Algorithms for Geospatial Data Matching

An algorithm for error detection, correction and updation of spatial data sets

Group Members

Kartik Bhadada	U101114FCS078
Surajnath Sidh	U101114FCS146
Shubham Singh	U101114FCS192
Mehak Bhatia	U101114FCS193

Project Mentor: Dr. Prosenjit Gupta

A little about our topic...

- The field of Geographical Information Systems (GIS) experienced a rapid growth of available sources of geospatial datasets.
- Data Matching aka linking/alignment/reconciliation is a process of gathering different data sources with pronounced similarity in a geospatial or semantic way.
- In layman terms, we consider it as a process of finding correspondences between two spatial datasets, like concepts, objects or their components.

Rationale of work

- Rapid growth of data sets resulted in the need to plot them.
- Since most of the data has been plotted, now there is a need to detect the errors and correct/update them by matching the data with the crowdsourced data.
- Crowdsourced data also known as retrieved data is the data collected by sources such as check-ins etc. Relevant data or authoritative data is the actual data that needs to be verified/corrected.

Continued...

- Since huge amount of retrieved and relevant data is available, matching it becomes a rather hectic task.
- Hence certain specific algorithms needs to be implemented in order to lighten the job.
- In our research, we will be working on the Foursquare check-In dataset and using it as a test dataset for our algorithm.

Objectives

- To bring attention towards data integration and data matching.
- To increase focus towards correcting erroneous data with updating routes.
- To create an efficient algorithm for dynamic error detection, correction and updating the location of different places(hotels, restaurants etc) using retrieved or crowdsourced data.

Review Of Literature (Research Papers)

A Survey of Measures and Methods for Matching Geospatial Vector Datasets.

OVERVIEW

In the literature, one can find many different approaches to solve the geospatial matching problem.

Two key concepts to organise these approaches are:

1) Measures

2) Methods

Continued...

- Measures can be single, or multiple, combined using normalized score, a weighted combination, the probabilistic theory, optimization processes or brief theory.

- Methods on the other hand have been classified according to different perspectives. They have evolved from being simpler methods to increasingly complex methods which include a more diverse perspective.

Measures

The matching measures are organised on the basis of nature of measured quantity:-

1) Geometry (shape, location, length, area)
2) Topology (geographic space, for instance network matching)
3) Attributes (non-geometric properties; numeric, list and text measures)
4) Context (geographic context, similarity on the basis of presence of another object)

Methods

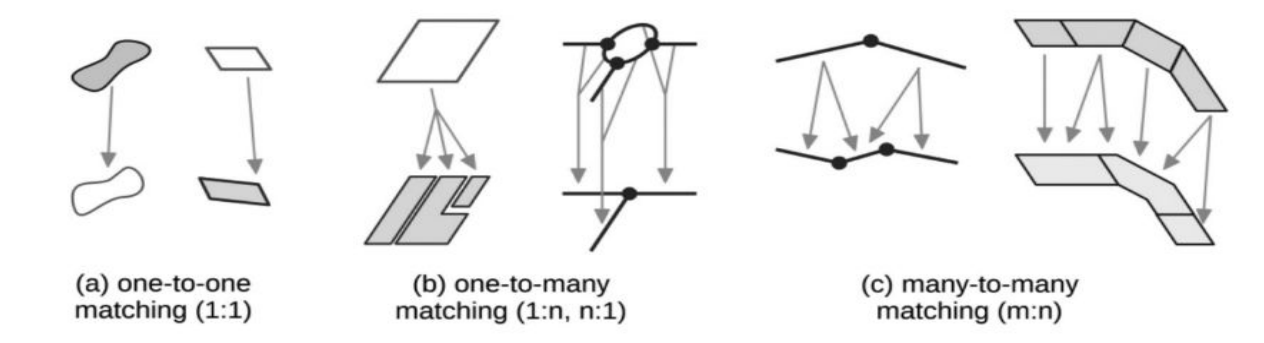
Matching methods are classified according to-

1) **Level of actuation**
- Schema (highest level, that is, modelling)
- Feature (object level)
- Internal (internal components)

Methods(cont.)

2) Case of correspondence:

- One to One
- One to Many
- Many to Many



Conclusion

This paper presented measures and methods applied to matching geospatial datasets.

The paper discusses

1) **Classification of measures and methods**
2) **Issues choosing measures and methods.**

Similarity of Spatial Scenes

- Similarity of scenes is a causal judgement people make in everyday life. It is intuitive, subjective and displays no strict mathematical models.
- This paper focusses on similarity measures of spatial scenes that are contained in geographic databases.
- A spatial scene is basically a set of geographic objects together with their spacial relations. These spatial relations are being discussed in this entire paper.

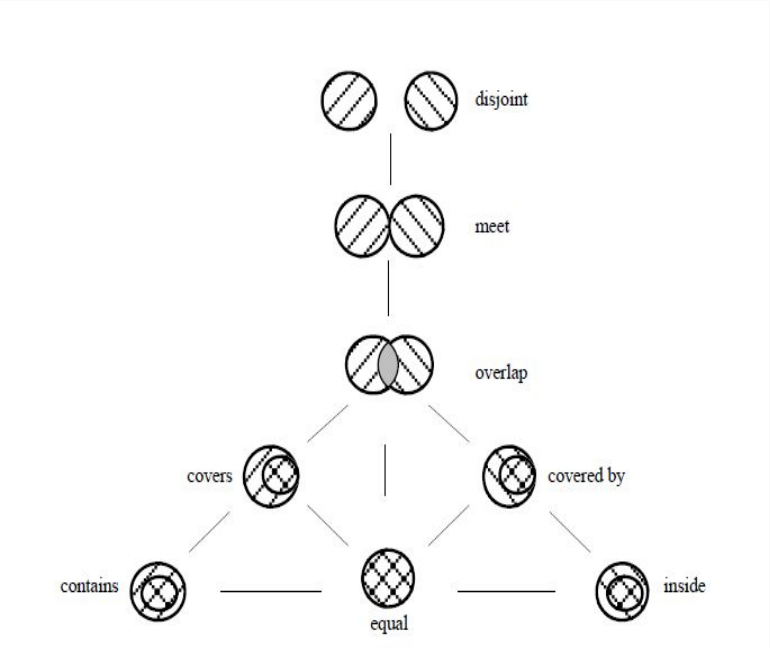
Overview

The paper deals with three major similarity measures of spatial relations

1) **Similarity of Topological Relations**
2) **Similarity of Distance Relations**
3) **Similarity of Direction Relations**

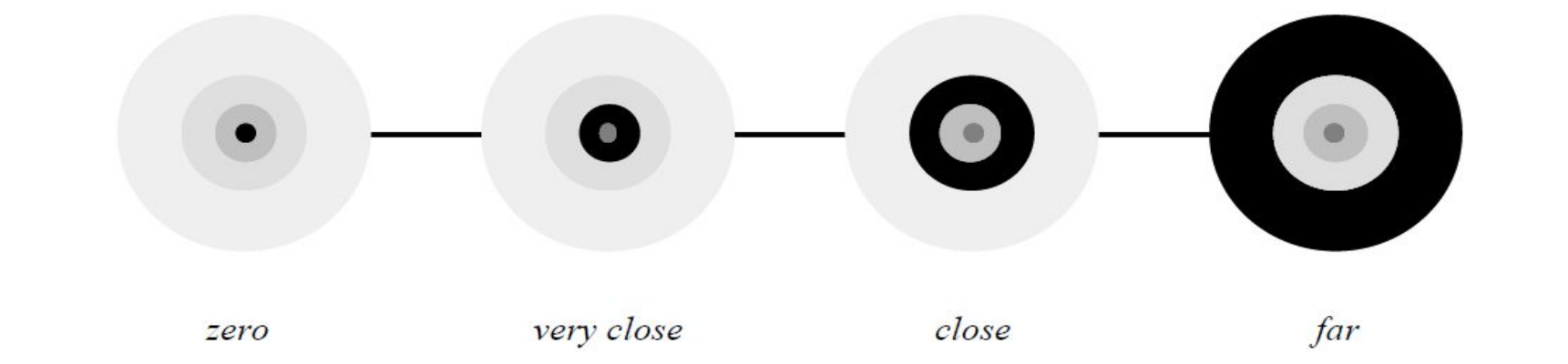
Similarity of Topological relations

Similarity of Topological relations can be explained based on the given diagram.

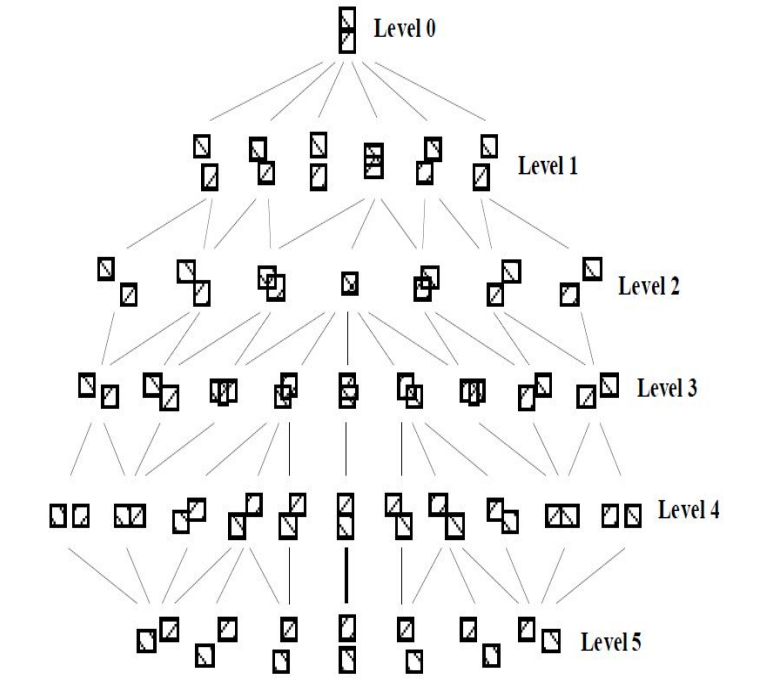


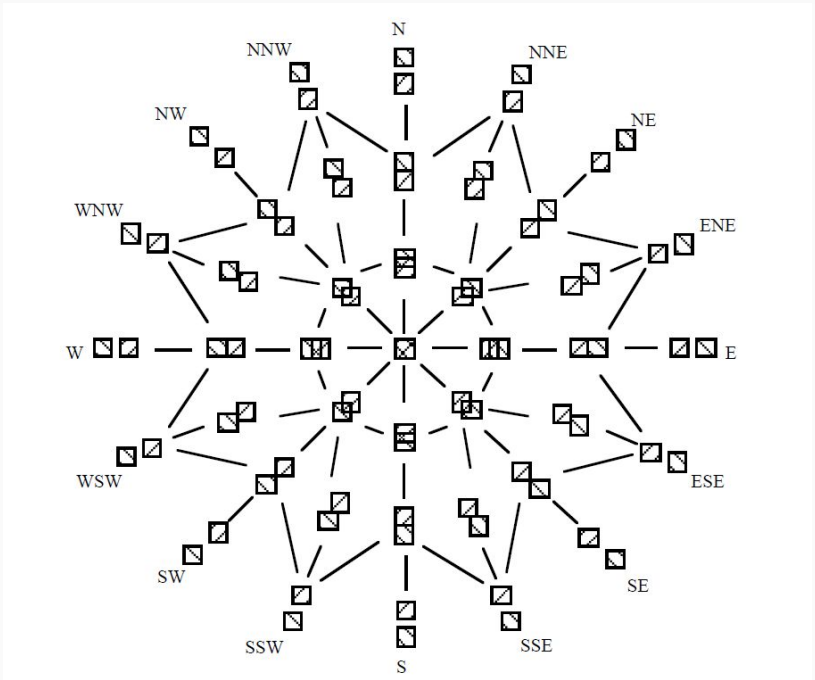
Similarity of Distance relations

Distance relations are matched according to the radial distance between two objects. Here one of the objects is assumed to be a reference object and the distance is measured by moving the other object.



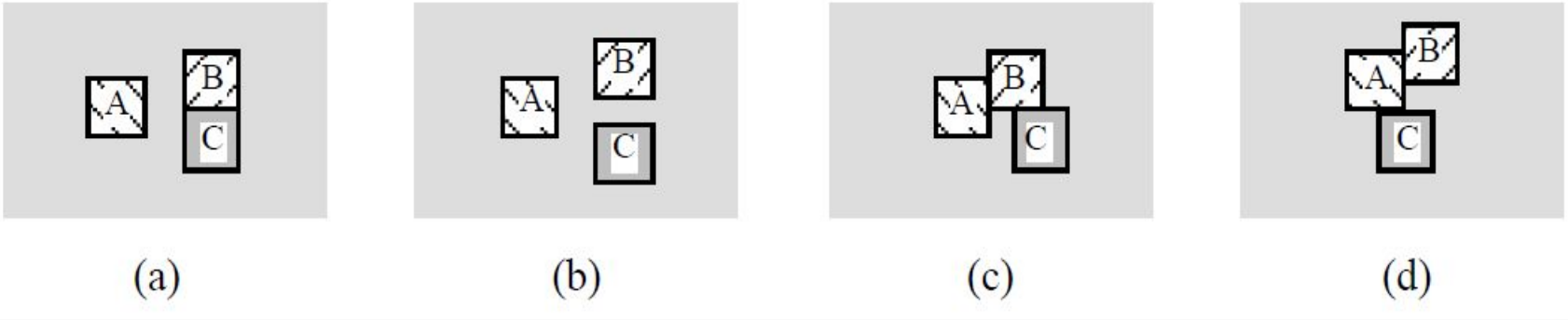
Similarity of Direction Relations





Assessing Spatial Similarity

- The three models of topology, direction and distance form the basis for the assessment of spatial similarity of scenes.
- The relative similarity of the sets of scenes can also be determined by assessing the spatial similarity. For instance-



Conclusion

This paper presents a method for assessing spatial similarity. It is based on the topological, directional and distance relations in the spatial scenes and their conceptual neighborhoods.

Geo MatchMaker: Automatic and Efficient Matching of Vector Data with Spatial Attributes in Unknown Geometry Systems

This paper discusses about the large amount of geospatial data that is available through different sources these days and integration and fusion of data from different sources.

Introduction

- Geospatial data fusion has been one of the central issues in GIS.
- There have been a number of efforts to automatically or semi-automatically detect matched features across different road vector datasets
- The paper proposes several approaches to handle the matching of diverse data sets automatically and efficiently.

Related Work

- Vector to vector conflation.
- The initial focus was to remove spatial inconsistency between two data sets. Once the inconsistency is removed it is easier to compare attributes and fuse datasets.
- Multiple attempts to solve this problem have been there.

Methods

1. PPM(point pattern matching)

This uses the brute force method comparing each attribute of dataset 1 (s1) to all the attributes of dataset 2 (s2) to find the matching point(intersection). It is accurate but very complex for big data sets and very time consuming as well.

Methods(Continued...)

2. Geo-PPM

Geo PPM focuses on eliminating the comparison of point pairs that can't be the candidate for matching pattern by using network properties. It is not possible for big data sets especially when datasets follow a pattern like grid or manhattan etc.

Methods(Continued...)

3. Prioritized Geo-PPM

The intuition in prioritized Geo-PPM is that the matching feature in one data set is likely to be a matching feature in another data set.This means that if we find the matching feature in one data set we can find it very quickly in another data set and hence it will be easy to find intersections.Prioritized Geo-PPM can provide a substantial improvement over Geo-PPM for large networks.
Matching features - angle between the points, distance between the points etc.

Conclusion

In this paper, we learned about problems with data integration of different data sets and more issues with GIS and the solutions to solve these problems .

- PPM
- Geo-PPM
- Prioritized Geo-PPM.

Geospatial Information Integration for Authoritative and Crowd Sourced Road Vector

This Paper explores the technical issues associated with integrating unstructured crowd sourced data with authoritative national mapping data.

Introduction

- Crowdsourced data may provide a rich source of complementary information with the benefit of often more recent and frequent update than is the case for authoritative data .
- Many experiments and researches are taking place with main aim to develop appropriate methodologies for geospatial information linking and merging from disparate sources
- New ontologies (data sets) are created using existing ones to attain consistency so that data integration is possible.

Related Work

- Ontology linking or alignment. - It means to find correspondences between concepts, which have the same meaning and are from different ontologies.
- Some work needs to be done to find correspondence among ontologies,and judge which concepts are similar, overlapping or unique.
- Ontology merging is creating a single coherent ontology which contains information from all the sources.

Method

- Shp2OWL
- Ontology Builder
- Generating Merged Consistent Ontology
- Integration

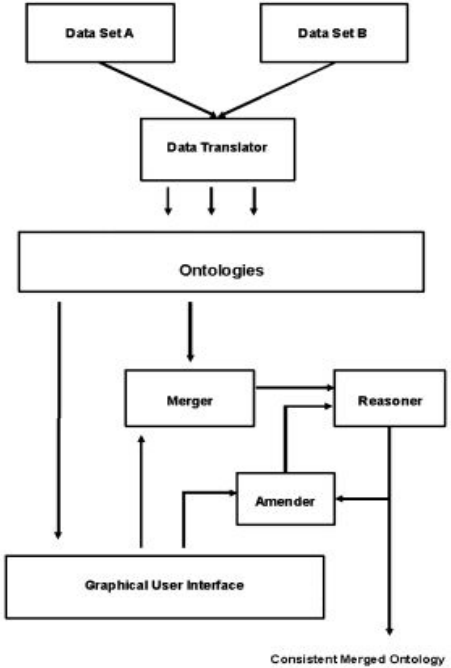


Figure 1 Architecture Design

Conclusion

This paper discusses the development of techniques for geographical information fusion from disparate sources for road vector data. An ontology based methodology was developed and implemented for geospatial information linking and merging.

Geo Crowdsourcing and accessibility for dynamic environments

This paper presents research on the development of tools to provide transient
obstacle information to blind, visually-impaired, and mobility-impaired individuals through crowdsourcing Geographic information.

Introduction

- Specially abled people have difficulties getting around due to obstacles found in our infrastructure.
- Accessibility mapping is mapping with knowledge of obstacles.
- That knowledge map of obstacle data is used to build an Navigation system for specially abled people
- This route calculation takes obstacles into account while suggesting route.

Methods

- Tactile Maps
- Notification system in Public Transit and places
- Using GIS Data and network Analysis for Routing
- Using VGI(volunteered geographic information) for Improved Routing

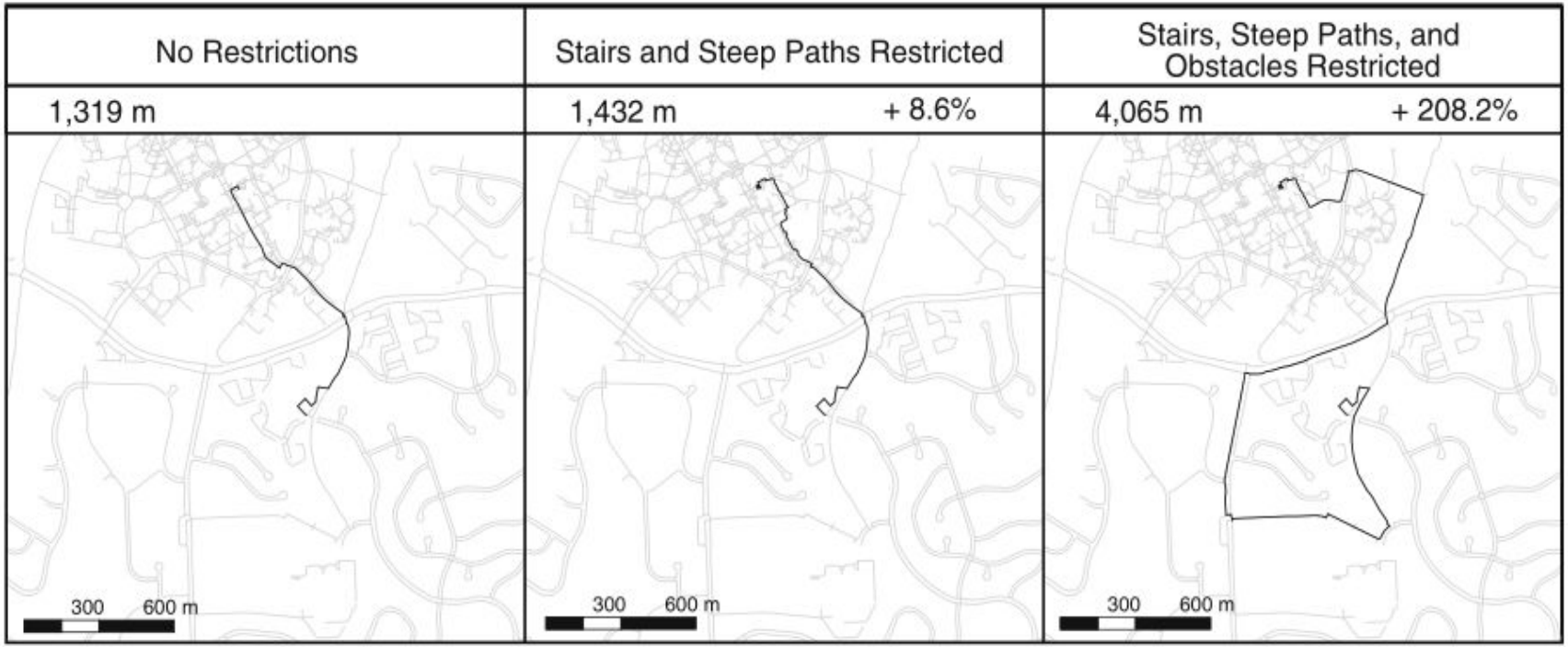


Fig:. Example routing scenario (Qin et. al. - 2015)

Conclusion

This paper discusses the development of techniques in Routing for Specially Abled people and Use of GIS and VGI data to improve life of Specially abled people

Some Application of Data Matching

- Disaster Relief
- Route Filtering
- Toll Plaza Markings
- Identity Verification
- and many-many more...

Our Problem

- We are working on a dataset (Foursquare) . It contains location of Hotels and Restaurants attained through check-ins of various visitors.
- Through this data set we are trying to correct or update locations of hotels that are marked inappropriately as of now.

Methodology to solve the problem

Our approach as of now:-

- Dataset extraction - To extract all the required information from the data set.
- Algorithm - To design an algorithm which will compare the information retrieved from data set and the relevant data to check if the retrieved data contains a relatively different coordinates for the same place.
- If yes, suggest an update in location.
- If not, no change.

Continued..

- The output will contain the previous locations of all the hotels and the updated locations will be marked using tags.
- It will also contain the weighted error or change in location of the Hotels.
- Algorithm is still in it’s early stages.

References:

[8-13] Xavier, Emerson, Francisco J. Ariza-López, and Manuel A. Ureña-Cámara. "A Survey of Measures and Methods for Matching Geospatial Vector Datasets." *ACM Computing Surveys (CSUR)* 49.2 (2016): 39.

[14-20] Bruns, Tom, and Max Egenhofer. "Similarity of spatial scenes." Seventh international symposium on spatial data handling. Delft, The Netherlands, 1996.

[21-27] Kolahdouzan, M. R., et al. "GeoMatchMaker: automatic and efficient matching of vector data with spatial attributes in unknown geometry systems." *Proc., UCGIS 2005 Summer Assembly* (2005).

References (Continued)

[28-32] Du, Heshan, et al. "Geospatial information integration for authoritative and crowd sourced road vector data." *Transactions in GIS* 16.4 (2012): 455-476.

[33-37] Geocrowdsourcing and accessibility for dynamic environments H Qin, RM Rice, S Fuhrmann, MT Rice, KM Curtin… - GeoJournal, 2016

Zook, Matthew, et al. "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake." World Medical & Health Policy 2.2 (2010): 7-33.

Questions ?

Thanks