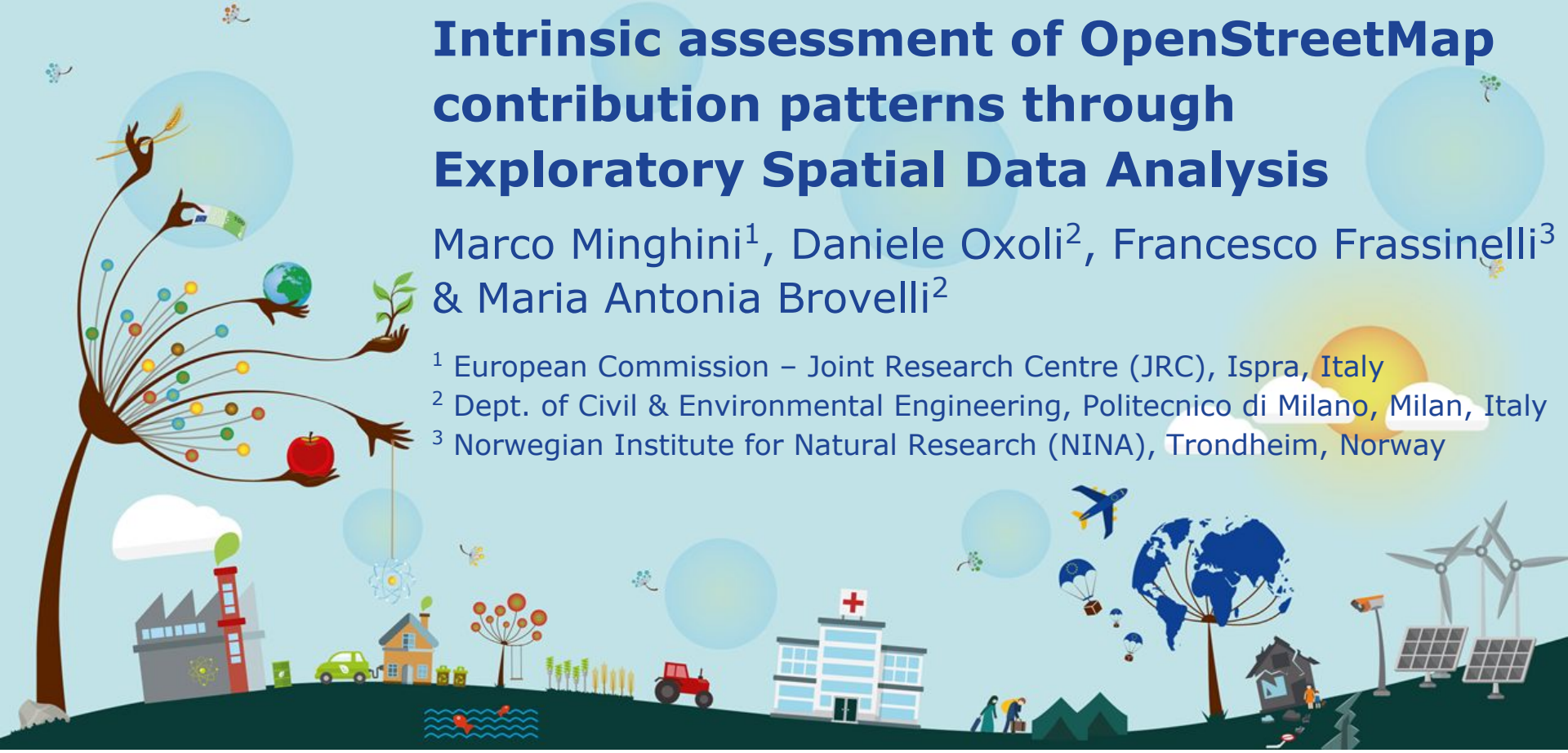
Intrinsic assessment of OpenStreetMap contribution patterns through Exploratory Spatial Data Analysis

Marco Minghini[1], Daniele Oxoli[2], Francesco Frassinelli[3] & Maria Antonia Brovelli[2]

[1] European Commission – Joint Research Centre (JRC), Ispra, Italy
[2] Dept. of Civil & Environmental Engineering, Politecnico di Milano, Milan, Italy
[3] Norwegian Institute for Natural Research (NINA), Trondheim, Norway



European Commission

Introduction

- Availability of the full history of OpenStreetMap (OSM) data
  - intrinsic analyses to assess:
    - OSM quality
    - OSM spatio-temporal evolution
    - OSM contribution patterns
  - focused on several history-based variables:
    - origin, history & profiling of contributors
    - number of contributors
    - number of object versions
    - location, amount, nature and frequency of edits

Purpose of the work

- Apply Exploratory Spatial Data Analysis (ESDA) techniques to study OSM contribution patterns.

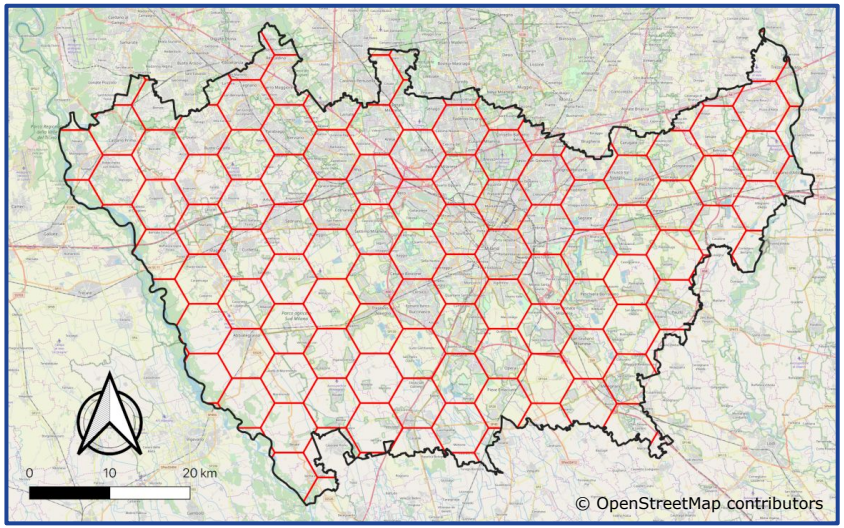
Purpose of the work

- Apply Exploratory Spatial Data Analysis (ESDA) techniques to study OSM contribution patterns.

- ESDA is a framework:
  - based on statistical methods, visualisation techniques & software tools
  - aimed at identifying spatial patterns/trends & discover spatial relations characterising geospatial datasets
    - spatial association: degree of similarity between neighbouring observations in the geographical space
    - Local Indicators of Spatial Association (LISA)

Methodology

- Assessment of spatial association (clusters & outliers):
  - in Milan, Northern Italy
    - total area ≈ 1.500 km^2
    - sampled through a regular hexagonal grid
    - hexagon side: 500m, 1000m, 2500m



European Commission

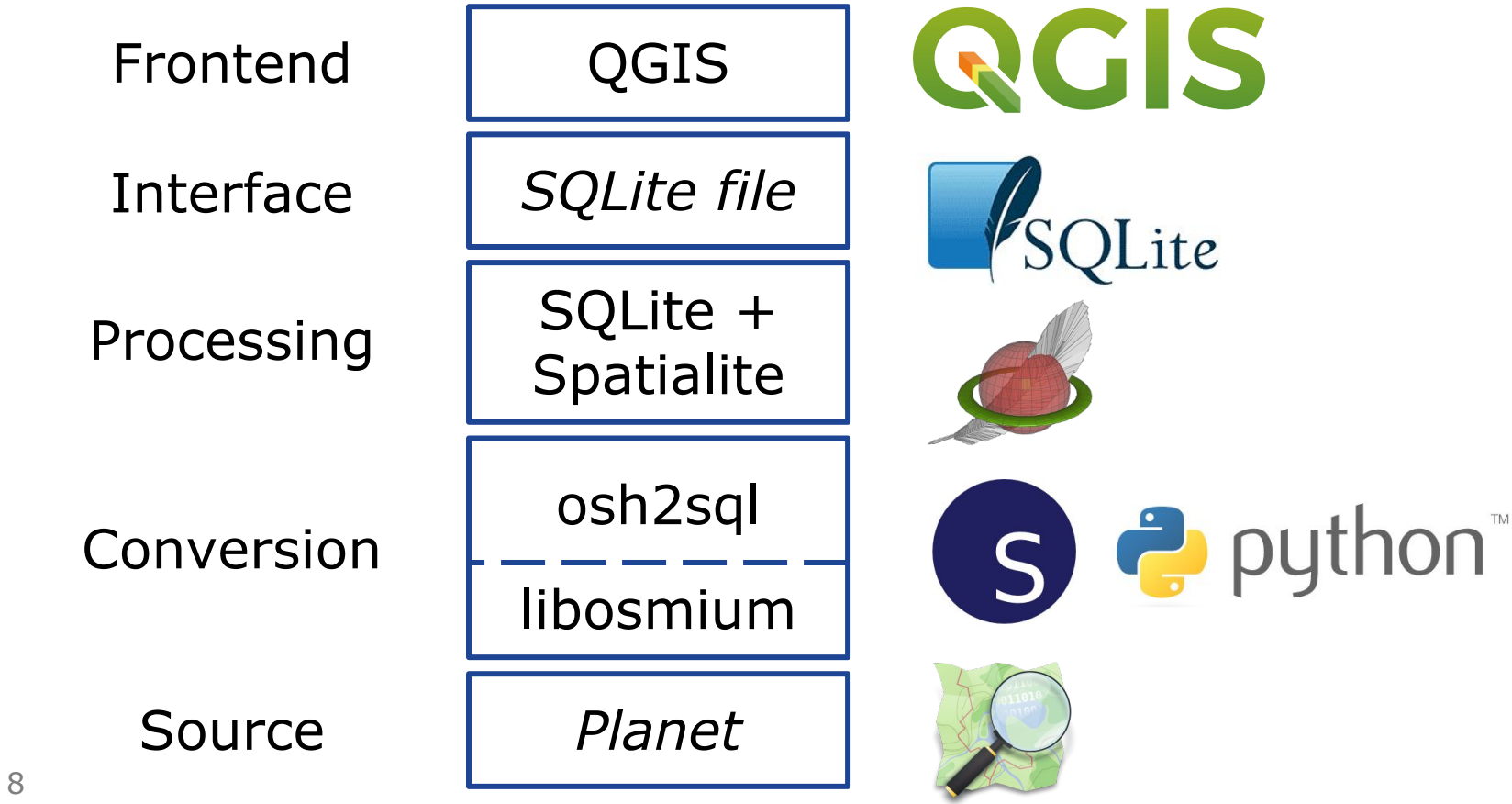
Methodology

- Assessment of spatial association (clusters & outliers):
  - in Milan, Northern Italy
  - considering only the history of OSM nodes, with some hypotheses
    - only nodes currently existing in the OSM database are considered
    - only nodes with at least one tag are considered
    - a new version of a node is counted only when there is a change in tags (not in geometry)

Methodology

- Assessment of spatial association (clusters & outliers):
  - in Milan, Northern Italy
  - considering only the history of OSM nodes, with some hypotheses
  - between the following history-related variables
    - total number of different contributors who edited OSM nodes
    - average number of different contributors who edited each OSM node
    - average date of creation of OSM nodes
    - average date of last edit of OSM nodes
    - average number of versions of OSM nodes
    - average frequency of update of OSM nodes

Software architecture



LISA – Univariate analysis

- Measures the local spatial association for 1 single variable.
- Based on Local Moran's I (Anselin, 1995):

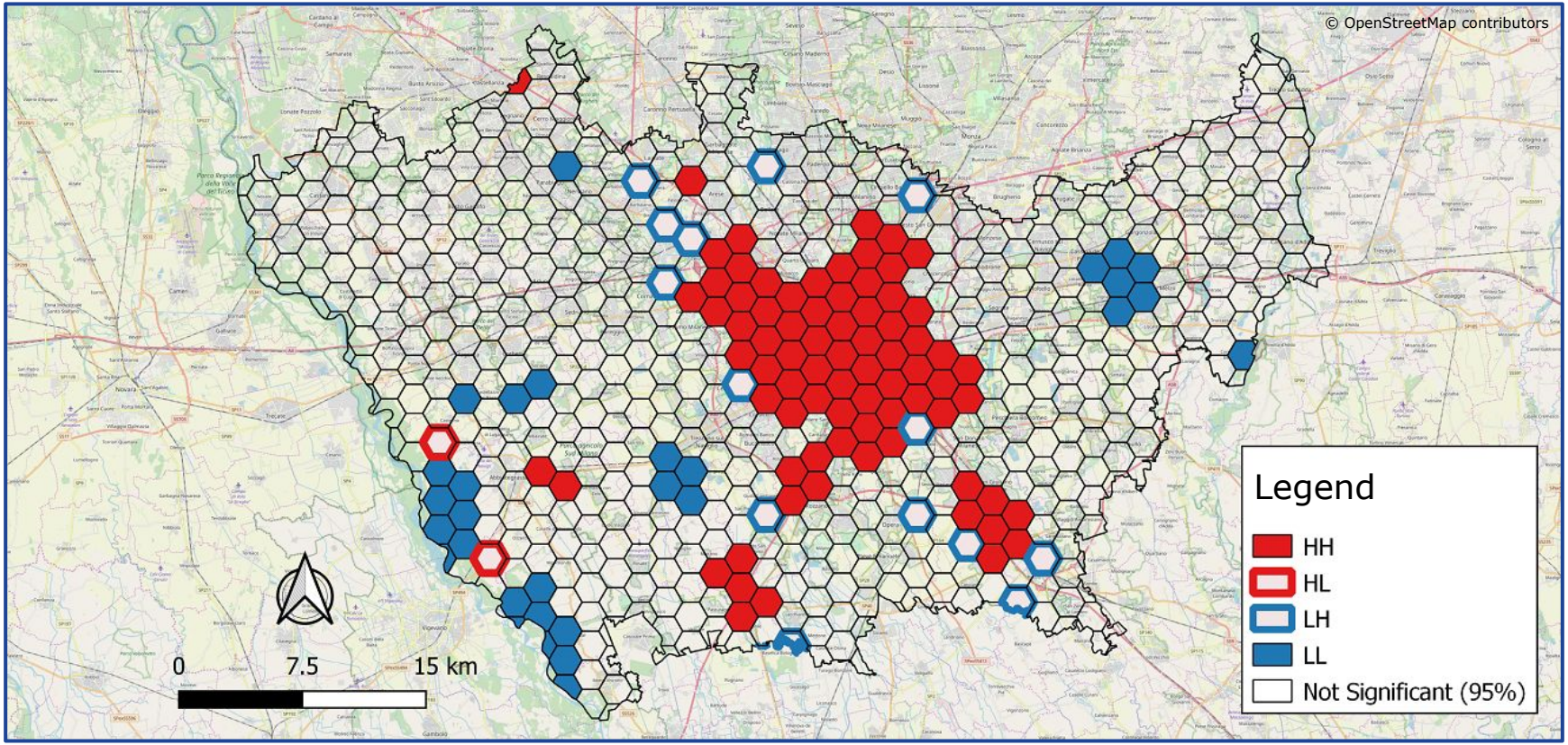
$$I_i = z_i \sum_{j=1}^{n} W_{i,j} z_j \; ; \; i \neq j$$

  - n = number of locations in the dataset
  - z_i, z_j = standardized observation values at locations i and j
  - $W_{i,j}$ = spatial weights matrix

- High values of $I_i \rightarrow$ clusters:
  - high values surrounded by high values (HH) and low values surrounded by low values (LL)
- Low values of $I_i \rightarrow$ outliers:
  - high values surrounded by low values (HL) and low values surrounded by high values (LH)
- Implemented in the QGIS Hotspot Analysis plugin.

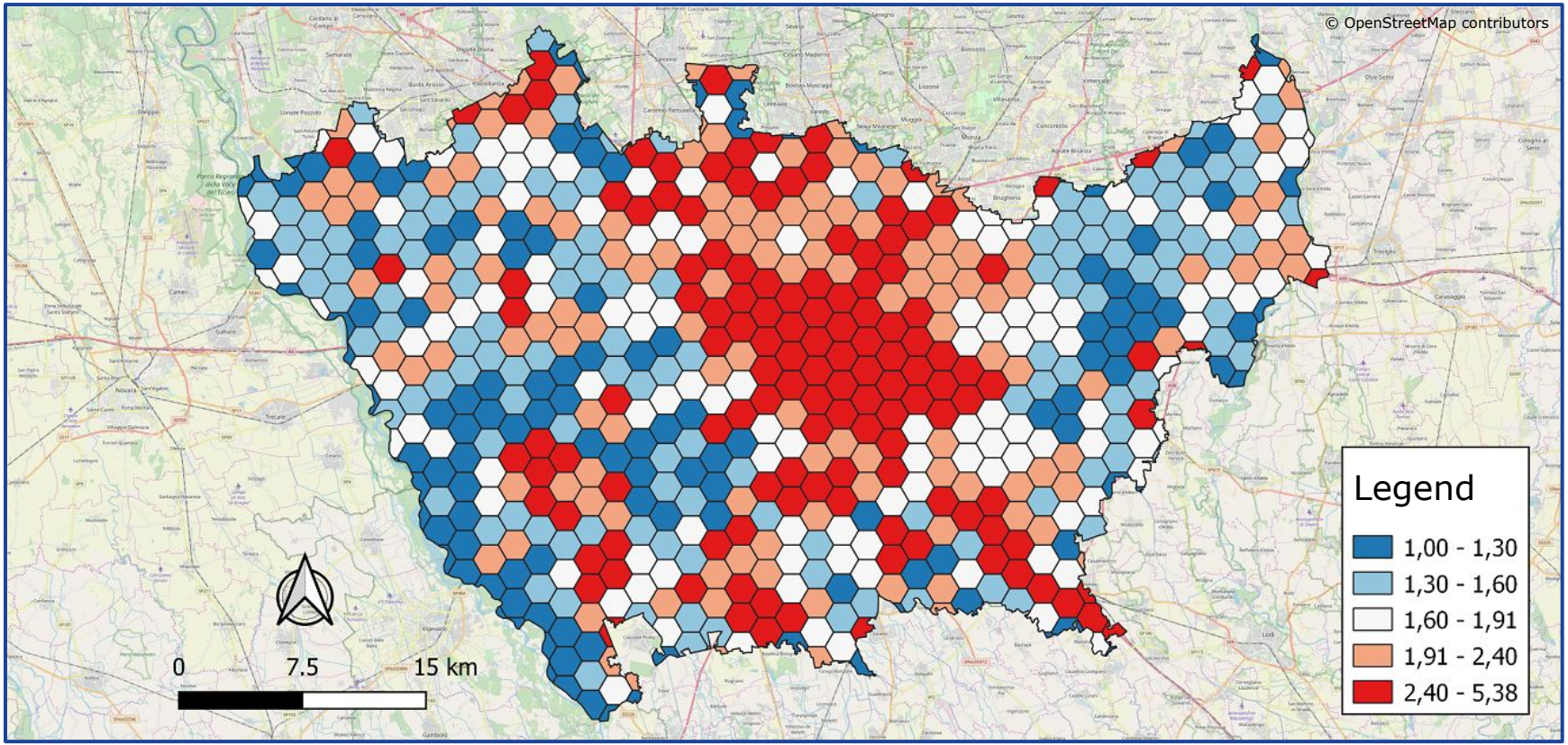
LISA – Univariate analysis [hexagon side: 1000m]

- Average number of different contributors who edited each OSM node



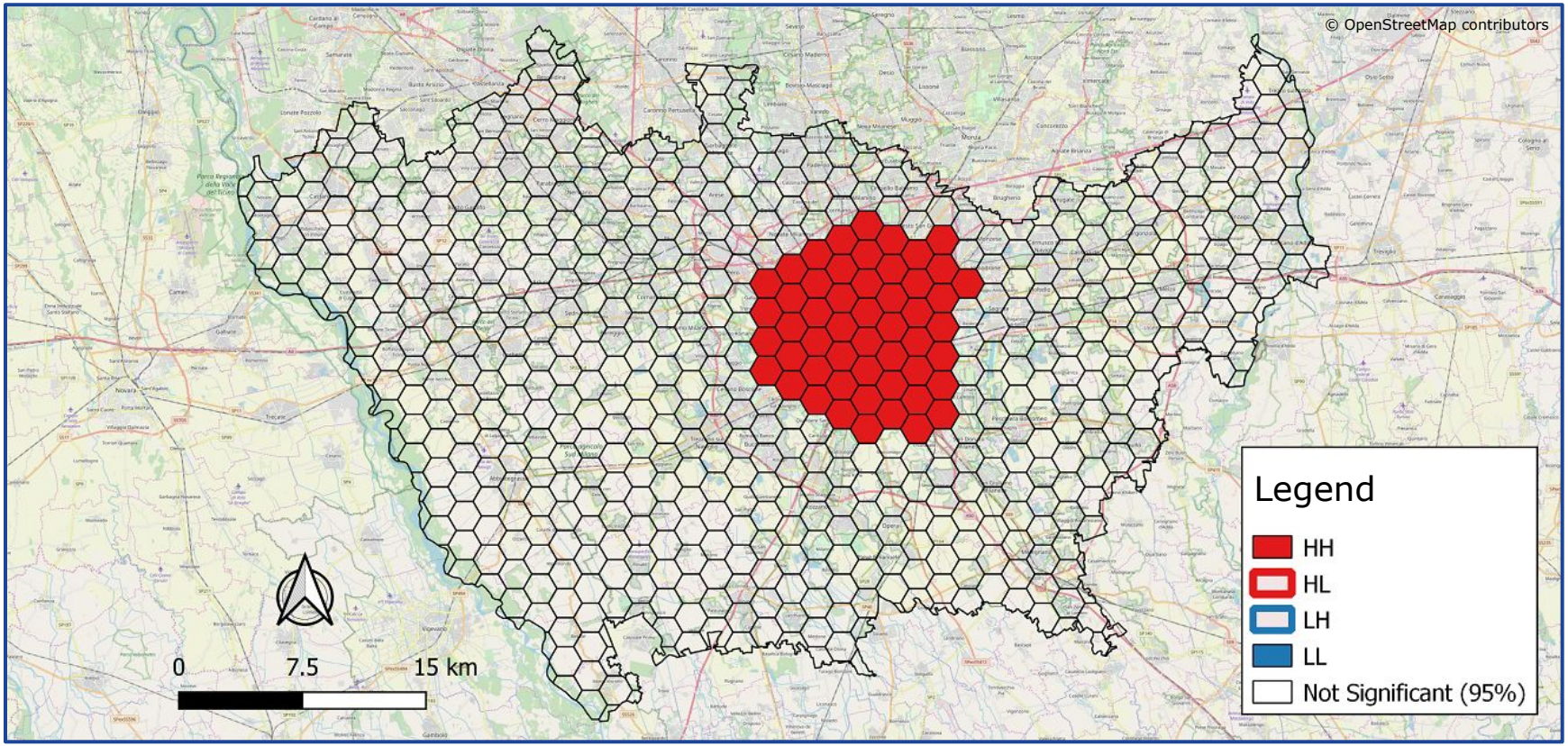
Quantile analysis [hexagon side: 1000m]

- Average number of different contributors who edited each OSM node



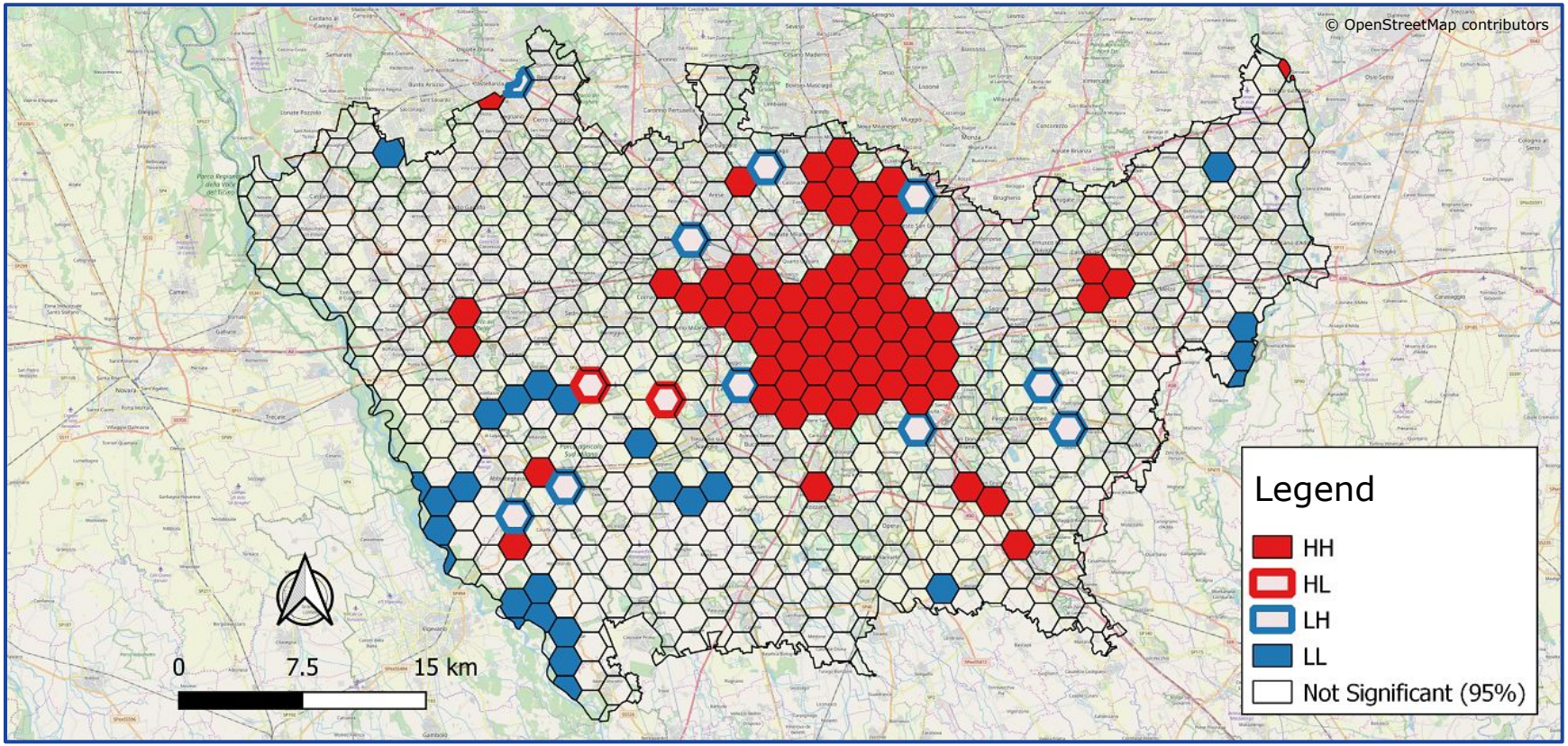
LISA – Univariate analysis [hexagon side: 1000m]

- Total number of different contributors who edited the OSM nodes



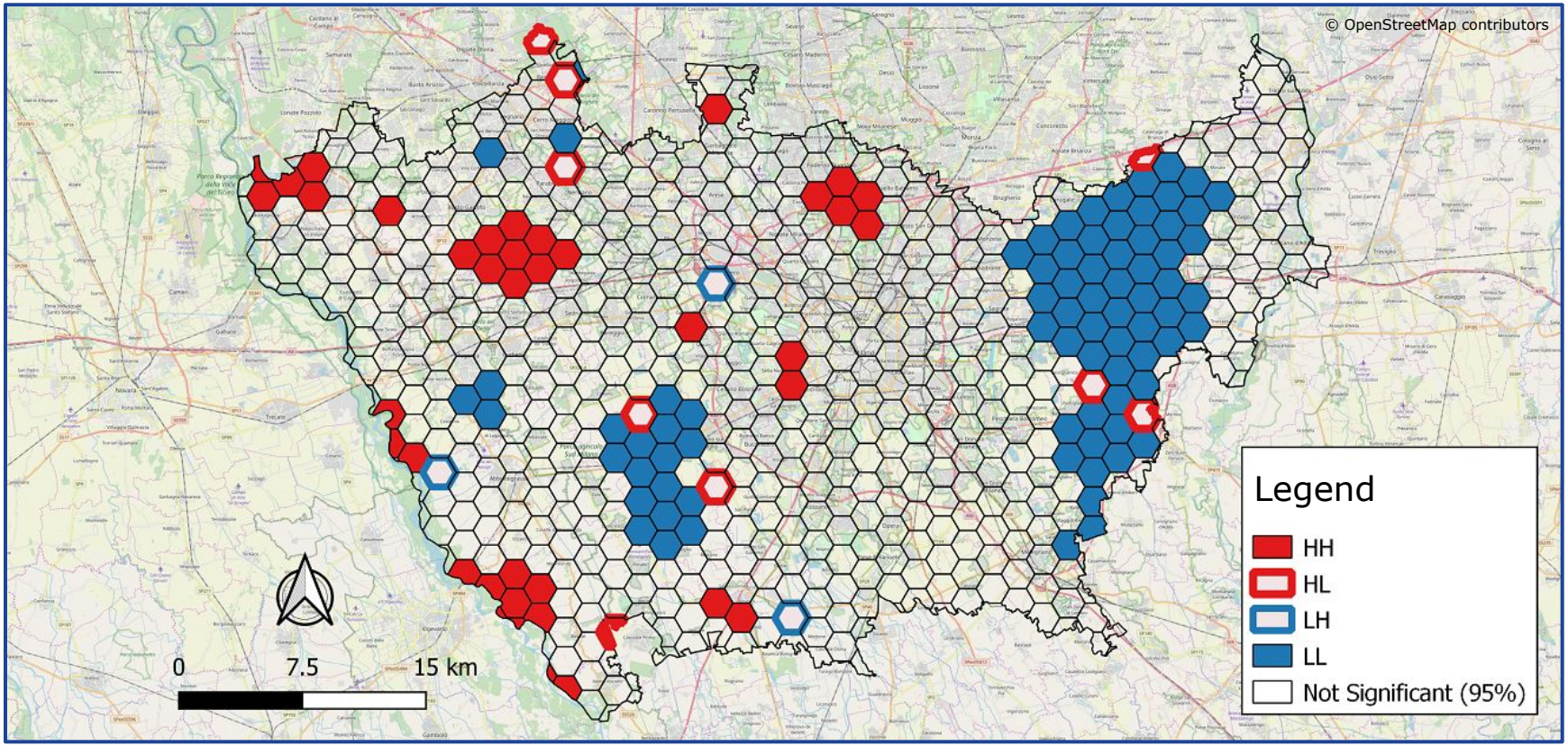
LISA – Univariate analysis [hexagon side: 1000m]

- Average number of versions of OSM nodes



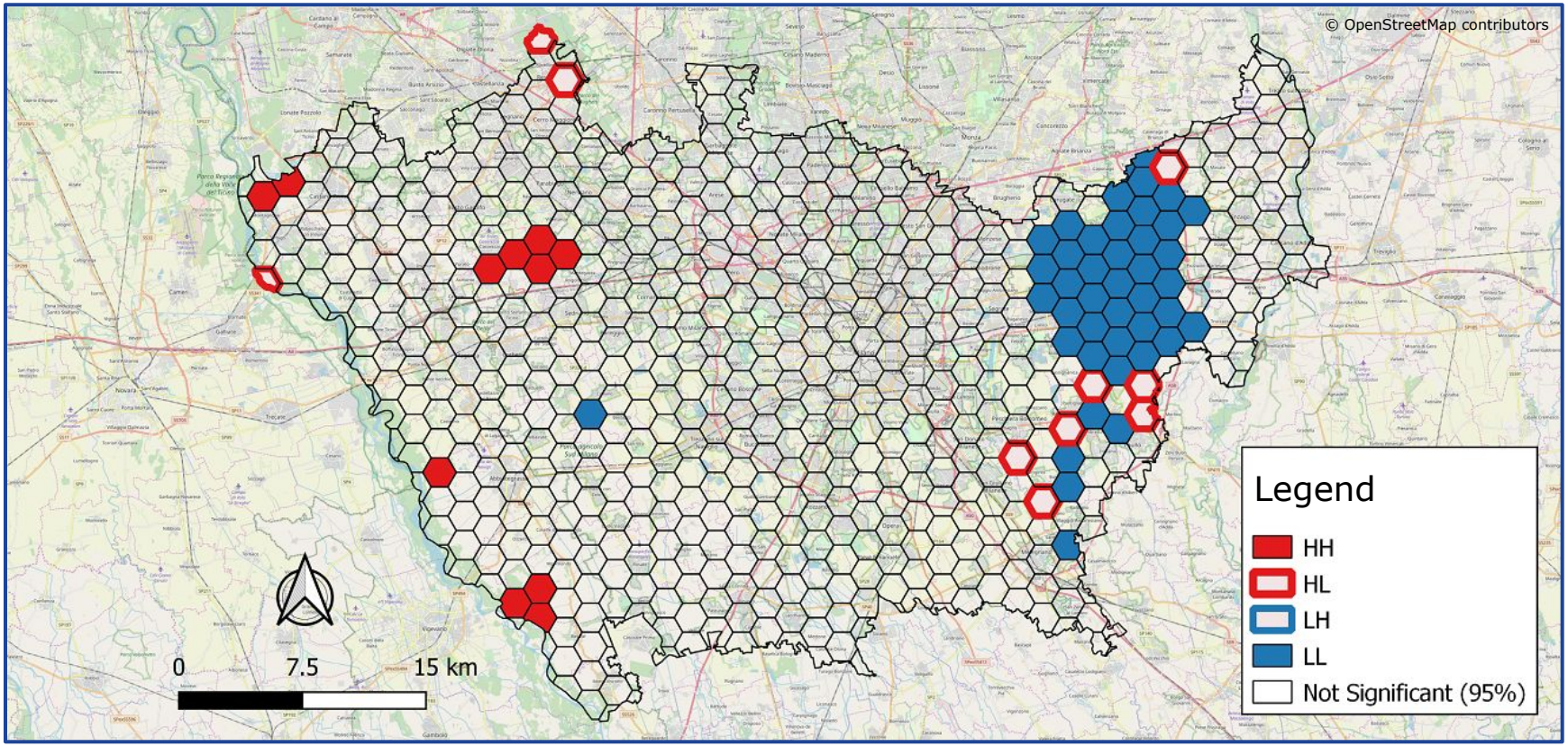
LISA – Univariate analysis [hexagon side: 1000m]

- Average date of creation of OSM nodes



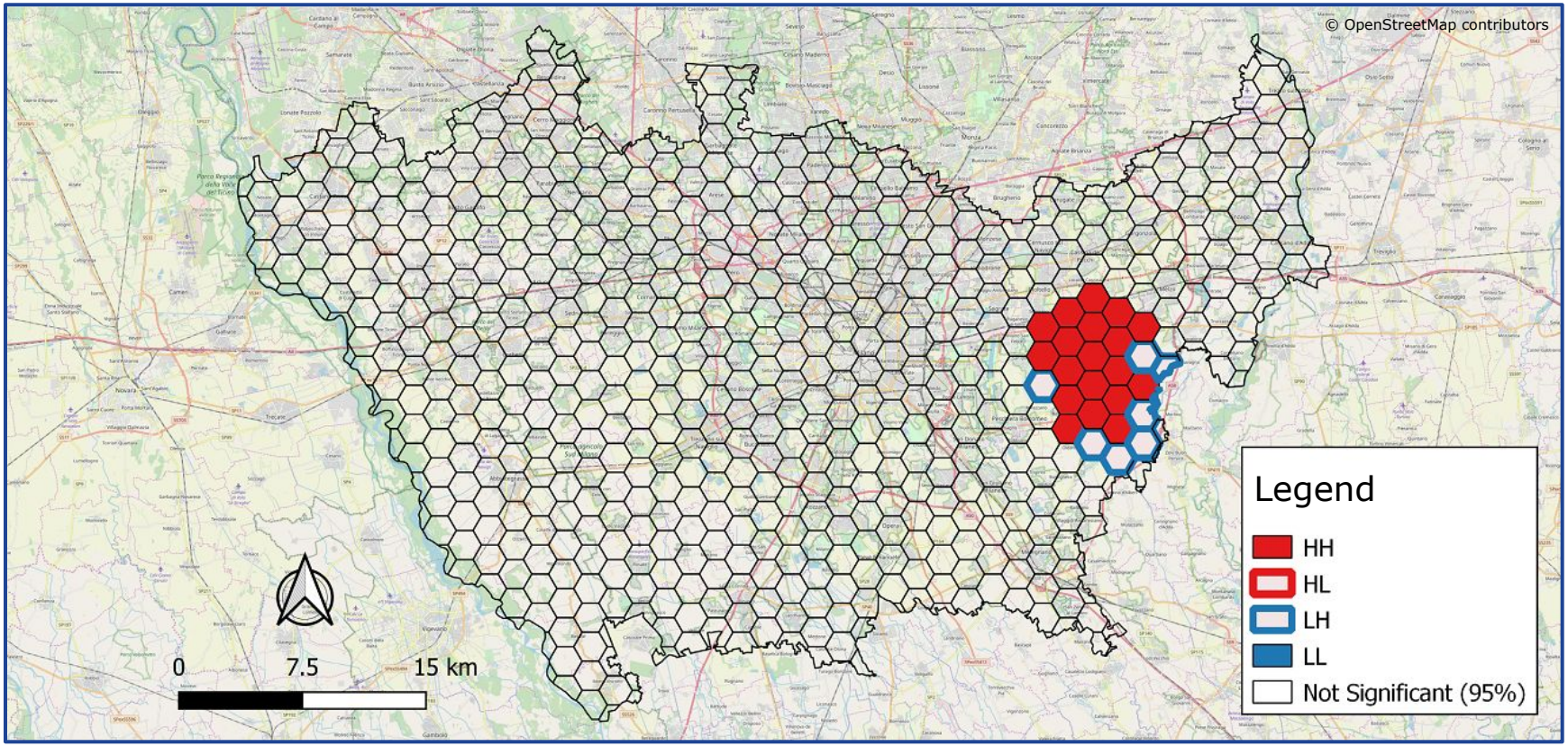
LISA – Univariate analysis [hexagon side: 1000m]

- Average date of last edit of OSM nodes



LISA – Univariate analysis [hexagon side: 1000m]

- Average frequency of update of OSM nodes



LISA – Bivariate analysis

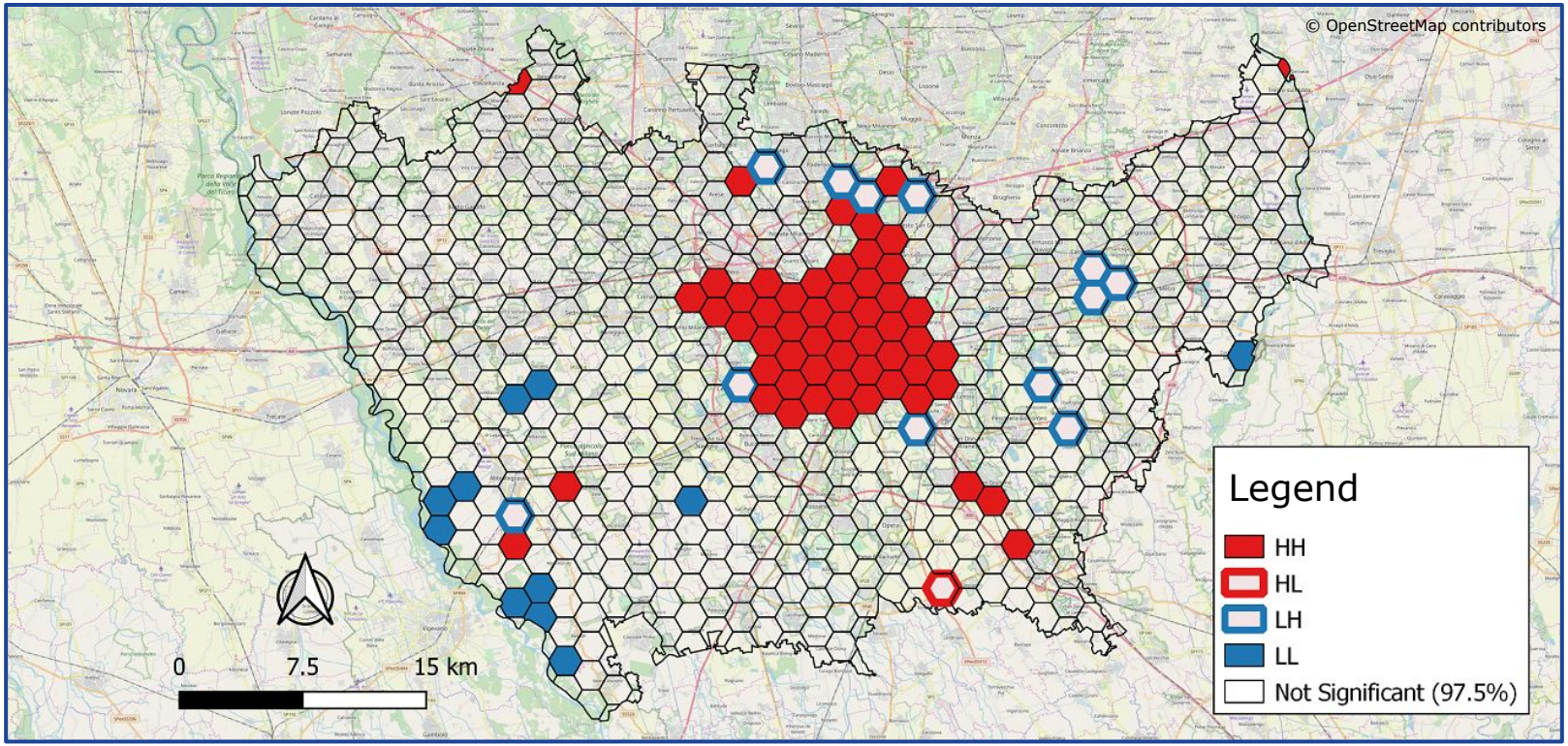
- Measures the local spatial association for 2 variables (k and l)
- Based on Local Moran's I (Anselin, 1995):

$$I^i_{k,l} = z^i_k \sum_{j=1}^{n} W_{i,j} z^j_l \; ; \; i \neq j$$

  - n = number of locations in the dataset
  - z^i_k = standardized observation value of k at location i
  - z^j_l = standardized observation values of l at locations j
  - $W_{i,j}$ = spatial weights matrix
- High values of $I^i_{k,l}$ → clusters:
  - high values of k surrounded by high values of l (HH) and low values of k surrounded by low values of l (LL)
- Low values of $I^i_{k,l}$ → outliers:
  - high values of k surrounded by low values of l (HL) and low values of k surrounded by high values of l (LH)
- Implemented in the QGIS Hotspot Analysis plugin.

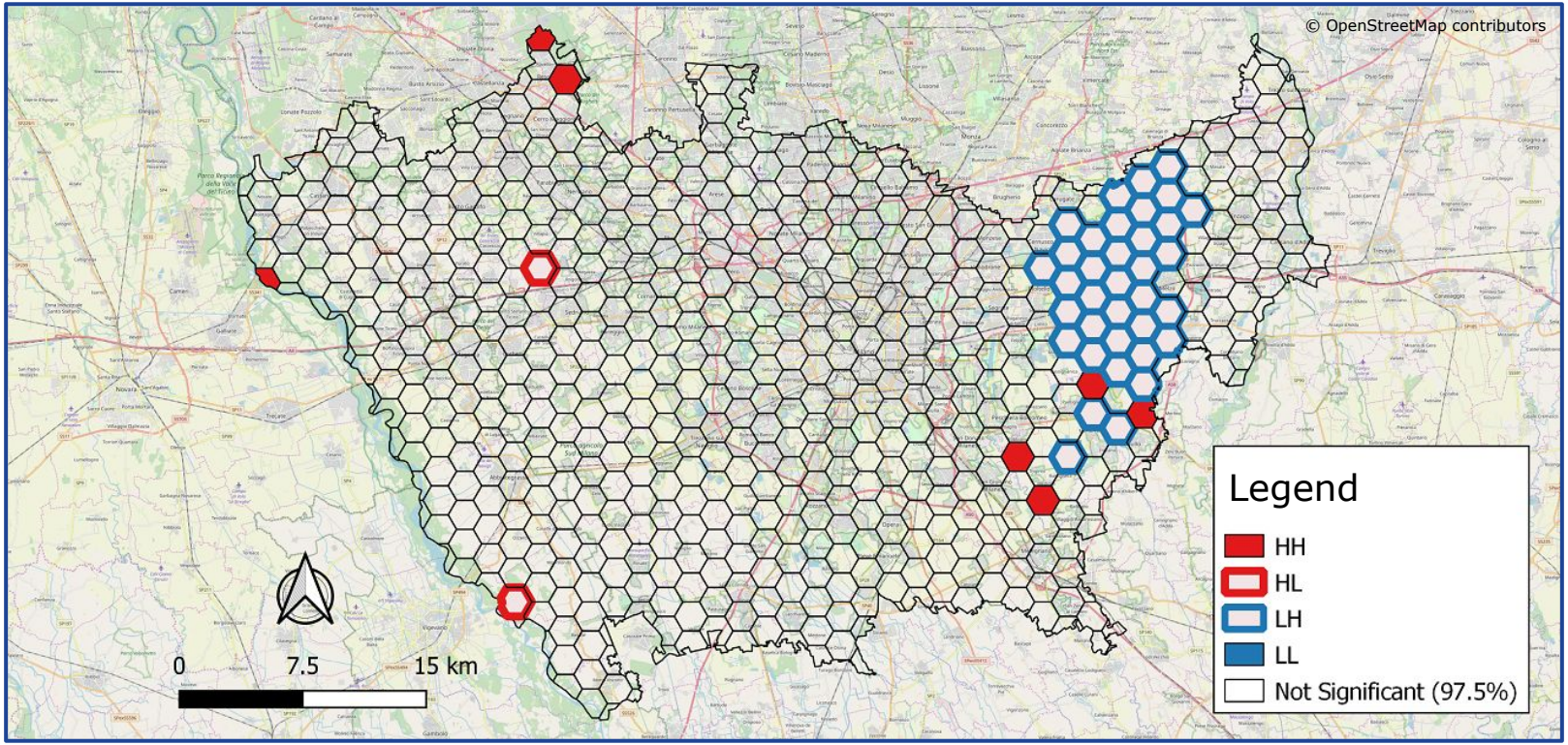
LISA – Bivariate analysis [hexagon side: 1000m]

- Average number of different contributors who edited each OSM node & average number of versions of OSM nodes



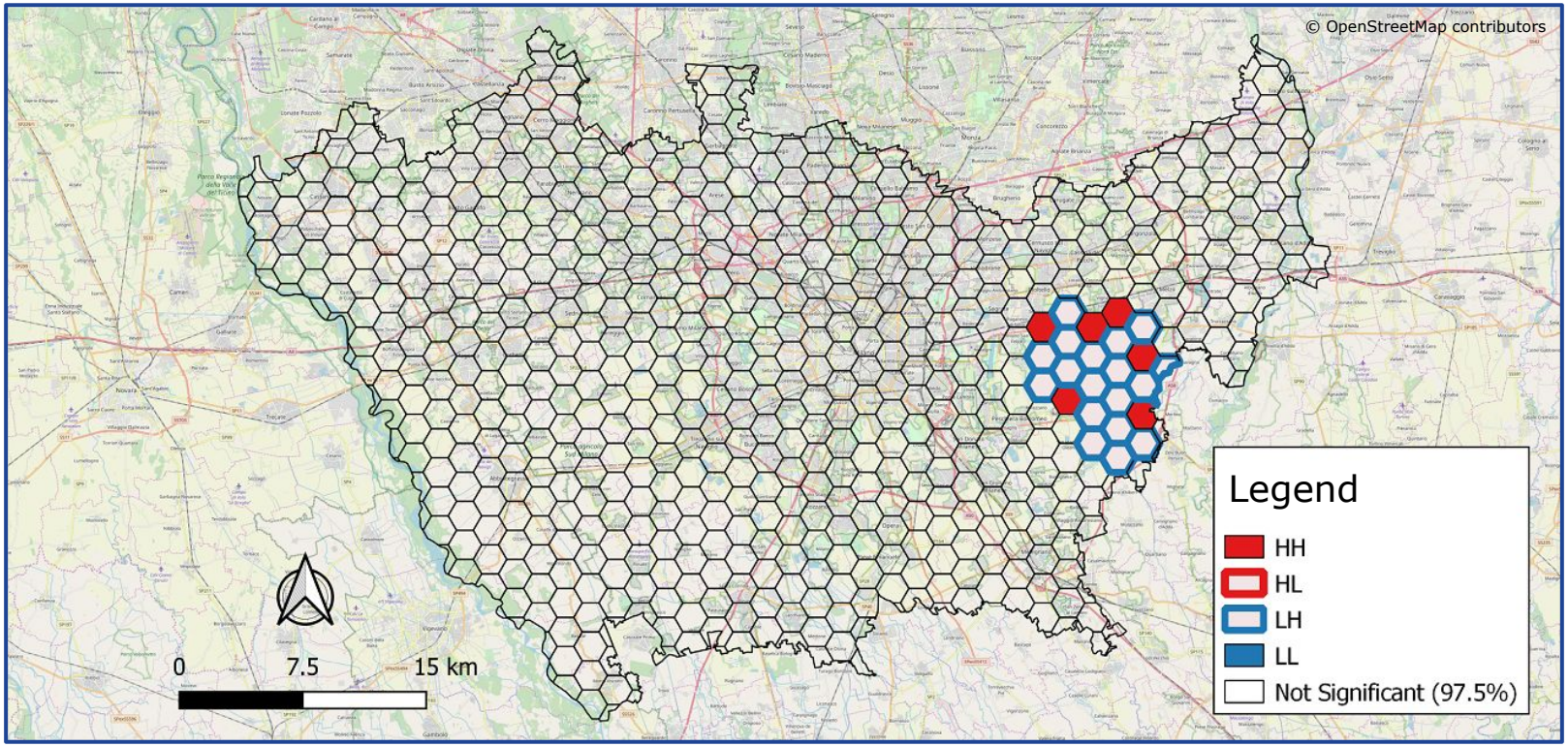
LISA – Bivariate analysis [hexagon side: 1000m]

- Average date of creation of OSM nodes & average date of last edit of OSM nodes (inverted)



LISA – Bivariate analysis [hexagon side: 1000m]

- Average number of versions of OSM nodes & average frequency of update of OSM nodes



LISA – Multivariate analysis

- Measures the local spatial association for multiple variables
- Based on Local Geary's *c* (Anselin, 2019):

$$c_{k,i} = \sum_{v=1}^{k} \sum_{j=1}^{n} W_{i,j} d^2_{v_{i,j}} \; ; \; d^2_{v_{i,j}} = (z_{1,i} - z_{1,j})^2 + \ldots + (z_{k,i} - z_{k,j})^2; \; i \neq j$$

  - k = number of variables, n = number of locations in the dataset
  - $d^2_{v_{i,j}}$ = k-dimensional squared distance between standardized z_v observation values at locations i and j
  - $W_{i,j}$ = spatial weights matrix

- Low values of $c_{k,i}$ → clusters
- High values of $c_{k,i}$ → outliers
  - not possible to determine the type of clusters/outliers!

LISA – Multivariate analysis

- Measures the local spatial association for multiple variables
- Based on Local Geary's *c* (Anselin, 2019) + 2 new indicators (Oxoli, 2019):

$$Mm_{c,i} = \mu_{X^M_{i,j}} - \mu_{X^M}, \rightarrow \begin{cases} Mm_{c,i} > 0, & \text{High values cluster} \\ Mm_{c,i} \leq 0, & \text{Low values cluster} \end{cases}$$

  - $\mu_{X^M_{i,j}}$ = mean of the medians of the standardized observation values at cluster location i and its geographical neighbours j
  - μ_{X^M} = mean of the medians of the standardized observation values at all locations

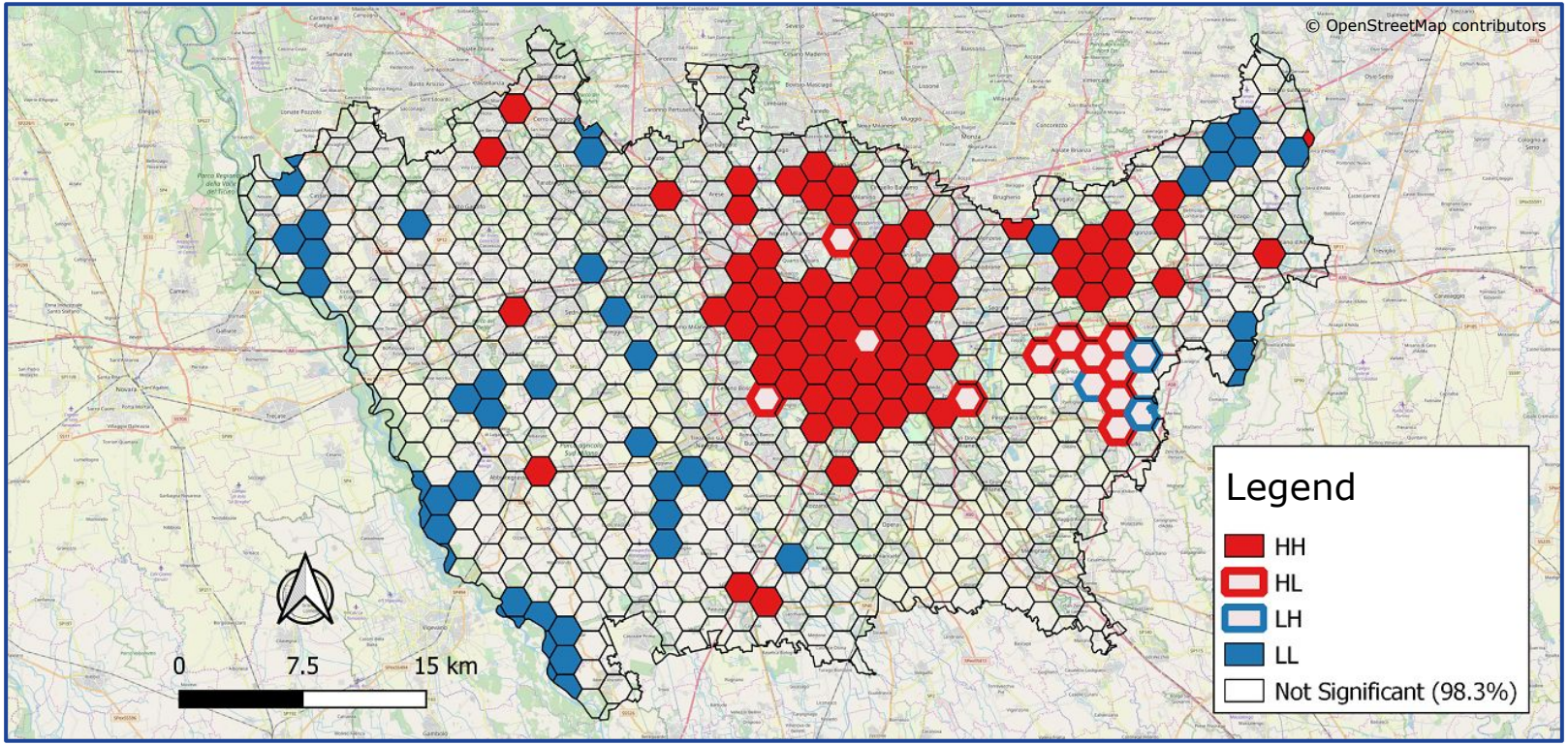
$$Mm_{o,i} = \mu_{X_i} - \mu_{X^M_j}, \rightarrow \begin{cases} Mm_{o,i} > 0, & \text{High-low outlier} \\ Mm_{o,i} \leq 0, & \text{Low-high outlier} \end{cases}$$

  - μ_{X_i} = mean of the standardized observation values at outlier location i
  - $\mu_{X^M_j}$ = mean of the medians of the standardized observation values at all geographical neighbours j

https://github.com/danioxoli/mvar_spatial_association_mapping

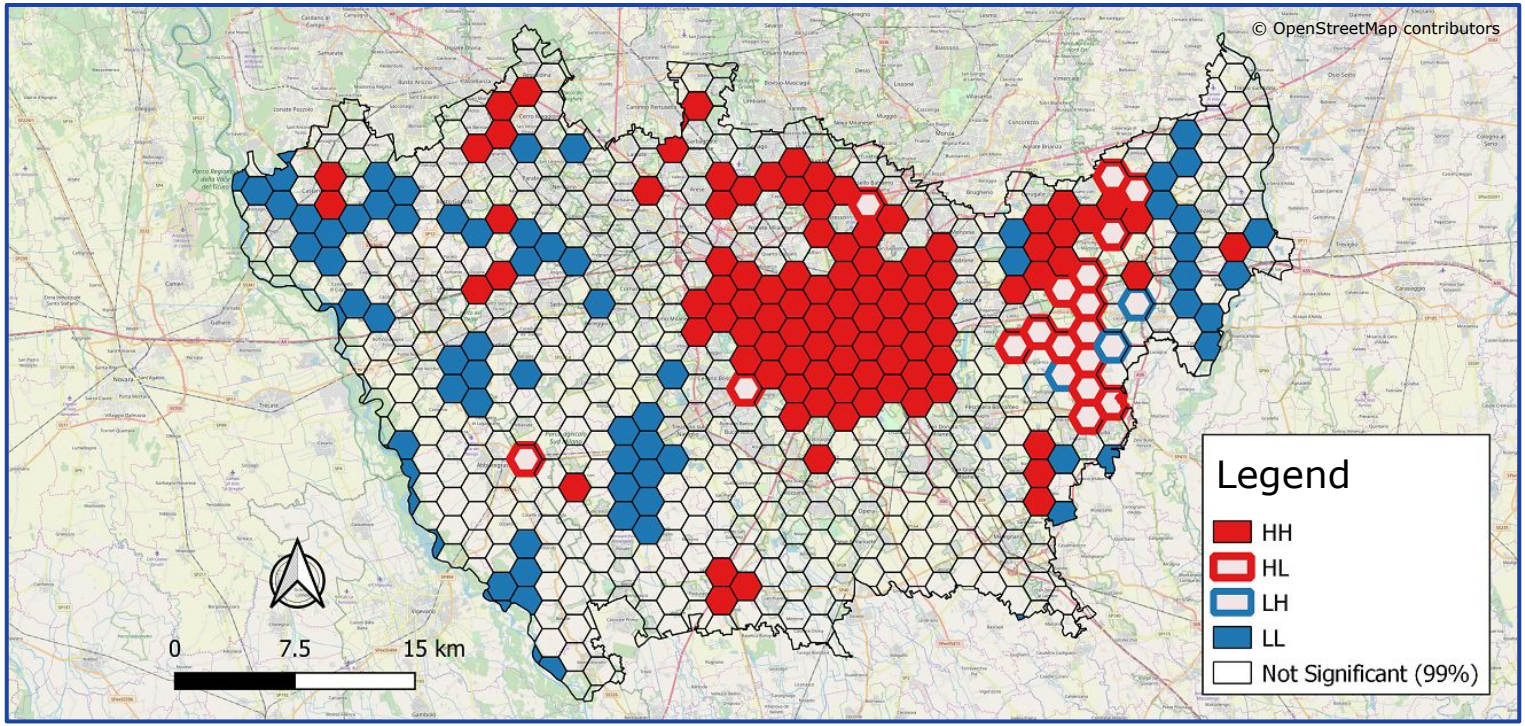
LISA – Multivariate analysis [hexagon side: 1000m]

- Average number of different contributors who edited each OSM node & average number of versions of OSM nodes & average frequency of update of OSM nodes



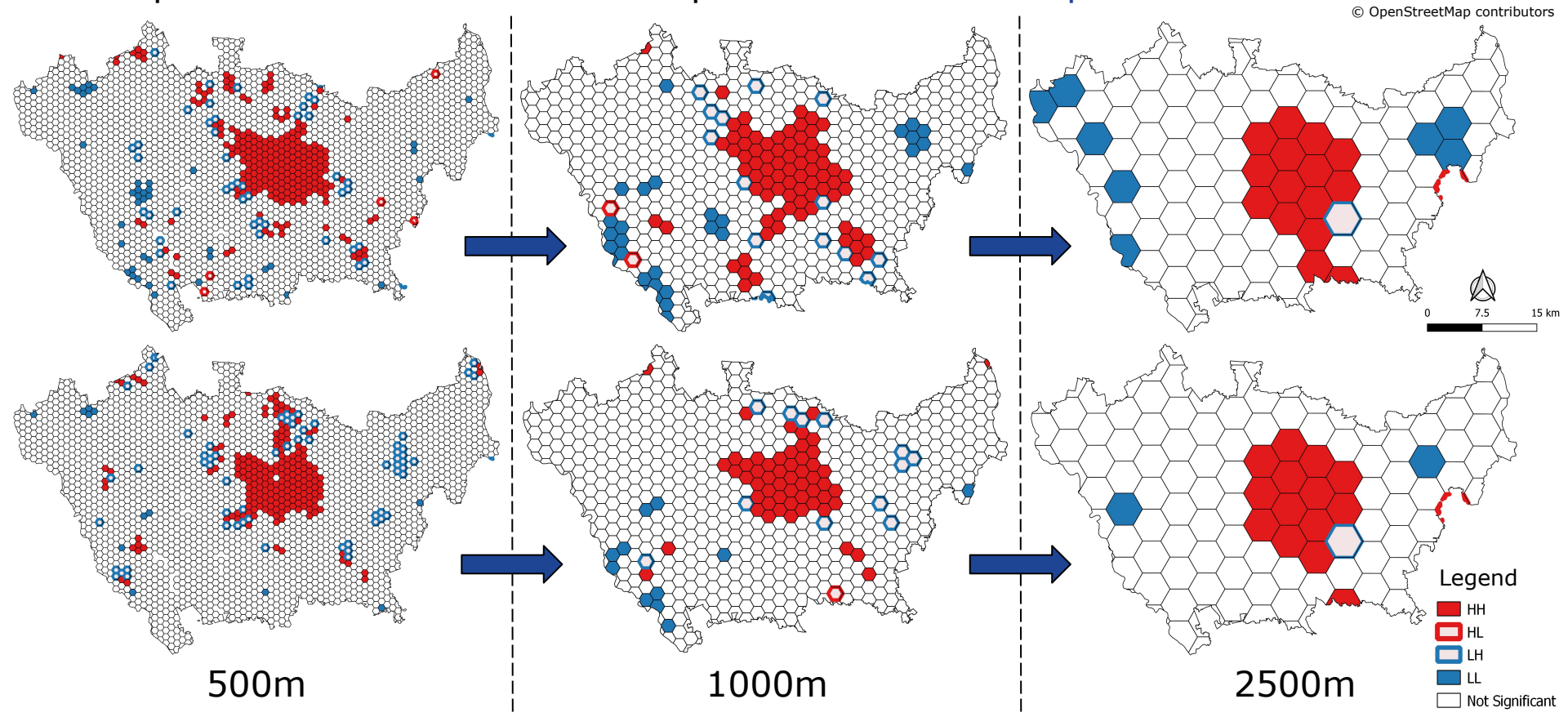
LISA – Multivariate analysis [hexagon side: 1000m]

- Average number of different contributors who edited each OSM node & average number of versions of OSM nodes & average frequency of update of OSM nodes & average date of creation of OSM nodes & average date of last edit of OSM nodes (inverted)



LISA – Changing the grid size

- Spatial associations detected depend on the chosen spatial unit!



Conclusions

- First-time application of ESDA to study OSM contributions patterns:
  - suitable to identify patterns resulting from specific local contributions
    - single contributors, mapping parties, imports, etc.
    - intrinsic assessment of OSM quality
  - outperforms quantile-based analysis, which
    - does not directly highlight clusters and outliers
    - is limited to the univariate case
  - choice of the grid impacts on the results
    - sensitivity analysis

Minghini M., Oxoli D., Frassinelli F. & Brovelli M.A. (2019). Intrinsic assessment of OpenStreetMap contribution patterns through Exploratory Spatial Data Analysis. In: *Proceedings of the Academic Track at the State of the Map 2019*, 13-14. doi: 10.5281/zenodo.3387683

Thank you!

marco.minghini@ec.europa.eu

@MarcoMinghini

https://wiki.openstreetmap.org/wiki/User:Mingo23

References

- Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, *27*(2), 93-115.
- Anselin, L. (2019). A local indicator of multivariate spatial association: extending Geary's C. *Geographical Analysis*, *51*(2), 133-150.
- Oxoli, D. (2019). *Exploratory approaches in spatial association analysis: methods, complements, and open GIS tools development*. Doctoral dissertation, Politecnico di Milano, Italy.

Stay in touch

EU Science Hub: ***ec.europa.eu/jrc***

Twitter: ***@EU_ScienceHub***

Facebook: ***EU Science Hub - Joint Research Centre***

LinkedIn: ***Joint Research Centre***

YouTube: ***EU Science Hub***

European Commission