Harnessing Heterogeneous Big Geospatial Data

Bo Yan, Gengchen Mai, Yingjie Hu, and Krzysztof Janowicz

Abstract The heterogeneity of geospatial datasets is a mixed blessing in that it theoretically enables researchers to gain a more holistic picture by providing different (cultural) perspectives, media formats, resolutions, thematic coverage, and so on, but at the same time practice shows that this heterogeneity may hinder the successful combination of data, e.g., due to differences in data representation and underlying conceptual models. Three different aspects are usually distinguished in processing big geospatial data from heterogeneous sources, namely geospatial data conflation, integration, and enrichment. Each step is a progression on the previous one by taking the result of the last step, extracting useful information, and incorporating additional information to solve specific questions. This chapter introduces and clarifies the scope and goal of each of these aspects, presents existing methods, and outlines current research trends.

1 Introduction

Among the often mentioned four characteristics, i.e., volume, variety, velocity, and veracity, of big data, variety is one of the most prominent in the geospatial domain. One grand challenge of consuming and utilizing big geospatial data is finding ways to utilize heterogeneous data, despite differences in their representations,

Bo Yan
University of California, Santa Barbara, CA, USA. e-mail: boyan@ucsb.edu

Gengchen Mai
University of California, Santa Barbara, CA, USA. e-mail: gengchen_mai@ucsb.edu

Yingjie Hu
University at Buffalo, NY, USA. e-mail: yhu42@buffalo.edu

Krzysztof Janowicz
University of California, Santa Barbara, CA, USA. e-mail: janowicz@ucsb.edu

resolution, data quality, semantics, data collection strategy, data cultures, and so forth (Janowicz, 2010). For example, remote sensing images are typically collected based on a field view, while most Points-of-Interest (POI) data are constructed using an object view. The vocabularies, often called feature type ontologies, used to categorize these POI vary between a handful and more than 1000 types making their integration challenging. In terms of data formats, geospatial data can be in the forms of unstructured data, semi-structured data, and structured data. A rich volume of geospatial data (such as place names and addresses) are contained in unstructured natural language texts, such as Wikipedia, news, books, and even in social media. Structured geospatial data have a well-defined schema and are contained in geospatial databases, shapefiles, gazetteers, and knowledge graphs, e.g., so-called Linked Data. Data will also show significant variation based on whether it is collected and maintained in the form of Volunteered Geographic Information (VGI) or by an authoritative source such as a government agency. In the age of Big Data, a research project often requires the use and integration of geospatial data from different sources which may have been collected using different approaches. The most common source of heterogeneity, however, are differences in semantics, i.e., in the conceptualizations related to used domain vocabulary such as `River`, `Poverty`, or `Neighborhood` (Harvey et al, 1999; Frank and Raubal, 1999; Bennett, 2001; Kuhn et al, 2014; Scheider and Kuhn, 2015) as well as cultural difference. For example, in Germany a bus stop is typically differentiated from other public spaces such as pavement because it usually has a distinct area with a roof. In some other countries, however, the concept of a bus stop may not exist at all (e.g., people in Turkey can stop the bus wherever they want to get on).

In this chapter, we review how big geospatial data can be conflated, integrated, and enriched (Kyriakidis et al, 1999; Arens et al, 1993; Samal et al, 2004; Lees and Ritman, 1991; Cobb et al, 1998; Fonseca et al, 2002). These terms themselves have different definitions across and even within communities. In the context of our overview, conflation is usually the initial step which involves combining and consolidating multiple instances of the same geographic entity with various lineage. To give an intuitive example from everyday experience, in order to gain a more comprehensive understanding of a Point of Interest, such as a particular restaurant, people may check multiple sources, e.g., website listings, social media reviews from multiple vendors such as Yelp, and even images, and cross-verify, combine, and mix them to provide more accurate and complete thematic (the type of restaurant, the food they serve, and other amenities for the restaurant), temporal (hours of operation), and spatial (coordinates and neighborhood) components. After conflating all this information, the integration step comes into play by which the data are combined, e.g. in the form of layers, into a larger project. One example for this integration step are map mashups, a term first made popular in 2004 (Batty et al, 2010). Almost all web maps we use today, e.g. Google Maps, are products of integration. These maps usually contain a base map and several thematic layers (such as the POI layer, terrain layer, satellite imagery layer, traffic layer, and transit layer). These layers can be in the form of vector data, raster data, or a combination of both. Different components of the map complement each other so that users can obtain a more comprehensive view of geographic entities from different perspectives. In a general geospatial data integration workflow, the conflation stage aims to retain accurate data, reconcile conflicting data, and minimize redundant data by considering different but overlapping sources; the integration stage aims to unify different aspects of the data after the initial conflation stage. Today, geospatial data enrichment plays an increasing role as an additional step following conflation and integration. With the current development of Web-accessible knowledge graphs, the barrier for interlinking and enriching geospatial data has become less severe. Geospatial data enrichment presents a convenient way of retrieving personalized, timely, and relevant geographic information. In this stage, methods in text mining or scene classification are also frequently used depending on whether text or image sources are considered. Machine learning and data mining models are essential to the success of geospatial data enrichment. For example, in semantic publishing[1,2], in order to link different articles to the same geographic entity, preprocessing steps typically include named entity recognition, place name disambiguation, and coreference resolution. Another example is event detection.[3,4] Such enriched data includes texts, temporal information, spatial footprints, and multimedia. Besides data mining approaches in obtaining events update in a geographic context, structured and standardized data markup guidelines can also facilitate the process. In this sense, geospatial data enrichment is the highlight of the marriage of top-down theory-driven and bottom-up data-driven approaches. It is worth noting that although there seems to be a linear order to which these three phases are conducted during the whole process, they do have some overlap and are not mutually exclusive, as shown in Fig. 1.

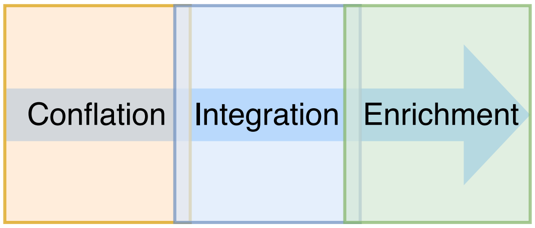

Fig. 1 The overlaps between the conflation process and the integration process as well as between the integration process and the enrichment process mean that there are no clear boundaries between them.

The content of this chapter is organized following the thread of conflation, integration, and enrichment. In Section 2, we lay out the major obstacles in conflating geospatial data from different sources, such as discrepancies in semantics, and examine previous research studies in tackling these challenges. In Section 3, we review

[1] https://en.wikipedia.org/wiki/Semantic_publishing

[2] http://now.ontotext.com

[3] http://eventregistry.org

[4] https://developers.google.com/search/docs/data-types/event

existing methods in spatial data integration, pointing out that the heterogeneous nature of geospatial data is, in fact, a blessing in disguise. In Section 4, we introduce a combination of top-down and bottom-up methods in enriching geospatial data and demonstrate the ways in which geospatial domain knowledge can benefit existing machine learning models. Throughout these sections, we discuss the Linked Data and geospatial semantics paradigm in which various knowledge graphs, such as DBpedia[5], Freebase[6], Wikidata[7], and LinkedGeoData[8], emerged in an attempt to faciliate the conflation, integration, and enrichment of geospatial data. While these semantically-rich datasets improve the interoperability, they also bring new challenges. In addition, we discuss the reciprocal relationship between geospatial data and various machine learning models: machine learning models can help integrate geospatial data, while geospatial data can be integrated into various other domains as complementary information sources for supporting cutting-edge models. In Section 5 , we summarize the three phases and conclude our chapter.

2 Geospatial Data Conflation

In most geographic information systems and services, geospatial data can be explored from both a map-centric view and a tabular-centric view (Mai et al, 2016). Research on facilitating geospatial data conflation can be organized from these two perspectives as well.

From a map view, geospatial data most often comes in two flavors: raster data and vector data. Accordingly, studies in geospatial data conflation have considered raster and raster conflation, raster and vector conflation, and vector and vector conflation (shown in Fig. 2). For raster and raster conflation, Lynch and Saalfeld (1985) described it as a problem of combining two raster maps to create a third map that is better than each of the two input maps in some regards, e.g., by reducing *NoData* cells. This definition considers geospatial data conflation as *map conflation* or *map compilation*. Lupien and Moreland (1987) decomposed the task into two generic problems, namely feature alignment and feature matching. In this context, raster pixels for points, lines, and polygons on the maps are referred to as features. Since different maps may have different projections and resolutions, feature alignment is applied to transform the coordinates of one map to fit another one. A common technique called *rubber-sheeting* is utilized to solve this problem, which is a transformation technique that preserves the topology of different features on the map. Typical rubber-sheeting algorithms utilize control points, triangulation, and other computational geometry concepts to provide computationally efficient ways to transform the maps and induce coincidence between different maps (Saalfeld, 1985; White Jr and

[5] https://wiki.dbpedia.org/

[6] https://en.wikipedia.org/wiki/Freebase

[7] https://www.wikidata.org

[8] http://linkedgeodata.org

Griffin, 1985; Gillman, 1985). Feature matching is then applied after feature alignment and the performance of feature matching depends on the strength of the feature alignment. Nearest neighbor pairings and intersection matching are commonly adopted for feature matching based on different criteria (Rosen and Saalfeld, 1985). The feature alignment and feature matching processes are often done in an iterative manner to increase the matched features. In order to solve the problem of positional discrepancy and the challenge of conflation maps with different levels of detail, Liu et al (2018) proposed a multiscale polygonal objectmatching approach, called the minimum bounding rectangle combinatorial optimization (MBRCO). This algorithm finds corresponding minimum bounding rectangles (MBRs) of matching pairs and aligns them to identify object-matching pairs.

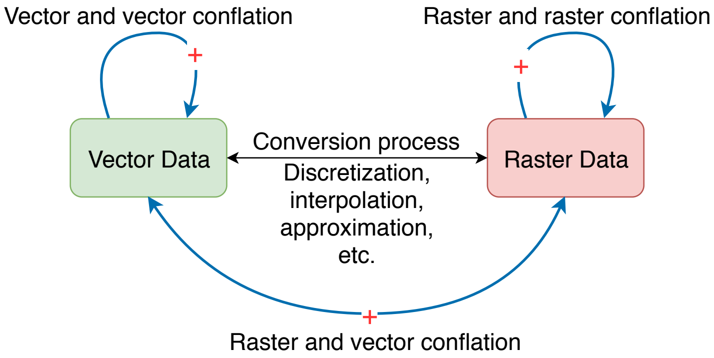

Fig. 2 Vector data and raster data are two commonly used types. Practically, a conversion process can be applied to switch between these two types. However, such a conversion is usually not lossless. As a result, three types of conflation, namely raster and raster conflation, vector and vector conflation, and raster and vector conflation, are studied in relevant research.

For raster and vector conflation, the core idea is to find the registration between the raster map data and the vector map data. For instance, Filin and Doytsher (2000) utilized linear features as seed entities for registering map data by detecting the counterpart elements, establishing correspondence between matched entities, and transforming the data. Chen et al (2004) used point pattern matching and exploited common vector datasets as ‘glue’ to automatically conflate street map imagery. Raster and vector data conflation is closely related to feature/object extraction from images and vector data update. As our geographic environment is constantly changing, updating vector maps automatically is of significance in order to provide the most relevant and accurate information. By successfully conflating satellite imagery with vector map data, street networks and other geographic features can be updated in a more timely manner. Baltsavias and Zhang (2005) automated the process of 3D road network reconstruction using aerial images and knowledge-based image analysis. Conflation of vector data and satellite imageries is also used for automatically geocoding satellite imageries (Hild and Fritsch, 1998).

For vector and vector conflation, existing research often examines three components, namely geometric component (Beeri et al, 2004; Foley, 1997), spatial relationship component (Fan et al, 2016), and attribute component (Hastings, 2008; Samal et al, 2004). Many conflation models also combine all of the three components, provided that such information is available in the dataset. For example, a hierarchical rule-based approach was proposed to take into account both geometric proximity and attribute information to match features (Cobb et al, 1998). A weighted average of positional measure, shape measure, directional measure, and topological measure was proposed as the criteria for point, linear, and areal feature matching and achieved better results compared with traditional distance-based counterparts (Fan et al, 2016). Li and Goodchild (2011) developed an optimization model to improve linear feature matching that could handle one-to-one, one-to-many, and one-to-none correspondence by making use of directed Hausdorff distance (an asymmetric dissimilarity metric). Instead of using a proximity-based matching approach, Song et al (2011) adopted a relaxation labeling approach by utilizing iterated local context updates to match the road intersections for vector road datasets. After initializing the point-to-point matching confidence matrix using the road connectivity information, in each iteration update, the relative distances between points are incorporated into the compatibility function. The proposed relaxation labeling approach yielded much better result than proximity matching approaches.

From a tabular view, geospatial data conflation can be performed based on place names, spatial and non-spatial relations, and other attributes. McKenzie et al (2014) used a weighted multi-attribute method that considered categorical information, activities, and topic similarity to match place entries in the gazetteers of Foursquare and Yelp. To reduce redundancy in a place database, Dalvi et al (2014) employed a language model that encapsulates domain knowledge (core words) and geographic knowledge (spatial context) to detect duplicate place entities. For example, in the place name “Fresca’s Peruvian Restaurant”, “Fresca’s” is the core word and “Peruvian Restaurant” is the description word. By accurately detecting core words and weighing them by their spatial context (such as the city or country these places are located in), their model outperformed other models. Another research area that is closely related to geospatial data conflation based on non-spatial attributes is place name disambiguation. The task is to identify the corresponding geographic entity given a place name in a text or other unstructured format. This problem is due to the one-to-many mapping between place names and geographic entities, which is also frequently encountered in geospatial data conflation. Hu et al (2014a) used Wikipedia and enhanced Term Frequency-Inverse Document Frequency (TF-IDF) with DBpedia terms to improve place name disambiguation. Ju et al (2016) integrated entity co-occurrence and topic modeling and outperformed benchmark systems such as DBpedia Spotlight and Open Calais in terms of F1 score and Mean Reciprocal Rank for place name disambiguation in short texts.

The recent marriage of geospatial data and the Linked Data paradigm (Kuhn et al, 2014) also increases the demand for data conflation from a tabular view. Linked Data uses a graph data model based on the Resource Description Framework (RDF) to describe both statements about the world and schema knowledge, called an ontology. RDF triples or statements consist of three parts: subject, predicate, and object. Subjects can be entities while objects can be both entities and literals. Predicates are the relationships between subjects and objects. For example, in the triple *:Santa_Barbara :isPartOf :California*, *:isPartOf* is the predicate that connects the subject *:Santa_Barbara* and the object *:California*. Geospatial Linked Data are often conflated using a tabular view. Related research has focused on reconciling data conflict, reducing data redundancy, and providing comprehensive data on both the ontology level and instance level. Geo-ontology alignment itself is a research topic that has attracted a lot of attention. Existing ontology matching or alignment systems include: Falcon (using a divide-and-conquer approach) (Hu and Qu, 2008), DSSim (using an agent-based framework) (Nagy et al, 2006), RiMOM (using a dynamic multi-strategy framework) (Li et al, 2009), AgreementMaker (Cruz et al, 2009), and so on. Although most ontology alignment systems are domain-agnostic, some research specifically took a geospatial perspective. Janowicz (2012) proposed an observation geo-ontology engineering framework that takes into account thematic, spatial, and temporal components. Zhu et al (2016a) implemented a feature engineering approach using spatial statistics in an attempt to align three major geo-ontologies, namely DBpedia Places[9], GeoNames[10], and Getty Thesaurus of Geographic Names (TGN)[11]. While these works emphasize aligning geographic concepts or place types from different data sources, Yan (2016) investigated how modeling bias would influence geo-ontologies and developed a data-driven method to detect these issues in order to harmonize the conflict between geo-ontology and the actual geographic entities that populate the ontology. Along the same line, Janowicz et al (2018) discussed the issue of bias in Linked Data from data, schema, and inferential perspectives, implying potential challenges for data conflation. On the instance level, Zhu et al (2016b) utilized spatial statistics and semantics to conflate entities in different geospatial Linked Datasets. Systems that focus particularly on the coreference resolution aspects of conflation include frameworks such as LIMES (Ngomo and Auer, 2011) and SILK (Volz et al, 2009).

3 Geospatial Data Integration

Geospatial data integration focuses on combining data about different themes or covering different geographic areas into a unified and semantically-consistent database for various geospatial applications (Abdalla, 2016). It should be differentiated from geospatial data conflation where the major goal is to reconcile the conflicts or duplications in datasets about the same theme and the same geographic areas (e.g., conflating the transportation network data from OpenStreetMap and Google Map within the same geographic area). In geospatial data integration, dif-

[9] http://mappings.dbpedia.org/server/ontology/classes/

[10] http://www.geonames.org/ontology/documentation.html

[11] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

ferent datasets often provide perspectives that are complementary to each other. For example, these datasets may provide different information, such as road network and POI, about the same region. The datasets to be integrated may also focus on the same theme but are about different geographic regions, and geospatial data integration can combine them and produce a whole dataset for the entire area. An example is to integrate the temperature measurements from different sensors which monitor the temperature in neighboring counties. Since the data from different sources can have varied interpretation (e.g. some sensors measure temperature in Celsius while others measure temperature in Fahrenheit), it is important to identify and accommodate data inconsistency during the integration process to achieve semantic interoperability. In other words, we need to ensure the semantic interoperability in order to produce a correctly integrated dataset.

Geospatial data integration is closely related to spatial data infrastructure (SDI) (Janowicz et al, 2010) and CyberGIS (Wang, 2010). One can distinguish five typical activities performed within SDIs in the context of Web-scale systems: finding, accessing, updating, processing, and visualizing geospatial data. In fact, these five activities are also among the most commonly used steps for geospatial data integration. In the following, we will discuss each of these five steps with an example of disaster mapping for Santa Barbara County after the 2017 Thomas Fire, the largest wildfire on record in California. In each step, we will emphasize the importance of semantic interoperability, and discuss how Semantic Web and Linked Data can help to ensure the propagation of semantics during the data integration process.

Discovering relevant data sources is the first step for geospatial data integration. In order to produce a disaster map for the Thomas Fire, multiple datasets for Santa Barbara county and Ventura County have to be retrieved, such as updated remote sensing images, transportation network data, wind direction, wind speed, and air pollution information from the sensor network, population data, and so on. Geographical information retrieval (GIR) systems (Jones and Purves, 2008) can support such a data discovery process. Query term recommendation and query expansion techniques (Delboni et al, 2007; Mai et al, 2018) are necessary to reformulate the search query in order to find relevant datasets. In this step, semantically-similar terms like similar geographic feature types can be suggested by using either ontology-based methods like SIM-DL or machine learning based method like Place2Vec (Yan et al, 2017).

Accessing the content and metadata of the retreived datasets is the next step for geospatial data integration. It should be noted that the information required by disaster mapping or other tasks are usually from different data sources, represented in different data models, or have different internal meanings. Accordingly, it is essential to have a clear semantic interpretation of the data to achieve semantic interoperability. For example, we have two datasets about wind directions. Dataset A has the wind direction for Goleta city with a *wind blow from* conceptualization while Dataset B has the wind direction for Santa Barbara city with a *wind blow to* conceptualization. In this data accessing step, we need to have a clear understanding on the semantics of the different datasets when pass these data to the following workflow. Otherwise, such semantic inconsistency can introduce serious error in the analysis results. Semantic annotation (Janowicz et al, 2010) on the data can help to clarify the semantic inconsistency and lead to a meaningful integration result, such as a wind direction map layer of Santa Barbara County.

Registration of geospatial data is another step for integration. Data conflicts or redundancy should be removed (if they were not done in data conflation). Also some new updates that have not made to the datasets need to be added. For example, some road segments were blocked during the Thomas Fire, and such real-time road connectivity information is typically not available in the original transportation network data. In the data registration process, semantically-supported integrity check also needs to be done to preserve data quality.

Processing geospatial data is a necessary step when the initial datasets do not directly satisfy the needs of an application. Consider the example of producing a map showing the air pollution in the next 24 hours of Thomas Fire. An air pollution dispersion model needs to be developed and executed based on the current wind direction, wind speed, the current air pollution distribution, the digital elevation map, and the fire locations. Let's assume that we have a geoprocessing service available for the air pollution dispersion model through an OGC Web Processing Service interface (WPS). The challenge here is not to understand the theory behind this service but to correctly interpret the intended meaning of the output of this service (Janowicz et al, 2010). Semantic inconsistency may occur when the semantics of the data in hand is not in line with the semantic definition of the input for the current service. For example, the wind speed data we have are measured in feet per second (ft/s) while the service requires the input wind speed data to be measured in miles per hour (mph). Using Semantic Web technologies to conceptualize the geoprocessing services (Scheider and Ballatore, 2018) can help to clarify the semantics of each service, improve their reusability, and achieve semantic interoperability among these services.

Visualizing the integrated dataset is the last step for geospatial data integration. After we have obtained various geographic layers such as the transportation layer, the predicted air pollution layer, the fire zones layer, the POI layer, we need to combine them to produce a visualization to end users. In this step, semantics also plays an important role because the visualization need to be aware of the semantics of different geographic features in order to select the appropriate styles and symbols for each element. For example, we cannot use the blue color to represent fire zones because they may be confused with water bodies which are also colored blue on maps.

Fig. 3 shows the five important steps in geospatial data integration which correspond to five activities in SDI. In conclusion, geospatial data integration combines heterogeneous data for addressing various spatial problems. Semantic interoperability and propagations are critical for effectively and correctly integrating geospatial datasets from different sources.

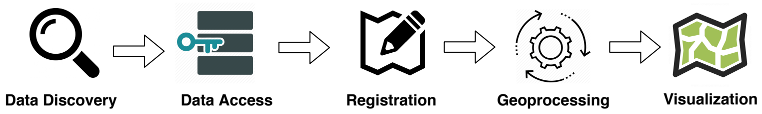

Fig. 3 Five important activities of spatial data infrastructure and geospatial data Integration

4 Geospatial Data Enrichment

Geospatial data enrichment aims to augment existing datasets with additional cross-domain information, typically streamed on-the-fly from an external API or knowledge graph endpoint. This can be seen as a way to contextualize data (Janowicz et al, 2019). Compared with geospatial data conflation and integration which are probably the bread and butter for harnessing heterogeneous big data, geospatial data enrichment is a relatively novel step that emerged within the last few years. However, it is playing an increasingly important role with the fast advancements in machine learning models as well as knowledge engineering in the context of global, Web-accessible knowledge graphs such as Linked Data that aims at breaking apart data silos. A simple example would be to access up-to-date demographic data from within a GIS while loading a shapefile about towns and cities. However, the term should be defined more broadly, e.g., including data about events, relevant research literature that uses the area currently loaded into a GIS as study area (Gahegan and Adams, 2014; Lafia et al, 2016), the biographies of historic figures and their travels, enriching 3D models with semantic annotations from social media (Jones et al, 2014), and so on.

There are many areas that may benefit from geospatial data enrichment. One is to enrich data streams with geosocial events that happened at certain locations during a particular time period. The challenge of event detection stems from the sheer amount of streaming data and the overwhelmingly large number of noise associated with them. Weng and Lee (2011) attempted to tackle these problems by proposing a clustering algorithm with wavelet-based signals using Twitter streams and showed promising result. In order to detect important geospatial events such as earthquakes, Sakaki et al (2010) examined Twitter streams, applied Kalman filtering and particle filtering, and developed a probabilistic spatiotemporal model to find the center and the trajectory of the event location. Pat and Kanza (2017) utilized geotagged posts in social media and developed a geosocial search system that effectively finds geospatial events. Zhu et al (2017) developed a deep learning framework to analyze geo-tagged videos in a real-time manner in order to recognize events and activities on the map. Balduini et al (2013) used their streaming Linked Data Framework to give city managers real-time access to event data for large-scale events and integrate the data with GIS functionality such as heatmaps.

Geospatial data enrichment is not limited to the domain of geography. The core idea of geospatial data enrichment lies in its intricate interplay between other domain areas or knowledge. In the following, we provide two examples that demonstrate the value of enriching datasets in other domains with geographic information. The first example is in scientometrics, and, in particular, spatial scientometrics which study the spatial aspect of science systems (e.g., scientific collaboration) (Frenken et al, 2009). By enriching scientometrics data with geospatial information such as the countries from which conference participants and authors came, the geographic distributions of co-authorship, the local or global scope of certain subdisciplines, and so on, researchers are able to explore spatial distributions of citations, spatial biases in collaborations, and differences between local and global citation impact. Frenken et al (2009) also pointed out that the affiliation information that is frequently used to provide the geographic knowledge has some issues. For example, it only reflects the home institute of a visiting scholar, and the granularity of this information is very coarse. Gao et al (2013) proposed a series of *s_indices* to evaluate the spatial impact of scientists and developed a framework that used the statistics of categorical places, spatiotemporal kernel density estimations, cartograms, distance distributions, and point-pattern analysis to identify spatiotemporal citation patterns. Hu et al (2013, 2014b) developed several visualization components using scientometrics and geospatial Linked Data to provide analysis functions for scientific knowledge discovery from a geographic perspective.

Semantic publishing provides a second example where geospatial data enrichment can benefit the analysis of the initial data. The idea of semantic publishing is to enhance online documents with linked metadata, which facilitates machines and softwares to understand the structure and consume the information in order to provide richer content. Many of the online documents contain spatial as well as temporal information. In order to extract semantics and create structured content, methods involving geographic information retrieval are frequently utilized. For instance, place name disambiguation is used to determine the corresponding geographic entity for geographic terms in the document and coreference resolution is used to connect different surface forms of the same geographic entity. By enriching these geographic entities with semantic content, users can either follow their nose to explore the information or the system can generate analytics and graphs to summarize the geographic knowledge. OpenCalais[12] is such a system that can highlight places mentioned in a document and link them to facts from an external knowledge graph.

Geospatial data enrichment can also improve machine learning models by enriching the input training data with additional geographic information. For example, aiming to provide better embeddings for map search and location recommendation, Yan et al (2017) devised an augmented spatial context-based algorithm that considered both local and global geographic context to learn embeddings for different place types and achieved better results based on three different evaluation schemes.

[12] http://www.opencalais.com/

Berg et al (2014) estimated the spatiotemporal priors given locations of bird species and developed an image classifier that can greatly improve the accuracy of categorizing highly similar species of birds. Tang et al (2015) explored different ways of encoding features extracted from the GPS information of images into Convolutional Neural Networks (CNN) and improved the mean average precision on classifying Flickr images by 7%. Along the same line, Yan et al (2018) incorporated location Bayesian priors based on spatial contexts into the state-of-the-art CNN models, such as ResNet and DenseNet, using different approaches, such as co-occurrence models and Long Short-Term Memory (LSTM), and improved the classification of the exterior and interior images of different places collected on Google Maps, Google Street View, and Yelp by over 40% in accuracy. In addition, Mai et al (2020) proposed a general-purpose location encoding model called Space2Vec. By combining it with the state-of-the-art image classification model, the hybrid model achieved better performances on fine-grained image recognition tasks. All these examples have shown that geospatial data and domain knowledge can further enhance machine learning models.

5 Summary

In this chapter, we discussed three aspects of harnessing the power heterogeneous geospatial data, namely conflation, integration, and enrichment. These three parts are not mutually exclusive and can overlap. Conflation deals with reconciling data from multiple sources to resolve inconsistencies and arrive at a new dataset that is improved in terms of spatial accuracy, feature completeness, logical consistency, and so on. Data integration focuses on combining datasets in meaningful ways, e.g., as part of larger workflows or to arrive at a new, more holistic data product. In many regards conflation can be considered as a strategy of data integration (Saalfeld, 1988), e.g., in the form of vector and raster conflation to correct street networks. However, conflation is just one such strategy, and, thus, we decided to address both separately here and also focus on *integration* as a driver of cross-thematic analysis. This view seems more in line with the recent thinking about heterogeneity in the context of big data. A similar concept that often occurs in discussions about the integration of geospatial data is semantic interoperability which studies how to ensure that services can exchange information meaningfully, i. e., in a way that preserves the intended interpretation of domain vocabularies. Finally, enrichment is the step of getting additional information, e.g., in the form of statements from a knowledge graph, about entities in the current dataset or project to provide additional contextual information. A typical example would be up-to-date demographics for a study area as well as events that happened in the past. This last enrichment step is part of ongoing research.

References

Abdalla R (2016) Geospatial data integration. In: Introduction to Geospatial Information and Communication Technology (GeoICT), Springer, pp 105–124

Arens Y, Chee CY, Hsu CN, Knoblock CA (1993) Retrieving and integrating data from multiple information sources. International Journal of Intelligent and Cooperative Information Systems 2(02):127–158

Balduini M, Della Valle E, DellAglio D, Tsytsarau M, Palpanas T, Confalonieri C (2013) Social listening of city scale events using the streaming linked data framework. In: International Semantic Web Conference, Springer, pp 1–16

Baltsavias E, Zhang C (2005) Automated updating of road databases from aerial images. International Journal of Applied Earth Observation and Geoinformation 6(3-4):199–213

Batty M, Hudson-Smith A, Milton R, Crooks A (2010) Map mashups, web 2.0 and the gis revolution. Annals of GIS 16(1):1–13

Beeri C, Kanza Y, Safra E, Sagiv Y (2004) Object fusion in geographic information systems. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment, pp 816–827

Bennett B (2001) What is a forest? on the vagueness of certain geographic concepts. Topoi 20(2):189–201

Berg T, Liu J, Woo Lee S, Alexander ML, Jacobs DW, Belhumeur PN (2014) Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2011–2018

Chen CC, Knoblock CA, Shahabi C, Chiang YY, Thakkar S (2004) Automatically and accurately conflating orthoimagery and street maps. In: Proceedings of the 12th annual ACM international workshop on Geographic information systems, ACM, pp 47–56

Cobb MA, Chung MJ, Foley III H, Petry FE, Shaw KB, Miller HV (1998) A rule-based approach for the conflation of attributed vector data. GeoInformatica 2(1):7–35

Cruz IF, Antonelli FP, Stroe C, Keles UC, Maduko A (2009) Using agreement maker to align ontologies for oaei 2009: overview, results, and outlook. In: Proceedings of the 4th International Conference on Ontology Matching-Volume 551, CEUR-WS. org, pp 135–146

Dalvi N, Olteanu M, Raghavan M, Bohannon P (2014) Deduplicating a places database. In: Proceedings of the 23rd international conference on World wide web, ACM, pp 409–418

Delboni TM, Borges KA, Laender AH, Davis Jr CA (2007) Semantic expansion of geographic web queries based on natural language positioning expressions. Transactions in GIS 11(3):377–397

Fan H, Yang B, Zipf A, Rousell A (2016) A polygon-based approach for matching openstreetmap road networks with regional transit authority data. International Journal of Geographical Information Science 30(4):748–764

Filin S, Doytsher Y (2000) A linear conflation approach for the integration of photogrammetric information and gis data. International archives of photogrammetry and remote sensing 33(B3/1; PART 3):282–288

Foley HA (1997) A multiple criteria based approach to performing conflation in geographical information systems. Tulane University

Fonseca FT, Egenhofer MJ, Agouris P, Câmara G (2002) Using ontologies for integrated geographic information systems. Transactions in GIS 6(3):231–257

Frank AU, Raubal M (1999) Formal specification of image schemata–a step towards interoperability in geographic information systems. Spatial Cognition and Computation 1(1):67–101

Frenken K, Hardeman S, Hoekman J (2009) Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics 3(3):222–232

Gahegan M, Adams B (2014) Re-envisioning data description using peirces pragmatics. In: International Conference on Geographic Information Science, Springer, pp 142–158

Gao S, Hu Y, Janowicz K, McKenzie G (2013) A spatiotemporal scientometrics framework for exploring the citation impact of publications and scientists. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, pp 204–213

Gillman D (1985) Triangulations for rubber sheeting. In: Proceedings of 7th International Symposium on Computer Assisted Cartography (AutoCarto 7), vol 199

Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C (1999) Semantic interoperability: A central issue for sharing geographic information. The annals of regional science 33(2):213–232

Hastings J (2008) Automated conflation of digital gazetteer data. International Journal of Geographical Information Science 22(10):1109–1127

Hild H, Fritsch D (1998) Integration of vector data and satellite imagery for geocoding. International Archives of Photogrammetry and Remote Sensing 32:246–251

Hu W, Qu Y (2008) Falcon-ao: A practical ontology matching system. Web semantics: science, services and agents on the world wide web 6(3):237–239

Hu Y, Janowicz K, McKenzie G, Sengupta K, Hitzler P (2013) A linked-data-driven and semantically-enabled journal portal for scientometrics. In: International Semantic Web Conference, Springer, pp 114–129

Hu Y, Janowicz K, Prasad S (2014a) Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In: Proceedings of the 8th workshop on geographic information retrieval, ACM, p 8

Hu Y, McKenzie G, Yang JA, Gao S, Abdalla A, Janowicz K (2014b) A linked-data-driven web portal for learning analytics: Data enrichment, interactive visualization, and knowledge discovery. In: LAK Workshops

Janowicz K (2010) The role of space and time for knowledge organization on the semantic web. Semantic Web 1(1, 2):25–32

Janowicz K (2012) Observation-driven geo-ontology engineering. Transactions in GIS 16(3):351–374

Janowicz K, Schade S, Bröring A, Keßler C, Maué P, Stasch C (2010) Semantic enablement for spatial data infrastructures. Transactions in GIS 14(2):111–129

Janowicz K, Yan B, Regalia B, Zhu R, Mai G (2018) Debiasing knowledge graphs: Why female presidents are not like female popes. In: Proceedings of the 17th International Semantic Web Conference

Janowicz K, Gao S, McKenzie G, Hu Y, Bhaduri B (2019) Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science 0(0):1–12, DOI 10.1080/13658816.2019.1684500

Jones C, Rosin P, Slade J (2014) Semantic and geometric enrichment of 3d geospatial models with captioned photos and labelled illustrations. In: Proceedings of the Third Workshop on Vision and Language, pp 62–67

Jones CB, Purves RS (2008) Geographical information retrieval. International Journal of Geographical Information Science 22(3):219–228

Ju Y, Adams B, Janowicz K, Hu Y, Yan B, McKenzie G (2016) Things and strings: improving place name disambiguation from short texts by combining entity cooccurrence with topic modeling. In: European Knowledge Acquisition Workshop, Springer, pp 353–367

Kuhn W, Kauppinen T, Janowicz K (2014) Linked data-a paradigm shift for geographic information science. In: International Conference on Geographic Information Science, Springer, pp 173–186

Kyriakidis PC, Shortridge AM, Goodchild MF (1999) Geostatistics for conflation and accuracy assessment of digital elevation models. International Journal of Geographical Information Science 13(7):677–707

Lafia S, Jablonski J, Kuhn W, Cooley S, Medrano FA (2016) Spatial discovery and the research library. Transactions in GIS 20(3):399–412

Lees BG, Ritman K (1991) Decision-tree and rule-induction approach to integration of remotely sensed and gis data in mapping vegetation in disturbed or hilly environments. Environmental Management 15(6):823–831

Li J, Tang J, Li Y, Luo Q (2009) Rimom: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and data Engineering 21(8):1218–1232

Li L, Goodchild MF (2011) An optimisation model for linear feature matching in geographical data conflation. International Journal of Image and Data Fusion 2(4):309–328

Liu L, Zhu X, Zhu D, Ding X (2018) M: N object matching on multiscale datasets based on mbr combinatorial optimization algorithm and spatial district. Transactions in GIS 22(6):1573–1595

Lupien AE, Moreland WH (1987) A general approach to map conflation. In: Proceedings of 8th International Symposium on Computer Assisted Cartography (AutoCarto 8), Citeseer, pp 630–639

Lynch MP, Saalfeld AJ (1985) Conflation: Automated map compilationa video game approach. In: Proceedings Auto-Carto, vol 7, pp 343–352

Mai G, Janowicz K, Hu Y, McKenzie G (2016) A linked data driven visual interface for the multi-perspective exploration of data across repositories. In: VOILA@ ISWC, pp 93–101

Mai G, Janowicz K, Yan B (2018) Combining text embedding and knowledge graph embedding techniques for academic search engines. In: Semdeep/NLIWoD@ ISWC, pp 77–88

Mai G, Janowicz K, Yan B, Zhu R, Cai L, Lao N (2020) Multi-scale representation learning for spatial feature distributions using grid cells. In: The Eighth International Conference on Learning Representations

McKenzie G, Janowicz K, Adams B (2014) A weighted multi-attribute method for matching user-generated points of interest. Cartography and Geographic Information Science 41(2):125–137

Nagy M, Vargas-Vera M, Motta E (2006) Dssim-ontology mapping with uncertainty. In: Proceedings of 1st International Workshop on Ontology Matching, Online

Ngomo ACN, Auer S (2011) Limes-a time-efficient approach for large-scale link discovery on the web of data. In: IJCAI, pp 2312–2317

Pat B, Kanza Y (2017) Where's waldo?: Geosocial search over myriad geotagged posts. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, p 37

Rosen B, Saalfeld A (1985) Match criteria for automatic alignment. In: Proceedings of 7th international symposium on computer-assisted cartography (Auto-Carto 7), pp 1–20

Saalfeld A (1985) A fast rubber-sheeting transformation using simplicial coordinates. The American Cartographer 12(2):169–173

Saalfeld A (1988) Conflation automated map compilation. International Journal of Geographical Information System 2(3):217–228

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web, ACM, pp 851–860

Samal A, Seth S, Cueto K (2004) A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science 18(5):459–489

Scheider S, Ballatore A (2018) Semantic typing of linked geoprocessing workflows. International Journal of Digital Earth 11(1):113–138

Scheider S, Kuhn W (2015) How to talk to each other via computers: Semantic interoperability as conceptual imitation. In: Applications of Conceptual Spaces, Springer, pp 97–122

Song W, Keller JM, Haithcoat TL, Davis CH (2011) Relaxation-based point feature matching for vector map conflation. Transactions in GIS 15(1):43–60

Tang K, Paluri M, Fei-Fei L, Fergus R, Bourdev L (2015) Improving image classification with location context. In: Proceedings of the IEEE international conference on computer vision, pp 1008–1016

Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Silk-a link discovery framework for the web of data. LDOW 538

Wang S (2010) A cybergis framework for the synthesis of cyberinfrastructure, gis, and spatial analysis. Annals of the Association of American Geographers 100(3):535–557

Weng J, Lee BS (2011) Event detection in twitter. ICWSM 11:401–408

White Jr MS, Griffin P (1985) Piecewise linear rubber-sheet map transformation. The American Cartographer 12(2):123–131

Yan B (2016) A Data-Driven Framework for Assisting Geo-Ontology Engineering Using a Discrepancy Index. University of California, Santa Barbara

Yan B, Janowicz K, Mai G, Gao S (2017) From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, p 35

Yan B, Janowicz K, Mai G, Zhu R (2018) xnet+sc: Classifying places based on images by incorporating spatial contexts. In: LIPIcs-Leibniz International Proceedings in Informatics, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, vol 114

Zhu R, Hu Y, Janowicz K, McKenzie G (2016a) Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. Transactions in GIS 20(3):333–355

Zhu R, Janowicz K, Yan B, Hu Y (2016b) Which kobani? a case study on the role of spatial statistics and semantics for coreference resolution across gazetteers. In: International Conference on GIScience Short Paper Proceedings, vol 1

Zhu Y, Liu S, Newsam S (2017) Large-scale mapping of human activity using geotagged videos. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, p 68