International Journal of
Geo-Information

Article

HGeoKG: A Hierarchical Geographic Knowledge Graph for Geographic Knowledge Reasoning

Tailong Li [1], Renyao Chen [2], Yilin Duan [2], Hong Yao [1,2,3,4], Shengwen Li [2,3,4] and Xinchuan Li [2,3,4,*]

[1] School of Future Technology, China University of Geosciences, Wuhan 430074, China; ltl@cug.edu.cn (T.L.); yaohong@cug.edu.cn (H.Y.)
[2] School of Computer Science, China University of Geosciences, Wuhan 430074, China; cryao@cug.edu.cn (R.C.); duanyl@cug.edu.cn (Y.D.); swli@cug.edu.cn (S.L.)
[3] State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan 430074, China
[4] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China
* Correspondence: lixinchuan@cug.edu.cn

Abstract: The Geographic Knowledge Graph (GeoKG) serves as an effective method for organizing geographic knowledge, playing a crucial role in facilitating semantic interoperability across heterogeneous data sources. However, existing GeoKGs are limited by a lack of hierarchical modeling and insufficient coverage of geographic knowledge (e.g., limited entity types, inadequate attributes, and insufficient spatial relationships), which hinders their effective use and representation of semantic content. This paper presents HGeoKG, a hierarchical geographic knowledge graph that comprehensively models hierarchical structures, attributes, and spatial relationships of multi-type geographic entities. Based on the concept and construction methods of HGeoKG, this paper developed a dataset named HGeoKG-MHT-670K. Statistical analysis reveals significant regional heterogeneity and long-tail distribution patterns in HGeoKG-MHT-670K. Furthermore, extensive geographic knowledge reasoning experiments on HGeoKG-MHT-670K show that most knowledge graph embedding (KGE) models fail to achieve satisfactory performance. This suggests the need to accommodate spatial heterogeneity across different regions and improve the embedding quality of long-tail geographic entities. HGeoKG serves as both a reference for GeoKG construction and a benchmark for geographic knowledge reasoning, driving the development of geographical artificial intelligence (GeoAI).

Keywords: geographic knowledge graph; knowledge reasoning; hierarchical structure; long-tail distribution; spatial heterogeneity

check for updates

Academic Editors: Wolfgang Kainz and Huayi Wu

Received: 30 October 2024
Revised: 23 December 2024
Accepted: 1 January 2025
Published: 3 January 2025

Citation: Li, T.; Chen, R.; Duan, Y.; Yao, H.; Li, S.; Li, X. HGeoKG: A Hierarchical Geographic Knowledge Graph for Geographic Knowledge Reasoning. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 18. https://doi.org/10.3390/ijgi14010018

1. Introduction

The Knowledge Graph (KG), as a structured form of knowledge, plays a pivotal role in enabling semantic interoperability across multi-source heterogeneous data [1,2], and has demonstrated significant capabilities in various artificial intelligence applications [3–5]. In recent years, the geographic knowledge graph (GeoKG) has been proposed, which organizes, links, and infers geospatial knowledge, and serves various geographical artificial intelligence (GeoAI) applications, such as geographic spatiotemporal question answering systems [6], economic indicator prediction [7], weather prediction [8], traffic forecasting [9], human activity trajectory mining [10], point of interest (POI) recommendation [11], geographic entity retrieval [12], and urban functional area detection [13].

The previous GeoKGs can be categorized into three types based on the differences in data sources used during the construction process. A detailed comparison of these GeoKGs is provided in Table 1.

Table 1. Detailed comparison of GeoKGs.

GeoKG	Data Source	Ontology Design	Entity Spatial-Type Coverage	Attributes	Spatial Relationships	Downstream Applications	Data Scale (Million)
YAGO2 [14]	Wikipedia, GeoNames, WordNet	YAGO2	Point	Common attributes from Wikipedia	/	/	447
Clinga [15]	Chinese Baidu Baike, DBpedia, GeoNames	Clinga	Point, Line, Polygon	Common attributes from Baidu Baike	/	/	75
NCGKB [16]	Chinese Wikipedia	NCGKB	Polygon (Administrative regions)	Common attributes from Chinese Wikipedia	Adjacent	/	0.1
YAGO2Geo [17]	YAGO2, referenced geospatial datasets (GAG, GADM, parts of OSM)	GAG Ontology	Point, Line, Polygon	Common attributes from Wikipedia	/	/	447
GeoKG [18]	Baidu Baike, vector spatial data	GeoKG	Point, Line, Polygon	Spatial attributes, Baike attributes	Adjacent	/	1
LinkedGeoData [19]	OSM, DBpedia, GeoNames	LinkedGeoData	Point, Line, Polygon	OSM attributes	/	/	300
CrowdGeoKG [20]	OSM, Wikidata	OSMonto [21]	Point, Line	OSM attributes, Baike attributes	/	/	5
WorldKG [22]	OSM, Wikidata, DBpedia Ontology	WorldKG	Point	OSM attributes, Baike category attributes	/	Knowledge graph query	100
GeoKG [23]	Limited specialized geographic texts	GeoKG	/	Geographic text semantic relations	/	Knowledge graph query	0.1
GEKG [24]	Limited specialized geographic texts	GEKG	/	Geographic text semantic relations	/	Knowledge graph query	0.1
AugKG [25]	Limited specialized geographic texts	AugKG	/	Geographic text semantic relations	/	Knowledge graph query	0.1
HGeoKG (Ours)	OSM, referenced geospatial data (census tracts, electoral districts, etc.)	HGeoKG	Point, Line, Polygon	OSM attributes, category attributes	Regional hierarchical relations, typed adjacency, intersection, containment, location	Knowledge graph reasoning	0.67

(1) GeoKGs based on general encyclopedias: These GeoKGs obtain geographic items from large-scale general-purpose internet encyclopedia data, such as YAGO [26], Wikidata [27], and Freebase [28]. They are rich in attribute information and provide common sense geographic knowledge. However, the geographic entities are sparsely distributed, with a lack of spatial relations between entities. Additionally, the coverage of geographic entity types and regions is limited, making it difficult to comprehensively represent geospatial semantics.

(2) GeoKGs extracted from geographic texts: These GeoKGs focus on specialized geographic concepts and the interactions between geographic features, such as GeoKG [23], GEKG [24], and AugKG [25]. This type of GeoKG offers in-depth theoretical support and covers semantic relationships found in geographic texts, making it useful for research and applications in specific fields. However, due to limitations in data acquisition and coverage, these GeoKGs tend to have a small number of items and limited coverage of entity types and regions, making it challenging to meet broader geographic knowledge demands.

(3) GeoKGs based on OpenStreetMap (OSM): These GeoKGs rely on abundant open geographic information resources, such as LinkedGeoData [19], CrowdGeoKG [20], and WorldKG [22], which cover a wide range of geographic entities and attribute information. They excel in terms of geographic entity coverage and the richness of attribute information, yet they still fall short in terms of relationships between geographic entities and the repre-

sentation of spatial semantics, lacking the comprehensive modeling of spatial relationships and hierarchical structures.

Overall, existing GeoKGs, to varying extents, cover certain aspects of geospatial semantic information, and each type of GeoKG has its strengths in representing geographic knowledge. However, these models still suffer from limited geographic entity coverage, insufficient attribute information, and a lack of spatial relationships. As a result, they fail to comprehensively model key geographic semantics, hindering the effective utilization and representation of the rich semantics and prominent patterns in geographic knowledge.

This paper proposes a hierarchical GeoKG (HGeoKG) that encompasses most types of geographic entities and relationships with rich attribute information. HGeoKG can be used to evaluate and advance geographic knowledge embedding techniques, and thus more effectively supports downstream GeoAI applications.

The contributions of this study are as follows:

1. This paper proposes the concept of HGeoKG, the first geographic knowledge graph that integrates rich attributes, spatial relationships, and regional hierarchical semantics, thereby providing a comprehensive representation of geographic knowledge.

2. This paper proposed a method for constructing HGeoKG and presented the dataset named HGeoKG-MHT-670K. Through statistical analysis of this dataset, we revealed significant regional heterogeneity and long-tail distribution patterns, providing valuable insights into the intrinsic structure and distribution characteristics of GeoKGs.

3. This paper conducted extensive knowledge graph reasoning experiments on HGeoKG-MHT-670K. The experimental results indicate that the regional heterogeneity of the dataset poses challenges for Knowledge Graph Embedding (KGE) models to achieve consistent performance across all regions, highlighting the necessity for differentiated modeling strategies tailored to regional differences. Additionally, the geographic long-tail distribution pattern leads to a decline in embedding quality when handling low-popularity entities, underscoring the urgent need to enhance model capabilities in managing such data. This study provides strong empirical support for the further optimization and application of GeoKGs.

2. Related Work

In this study, a new classification framework for GeoKGs is presented, based on the geographic data sources they use: GeoKGs based on internet encyclopedias, GeoKGs extracted from geographic texts, and GeoKGs based on OSM.

2.1. Internet Encyclopedias-Based GeoKGs

With the development of large-scale general-purpose internet encyclopedic data, some studies have highlighted the rich geographic semantic information embedded in general encyclopedic data, such as geographic entities and spatial location information. These GeoKGs are derived from subsets of large general knowledge bases, including YAGO [26], Wikidata [27], and Freebase [28], which contain geographic knowledge. For the representation of geospatial data, DBpedia [29] offers latitude and longitude values for various geographic entities. YAGO2 [14], Clinga [15], and NCGKB [16] are knowledge bases with human geography knowledge derived from Wikipedia, Baidu Baike, and Chinese Wikipedia, respectively. Additionally, GeoKG [18] incorporates vector geographic datasets into Baidu Baike, adding precise coordinates and spatial relationships to the general GeoKG. YAGO2geo [17], based on YAGO2 and reference geospatial datasets such as Greek administrative geography (GAG), the global administrative areas database (GADM), and OSM, focuses primarily on administrative regions and reuses existing ontologies from the GAG dataset, leading to limited coverage of geographic entity types.

2.2. Geographic Text-Based GeoKGs

Geographic semantic information in geographic knowledge is complex and diverse. Specialized geographic texts encompass detailed semantic information regarding the interactions between geographic entities. Some studies, leveraging these semantic characteristics, have designed conceptual models of GeoKGs that theoretically represent geographic knowledge more effectively. Among these, GeoKG [23] is a formalized representation of geographic knowledge, extending Attribute Language with Complements (ALC) description logic. It focuses on spatiotemporal knowledge, using entity states to represent changes in each geographic object. Zheng proposed a Geographic Evolution Knowledge Graph (GEKG), which is based on spatiotemporal processes and establishes a hierarchical cube model structure [24]. AugGKG [25], an augmented GeoKG, utilizes the GeoSOT global subdivision grid model and time-slice subgraph architecture to discretize and normalize spatiotemporal data within the knowledge graph. These models, through case studies based on knowledge graph queries, have demonstrated their capability to represent the spatiotemporal characteristics of geographic knowledge.

2.3. OpenStreetMap-Based GeoKGs

OSM is a rich source of open geographic information, encompassing a vast array of geographic entities. The representation of these entities (e.g., buildings, mountains, rivers) is characterized by high heterogeneity, diversity, and incompleteness. With the growth of large-scale open crowdsourced geographic data like OSM, some research has focused on utilizing the geographic information from OSM to construct GeoKGs.

Early studies developed ontologies suited to the structure of OSM data: OSMOnto [21] describes an ontology for OSM tags (e.g., (building, yes)), representing a class hierarchy extracted from OSM keys and values. OSM Semantic Network [30] contains RDF triples extracted from OSM tags available on the OSM Wiki website. Although OSMOnto and the OSM Semantic Network extracted a significant number of concepts, they did not include any geographic entity instances. Subsequently, LinkedGeoData [19] converted OSM data into an RDF knowledge graph. This is based on a formal ontology created using OSM tags and keys, offering simplified mappings between OSM data and classes and attributes from other data sources. CrowdGeoKG [20] extracted different types of entities from OSM and enriched them with human geographic knowledge from Wikidata. WORLDKG [22], by analyzing a large set of heterogeneous OSM data tags, distilled a class hierarchy of OSM elements. After a degree of manual filtering, geographic entities in OSM were classified into a top-down hierarchical structure, covering various geographic categories and linking geographic entities to specific classes in Wikidata and the DBpedia ontology. However, WORLDKG mainly utilizes point-type entities from OSM and their attributes to construct GeoKGs, lacking coverage of other geographic entity types such as line and polygon entities.

In summary, the existing GeoKG has limitations in several key areas. First, GeoKG based on internet encyclopedias has deficiencies in geographic entity types and spatial coverage, typically including only common attributes and lacking in-depth descriptions of geographically specific attributes. Second, the conceptual model of GeoKG extracted from geographic texts requires high precision and a breadth of geographic data, which is often dispersed across specialized texts in the geographic domain. These texts contain fewer items, and the extraction of entities and relationships is difficult, greatly limiting its scalability and applicability. Furthermore, some knowledge graphs have incomplete coverage of spatial types, typically supporting only one or two spatial types, such as points, lines, or polygons. Additionally, most knowledge graphs do not explicitly model spatial relationships, with only a few providing basic adjacency relations and lacking support for complex spatial relationships, such as inclusion or intersection. Finally, many GeoKGs fail to effectively model hierarchical relationships between geographic entities (e.g., the hierarchical structure of administrative divisions), which limits their performance in applications requiring hierarchical reasoning.

3. HGeoKG

This section introduces the schema design and data construction methods of HGeoKG. First, Section 3.1 provides an overview of HGeoKG. Then, Section 3.2 discusses the ontology schema, regional hierarchical structure, and multi-granular relationships in HGeoKG. Finally, Section 3.3 details the specific construction process of HGeoKG's data layer.

3.1. Overview

Figure 1 presents the overall framework of HGeoKG, comprising two core layers: Schema and Data, organized according to the workflow from data to knowledge graph construction. The Schema layer consists of three components: ontology design, spatial relationship hierarchical structure design, and regional hierarchical structure design. This layer defines the overall structure and organizational rules of the geographic knowledge graph, providing theoretical guidance and framework support for subsequent data processing. The Data layer illustrates the specific construction process of the geographic knowledge graph, with data sources including administrative boundary data, OSM polygon data, OSM line data, and OSM point data. Data processing is primarily divided into three steps: first, the extraction of geographic entities and attributes; second, the extraction of spatial relationships; and finally, the extraction of regional layering and partitioning. This series of steps ultimately achieves the data generation and construction of the geographic knowledge graph.

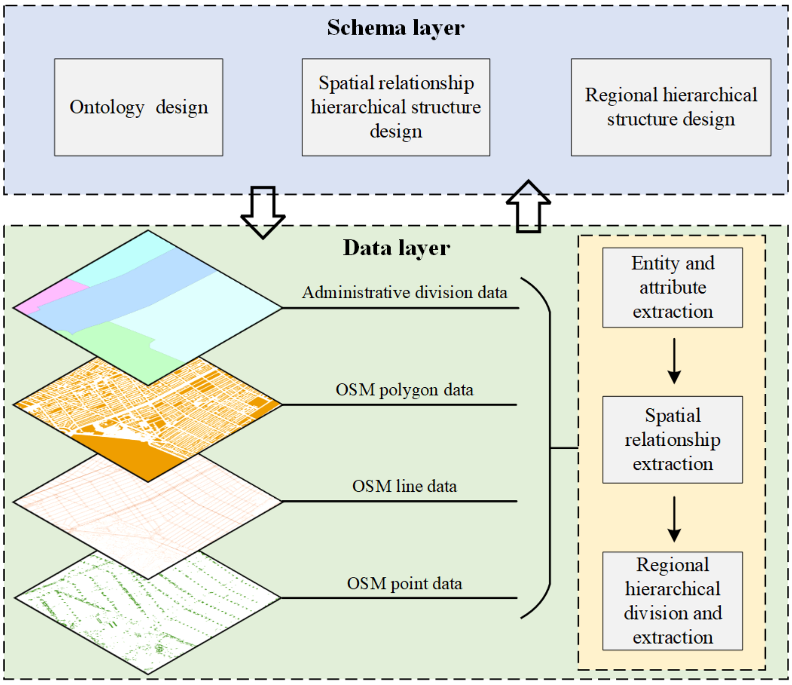

Schema layer
Ontology design
Spatial relationship hierarchical structure design
Regional hierarchical structure design
Data layer
Administrative division data
OSM polygon data
OSM line data
OSM point data
Entity and attribute extraction
Spatial relationship extraction
Regional hierarchical division and extraction

Figure 1. Overview of HGeoKG.

3.2. Schema Layer

This subsection conceptualizes and implements the schema layer to integrate geographic knowledge.

3.2.1. HGeoKG Ontology

The ontology of HGeoKG, as shown in Figure 2, illustrates the attribute information of geographic entities and their spatial relationships. In our ontology design, the attributes of geographic entities are categorized into two types: general attributes and heterogeneous attributes. General attributes include the spatial types of geographic entities and the common sense categories. The definitions of geographic entity categories in Table 2 are based on the official OSM documentation (https://download.geofabrik.de/osm-data-in-gis-formats-free.pdf (accessed on 1 January 2024)). Specifically, "Point", "Line", and "Polygon" represent point, line, and polygon geometric shapes, respectively. The subclasses within each category (such as roads, railways, buildings, etc.) are directly derived from the OSM classification system to ensure their broad applicability. Heterogeneous attributes provide unique descriptive features for each geographic entity, which are not shared by all entities. The types of spatial relationships in Table 3 are derived from the topological spatial relationship model known as the Dimensionally Extended 9-Intersection Model (DE-9IM) (http://docs.geotools.org/latest/userguide/library/jts/dim9.html (accessed on 1 January 2024)). The spatial relationships between geographic entities describe the spatial semantic connections between them. This ontology design enables HGeoKG to represent geographic knowledge with greater accuracy and comprehensiveness.

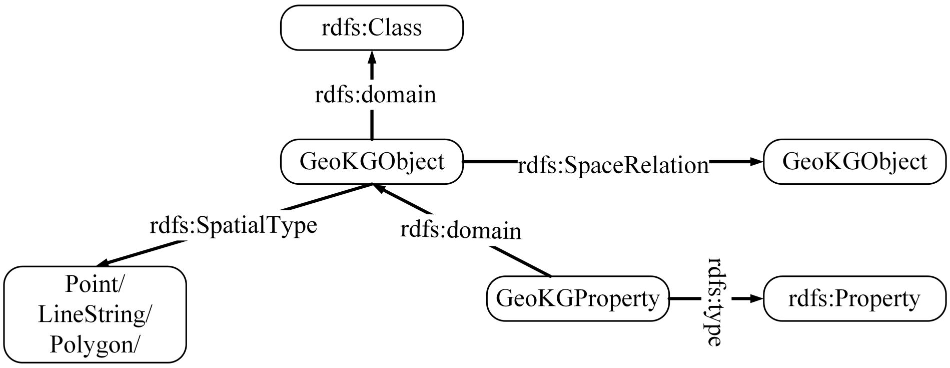


Figure 2. The HGeoKG ontology.

Table 2. Categories of geographic entities.

Point	Line	Polygon
places, pois, pofw, natural, traffic, transport	roads, railways, waterways	buildings, landuse, water

Table 3. Types of spatial relationships

Spatial Relationship	Point	Line	Polygon
Point	Adjacent	Adjacent	Contains, Contained by, Adjacent
Line	/	Adjacent, Intersects	Contains, Contained by, Adjacent, Intersects
Polygon	/	/	Contains, Contained by, Adjacent

3.2.2. Spatial Relationship Hierarchical Structure

Considering the spatial relationships based on entity types enables the more effective modeling of potential human, commercial, and economic semantic connections between geographic entities, which are not easily revealed by distance-based spatial relationships alone. In this study, we explicitly model these latent semantics, with spatial relationships serving as the bridge that carries these hidden meanings. A straightforward example illustrates this: typically, stationery stores are located near primary and secondary schools. The spatial relationship of "school-adjacent-stationery store" not only uncovers the commercial connection between schools and stationery stores but also, through the explicit modeling of this relationship, allows for the more effective use of the semantic information inherent in the geographic entities themselves. These type-based spatial relationships can be viewed as prior semantic rules extracted from the data, revealing semantic content that pure distance-based spatial relationships cannot express. This modeling approach plays a crucial role in achieving a comprehensive semantic representation of geographic knowledge.

The general spatial relationship types, as shown in Table 3, simply reflect the spatial semantics between geographic entities. In this study, the spatial types of geographic entities (point, line, polygon) and their common sense types, as shown in Table 2, are integrated into the representation of spatial relationships. This has led to an extension of these relationships at different levels of granularity, and the construction of a hierarchical structure, as shown in Figure 3. As shown in Figure 3a, the coarse-grained spatial relationships integrate the spatial type semantics of two geographic entities. Correspondingly, as illustrated in Figure 3b, the fine-grained spatial relationships incorporate the common sense type semantics of the same entities. This explicit modeling approach of hierarchical spatial relationships, which combines entity types of different granularities, enables a more specific and accurate expression of spatial semantics in geographic knowledge.

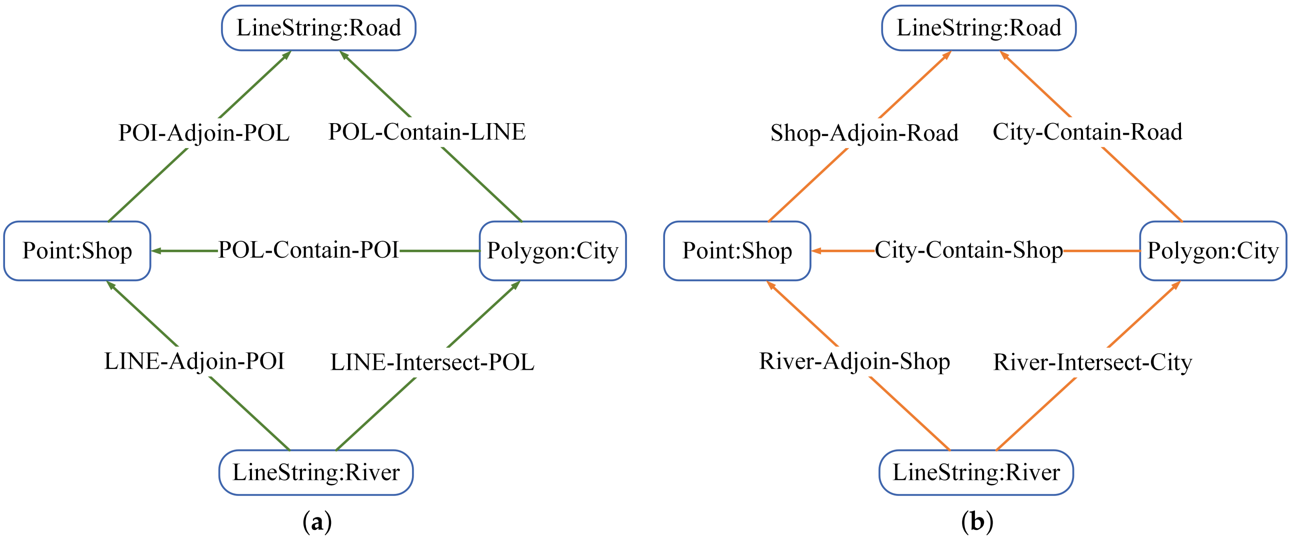


Figure 3. Hierarchical structure of spatial relationships. (**a**) Coarse-grained spatial relationship; (**b**) fine-grained spatial relationship.

We incorporated geographic entity type information into the spatial relationship modeling. Based on the richness of the entity type information, spatial relationships were categorized into different granular hierarchical structures. In the subsequent experiments detailed in Section 4.3.5, we evaluated the impact of spatial relationships at various hierarchical levels on geographic knowledge embedding learning, further validating the expressive capability of the knowledge graph.

3.2.3. Regional Hierarchical Structure

Spatial heterogeneity reflects the variation and diversity of geographic phenomena, exhibiting inherently uncontrollable spatial patterns. To promote the study of geographic knowledge heterogeneity across regions, this paper proposes a hierarchical regional structure based on real-world administrative division data, as illustrated in Figure 4. HGeoKG first partitions OSM data into larger regions using coarse-grained administrative division data and subsequently further subdivides these large regions with finer-grained administrative division data, thereby forming a more detailed regional hierarchy. This hierarchical regional structure not only reveals spatial heterogeneity within each level and the interactions between coarse- and fine-grained regions but also enables HGeoKG to accurately analyze geographic entities and their relationships within each region. This facilitates more refined modeling, thereby comprehensively enhancing the accuracy and granularity of geographic knowledge representation.

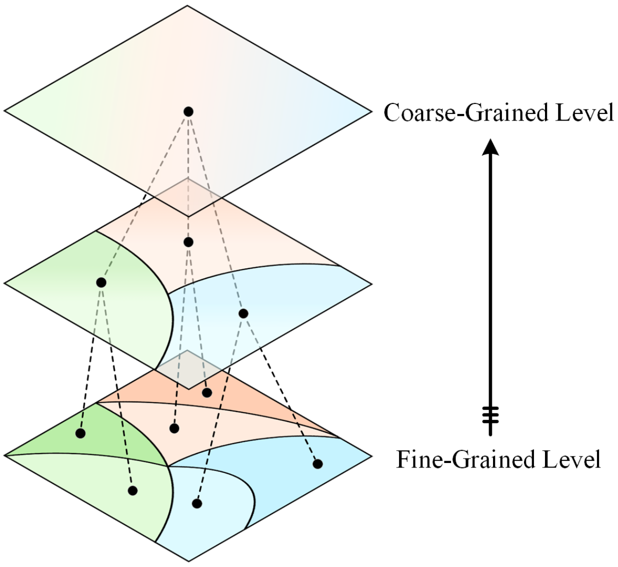


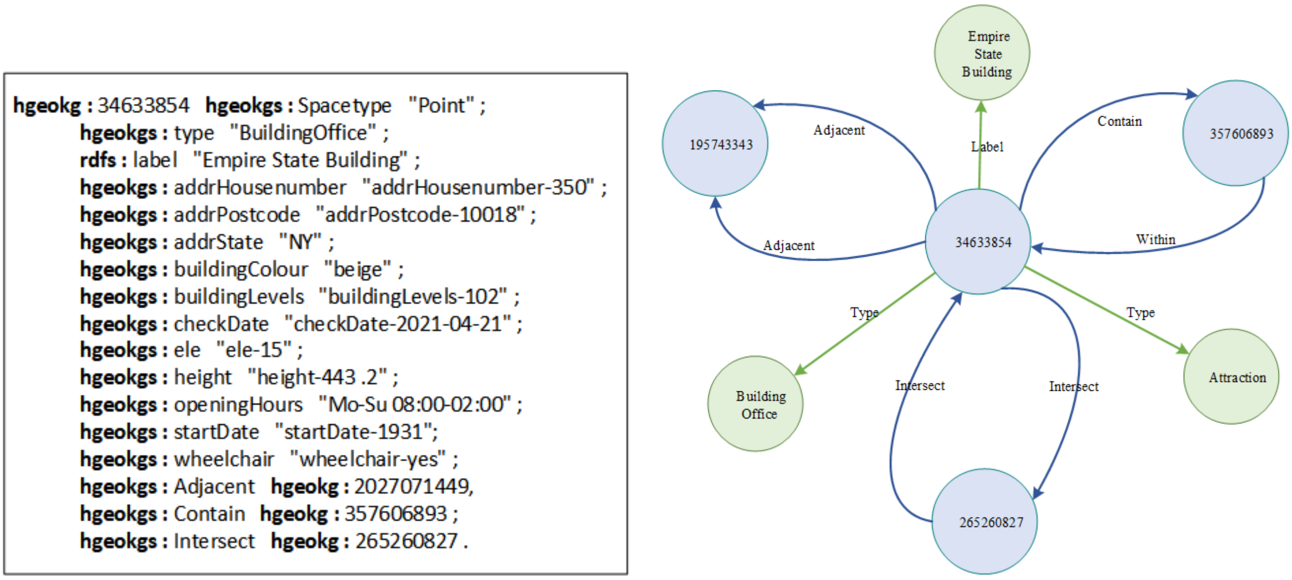
Figure 4. Hierarchical structure of regions.

In existing GeoKGs, geographic entities are typically assigned discrete spatial locations, usually represented by a single latitude and longitude coordinate. However, the regional distribution of geographic entities and the spatial heterogeneity between regions are crucial for their semantic representation across different areas. Current GeoKGs have not sufficiently accounted for the influence of these factors on the comprehensive semantic representation of geographic entities. The set of geographic entities within a specific region reflects the region's human, economic, ecological, and transportation conditions, and the distributional differences of these factors between regions are of significant importance for cross-regional studies. Therefore, incorporating regional distribution constraints and prior knowledge into GeoKGs, as well as constructing benchmark datasets for spatial heterogeneity research, is essential for exploring the homogeneous and heterogeneous patterns and rules across different regions.

HGeoKG constructs a regional hierarchical structure based on real administrative divisions. This hierarchical structure effectively captures spatial heterogeneity through multi-level regional representations. For example, fine-grained regions at various levels are used to characterize the local features of geographic spaces, while coarse-grained regions reflect global characteristics. This approach provides a more comprehensive depiction of spatial heterogeneity and distribution differences.

3.2.4. Meta-Analysis and Geographic Entity Examples

For each entity, we use the unique id of the original OSM element as its name and use the tags from the OSM element as the entity's attributes. Geographic entities are connected through spatial relationships. Figure 5a provides an example of a resource description framework (RDF) triple file in Turtle format for a GeoKG. This example includes information about the geometric spatial type of the entity, its common sense type, and various heterogeneous attributes such as name tags, business hours, and more. In addition to the attributes of the geographic entities themselves, the example also includes spatial semantic relationships between entities, such as adjacency and intersection. Figure 5b also shows a visualization of the GeoKG, intuitively displaying the attributes of geographic entities and the spatial relationships between them.



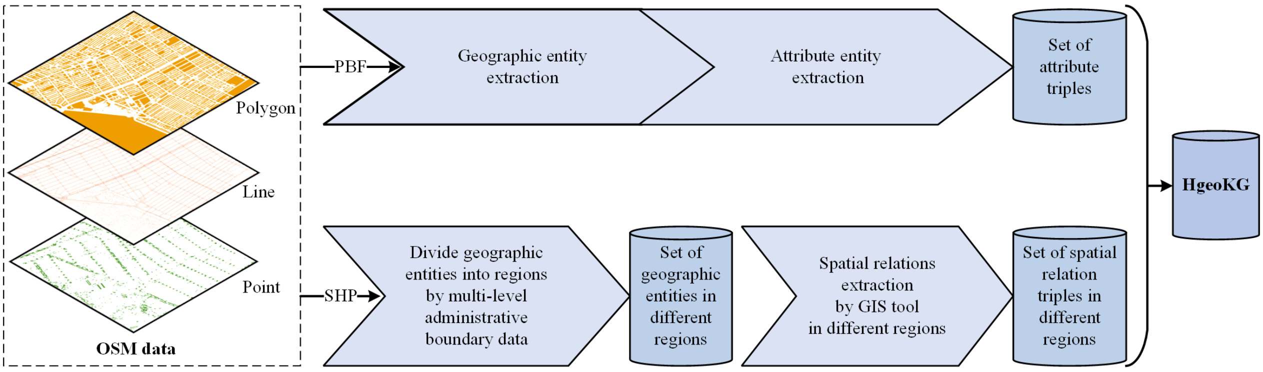
```
hgeokg : 34633854   hgeokgs : Spacetype  "Point" ;
        hgeokgs : type  "BuildingOffice" ;
        rdfs : label  "Empire State Building" ;
        hgeokgs : addrHousenumber  "addrHousenumber-350" ;
        hgeokgs : addrPostcode  "addrPostcode-10018" ;
        hgeokgs : addrState  "NY" ;
        hgeokgs : buildingColour  "beige" ;
        hgeokgs : buildingLevels  "buildingLevels-102" ;
        hgeokgs : checkDate  "checkDate-2021-04-21" ;
        hgeokgs : ele  "ele-15" ;
        hgeokgs : height  "height-443 .2" ;
        hgeokgs : openingHours  "Mo-Su 08:00-02:00" ;
        hgeokgs : startDate  "startDate-1931";
        hgeokgs : wheelchair  "wheelchair-yes" ;
        hgeokgs : Adjacent  hgeokg : 2027071449,
        hgeokgs : Contain  hgeokg : 357606893 ;
        hgeokgs : Intersect  hgeokg : 265260827 .
```

(a) (b)

Figure 5. An example of HGeoKG. (**a**) is an example of an RDF triple in turtle format, and (**b**) is a partial visualization of the GeoKG. Blue nodes represent geographic entities, green nodes represent attribute values, blue edges represent spatial relationships, and green edges represent attribute relationships.

3.3. Data Layer

This subsection presents the data layer of HGeoKG for extracting, processing, and integrating geographic entities, attributes, spatial relationships from OSM data, and constructing the regional hierarchical structure based on reference spatial region data. Figure 6 illustrates the complete process of building the hierarchical GeoKG from OSM data. First, Section 3.3.1 describes the extraction of attribute information for geographic entities. Then, in Section 3.3.2, GIS tools are used to compute the regional hierarchical divisions of geographic entities and their spatial relationships within the regions. Finally, Section 3.3.3 details the integration of geographic entities' attribute information and spatial relationships to construct the complete GeoKG, alongside data storage and visualization examples.



Polygon
Line
Point
OSM data
PBF
SHP
Geographic entity extraction
Attribute entity extraction
Set of attribute triples
HgeoKG
Divide geographic entities into regions by multi-level administrative boundary data
Set of geographic entities in different regions
Spatial relations extraction by GIS tool in different regions
Set of spatial relation triples in different regions

Figure 6. HGeoKG construction process.

3.3.1. Entity and Attribute Extraction

This subsection focuses on extracting attribute information for geographic entities from OSM data. First, specific geographic regions are identified, and the OSM data for these regions, including protocolbuffer binary format (PBF) and shapefile (SHP) format files, are downloaded. PBF files contain the complete attribute information for each geographic entity, while SHP files primarily provide the spatial types and spatial information of geographic entities. Osmium is an efficient library specifically designed for processing OSM data, particularly adept at handling large-scale datasets. By utilizing the Osmium library, attribute information for geographic entities is extracted from PBF files. This extraction results in triples formatted as (geographic entity, attribute, attribute value), with attribute names converted to camel case. Inferring the line and polygon types of geographic entities from latitude and longitude information in raw PBF data poses significant challenges. Consequently, GIS tools are employed to extract spatial-type information from SHP files, resulting in spatial-type triples formatted as geographic entity, spatial type, and point/line/polygon. It is essential that the extracted geographic entities contain at least one attribute; those with only latitude and longitude and lacking additional attributes will be filtered out. The triples extracted in this subsection represent only the relationships between geographic entities and attribute entities, without establishing direct connections among geographic entities.

3.3.2. Spatial Relationship Extraction

After the extraction of entities and attributes, the KG still lacks spatial relationships between geographic entities. This subsection introduces the data-processing methods for extracting spatial relationships between geographic entities. Utilizing the neighborhood analysis and spatial join functions of GIS tools, spatial relationships among geographic entities within the same region are extracted, resulting in triples formatted as geographic entity, spatial relationship, and geographic entity. Table 3 presents the spatial relationships between point, line, and polygon geographic entities, including relationships such as containment, adjacency, and intersection. Based on the spatial relationships designed in Section 3.2.2, this study explicitly enhances the common sense semantic information of spatial relationships by incorporating geographic entities at different hierarchical levels. In the following Section 4.3.5, we will discuss, through experiments, the impact of explicitly integrating entity type information of varying granularity into spatial relationships.

3.3.3. Regional Hierarchical Division and Extraction

As illustrated in Figure 6, this study employs the Clip operation within GIS tools to partition geographic entities from OSM data into administrative regions of varying hierarchical levels, thereby obtaining collections of geographic entities within coarse-grained or fine-grained regions at each level. Specifically, we first utilize coarse-grained administrative region data to perform an initial division of the OSM data, generating several coarse-grained regions. Subsequently, based on fine-grained administrative division data, we further subdivide the OSM data within these coarse-grained regions to form fine-grained regions. As shown in Figure 4, the regional hierarchy progresses from coarse to fine, with each coarse-grained region encompassing multiple fine-grained regions, and as the granularity increases, the number of fine-grained regions progressively increases. This multi-level regional partitioning method effectively reflects differences in data distribution and other aspects across regions.

Through this approach, HGeoKG is capable of effectively managing geographic entities within administrative regions of varying hierarchical levels and facilitates the computation of spatial relationships between entities within each region. Compared to other GeoKG methods that typically employ a unified global hierarchical structure or simple planar models, HGeoKG excels in capturing the spatial structures and regional differences at each hierarchical level. GeoKG methods that lack an effective modeling of regional hierarchical structures fail to adequately differentiate the details of various hierarchical regions and do not fully consider regional differences, which can lead to insufficient spatial semantic representation. By implementing multi-level regional divisions, HGeoKG not only meticulously reflects the characteristics of each hierarchical region but also comprehensively enhances the accuracy and granularity of geographic knowledge representation.

Furthermore, the dataset files of HGeoKG are stored separately for each hierarchical level of the regions, with each region's data organized into individual files. This file structure allows the knowledge of each region to be used and studied independently, further enhancing the flexibility and operability of HGeoKG in geographic knowledge processing.

Moreover, for geographic entities that span multiple regions (such as streets crossing multiple Census Tracts), our approach is to retain information about the entity in each relevant regional dataset. This method ensures that each regional dataset fully reflects the entities it contains. For cross-regional spatial relationships (such as linear spatial relationships connecting different regions), we choose to store them separately rather than directly integrating them into the regional datasets. These cross-regional relationships have been organized into independent data files and are included with the project files.

3.4. Generalizability and Scalability of HGeoKG

This subsection discusses the generalizability and scalability of HGeoKG. Firstly, regarding generalizability, HGeoKG is constructed using multi-source heterogeneous geographic data, including OSM point, line, and polygon data, as well as administrative boundary data. This enables it to cater to the geographic knowledge representation needs of different regions. Our hierarchical structure design, which includes spatial relationship hierarchy and regional hierarchy, is not only applicable to the data of the current experimental area, but also provides a transferable modeling framework for other geographic regions. Additionally, the construction process and methods of HGeoKG can be applied to various types of geographic datasets, demonstrating good generalizability.

Secondly, in terms of scalability, HGeoKG adopts a modular design, separating the Schema layer from the Data layer. This design ensures the convenient incorporation of new data and new relationships. Specifically, the Data layer can be dynamically expanded based on different regions or larger-scale data sources, while the ontology structure and hierarchical design of the Schema layer can be reused, supporting efficient knowledge updates and expansions.

4. Case Study: HGeoKG-MHT-670K

This section provides a comprehensive case analysis of the HGeoKG-MHT-670K through statistical and experimental methods. Section 4.1 introduces the data sources and spatiotemporal distribution of the dataset. Section 4.2 conducts statistical analysis to reveal the regional heterogeneity and long-tail distribution patterns within HGeoKG. Section 4.3 performs knowledge reasoning experiments to demonstrate how the data characteristics of HGeoKG impact the quality of knowledge graph embeddings and present challenges to existing knowledge reasoning models.

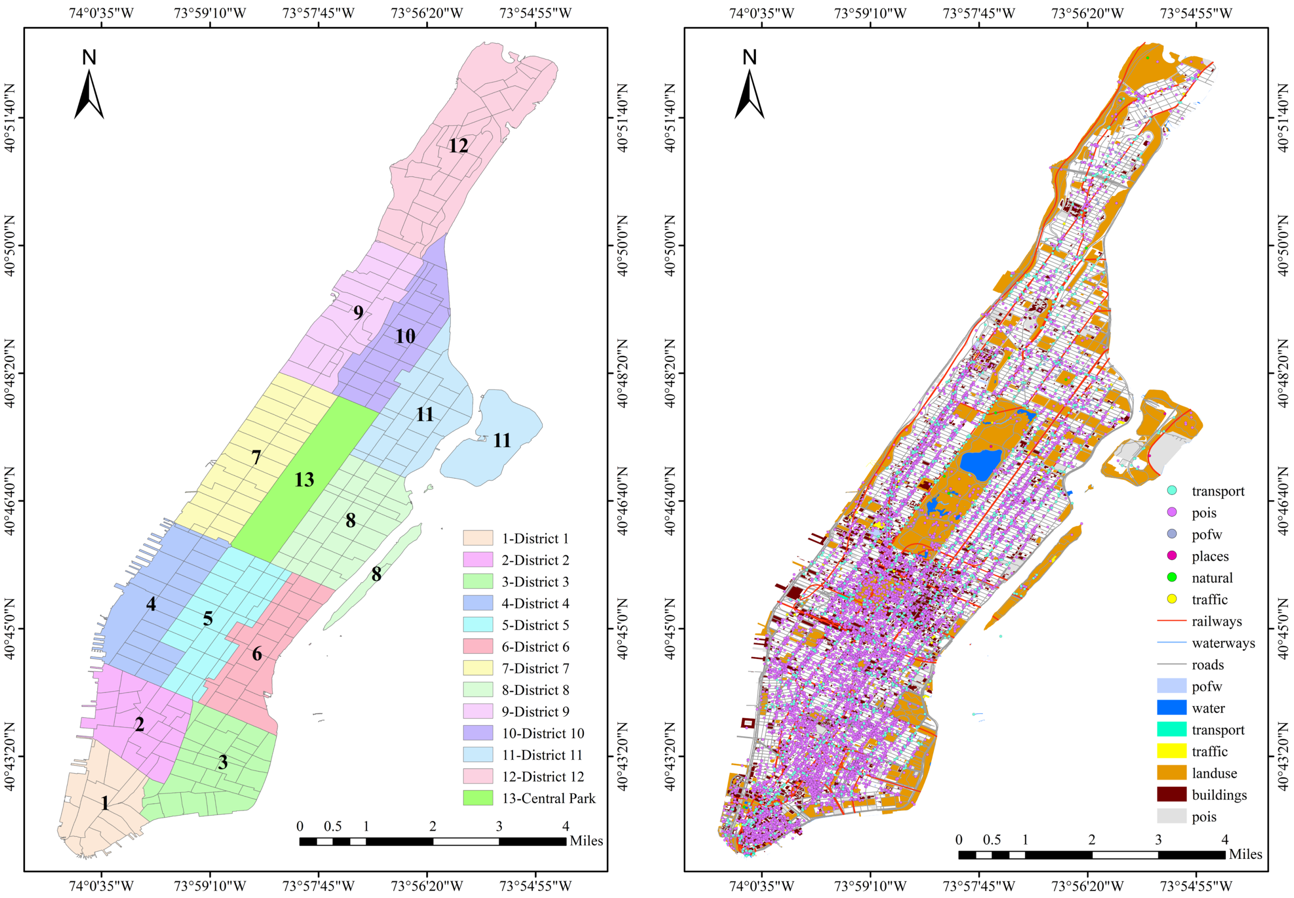
4.1. Data Sources and Study Area

This subsection presents the new geographic dataset HGeoKG-MHT-670K used for statistical analysis, with its spatial distribution shown in Figure 7. To construct HGeoKG-MHT-670K, we selected geographic data for Manhattan, New York (https://download.geofabrik.de/north-america/us/newyork.html (accessed on 1 January 2024)), from the open data platform OSM, including three types of geographic entities: points, lines, and polygons. Using the HGeoKG data-processing method proposed in this paper, discrete geographic entities and their associated attribute information were extracted. Based on the latitude and longitude information of the geographic entities, spatial relationships between them were calculated, and the spatial relationships were refined according to the entity types. By adding spatial relationships between geographic entities, a connected and continuous GeoKG was formed. The statistics of the dataset are presented in Table 4.

Considering the spatial hierarchical characteristics of GeoKGs, the Manhattan data was divided into 13 coarse-grained regions based on Community Districts (https://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4 (accessed on 1 January 2024)), and these 13 coarse-grained regions were further refined into 286 fine-grained regions based on Census Tracts (https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku (accessed on 1 January 2024)). It is important to note that the statistical data is arranged in descending order based on the number of triples in each region.

$$\text{Density} = \frac{|\text{triples}|}{|\text{region\_area}|}, \tag{1}$$

where |region_area| represents the area of the region, measured in square kilometers (km^2). |triples| represents the number of triples within the corresponding region.



(**a**) (**b**)

Figure 7. Spatial distribution of HGeoKG-MHT-670K. (**a**) Regional distribution; (**b**) geographic entity distribution.

Table 4. Statistics of HGeoKG-MHT-670K.

Region	Triples	Entity	Geo_ent	Atrr_ent	Relation	Geo_rel	Atrr_rel	Classes	Density (Triples/km^2)
All	669,222	61,784	22,362	39,422	600	194	406	12,226	8618.52
R1	110,802	11,751	3130	8621	334	93	241	2132	20,515.77
R2	98,271	7292	2441	4851	359	129	230	1333	19,091.84
R3	86,142	8117	2535	5582	287	74	213	1548	18,523.37
R4	74,527	7302	2345	4957	289	83	206	1345	12,885.54
R5	56,249	6939	1953	4986	286	77	209	1204	9247.36
R6	54,285	5860	1805	4055	298	90	208	1010	11,364.00
R7	49,441	6412	1745	4667	299	88	211	1130	7545.29
R8	43,943	7657	1854	5803	295	84	211	1203	6456.06
R9	26,393	3925	1208	2717	274	94	180	722	3238.67
R10	2252	3133	1134	1999	257	91	166	536	2370.18
R11	19,571	2867	878	1989	232	78	154	544	3791.49
R12	17,423	2527	868	1659	194	63	131	473	3618.31
R13	9423	1351	466	885	220	84	136	283	1995.28

In the table, "Geo_ent" denotes geographic entity, "Atrr_ent" denotes attribute entity, "Geo_rel" denotes spatial relationship, and "Atrr_rel" denotes attribute relationship.

4.2. Data Statistical Analysis

4.2.1. Spatial Heterogeneity Statistical Analysis

Regional heterogeneity refers to the differences exhibited by various regions in geographic spaces across natural, social, and economic dimensions. These differences are typically manifested in the distribution density of geographic entities, the complexity of spatial relationships, the diversity of attribute features, and the unevenness of data coverage. Regional heterogeneity is primarily reflected in the contrast between densely and sparsely populated areas, including significant disparities in the number of entities, relationships, density, and category distributions. This pattern poses unique challenges to the construction of geographic knowledge graphs and to representation learning.

The horizontal axis in Figure 8 is divided into 13 coarse-grained regions based on Community Districts. Since these coarse-grained regions are named with numbers, we have labeled them sequentially as R1 to R13 according to their actual names. Additionally, we used Census Tracts to subdivide the coarse-grained regions into multiple fine-grained areas, and the order of the horizontal axis is arranged based on the entity density of each fine-grained region, from largest to smallest. According to the statistical results shown in Table 4 and Figure 8, several key findings can be observed:

Firstly, at both coarse and fine granular levels, different regions exhibit significant distributional differences in the number of triples, entities, relationships, geographic entity categories, and density. The statistical results indicate that geographic semantic information in different regions displays obvious spatial heterogeneity.

Secondly, regarding geographic density, as shown in Table 4, there are significant differences in geographic density between regions, with dense and sparse areas showing distinct distribution characteristics. For example, regions 1 to 8 are relatively dense, while regions 9 to 13 are comparatively sparse. This density variation not only affects the data distribution of the geographic knowledge graph but also presents different challenges for subsequent model representation learning.

Thirdly, the number of attribute entities is significantly higher than that of geographic entities. This is because the attributes of different entities exhibit considerable variability. Additionally, relationships are primarily composed of spatial and attribute relationships, with the number of attribute relationship types being significantly greater than that of spatial relationship types. This is because spatial relationships describe the spatial connections between two geographic entities, which are usually more homogeneous and limited in type. In contrast, attribute relationships describe the characteristics of geographic entities, and the high heterogeneity of these characteristics leads to greater diversity in attribute relationship types.

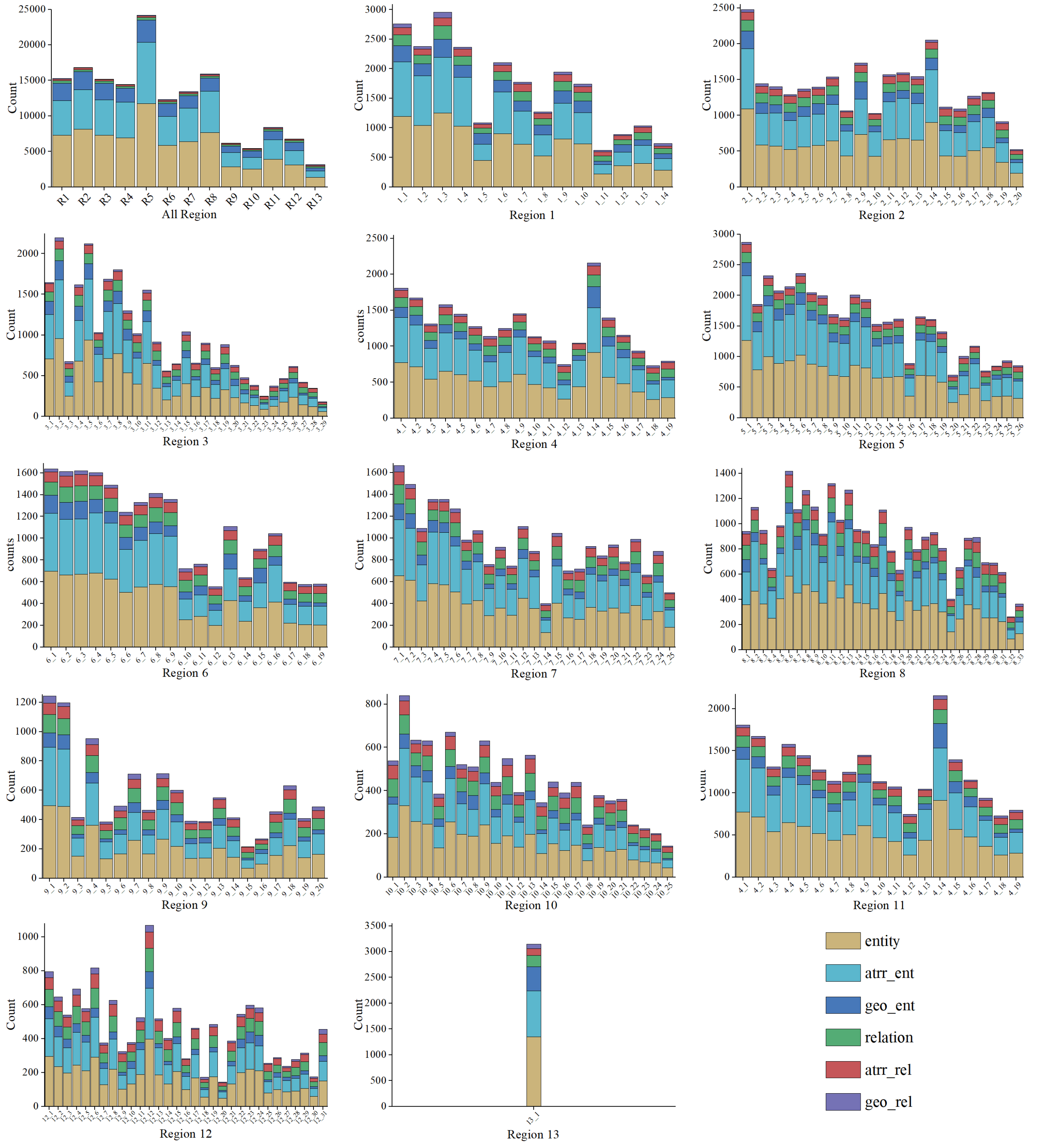


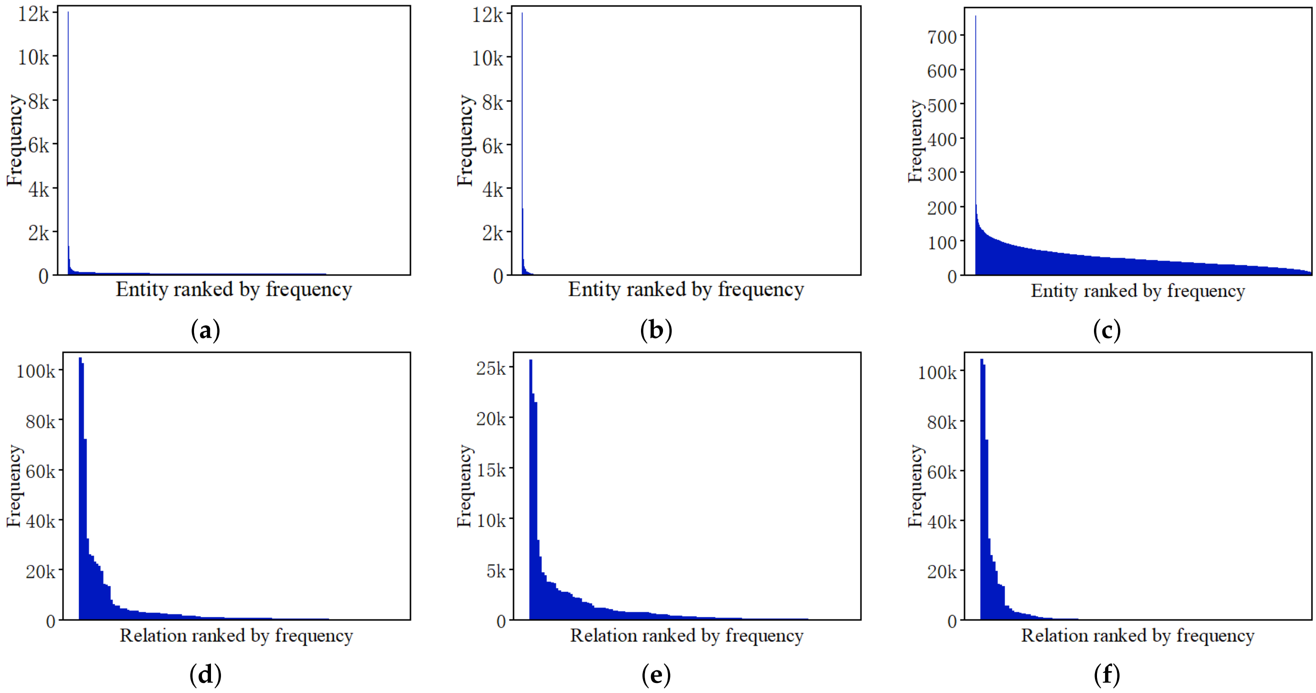
Figure 8. Data statistics in the coarse and fine-grained regions of HGeoKG-MHT-670K.

These findings indicate that although different regions have certain similarities in the composition of entities and relationships, there are significant differences in sparsity and spatial heterogeneity between regions. Such pronounced spatial heterogeneity and sparsity differences may potentially impact the representation learning of geographic knowledge. Therefore, when applying this data, these characteristics should be fully considered, and methods that adapt to these regional features should be designed to improve model performance.

4.2.2. Popularity Bias in HGeoKG

Many general-purpose knowledge graphs are primarily constructed by automatically extracting information from online resources like Wikipedia [31], leading to popularity bias: a small number of well-known entities possess rich information, while the majority of entities have sparse data [32]. Knowledge graphs built through crowdsourcing may exacerbate this issue due to contributors' implicit biases, affecting representation learning and potentially resulting in erroneous rule learning or poor embedding performance [33]. Similarly, GeoKGs constructed based on crowdsourced geographic data (e.g., OpenStreetMap) are also susceptible to contributor biases, impacting their quality. To enhance the comprehensiveness and accuracy of GeoKGs, this paper conducts a case study to analyze the issue of popularity bias within GeoKGs, exploring its characteristics and the modeling challenges it presents.

The long-tail distribution pattern describes the characteristic where a small number of high-frequency categories account for the majority of the data, while a large number of low-frequency categories form the long tail. In HGeoKG-MHT-670K, this pattern is primarily reflected in the distribution of entity and relationship categories. To reveal this phenomenon, we counted the frequency of occurrence for entity and relationship categories and plotted frequency distribution curves. The x-axis represents entities or relationships sorted in descending order of frequency, and the y-axis represents the frequency of category occurrences. As clearly shown in Figure 9, a few high-frequency entities and relationships account for the vast majority of the data, while a large number of low-frequency categories form a noticeable long-tail distribution.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 9. Frequency distribution of entities and relationships in the Triples of HGeoKG-MHT-670K. The entities or relations on the horizontal axis are ranked in descending order of frequency. (**a**) All entities; (**b**) attribute entities only; (**c**) geographic entities only; (**d**) all relations; (**e**) attribute relations only; (**f**) spatial relations only.

Further analysis indicates that, compared to relationship categories, the frequency distribution of entity categories is more skewed, as the number of entities in the knowledge graph far exceeds the number of relationships. Additionally, the frequency distribution of attribute entities is even more skewed than that of geographic entities. Geographic entities typically participate in spatial relationship triples with other geographic entities, as well as attribute triples related to their own attribute labels. Therefore, the frequency cap of a single geographic entity is limited by its role in the knowledge graph. In contrast, some common attribute labels can form triples with almost all geographic entities, resulting in the highest frequency of attribute entities approaching the total number of geographic entities.

4.3. Geographic Knowledge Reasoning

In this section, we start with Section 4.3.1, which introduces the geographic knowledge reasoning task and its evaluation metrics. We then provide a detailed description of the HGeoKG-MHT-670K dataset statistics, followed by an overview of the baseline models and experimental environment. The subsequent subsections further explore the challenges that the data characteristics of HGeoKG pose to existing baseline models.

4.3.1. Experimental Setup

Evaluation task and metrics. This study uses geographic knowledge reasoning as the evaluation task. The goal of geographic knowledge reasoning is to predict the missing entity in a triple. For example, given (?, r, t), the task is to predict the head entity h, or given (h, r, ?), the task is to predict the tail entity t. In the test set, the results are computed by ranking the scores predicted by a scoring function. The performance of geographic knowledge reasoning is evaluated using the Hit@k metric and the Mean Reciprocal Rank (MRR). Hit@k measures the number of times the correct entity appears in the top k predictions, while MRR represents the mean reciprocal rank of the correct predictions. Higher values of Hit@k and MRR indicate better performance. Considering the spatial regions to which geographic knowledge belongs, the Hit@k and MRR metrics can be extended into both macro and micro forms to better evaluate the impact of regional spatial heterogeneity on geographic knowledge reasoning task performance. The specific calculations for these metrics are as follows:

$$\text{Micro\_Hit@k} = \frac{1}{|N|}\sum_{i=1}^{|N|}\Pi(\text{rank}_i \le k), \tag{2}$$

$$\text{Micro\_MRR} = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{1}{\text{rank}_i}, \tag{3}$$

$$\text{Macro\_Hit@k} = \frac{1}{|\text{region}|}\sum_{i=1}^{|\text{region}|}\left(\frac{1}{|N_i|}\sum_{j=1}^{|N_i|}\Pi\left(\text{rank}_j \le k\right)\right), \tag{4}$$

$$\text{Macro\_MRR} = \frac{1}{|\text{region}|}\sum_{i=1}^{|\text{region}|}\left(\frac{1}{|N_i|}\sum_{j=1}^{|N_i|}\frac{1}{\text{rank}_j}\right), \tag{5}$$

where $|\text{region}|$ represents the number of geographic regions. $|N_i|$ represents the total number of prediction sets in region i. Metric_{ij} refers to the metric (such as Hit@k or MRR) for the j-th triple entity in region i. Π is a conditional function that equals 1 if the condition is true, otherwise 0.

The macro metrics used in this paper differ from traditional macro metrics used in classification tasks. Traditional category-based macro metrics are designed to handle class imbalance problems, typically performing well in large classes with abundant labels, while small classes tend to perform poorly due to sparse label data. The regional macro metrics proposed in this paper are designed to address differences in spatial heterogeneity across regions. The modeling difficulty for each region is determined not only by the number of samples in the region but also by the complex geographic distribution patterns and rules within each region (such as sample quantity, density, spatial range, and connectivity).

Experimental data. We carefully considered regional distribution when dividing HGeoKG-MHT-670K. The training set, validation set, and test set for each sub-region were divided in proportions of 80%, 10%, and 10%, respectively. The sum of the training, validation, and test sets of each sub-region constitutes the training, validation, and test sets of the parent region, maintaining the same 8:1:1 ratio for the entire dataset. This region-based hierarchical data-partitioning method effectively balances the distribution of the training, validation, and test sets across regions. Traditional random partitioning methods may lead to an imbalanced regional sample distribution in the training, validation, and test sets, introducing significant regional bias into the data. The results of the dataset partitioning are shown in Table 5.

Table 5. Dataset partition statistics.

Region	Triples	Entity	Relation	Train	Valid	Test
All	669,222	61,784	600	535,492	66,932	66,798
R1	110,802	11,751	334	88,651	11,080	11,071
R2	98,271	7292	359	78,622	9827	9822
R3	86,142	8117	287	68,922	8618	8602
R4	74,527	7302	289	59,635	7455	7437
R5	56,249	6939	286	45,007	5624	5618
R6	54,285	5860	298	43,436	5429	5420
R7	49,441	6412	299	39,564	4945	4932
R8	43,943	7657	295	35,165	4396	4382
R9	26,393	3925	274	21,125	2639	2629
R10	22,752	3133	257	18,215	2277	2260
R11	19,571	2867	232	15,666	1956	1949
R12	17,423	2527	194	13,945	1744	1734
R13	9423	1351	220	7539	942	942

Baseline models. We selected classic KGE models as baselines to perform experimental analysis on the dataset proposed in this paper.

TransE [34]: A translation-based KGE model and one of the most widely used KGE models, which interprets relationships as translation operations between low-dimensional entity embeddings.

DistMult [35]: A semantic matching-based KGE model and a popular tensor factorization approach that uses a bilinear scoring function to evaluate knowledge triples.

ConvE [36]: A convolutional neural network-based KGE model, utilizing a multi-layer convolutional network. The embedding vectors of the subject entity and relationship are reshaped into matrices and concatenated, followed by a global 2D convolution to learn deeper features.

R-GCN [37]: A graph neural network-based KGE model and an extension of the graph convolutional network (GCN), designed to handle highly multi-relational knowledge graph data. It aggregates contextual information into entities through relation-specific transformations to capture neighborhood information.

Experimental environment. The baseline models were evaluated on a server equipped with an Intel® Core™ i7-10700 CPU and an NVIDIA GeForce RTX 3090 GPU (24GB VRAM). The experimental environment was based on Ubuntu 18.04 and CUDA 11.1. All methods were implemented in Python 3.7 and PyTorch 1.12.0.

4.3.2. Macro and Micro Comparison Analysis

HGeoKG-MHT-670K provides spatial semantic connections between geographic entities in the Manhattan region, as well as highly heterogeneous geographic entity attributes. Given the high connectivity among geographic entities and the high sparsity and heterogeneity of attribute entities, we divided the evaluation metrics for the geographic knowledge reasoning task of baseline models into three categories: all triples, attribute triples only, and spatial triples only. This was carried out to analyze how the characteristics of spatial and attribute relationships affect the baseline models' metrics. In this subsection, we provide a more comprehensive comparison of the benchmark methods.

From the results in Table 6, we can observe the following:

Table 6. Comparison of macro and micro performances of different baseline models.

Model		**All Triples**				**Attribute Triples Only**				**Spatial Triples Only**			
		MRR	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
TransE	micro	28.26	8.23	38.35	69.30	35.08	29.49	38.47	44.96	25.63	0.03	38.30	78.68
	macro	29.53	10.01	39.45	70.16	36.66	30.87	40.31	46.72	25.99	0.03	38.90	81.09
DisMult	micro	45.69	34.51	53.45	65.66	0.95	0.54	0.96	1.64	62.94	47.62	73.70	90.35
	macro	42.99	32.63	50.34	61.21	0.84	0.46	0.86	1.49	62.60	47.65	73.38	88.78
ConvE	micro	43.08	31.67	49.26	63.25	32.47	27.24	32.81	41.70	52.23	37.78	60.66	77.78
	macro	46.16	34.46	51.91	67.57	33.30	28.02	33.73	42.53	51.98	37.27	60.18	79.29
RGCN	micro	47.16	30.58	58.58	78.47	26.45	21.43	29.24	35.31	55.17	34.14	69.90	95.11
	macro	46.62	30.70	57.48	76.84	27.61	22.34	30.58	36.96	55.34	34.56	69.83	95.06

In this table, **Micro_Hit@k** represents the proportion of hits within the top k rankings at the micro level. By averaging across all samples, it measures the overall hit rate (corresponding to Equation (2)). **Micro_MRR** refers to the micro-level Mean Reciprocal Rank (MRR). It assesses the quality of the model's ranking results by calculating the average of the reciprocal ranks across all samples (corresponding to Equation (3)). **Macro_Hit@k** calculates the hit rate within the top k rankings independently for each region at the macro level and then averages the results across all regions to evaluate the overall performance between regions (corresponding to Equation (4)). **Macro_MRR** is the macro-level Mean Reciprocal Rank. It first computes the average reciprocal rank for samples within each region and then averages these across all regions, reflecting the ranking accuracy at the regional level (corresponding to Equation (5)).

In the TransE and ConvE models, macro-averaged metrics are higher than micro-averaged metrics, indicating that these models perform better in smaller regions by focusing on individual triple samples and localized data distributions. These models can handle data from regions of varying sizes relatively consistently. In contrast, in the DistMult and R-GCN models, micro-averaged metrics outperform macro-averaged metrics, showing that these models excel in larger regions by focusing on the neighborhood of triples and the overall regional data distribution. They achieve the best and second-best performance in the micro metrics, demonstrating their effectiveness in handling neighborhood and contextual information in large-scale, data-rich regions.

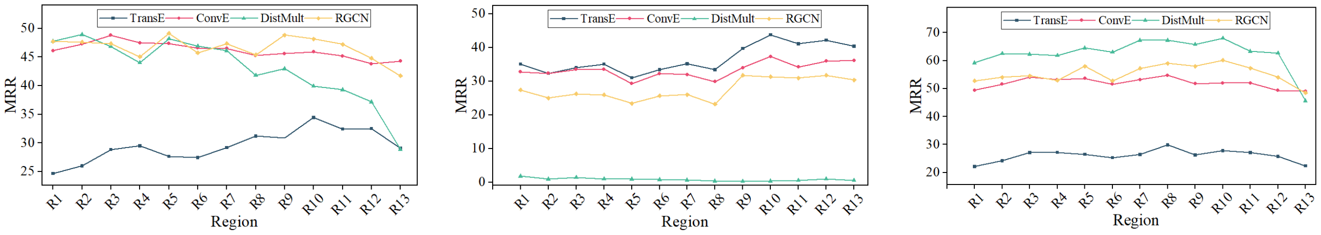
Due to spatial heterogeneity, geographic distributions across regions exhibit varying complexities. The TransE and ConvE models, which focus on local information, maintain relatively balanced performance across both small and large regions. On the other hand, the DistMult and R-GCN models, which emphasize overall data distribution, perform better in larger regions but worse in smaller ones. Therefore, given the complex geographic distribution patterns across different regions, our experimental analysis suggests that a single model type cannot effectively capture the intricate geographic patterns of all regions. It may be more effective to use different modeling strategies for different regions based on the characteristics of the data.

In response to the specific phenomena presented in Table 6, we have conducted an in-depth analysis of the following two significant issues. Firstly, the TransE model almost completely fails (approaching 0) in Hits@1 for the Spatial triples task. This observation reflects the notable limitations of TransE in handling spatial relationship types. TransE's embedding mechanism is primarily based on translational operations, which inadequately models complex geometric relationships, making it difficult to effectively capture the spatial semantics between entities. Secondly, the DisMult model exhibits significantly lower MRR and Hits@k values on the Attribute triples task compared to other models, indicating substantial difficulties in handling attribute relations. Although DisMult performs adequately on the Spatial triples task, its scoring mechanism fails to fully capture the joint features of attribute and spatial relations in the Attribute triples task. Our analysis suggests that this issue arises from DisMult's reliance on a bilinear scoring function, which demonstrates clear limitations when handling highly heterogeneous and sparse attribute data. This limitation hinders the effective modeling of complex interactions between attribute features and spatial information, leading to a significant decline in predictive performance.

4.3.3. Performance Comparison Across Regions with Different Levels of Sparsity

This subsection analyzes the performance of baseline models across the 13 coarse-grained regional datasets, examining the models' learning tendencies for different types of triples and the impact of data density. Specifically, the experiments were first conducted using the complete HGeoKG dataset for model training and evaluation. Then, the MRR performance for each region was calculated by dividing the dataset into coarse-grained regions. According to the regional density information in Table 4, R1 to R8 are denser regions, while R9 to R13 are sparser, with R13 being the sparsest.

From the results in Figure 10, it can be seen that the DistMult model performs normally on spatial triples but almost completely fails to predict attribute triples. This indicates that DistMult's bilinear scoring mechanism cannot effectively learn both types of triples simultaneously, with the model tending to focus on the spatial triples, which constitute the majority of the data. Further analysis reveals that DistMult's scoring mechanism cannot adequately capture the joint features of attributes and spatial relationships, which is closely related to the bilinear scoring function it relies on. This function shows clear limitations when dealing with highly heterogeneous and sparse attribute data, making it difficult to model the complex interactions between attribute features and spatial information effectively.



(**a**) (**b**) (**c**)

Figure 10. Experimental results in coarse-grained regions with different levels of sparsity. (**a**) All triples; (**b**) attribute triples only; (**c**) spatial triples only.

In addition, Figure 10b shows that, aside from DistMult, the other three baseline models (TransE, ConvE, and RGCN) exhibit significant regional differences in performance on attribute triples: their performance is better in sparser regions than in denser ones. This phenomenon is likely because the proportion of spatial triples decreases in sparse regions, while the proportion of attribute triples increases. This enables the model to focus more on learning attribute relationships during training, leading to better performance.

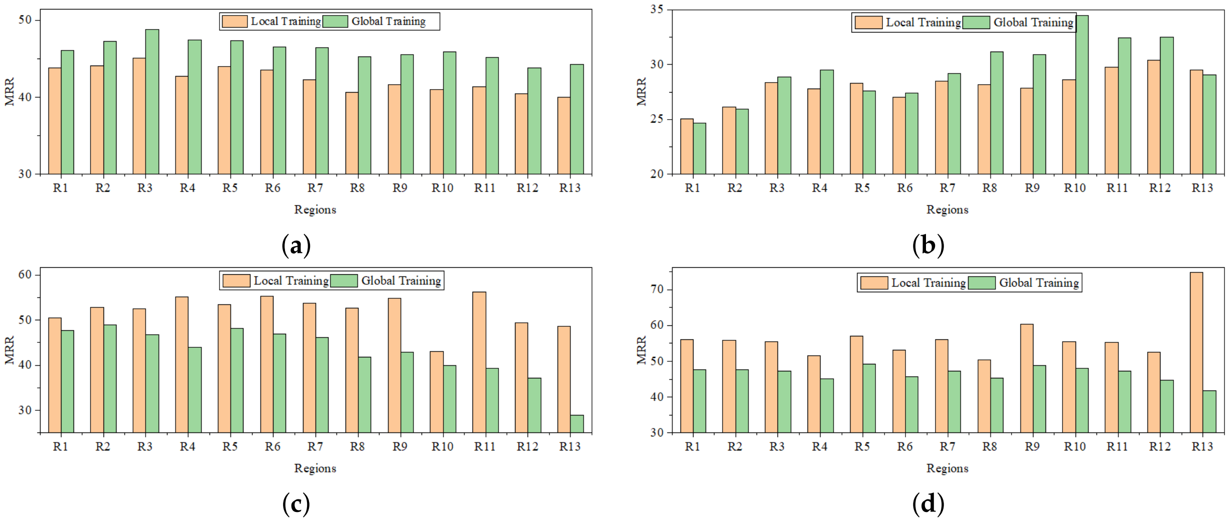
Finally, from Figure 10c, it can be observed that the four baseline models exhibit stable learning performance on spatial triples, with the MRR metric showing little variation across regions. However, the DistMult model experiences a significant performance drop in the sparsest R13 region, further highlighting its limitations when handling data sparsity.

The density differences between regions are one manifestation of spatial heterogeneity. Our experimental results reveal the learning characteristics of the baseline models on different types of triples, as well as the challenges posed by spatial heterogeneity in geographic knowledge representation learning and the significant impact of data density on model performance.

4.3.4. Comparative Analysis of Global and Local Training

In this subsection, we employed both global and local training strategies to evaluate the model's performance across different geographic regions. For global training, we used the complete HGeoKG dataset to ensure that the model could learn general geographic knowledge across regions. For local training, we divided the HGeoKG dataset by different geographic regions and performed model training on specific regions. This division allowed the model to better capture regional geographic semantics and improve its performance in scenarios with strong regional heterogeneity. Specifically, global training involved splitting the dataset into training, validation, and test sets and training the model on the entire global dataset. In local training, we divided the data based on geographic regions (such as administrative divisions or geographic units) and ensured that the model was trained and evaluated separately for each local region. This approach helps improve the model's adaptability and generalization across different geographic environments.

From the experimental results in Figure 11, it can be seen that ConvE and TransE perform better in global training than in local training, whereas DistMult and RGCN show better performance in local training compared to global training. The reasons for this may be related to the model mechanisms, data distribution characteristics, and the impact of spatial heterogeneity. ConvE and TransE rely more on global data distributions to learn general semantic features. In global training, these models are exposed to rich relational information from different regions, which helps them capture broader semantic patterns, particularly in cross-regional spatial relationship modeling. For example, TransE learns simple relational features through translation embeddings, and the rich global data help to generate more accurate relationship representations. Similarly, ConvE captures complex contextual relationships through convolution operations, and global data provide more samples and context, thereby enhancing model performance.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 11. Comparison between global and local training. (**a**) ConvE; (**b**) TransE; (**c**) DistMult; (**d**) RGCN.

In contrast, DistMult and RGCN perform better in local training because local training reduces data heterogeneity and complexity, allowing these models to focus more on relationship modeling within specific regions. DistMult relies on a bilinear scoring function, which struggles to model the diverse semantic relationships in complex global data. However, in local data, where heterogeneity is reduced, the model can more accurately learn regional characteristics. RGCN captures neighborhood features through graph neural networks, and local training offers a more compact data distribution, enabling it to model local semantic relationships more efficiently within each region.

Spatial heterogeneity leads to significant differences in data distributions across regions. In global training, the model has to handle the differences in data distribution between dense and sparse regions, which may cause the insufficient learning of specific regional characteristics. Local training, by partitioning the data by region, reduces this heterogeneity and allows the model to focus on learning the semantic relationships specific to each region. As a result, DistMult and RGCN perform better in local training.

The advantage of ConvE and TransE in global training lies in their ability to capture global general patterns, while the better performance of DistMult and RGCN in local training reflects their ability to adapt to regional characteristics once heterogeneity is reduced. These experimental results further reveal the profound impact of spatial heterogeneity on model performance and also suggest that model selection should be optimized based on the data distribution characteristics and the specific requirements of the application scenario.

4.3.5. The Impact of Relationship Hierarchy

This subsection explores the various challenges that explicit modeling of spatial semantic relationships at different granularities brings to existing KGE models. We focus on analyzing the impact of explicitly modeling spatial semantic relationships, the effect of integrating entity-type information into these relationships, and the influence of incorporating entity-type information at different granularity levels. Table 7 explains the types of spatial semantic relationships at different granularities and indicates what information is explicitly modeled.

Table 7. Different granularities of spatial relationships.

Example	**Spatial Relationship-Type Description**
Point_pois-Adjacent-Polygon_water	2, spatial relationship explicitly models fine-grained entity types
Point-Adjacent-Polygon	1, spatial relationship explicitly models coarse-grained entity types
Adjacent	0, spatial relationship does not consider any entity-type information
None	−1, no spatial relationship

Table 7 shows the explicit modeling of spatial semantic relationships at different levels of granularity. A granularity level of −1 indicates no spatial relationship semantics are included in the dataset, only attribute triples for geographic entities. A granularity level of 0 indicates that spatial relationship semantics are explicitly modeled between entities, but no entity-type information is integrated. In this case, these spatial relationships are defined solely based on the calculation of geometric distance or spatial topology. A granularity level of 1 incorporates the coarse-grained spatial types of entities (point, line, polygon) into the spatial relationship semantics, while a level of 2 incorporates both coarse-grained spatial types and fine-grained entity-type information into the spatial relationship modeling.

The results in Table 8 reveal the performance of different models on HGeoKG-MHT-670K with spatial semantic relationships of varying granularity. This phenomenon indicates that the inclusion of spatial semantic relationships enriches the semantic information and enhances the connectivity of the knowledge graph, allowing models to more effectively learn the important connections between geographic entities through spatial semantic edges. Additionally, the attribute information of entities can be more efficiently propagated to neighboring entities' embeddings through spatial semantic edges.

Table 8. Performance comparison of different baseline models on HGeoKG-MHT-670K.

Model	**Level**	**All Triples**				**Attribute Triples Only**				**Spatial Triples Only**			
		MRR	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
TransE	2	28.26	8.23	38.35	69.30	35.08	29.49	38.47	44.96	25.63	0.03	38.30	78.68
	1	28.27	8.11	38.37	69.32	34.76	29.06	38.46	44.65	25.77	0.03	38.34	78.84
	0	32.21	7.99	47.91	79.51	34.13	28.69	37.32	43.65	31.47	0.00	51.99	93.34
	−1	/	/	/	/	35.44	30.65	38.04	43.94	/	/	/	/
DistMult	2	45.69	34.51	53.45	65.66	0.95	0.54	0.96	1.64	62.94	47.62	73.70	90.35
	1	42.89	30.68	51.31	64.63	0.69	0.37	0.64	1.25	59.16	42.36	70.86	89.08
	0	36.78	24.97	43.83	58.97	0.72	0.38	0.69	1.24	50.68	34.45	60.46	81.23
	−1	/	/	/	/	22.24	19.34	23.75	27.36	/	/	/	/
ConvE	2	43.08	31.67	49.26	63.25	32.47	27.24	32.81	41.70	52.23	37.78	60.66	77.78
	1	42.96	30.73	50.06	63.90	33.08	27.50	33.08	43.11	51.78	36.31	61.59	78.06
	0	43.93	26.26	56.07	77.08	33.69	27.04	34.23	47.16	53.18	30.39	69.84	95.78
	−1	/	/	/	/	29.73	27.46	31.24	33.61	/	/	/	/
RGCN	2	47.16	30.58	58.58	78.47	26.45	21.43	29.24	35.31	55.17	34.14	69.90	95.11
	1	41.12	23.70	52.18	75.20	23.69	18.93	26.08	32.18	47.76	25.45	62.17	91.71
	0	29.65	9.41	41.41	71.15	23.01	17.83	26.01	31.94	32.24	6.23	47.36	86.22
	−1	/	/	/	/	24.92	20.11	27.44	32.96	/	/	/	/

In this table, **Micro_Hit@k** represents the proportion of hits within the top k rankings at the micro level. By averaging across all samples, it measures the overall hit rate (corresponding to Equation (2)). **Micro_MRR** refers to the micro-level Mean Reciprocal Rank (MRR). It assesses the quality of the model's ranking results by calculating the average of the reciprocal ranks across all samples (corresponding to Equation (3)).

When analyzing the different levels of spatial semantic edges, the three granularities—0, 1, and 2—represent the absence of entity-type information, the inclusion of coarse-grained entity-type information, and the inclusion of fine-grained entity-type information, respectively. The experimental results show that datasets incorporating entity-type information perform better in terms of metrics, demonstrating that richer entity-type information aids in achieving more comprehensive entity and relationship embedding learning. Furthermore, the results with fine-grained entity-type information outperform those with coarse-grained entity-type information, suggesting that more detailed entity-type semantics provide additional categorical details, helping models to learn and uncover potential patterns through the joint effect of spatial semantics and entity-type information. This finding emphasizes the importance of considering detailed entity-type information when modeling spatial semantic relationships.

In response to the unusual phenomena observed in Table 8, we conducted the following analyses. First, the TransE model performs poorly in all levels of the Spatial triples task, especially on the Hits@1 metric, which is almost zero. This result is consistent with the analysis in Table 6, further confirming the significant limitations of the TransE model in handling spatial relations. TransE's embedding mechanism is relatively simple and struggles to effectively model complex spatial semantic features, resulting in poor performance across different levels of spatial relations. Second, when handling the Attribute triples task, the DisMult model experiences a significant performance drop when spatial relations are incorporated, nearly failing entirely. Although DisMult performs adequately on the Spatial triples task, its scoring mechanism fails to fully capture the joint features of attribute and spatial relations. When faced with highly heterogeneous attribute data, the bilinear scoring function used by DisMult shows clear limitations in modeling the complex interactions between attribute features and spatial information, leading to a substantial decline in predictive performance. These analyses are consistent with the conclusions in Table 6 and highlight the differences in performance of these models when handling different types of relations.

4.3.6. Representation Challenges Caused by Popularity Bias

The previous subsection statistically analyzed the phenomenon of popularity bias in the GeoKG dataset, which exhibits a pronounced long-tail distribution pattern. This subsection evaluates the impact of such popularity bias on KGE quality, focusing on the geographic knowledge reasoning task.

We examined how popularity bias in GeoKGs affects the performance of knowledge graph embedding models. The experimental results as illustrated in Figure 12 show that popular entities and relationships achieve better performance metrics compared to their less popular counterparts. Additionally, as the proportion of popular entities decreases, the performance of embedding models also declines. This indicates that classic knowledge graph embedding models tend to generate more accurate embeddings for well-represented, popular entities and relationships, while neglecting less popular ones.

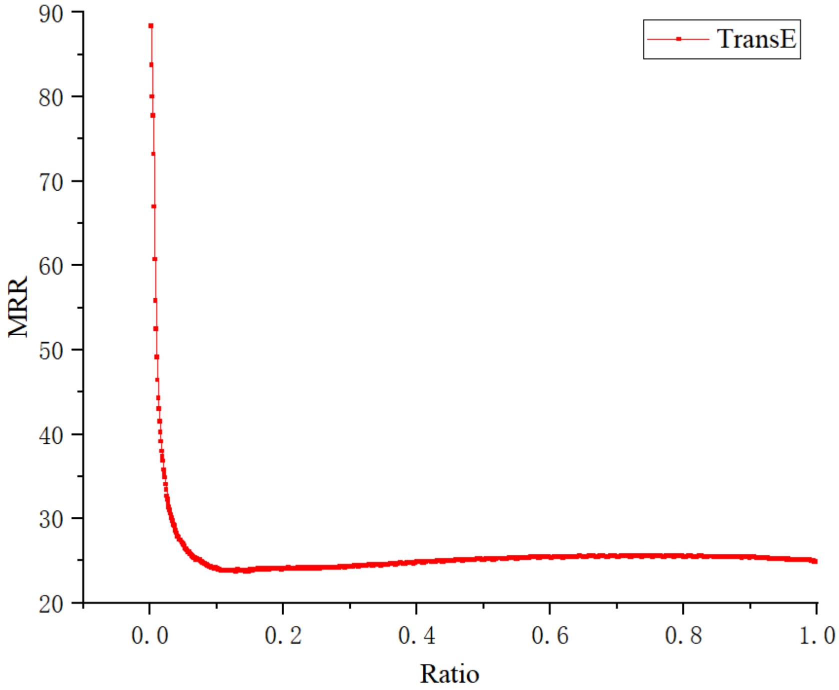


Figure 12. Performance of entity embeddings with different ratio on geographic knowledge reasoning tasks.

In GeoKGs, facts can be categorized into popular and unpopular groups. The high-frequency head items belong to the popular category, while unpopular items can be further divided into two subcategories:

Far-tail items: These appear in very few triples, sometimes only one, making it challenging to predict them reliably. Long-tail items: These appear in enough triples that a good model should be able to learn meaningful embeddings for them. Classic embedding models tend to perform well on both popular and long-tail items.

The notable bias in embedding models can be explained by their training process. During training, popular entities and relationships have more contextual information and background in the data and receive more attention during optimization. Additionally, popular entities and relationships appear in more triples, leading to more frequent updates. As a result, the model infers new facts about these entities with higher accuracy, while less popular entities and relationships are overlooked. This means that the accuracy of knowledge graph embedding models is largely driven by their ability to handle popular entities well, but they fail to effectively represent less popular entities. However, in geographic knowledge reasoning tasks, less popular entities are often the ones of greater interest. The dataset exhibits significant spatial heterogeneity and a long-tail distribution pattern, which poses challenges for the effectiveness of KGE.

5. Conclusions

This paper proposes a hierarchical geographic knowledge graph, HGeoKG, which provides a comprehensive semantic representation of geographic knowledge, encompassing rich attributes and spatial relationships, while featuring both regional and spatial relationship hierarchies. It offers theoretical and methodological references for constructing GeoKGs. Extensive geographic knowledge reasoning experiments on HGeoKG demonstrate that the performances of most knowledge graph embedding (KGE) models are significantly affected by the marked regional heterogeneity and long-tail distribution patterns in the HGeoKG dataset, resulting in unsatisfactory embedding quality. This highlights the importance of considering different modeling strategies for different regions and improving the embedding quality of long-tail geographic entities when designing or deploying HGeoKG embedding algorithms in practice. Current evaluation metrics do not adequately capture the effects of spatial heterogeneity, and designing suitable metrics specifically for geographic datasets remains a crucial direction for future research.

We believe that HGeoKG can serve as a valuable new benchmark for studying the characteristics of geographic knowledge and evaluating geographic knowledge representation learning. However, the open-source geographic information used in this study (e.g., OSM) may suffer from issues such as incompleteness and inconsistency. These data deficiencies could have a significant impact on the results, particularly in regions with imbalanced entity types or incomplete annotations. Additionally, the use of administrative boundaries as the basis for geographic unit division in this study could introduce certain biases. The administrative divisions were not specifically designed for this study, and their spatial distribution may not be fully compatible with the model's requirements. Finally, as this study focuses on hierarchical geographic knowledge graph embedding and inference, the model's ability to handle extreme conditions in specific tasks (e.g., sparse or heterogeneous data distributions) is still limited. Future work will further explore methods to improve model performance under these conditions.

Author Contributions: Conceptualization, Tailong Li and Hong Yao; methodology, Tailong Li, Shengwen Li, and Renyao Chen; validation, Tailong Li; formal analysis, Tailong Li and Xinchuan Li; data curation, Tailong Li; writing—original draft preparation, Tailong Li; writing—review and editing, Tailong Li, Renyao Chen, Yilin Duan, Hong Yao, Shengwen Li, and Xinchuan Li; visualization, Tailong Li and Yilin Duan. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) (No. 42071382).

Data Availability Statement: The data are available at https://github.com/Tailong-Li/HGeoKG (accessed on 30 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Peng, C.; Xia, F.; Naseriparsa, M.; Osborne, F. Knowledge Graphs: Opportunities and Challenges. *Artif. Intell. Rev.* **2023**, *56*, 13071–13102. [CrossRef] [PubMed]
2. Shen, T.; Zhang, F.; Cheng, J. A Comprehensive Overview of Knowledge Graph Completion. *Knowl.-Based Syst.* **2022**, *255*, 109597. [CrossRef]
3. Dietz, L.; Kotov, A.; Meij, E. Utilizing Knowledge Graphs for Text-Centric Information Retrieval. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 1387–1390.
4. Huang, X.; Zhang, J.; Li, D.; Li, P. Knowledge Graph Embedding Based Question Answering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, VIC, Australia, 11–15 February 2019; pp. 105–113.
5. Cao, Y.; Wang, X.; He, X.; Hu, Z.; Chua, T.-S. Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 151–161.
6. Mai, G.; Yan, B.; Janowicz, K.; Zhu, R. Relaxing Unanswerable Geographic Questions Using a Spatially Explicit Knowledge Graph Embedding Model. In *Geospatial Technologies for Local and Regional Development*; Kyriakidis, P., Hadjimitsis, D., Skarlatos, D., Mansourian, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 21–39.
7. Zhou, Z.; Liu, Y.; Ding, J.; Jin, D.; Li, Y. Hierarchical Knowledge Graph Learning Enabled Socioeconomic Indicator Prediction in Location-Based Social Network. In Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–1 May 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 122–132.
8. Han, J.; Liu, H.; Xiong, H.; Yang, J. Semi-Supervised Air Quality Forecasting via Self-Supervised Hierarchical Graph Neural Network. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 5230–5243. [CrossRef]
9. Guo, S.; Lin, Y.; Feng, N.; Song, C.; Wan, H. Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 922–929.
10. Liu, X.; Liu, Y.; Li, X. Exploring the Context of Locations for Personalized Location Recommendations. In Proceedings of the 2016 International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 1188–1194.
11. Zhao, S.; Lyu, M.R.; King, I. Geo-Teaser: Geo-Temporal Sequential Embedding Rank for POI Recommendation. In *Point-of-Interest Recommendation in Location-Based Social Networks*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 57–78.
12. Grbovic, M.; Cheng, H. Real-Time Personalization Using Embeddings for Search Ranking at Airbnb. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 311–320.
13. Zhai, W.; Bai, X.; Shi, Y.; Han, Y.; Peng, Z.-R.; Gu, C. Beyond Word2vec: An Approach for Urban Functional Region Extraction and Identification by Combining Place2vec and POIs. *Comput. Environ. Urban Syst.* **2019**, *74*, 1–12. [CrossRef]
14. Hoffart, J.; Suchanek, F.M.; Berberich, K.; Weikum, G. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artif. Intell.* **2013**, *194*, 28–61. [CrossRef]
15. Hu, W.; Li, H.; Sun, Z.; Qian, X.; Xue, L.; Cao, E.; Qu, Y. Clinga: Bringing Chinese Physical and Human Geography in Linked Open Data. In Proceedings of the 15th International Semantic Web Conference, Part II, Kobe, Japan, 17–21 October 2016; Springer: Kobe, Japan, 2016; pp. 104–112.
16. Li, J.; Liu, R.; Xiong, R. A Chinese Geographic Knowledge Base for GIR. In Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 21–24 July 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 361–368.
17. Karalis, N.; Mandilaras, G.; Koubarakis, M. Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge. In Proceedings of the 18th International Semantic Web Conference, Part II, Auckland, New Zealand, 26–30 October 2019; Springer: Auckland, New Zealand, 2019; pp. 181–197.
18. Guo, X.; Qian, H.; Wu, F.; Liu, J. A Method for Constructing Geographical Knowledge Graph from Multisource Data. *Sustainability* **2021**, *13*, 10602. [CrossRef]
19. Auer, S.; Lehmann, J.; Hellmann, S. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In Proceedings of the 8th International Semantic Web Conference, Washington, DC, USA, 25–29 October 2009; Springer: Chantilly, VA, USA, 2009; pp. 731–746.
20. Chen, J.; Deng, S.; Chen, H. CrowdGeoKG: Crowdsourced Geo-Knowledge Graph. In *Knowledge Graph and Semantic Computing: Language, Knowledge, and Intelligence*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 165–172.
21. Codescu, M.; Horsinka, G.; Kutz, O.; Mossakowski, T.; Rau, R. OSMonto—An Ontology of OpenStreetMap Tags. In *State of the Map Europe (SOTM-EU)*; University of Bremen: Bremen, Germany, 2011; pp. 23–24.
22. Dsouza, A.; Tempelmeier, N.; Yu, R.; Gottschalk, S.; Demidova, E. WorldKG: A World-Scale Geographic Knowledge Graph. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, QLD, Australia, 1–5 November 2021; pp. 4475–4484.

23. Wang, S.; Zhang, X.; Ye, P.; Du, M.; Lu, Y.; Xue, H. Geographic Knowledge Graph (GeoKG): A Formalized Geographic Knowledge Representation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 184. [CrossRef]
24. Zheng, K.; Xie, M.H.; Zhang, J.B.; Xie, J.; Xia, S.H. A Knowledge Representation Model Based on the Geographic Spatiotemporal Process. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 674–691. [CrossRef]
25. Han, B.; Qu, T.; Tong, X.; Wang, H.; Liu, H.; Huo, Y.; Cheng, C. AugGKG: A Grid-Augmented Geographic Knowledge Graph Representation and Spatio-Temporal Query Model. *Int. J. Digit. Earth* **2023**, *16*, 4934–4957. [CrossRef]
26. Suchanek, F.M.; Kasneci, G.; Weikum, G. YAGO: A Core of Semantic Knowledge. In Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 697–706.
27. Vrandečić, D.; Krötzsch, M. Wikidata: A Free Collaborative Knowledgebase. *Commun. Acm* **2014**, *57*, 78–85. [CrossRef]
28. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 9–12 June 2008; pp. 1247–1250.
29. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; et al. DBpedia—A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
30. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. *Knowl. Inf. Syst.* **2013**, *37*, 61–81. [CrossRef]
31. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the International Semantic Web Conference, Busan, Republic of Korea, 11–15 November 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
32. Mohamed, A.; Parambath, S.; Kaoudi, Z.; Aboulnaga, A. Popularity Agnostic Evaluation of Knowledge Graph Embeddings. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, Toronto, ON, Canada, 3–6 August 2020; PMLR: New York, NY, USA, 2020; pp. 1059–1068.
33. Janowicz, K.; Yan, B.; Regalia, B.; Zhu, R.; Mai, G. Debiasing Knowledge Graphs: Why Female Presidents Are Not like Female Popes. In Proceedings of the ISWC (P&D/Industry/BlueSky), Monterey, CA, USA, 8–12 October 2018.
34. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, 5–8 December 2013; Volume 26.
35. Yang, B.; Yih, S.W.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
36. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2D Knowledge Graph Embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
37. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; van den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In Proceedings of the 15th International Conference on The Semantic Web (ESWC), Heraklion, Greece, 3–7 June 2018; Springer: Heraklion, Greece, 2018; pp. 593–607.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.