Statistical disclosure control on visualising geocoded population data using a structure in quadtrees

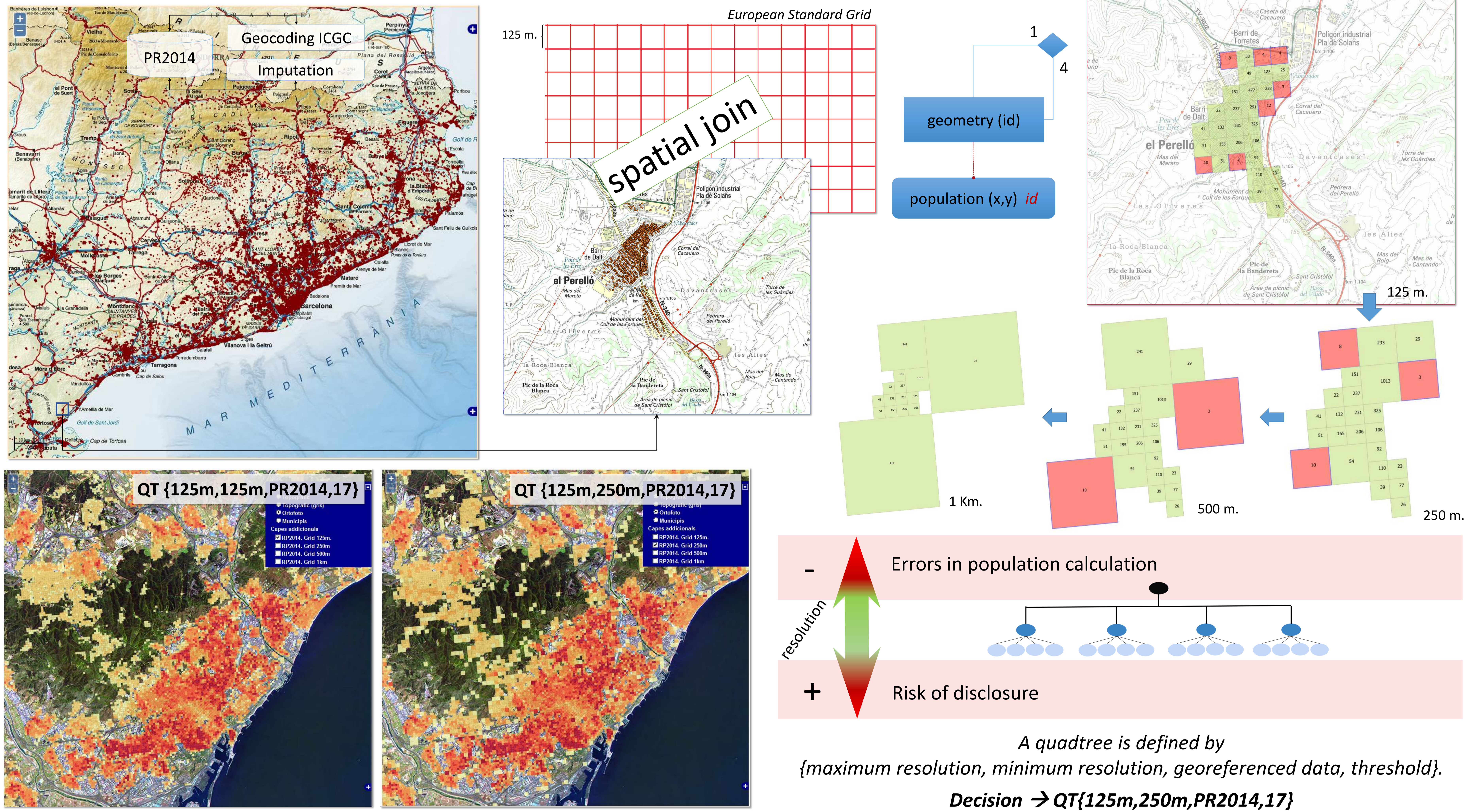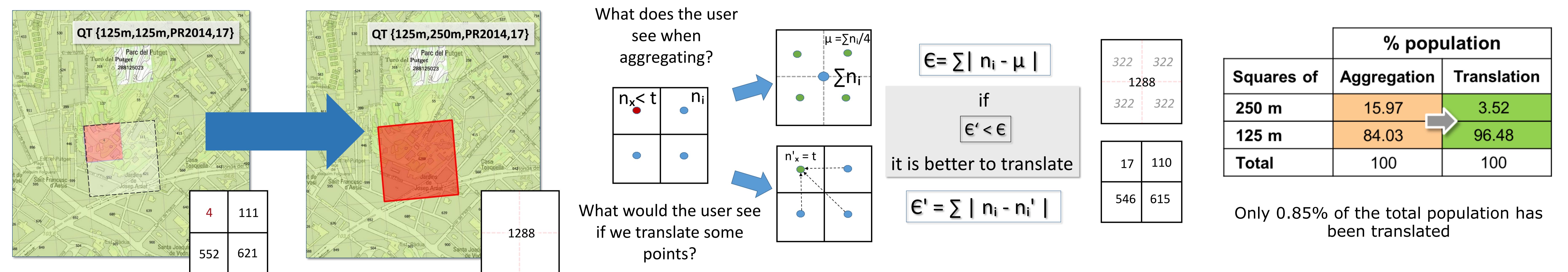Eduard Suñé, Cristina Rovira, Daniel Ibáñez, Mireia Farré

- The use of **quadtrees** for the dissemination of georeferenced data is a good method for the preservation of statistical confidentiality, as a certain balance between security and accuracy is achieved.
- This preservation method may lead to undesirable aggregations in areas which correspond to siblings in the hierarchy, due to the high values of population variance (**border effect**). A solution to the border effect consists of translating microdata under the condition that the absolute error of the aggregation is greater than that of the translation.
- **Monte Carlo** techniques allow the estimation of the relative error distribution for the population calculated within the quadtree structure **QT{125m,250m,PR2014,17,t}**. We have obtained a value of 5.3% for the median of these errors.

➢ Disclosure Control by Spatial Aggregation Using Quadtrees

European Standard Grid

125 m.

spatial join

PR2014 — Geocoding ICGC — Imputation

el Perelló

geometry (id)

population (x,y) *id*

125 m.

1 Km.   500 m.   250 m.

QT {125m,125m,PR2014,17}   QT {125m,250m,PR2014,17}

– Errors in population calculation

resolution

+ Risk of disclosure

A quadtree is defined by {maximum resolution, minimum resolution, georeferenced data, threshold}.

Decision → QT{125m,250m,PR2014,17}

➢ Border Effect. Proposed Solution

QT {125m,125m,PR2014,17}   QT {125m,250m,PR2014,17}

| 4 | 111 |
| 552 | 621 |

| 1288 |

What does the user see when aggregating?

$\mu = \sum n_i / 4$

$\sum n_i$

$n_x < t$ n_i

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

if $\epsilon' < \epsilon$ it is better to translate

$n'_x = t$

What would the user see if we translate some points?

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

322	322
	1288
322	322

| 17 | 110 |
| 546 | 615 |

	% population	
Squares of	Aggregation	Translation
250 m	15.97	3.52
125 m	84.03	96.48
Total	100	100

Only 0.85% of the total population has been translated

➢ Estimation of Errors. Monte Carlo Experiment

= 50,000 random polygons

For each polygon S_i, relative error is

$$\varepsilon_i = \frac{|n'_i - n_i|}{n_i}$$

n_i = Population within the X geometry S_i

$n'_i = \sum n_r * AREA(Q_r \cap S_i) / AREA(Q_r)$

	Quartile 1	Median	Quartile 3	Mean
QT{125m, 250m, PR2014, 17, t}	0.02	0.05	0.19	0.28
QT{125m, 250m, PR2014, 17}	0.02	0.07	0.22	0.33
QT{125m, 125m, PR2014, 17}	0.01	0.04	0.14	0.23

Group	Area	Average sqrt (area)	Cases
1	< 3906.25	43.201	906
2	3906.25- 15625	93.300	1848
3	15625 - 62500	188.543	3955
4	62500 - 250000	375.363	8517
5	250000 - 1000000	741.249	16242
6	1000000-4000000	1319.950	13407
7	> 4000000	2128.030	245
All		762.151	45120

QT {125m,250m,PR2014,17,t}
QT {125m,250m,PR2014,17}
QT {125m,125m,PR2014,17}

ϵ

QT {125m,125m,PR2014,17,t}

$f\left(\frac{n'_i - n_i}{n_i}\right)$