.118	7.478.797

Total habitantes trasladados: 64.056 (0,85%)

Total habitantes con posición imputada: 334.256 (4,42%)

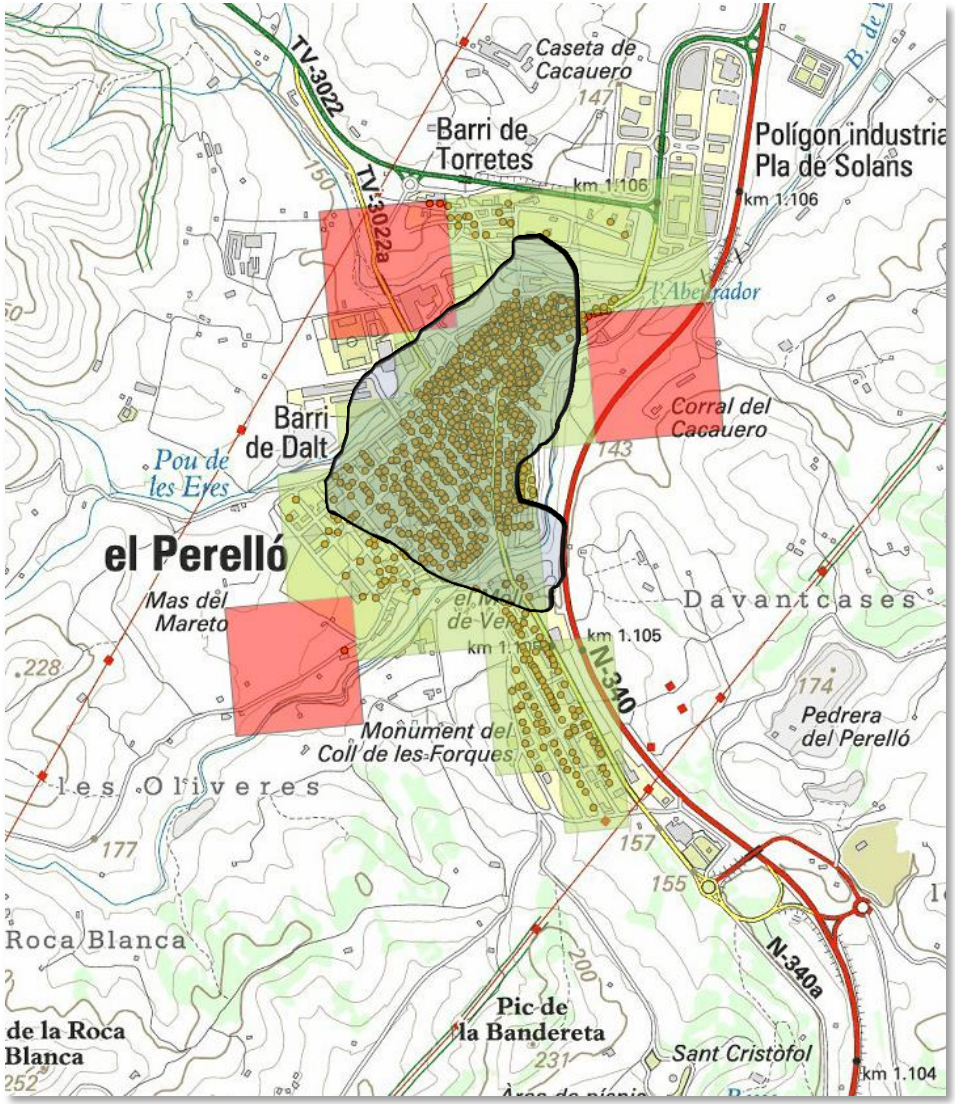
Quadtrees. RP2014. Ratio traslaciones.

n_q bigint	f_inf bigint	f_sup bigint	id_pare integer	level integer	allinfoquadtree double precision[]	matriu_translacio integer[]
4	3	1	426153	3	{88.5,22.125,42,2,3,4,62,53,54,61,62,2270814,2270815,2270822,2270823}	{15,14,13,-42,0}



Quadtree. Estimación de errores.

Partiendo de la capa original de puntos y los diferentes quadtrees, utilizaremos zonas per calcular poblaciones según ① i ② i calcular los errores relativos.



① p_x =número de puntos dentro de la geometría S_x

② $p_x = \sum p_i * ST\_AREA(Q_i \cap S_x) / ST\_AREA (Q_i)$

Para todas les secciones del año 2010 (5.006):

	Distribución de los errores relativos (%) de los quadtrees			
Percentil	**125**	**125t**	**250**	**250t**
10	0,22	0,25	0,40	0,35
20	1,33	1,39	1,86	1,60
30	3,13	3,21	3,77	3,32
40	5,23	5,33	6,06	5,42
50	7,62	7,69	8,62	7,74
60	10,73	10,77	11,73	10,84
70	14,44	14,52	15,61	14,54
80	20,02	20,05	21,13	20,03
90	29,40	29,5	31,69	29,50
100	5.093	5.093	5.568	5.094

Quadtree	**Media error relativo (%)**	**Desviación típica**
hasta 125	12,66	17,28
hasta 250	13,73	18,24
hasta 125 trasladados	12,74	17,33
hasta 250 trasladados	13,05	17,27

Conclusiones.

- La utilización de quadtrees como base espacial de agregación para preservar el secreto estadístico es un método reconocido i ampliamente utilizado.

- No obstante este método puede producir agregaciones no deseables en zonas de frontera, lugares en donde la población decae abruptamente.

- En los casos en que el error absoluto en les translaciones sea más pequeño que el error absoluto inherente a la agregación, se ha procedido a perturbar las posiciones evitando así las agregaciones.

- La mediana de los errores relativos al calcular poblaciones dentro de polígonos crece a medida que la resolución del quadtree baja. Con los criterios expuestos, las traslaciones de población entre hermanos de la jerarquía puede evitar la agregación espacial, disminuyendo el error relativo en los cálculos de población.