The GeoLOD catalog and recommender for spatial linked data – A demonstration

Vasilis Kopsachilis and Michail Vaitis

Department of Geography, University of the Aegean, Greece

UNIVERSITY OF THE AEGEAN

30 May 2022 in Hersonissos, Greece

GeoLOD
http://geolod.net

- A web catalog of semantic web spatial datasets and classes (datasets and classes that contain georeferenced instances)

- A recommender for relevant datasets and classes for link discovery processes (datasets and classes that contain related instances, such as sameAs)

GeoLOD highlights

- Focuses on the geographic semantic web exploration

- Exploits dataset spatial characteristics to aid exploration

- Creates and preserves spatial metadata for datasets, exposed in GeoVoID, a dataset metadata vocabulary that extends VoID

- Offers text and map-based dataset & classes search (e.g., find classes that contain instances in Crete)

- Maintains pre-computed recommendation lists for classes that may contain related instances (e.g., sameAs) to aid link discovery

- Offers on-the-fly recommendations for uncatalogued datasets and GIS spatial datasets (Shapefile, GeoJSON)

- Exports of Silk and LIMES configuration files for recommended pair of classes for direct use in link discovery processes

- Offers a REST API for software-based consumption of its content

GeoLOD Catalog

- Parses the LOD cloud and the Datahub & locates SPARQL Endpoints (datasets) that contain georeferenced instances

- Extracts datasets and classes basic metadata and generates additional spatial metadata (Bounding Box, number of georeferenced instances, spatial vocabularies)

- Provide access to lists of spatial datasets and classes through a user interface that allows text and map-based search

Spatial Datasets list

Dataset Info Snippet

Text & map search menu

Spatial Dataset Collector

1. Finds available SPARQL Endpoints by parsing:
    *   The list of LOD datasets exposed in JSON (https://lod-cloud.net/lod-data.json)
    *   The list of DatatHub datasets exposed in CKAN API (https://old.datahub.io/)

2. Identifies SPARQL Endpoints that contain georeferenced point instances with one of the W3C Basic Geo, GeoSPARQL, GeoVocab, Geonames and GeoRSS spatial vocabularies by sending SPARQL ASK queries.

W3C Basic Geo ASK query

```
ASK {
    ?s <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?x.
    ?s <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?y.
}
```
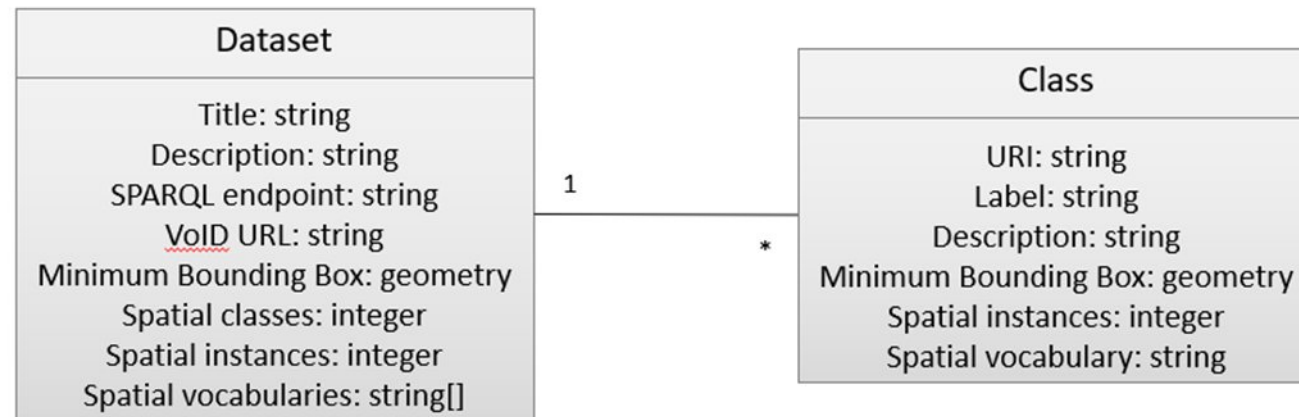
3. Retrieves the spatial classes found in each SPARQL Endpoint by sending SPARQL SELECT queries

List of classes containing instances that use W3C Basic Geo

```
SELECT DINSTINCT ?class {
    ?s <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?x.
    ?s <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?y.
    ?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?class.
}
```

Dataset & Class Metadata

- Basic metadata extracted by LOD and Datahub:
  - For datasets ☾ title, description, SPARQL endpoint URL and VoID URL
  - For classes ☾ URI, label, description

- Additional "spatial" metadata extracted by sending SPARQL SELECT queries to endpoints:
  - For datasets ☾ Bounding Box, # of spatial classes & instances, spatial vocabularies
  - For classes ☾ Bounding Box, # of instances, spatial vocabularies

- Schema of GeoLOD metadata database:

Dataset
Title: string
Description: string
SPARQL endpoint: string
VoID URL: string
Minimum Bounding Box: geometry
Spatial classes: integer
Spatial instances: integer
Spatial vocabularies: string[]

1 ——— *

Class
URI: string
Label: string
Description: string
Minimum Bounding Box: geometry
Spatial instances: integer
Spatial vocabulary: string

Dataset View

GeoVoID

A vocabulary that extends VoID for describing spatial datasets and classes.

It defines 3 datatype properties:

- **geovoid:entities** ☾ `The total number of spatial entities that are described in a void:Dataset or void:Class`

- **geovoid:classes** ☾ `The total number of distinct spatial classes in a void:Dataset`

- **geovoid:vocabulary** ☾ `A spatial vocabulary that is used in the dataset`

Schema definition is online at:
http://83.212.121.124/schemas/geovoid#

GeoVoID description of the AEMET dataset

```
@prefix geovoid: <http://83.212.121.124/schemas/geovoid#>.
@prefix void: <http://rdfs.org/ns/void#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix locn: <https://www.w3.org/ns/locn#>.

<AEMET> a void:Dataset;
    dcterms:title "AEMET metereological dataset";
    dcterms:description "AemetLinked Data (.es) is an open initiative of
geospatial data. This initiative started off by publishing diverse infor
    void:sparqlEndpoint <http://aemet.linkeddata.es/sparql>;
    dcterms:spatial [
    a dcterms:Location;
    locn:geometry "POLYGON((-17.9161111111 27.7355555556,-17.9161111111 43
];
    geovoid:vocabulary <http://www.w3.org/2003/01/geo/wgs84_pos>;
    geovoid:entities "1300"^^xsd:integer;
    geovoid:classes "1"^^xsd:integer;
    void:classPartition [
    void:class <http://www.w3.org/2003/01/geo/wgs84_pos#Point>;
    dcterms:spatial [
    a dcterms:Location;
    locn:geometry "POLYGON((-17.9161111111 27.7355555556,4.21555555556 27.
];
    geovoid:vocabulary <http://www.w3.org/2003/01/geo/wgs84_pos>;
    geovoid:entities "260"^^xsd:integer
].
```

Class View

Class Metadata

Various actions

Class Recommendations

On-the-fly class instances preview on map
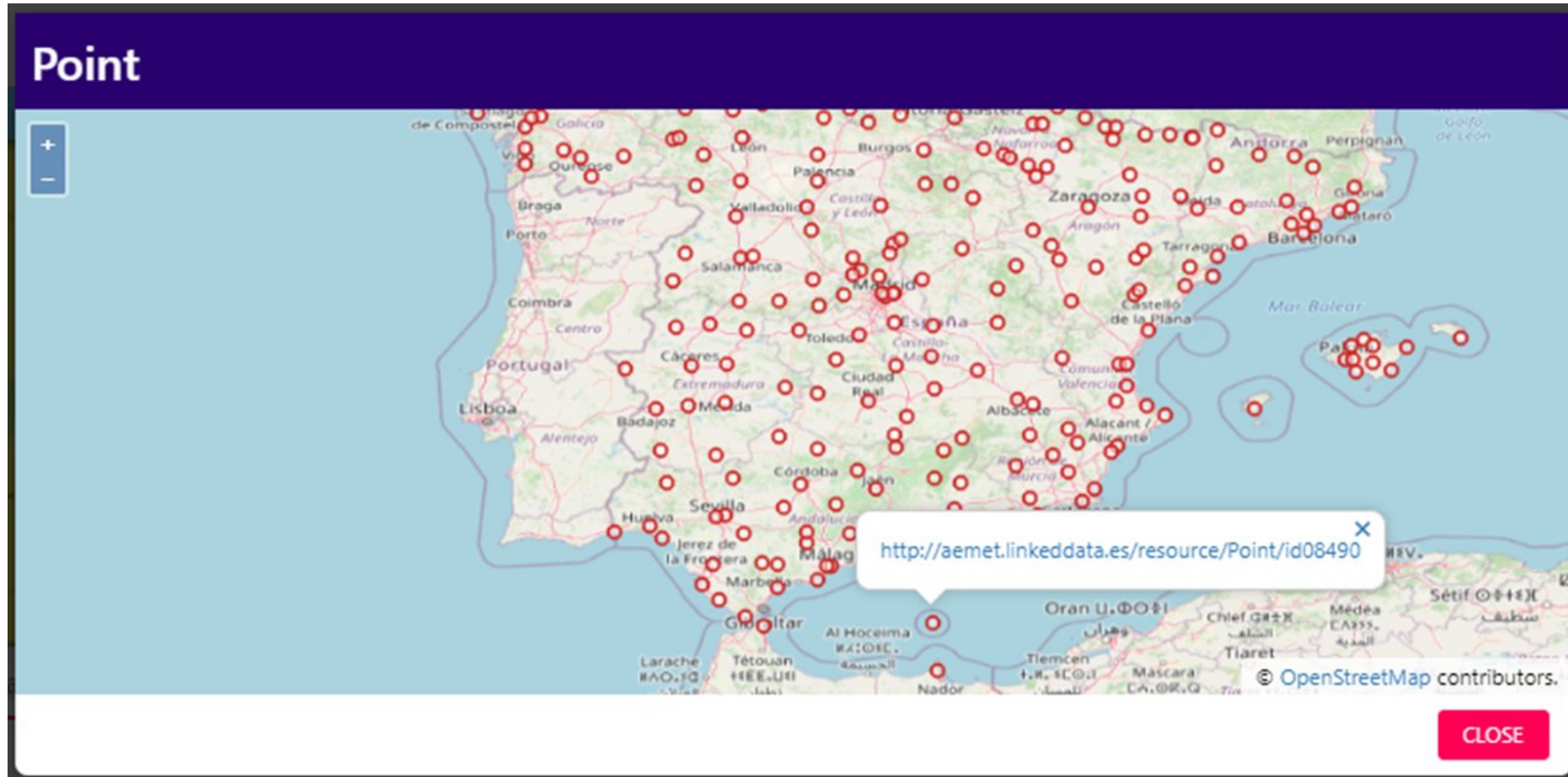
GeoLOD Recommender

- Recommends to a given class other classes that may contain related instances, such as sameAs.

- Maintains pre-computed recommendation lists for all classes in the catalog.

- Supports on-the-fly recommendations for uncatalogued SPARQL endpoints and GIS datasets (Shapefiles and GeoJSON).

Recommendation Algorithm

Based on the spatial distribution similarity of classes:

"pairs of classes whose instances present similar spatial distribution are more related than pairs of classes whose instances present dissimilar spatial distribution, in the sense that the former are more likely to contain semantically related instances"

similar spatial distribution

■ "Squares" class   ▲ "Triangles" class

not similar spatial distribution

● "Circles" class   ◆ "Diamonds" class
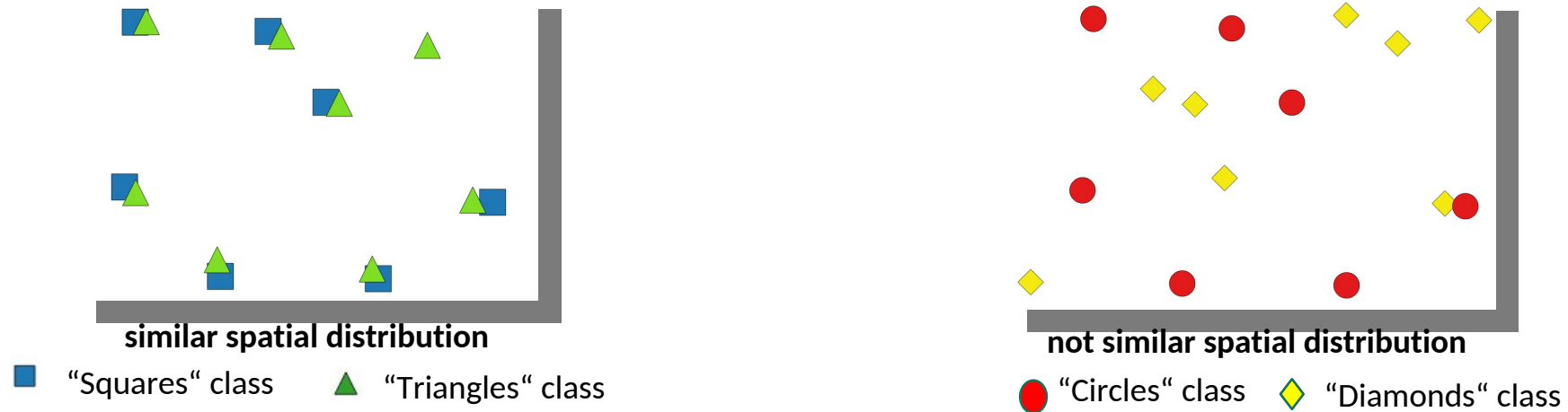
In a nutshell, the algorithm builds and stores summaries for all identified spatial classes that capture their spatial extent and index the location of their spatial instances on a QuadTree, and then apply metrics that compute the similarity of the class summaries.

More details in: V. Kopsachilis, M. Vaitis, N. Mamoulis, D. Kotzinos, Recommending geo-semantically related classes for link discovery, Journal on Data Semantics 9 (2020) 151–177. URL: https://doi.org/10.1007/s13740-020-00117-4. doi:10.1007/s13740- 020- 00117- 4

Recommendation List

Relevant classes to a "Meteorological Stations" class provided by the AEMET dataset

Silk & LIMES linkage specification export

Default matching rules based on label Levenshtein and geometry Euclidean distances

Silk configuration file example

LIMES configuration file example

On-the-fly recommender

REST API

GeoLOD services (the left part of the Request URI is http://snf-661343.vm.okeanos.grnet.gr).

Service Name	Request URI	Description
GeoLOD Description	/api/download/dcat	Returns a DCAT-compliant turtle file that contains general information about GeoLOD and the list of the datasets in the Catalog
Dataset List	/api/datasets	Returns, in JSON, the list of datasets with their metadata (including internal dataset IDs) in the GeoLOD Catalog
Dataset Description	/api/datasets/<ID>	Returns, in JSON, the specified dataset metadata with the list of its classes. The dataset ID is a variable corresponding to the internal dataset ID. (e.g., http://snf-661343.vm.okeanos.grnet.gr/api/datasets/915 returns the metadata for the *AEMET* dataset)
Class List	/api/classes	Returns, in JSON, the list of classes with their metadata (including internal classes IDs) in the GeoLOD Catalog.
Class Description	/api/classes/<ID>	Returns, in JSON, the specified class metadata with the list of its recommended classes. The class ID is a variable corresponding to the internal class ID. (e.g., http://snf-661343.vm.okeanos.grnet.gr/api/classes/139090 returns the metadata for the *CaveEntrance* class of *Linklion* dataset).
Dataset GeoVoID	/api/download/geovoid/<ID>	Returns, in turtle format, the GeoVoID description of the specified dataset.
Dataset Recommendations	/api/download/datasetrecommendations/<ID>	Returns, in JSON, the list of recommendations for all specified dataset classes.
Class Recommenations	api/download/classesrecommendations/<ID>	Returns, in JSON, the list of recommendations for the specified class.

Implementation

- The GeoLOD backend API was developed in **Node.js** and the frontend application in **React**

- The GeoLOD database is a **PostgreSQL** with the **PostGIS** extension for spatial data management

- Queries to the SPARQL endpoints were sent with the **Fetch SPARQL endpoint** node.js module

- Bounding box thumbnails depicting were generated with the **Static Image Mapbox API**, and the interactive maps were built on **Leaflet** and **OpenLayers**

- GeoLOD is hosted in a Ubuntu 18 LTS 4GB Virtual Machine, provided by **okeanos**, a GRNET cloud Infrastructure as a Service (IaaS) for Greek academic institutes

- GeoLOD content, that is the list of spatial datasets and classes with their metadata and the recommendation lists for all classes, is **updated automatically every two months**, as a background process

Findings & Discussion

- GeoLOD contains 82 spatial datasets (out of 629 unique and 152 active SPARQL endpoints) and 5,218 spatial classes.

- They spatially cover all the world. Most datasets are "global".

- In many datasets, georeferencing errors were found.

- GeoLOD provides approximately 90,000 class recommendations and on average it recommends 25 relevant classes per class.

- It requires on average 18 minutes to generate the class recommendation list for each class (although the execution time varies from a few seconds to several minutes), while LIMES would require approx. 5 hours for pair-wise class comparisons.

Legend
Dataset MBR
Datasets Density
<=10
11-15
16-19
20-24
>=25
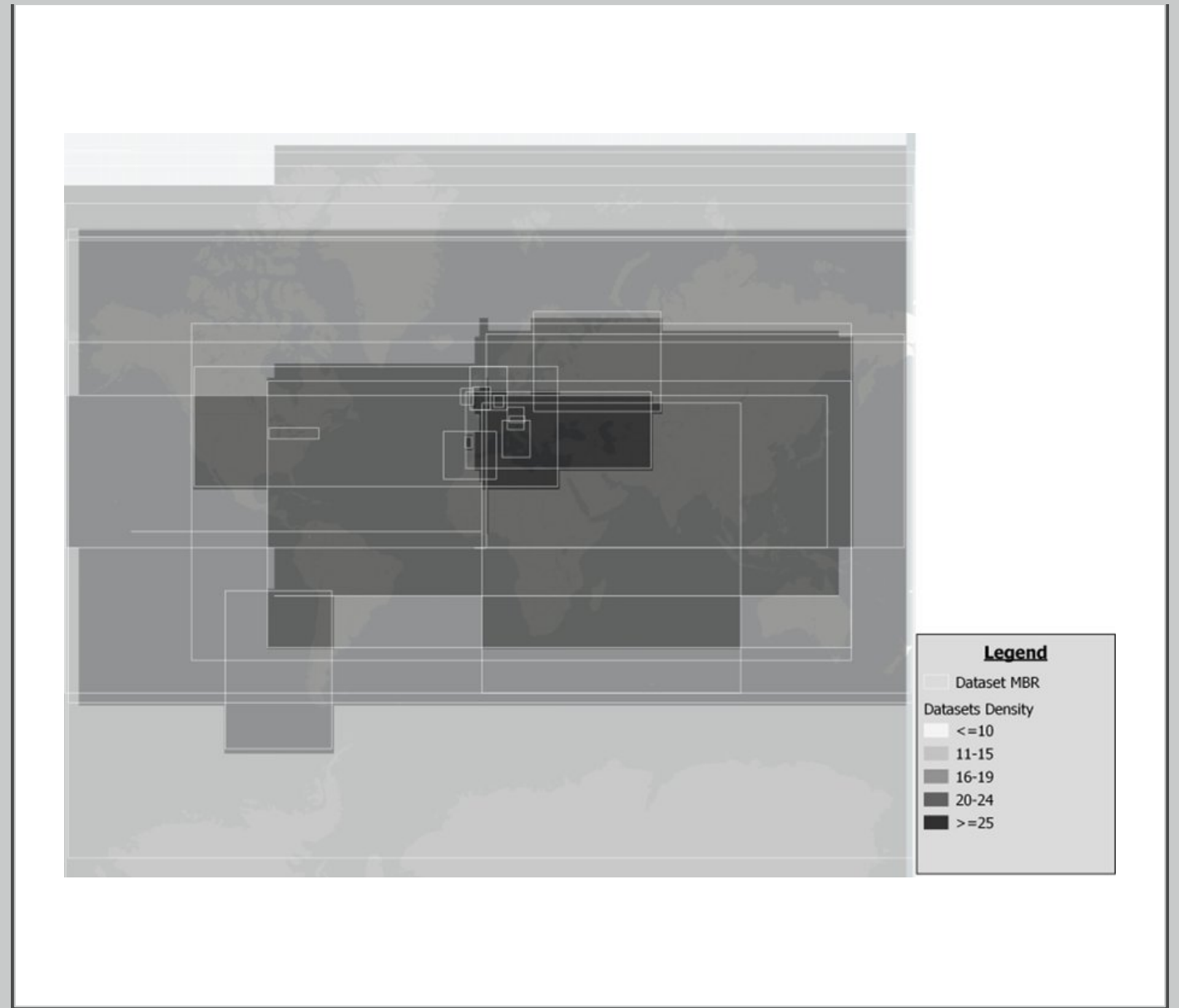
Pointers for Future Work

- Inclusion of datasets provided as RDF dumps and containing line and polygon geometries

- Extension of the on-the-fly recommender to support SPARQL endpoints that use additional spatial vocabularies (other than W3C Basic Geo) and additional spatial data formats, such as the Web Feature Service (WFS)

- Further improvement of the recommendation algorithm in terms of efficiency and effectiveness (e.g., integration with textual similarity)

- Integration with Silk or LIMES web services for the instant generation of instance link matching

Thank you!

ESWC

Heraklion | Crete | Greece

30 May 2022 in Hersonissos, Greece

SELECT Queries

Spatial Vocabulary	SELECT Query

GeoVocab	`SELECT DISTINCT ?class {`
	`?geom <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?x.`
	`?geom <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?y.`
	`?s <http://geovocab.org/geometry#geometry> ?geom.`
	`?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?class}`

GeoSPARQL	`SELECT DISTINCT ?class {`
	`?s <http://www.opengis.net/ont/geosparql#hasGeometry> ?geom.`
	`?geom <http://www.opengis.net/ont/geosparql#asWKT> ?wkt.`
	`?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?class}`

GeoNames	`SELECT DISTINCT ?class {`
	`?s <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?x.`
	`?s <http://www.w3.org/2003/01/geo/wgs84_pos#lat>?y.`
	`?s <http://www.geonames.org/ontology#featureClass> ?class.}`

W3C Basic Geo	`SELECT DISTINCT ?class {`
	`?s <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?x .`
	`?s <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?y.`
	`?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?class.}`

GeoRSS	`SELECT DISTINCT ?class {`
	`?s <http://www.georss.org/georss/point> ?point.`
	`?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?class}`

QuadTree

A global Quadtree index with not equally sized cells that capture global spatial instances distribution
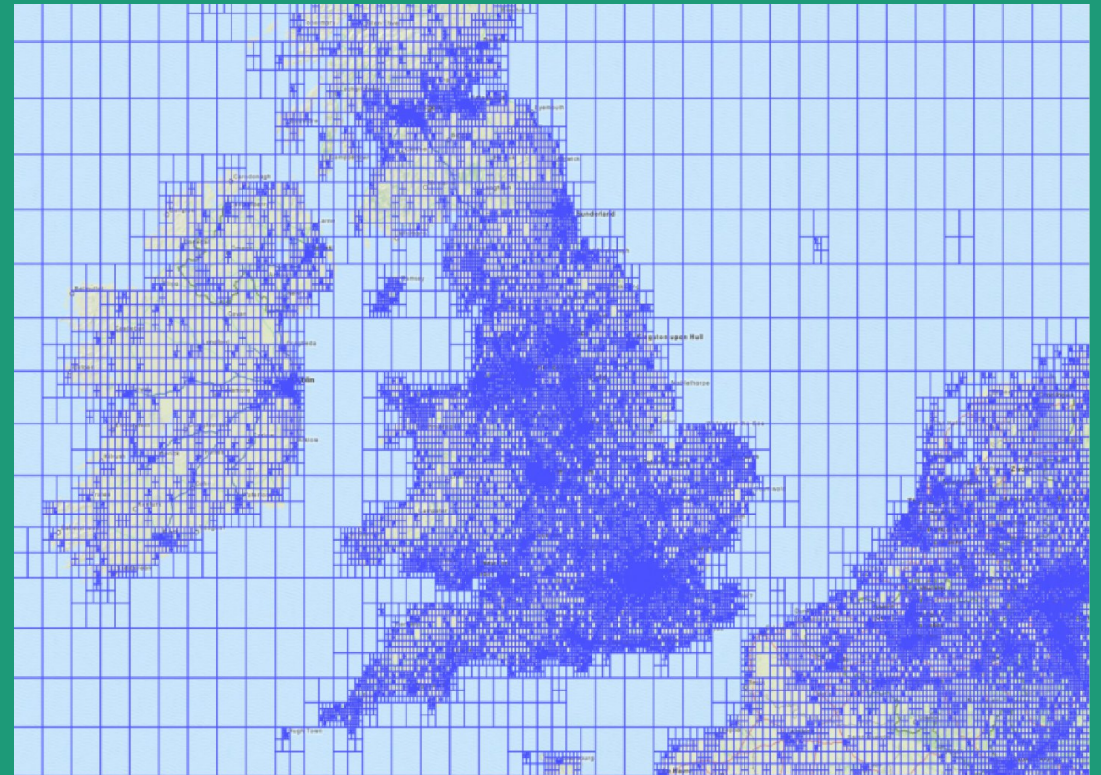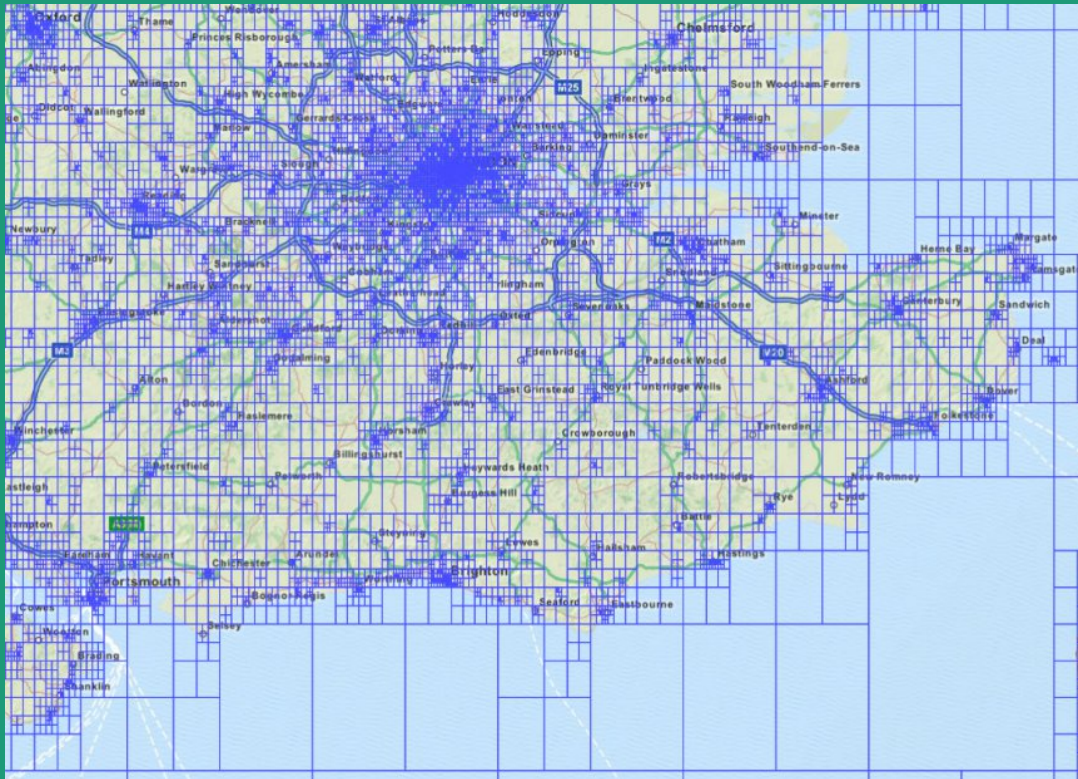
Class Summarization
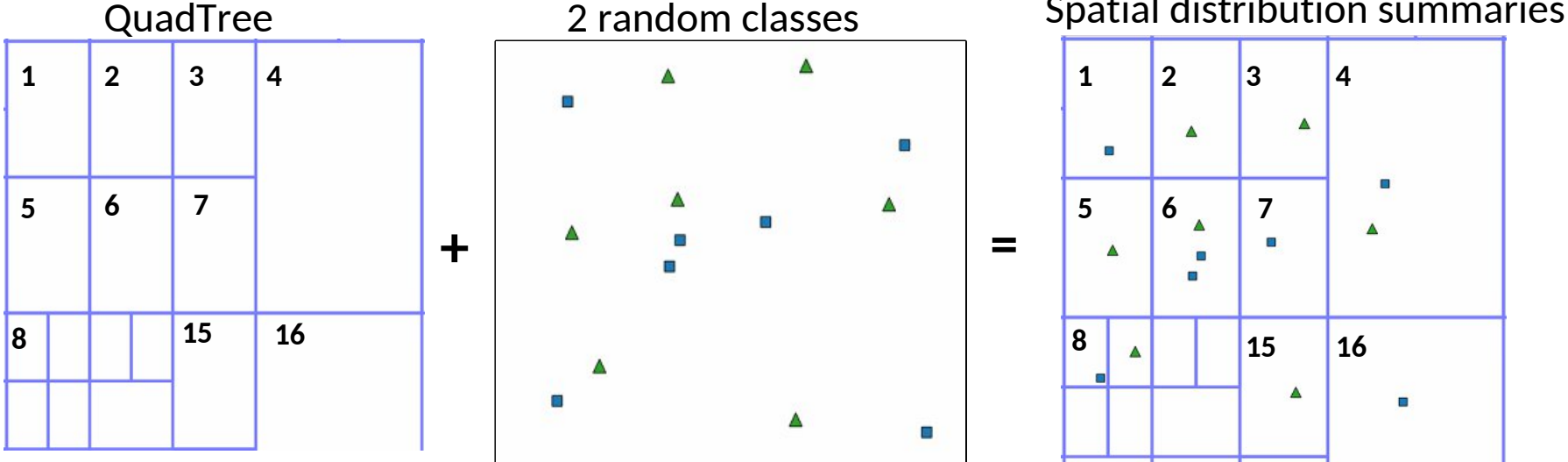
Summary of Squares class = {1,4,6,7,8,16}
Summary of Triangles class = {2,3,4,5,6,9,15}

....

Class Relevancy

The more similar the summaries of classes the more similar the spatial distribution

Measurement of class summaries:

- Number of common cells
- Jaccard index
- Overlap Coefficient
- (Pointwise) Mutual Information
- Poisson Distribution
- …

The algorithm returns a ranked list of relevant classes based on the values of the above metrics

Measurement of geographic similarity

Aim ☾ measure the similarity of classes spatial distribution. Based on their summaries

Similar spatial summaries ☾ similar spatial distribution ☾ classes with related instances

☐ (S) http://greek-lod.math.auth.gr/fire-brigade/resource/subdivisions (Hellenic Fire Brigade)
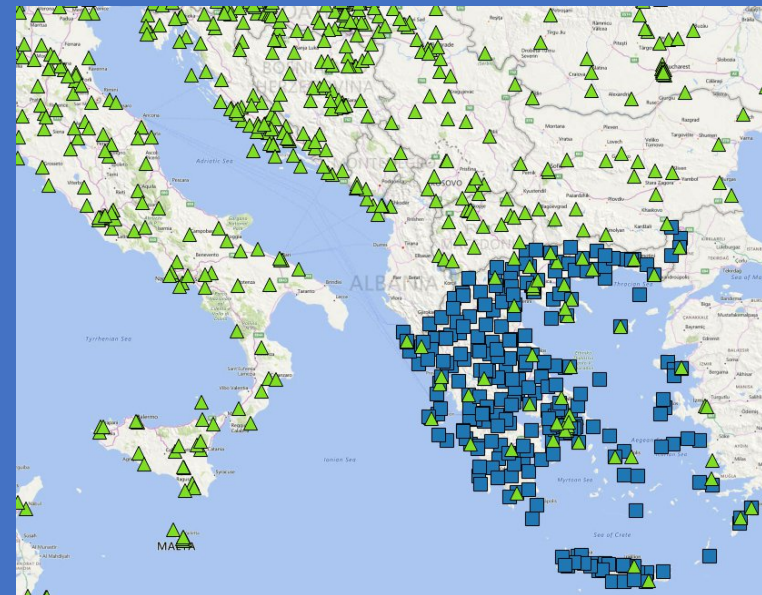
▲ (T) http://linkedgeodata.org/ontology/FireStation (LinkedGeoData)

Μέτρηση γεωγραφικής ομοιότητας

➢ Κλάσεις που εκτείνονται σε διαφορετικές περιοχές (μη αλληλεπικαλυπτόμενο ConvexHull) δεν παρουσιάζουν όμοια γεωγραφική κατανομή και δεν είναι πιθανό να περιέχουν σχετικές οντότητες.

➢ Μέτρηση της γεωγραφικής ομοιότητας των συνόψεων 2 κλάσεων, S και T, στον αλληλεπικαλυπτόμενο χώρο τους (intersecting ConvexHull).
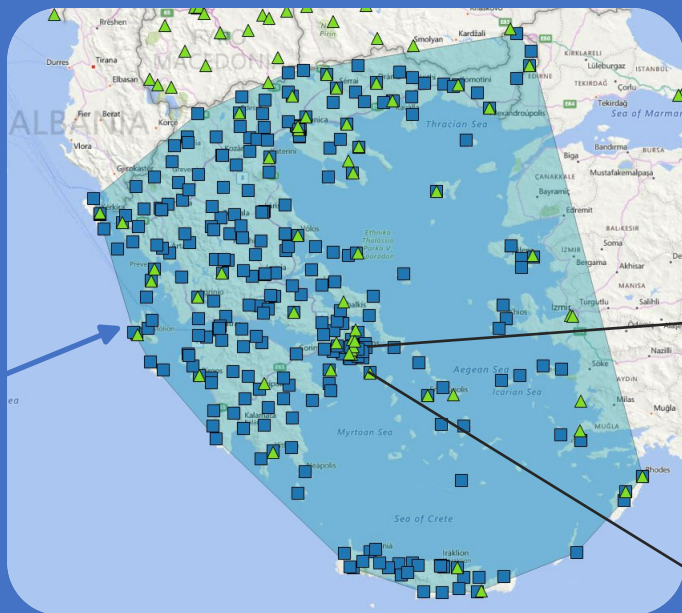
☐ (S) http://greek-lod.math.auth.gr/fire-brigade/resource/subdivisions (Hellenic Fire Brigade)

🔺 (T) http://linkedgeodata.org/ontology/FireStation (LinkedGeoData)

Intersecting
ConvexHull
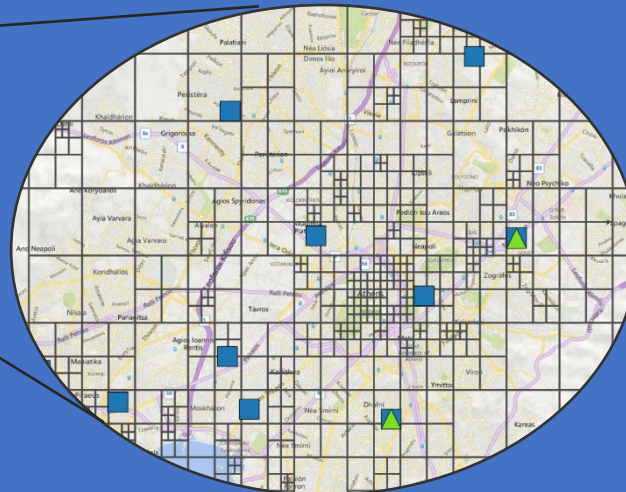
Υπολογισμοί από συνόψεις
|S| = 275 |T| = 54
|S ∩ T| *or* c = 29
|Q| = 13.691

Spatial Similarity Metrics I

- **Number of Common Cells (C)**: Ο αριθμός των στοιχείων της τομής των δύο συνόψεων.

$$C = |S \cap T|$$

Παρατήρηση: Δεν λαμβάνεται υπόψη το μέγεθος των συνόψεων.

- **Jaccard Similarity (JS):** Ο αριθμός της τομής (αριθμός κοινών κελιών) διά του αριθμού της ένωσης των δύο συνόψεων.

$$JS = \frac{|S \cap T|}{|S \cup T|}$$

Παρατήρηση: Δεν λαμβάνεται υπόψη η διαφορά του μέγεθος των συνόψεων.

- **Overlap Coefficient (OC):** Ο αριθμός της τομής (αριθμός κοινών κελιών) διά του αριθμού του μικρότερου συνόλου.

$$OC = \frac{|S \cap T|}{\min(|S|,|T|)}$$

Παρατήρηση: Δεν λαμβάνεται υπόψη η συχνότητα εμφάνισης των οντοτήτων στο χώρο μελέτης.

Υπολογισμοί από συνόψεις

$|S| = 275$ $|T| = 54$

$|S \cap T|$ *or* c = 29

$|S \cup T| = 300$

$|Q| = 13.691$

Υπολογισμός δεικτών

JS = 0.09   OC = 0.53

Η συχνότητα (άρα και πιθανότητα) εμφάνισης των οντοτήτων μίας κλάσης στο χώρο μελέτης υπολογίζεται ως το μέγεθος της σύνοψης της κλάσης διά τον αριθμό των συνολικών κελιών στο χώρο μελέτης.

$$P(s) = \frac{|S|}{|Q|}$$

Αν οι συνόψεις 2 κλάσεων είναι ανεξάρτητες, σύμφωνα με τη συνθήκη ανεξαρτησίας ενδεχομένων, ο αναμενόμενος αριθμός των κοινών κελιών τους είναι:

$$P(S \cap T) = P(S)P(T) \Rightarrow \hat{c} = \frac{|S||T|}{|Q|}$$

Αν ο πραγματικός αριθμός των κοινών κελιών των συνόψεων 2 κλάσεων (C) διαφέρει από τον αναμενόμενο αριθμό των κοινών κελιών των κλάσσεων σε περίπτωση που ήταν ανεξάρτητες (ĉ), τότε οι κλάσεις σχετίζονται.

Ο λόγος συσχέτισης (R) μπορεί να υπολογιστεί ως ο αριθμός των κοινών κελιών διά τον αναμενόμενο αριθμό των κοινών κελιών στην περίπτωση που τα δύο σύνολα είναι ανεξάρτητα:

$$R = \frac{C}{\hat{c}}$$

- **Pointwise Mutual Information (PMI):** Ο λογάριθμός της πιθανότητας της ταυτόχρονης εμφάνισης δύο ενδεχομένων διά το γινόμενο των πιθανοτήτων τους:

$$PMI = \log R$$

Υπολογισμοί από συνόψεις
|S| = 275 |T| = 54
|S ∩ T| *or* c = 29
|S ∪ T| = 300
|Q| = 13.691

Υπολογισμοί Δεικτών
PMI=1.42

Spatial Similarity Metrics III

- **Hypergeometric Distribution Culmulative Probability (HG):** Η πιθανότητα μία κλάση (S) να έχει c ή παραπάνω κοινά κελιά με μία κλάση (T) σε μία περιοχή που καλύπτεται από (Q) κελιά:

$$P(X \geq c) = \sum_c^{\min(|S|,|T|)} P(X = c) \, όπου \, P(X = c) = \frac{\binom{|S|}{c}\binom{|Q|-|S|}{|T|-c}}{\binom{Q}{T}}$$

- **Poisson Distribution Culmulative Probability (PD):** Η πιθανότητα να εμφανίζονται c ή παραπάνω κελιά σε μία περιοχή που αναμένεται να εμφανίζονται λ = ĉ κελιά:

$$P(X \geq x) = \sum_{x=|C|}^{Min(|S|,|T|)} P(x) \qquad όπου \qquad P(x) = \frac{e^{-\hat{c}}\hat{c}^x}{x!}$$

- **Δείκτης συσχέτισης δυαδικών μεταβλητών (*Phi Coefficient*).** Μία κλάση αναπαρίσταται από μία δυαδική μεταβλητή που παίρνει την τιμή 1 για τα κελιά που υπάρχουν στην σύνοψη μίας κλάσης και την τιμή 0 για τα κελιά που δεν υπάρχουν.

	T_1	T_0	Total
S_1	|C|	|S|-|C|	|S|
S_0	|T|-|C|	|Q|-|S|-|T|+|C|	|Q|-|S|=|S'|
Total	|T|	|Q|-|T| = |T'|	|Q|

2X2 πίνακας συνάφειας των κλάσεων S και T

$$\Phi = \frac{|C|*|Q|-|S|*|T|}{\sqrt{|S||S'||T||T'|}}$$

Υπολογισμοί από συνόψεις
\|S\| = 275   \|Ti\| = 54
\|S ∩ Ti\| *or* c = 29
\|S ∪ Ti\| = 300
\|Q\| = 13.691

Υπολογισμος δεικτών
HG = 2.37E-41
PD=0   Φ=0.23