remote sensing

MDPI

Article

A Comprehensive Study of Geochemical Data Storage Performance Based on Different Management Methods

Yinyi Cheng [1,2,3,4,5,6,7], Kefa Zhou [1,2,3,4], Jinlin Wang [1,2,3,4,*], Philippe De Maeyer [4,5,6,7], Tim Van de Voorde [4,5,6,7], Jining Yan [8] and Shichao Cui [1,2,3,4]

1 State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China; chengyinyi17@mails.ucas.ac.cn (Y.C.); zhoukf@ms.xjb.ac.cn (K.Z.); cuishichao147@mails.ucas.edu.cn (S.C.)
2 Xinjiang Key Laboratory of Mineral Resources and Digital Geology, Urumqi 830011, China
3 Xinjiang Research Center for Mineral Resources, Chinese Academy of Sciences, Urumqi 830011, China
4 University of Chinese Academy of Sciences, Beijing 100049, China; Philippe.DeMaeyer@UGent.be (P.D.M.); Tim.VandeVoorde@UGent.be (T.V.d.V.)
5 Department of Geography, Ghent University, 9000 Ghent, Belgium
6 Sino-Belgian Joint Laboratory for Geo-Information, Urumqi 830011, China
7 Sino-Belgian Joint Laboratory for Geo-Information, 9000 Ghent, Belgium
8 School of Computer Science, China University of Geosciences, Wuhan 430074, China; yanjn@cug.edu.cn
* Correspondence: wangjinlin@ms.xjb.ac.cn

Abstract: The spatial calculation of vector data is crucial for geochemical analysis in geological big data. However, large volumes of geochemical data make for inefficient management. Therefore, this study proposed a shapefile storage method based on MongoDB in GeoJSON form (SSMG) and a shapefile storage method based on PostgreSQL with open location code (OLC) geocoding (SSPOG) to solve the problem of low efficiency of electronic form management. The SSMG method consists of a JSONification tier and a cloud storage tier, while the SSPOG method consists of a geocoding tier, an extension tier, and a storage tier. Using MongoDB and PostgreSQL as databases, this study achieved two different types of high-throughput and high-efficiency methods for geochemical data storage and retrieval. Xinjiang, the largest province in China, was selected as the study area in which to test the proposed methods. Using geochemical data from shapefile as a data source, several experiments were performed to improve geochemical data storage efficiency and achieve efficient retrieval. The SSMG and SSPOG methods can be applied to improve geochemical data storage using different architectures, so as to achieve management of geochemical data organization in an efficient way, through time consumed and data compression ratio (DCR), in order to better support geological big data. The purpose of this study was to find ways to build a storage method that can improve the speed of geochemical data insertion and retrieval by using excellent big data technology to help us efficiently solve problem of geochemical data preprocessing and provide support for geochemical analysis.

Keywords: geochemical data; data storage; retrieval; database

Citation: Cheng, Y.; Zhou, K.; Wang, J.; Maeyer, P.D.; Voorde, T.V.d.; Yan, J.; Cui, S. A Comprehensive Study of Geochemical Data Storage Performance Based on Different Management Methods. *Remote Sens.* **2021**, *13*, 3208. https://doi.org/10.3390/rs13163208

Academic Editors: Prem Prakash Jayaraman, Federico Montori, Charith Perera and Felip Marti

Received: 13 July 2021
Accepted: 10 August 2021
Published: 13 August 2021

1. Introduction

Geochemical mapping plays an important role in both mineral exploration and environmental studies [1]. Geochemical data have the characteristics of complexity, region, and space. The traditional data management model cannot reflect the correlation characteristics of geochemical data, let alone preprocess the geochemically original sampling point data efficiently. Due to the complexity of geochemical data, it is difficult to ensure the integrity of the data in electronic form [2]. At the same time, floating-point-based main geochemical data types consume a lot of computer resources. Moreover, the increase in the amount of geochemical data makes the correlation analysis between elements more and more complicated. It is difficult to meet the needs of scientific research by using only

electronic forms of data. With the science-intensive fourth paradigm [3] becoming the main approach to scientific research, big data technology has provided new research ideas for geochemical research. As a typical data-intensive discipline, geology has abundant multisource heterogeneous data, including geochemical data, etc. The spatiotemporal and multivariate nature of geochemical data derives a series of data characteristics [4], which enables the accurate description of geological data. Considering the various, multisource observation modes existing in the quantitative description of geological data, they are divided into continuous data and discrete data. Discrete data are acquired via numerical measurement, while continuous data are acquired via sampling and testing, and vector data are an important data source in geological discrete data. Geochemical data are the most representative geological data in vector form. Because geological problems have multiple solutions, the same data can be analyzed from different perspectives and different conclusions can be drawn. In a sense, data themselves contain more value than conclusions [5]. With the improvement of test methods and test accuracy, the total amount of geochemical data increases rapidly. The method based on big data involves the processing of the whole geochemical dataset; it reveals the correlation between geochemical model and known mineralization, and provides a new method for finding geochemical anomalies in mineral exploration [6]. Therefore, creating an effective method to manage geochemical data helps us to achieve geochemical analysis based on big data technology.

Vector data play a very important role in geological research, and the geochemically original sampling point data are the most representative. Moreover, vector data are used to make hydrogeological maps of Europe, and to evaluate geological surface processes across the continent [7]. Vector data can also be used to simulate the surface geological model of coastal zones [8], as well as to calculate discharge density, frequency, bifurcation rate, and other parameters in river basins [9]. Furthermore, island paleogeography can be reconstructed by building archipelagic databases in vector data format [10].

Advances in observation instruments and storage technology have led to the upgrading of vector data from MB level to GB level, and the geochemically original sampling point data management model has also changed. Additionally, new methods can provide advantages for research work. Vector databases are established to study the ecological evaluation and correlation index of trees and forests, so as to calculate important parameters such as geological statistical analysis and ecological restoration ability [11]. Moreover, data analysis and protection of river basins can also be strongly supported [12]. Meanwhile, geochemical databases are used to monitor the nutrient content in lakes and rivers, which can help map the chemical spatial patterns related to atmospheric deposition and other environmental pressure sources [13]. Moreover, thematic geochemical databases have also been established in many countries [14–16]. A variety of spatial methods can be applied to produce geochemical patterns using original data stored in a geospatial database [17–19]. Shapefile is a geographic information system (GIS) file format developed by the Environmental Systems Research Institute (ESRI), and is the most widely used vector data format to store the location, shape, and attributes of geographical features [20]. Geochemical data in shapefile format will be used as experimental data.

Recently, the emergence of many advanced data storage technologies has brought more choices for spatial unstructured data management, and also provided a new storage method for the management of geochemical data. Han [21] proposed a spatial data index method based on the HBase database, which makes it possible to deploy an environment with fewer computer resources. Bigtable provides a flexible and high-performance solution for real-time processing of unstructured geological data [22]. Zheng [23] proposed that vector and raster data can be stored and managed uniformly using the Oracle database. At present, with the development of spatial information technology, the amount of vector data increases rapidly. It is very difficult for traditional file management systems to manage vector data at the PB level. Whether it is the column-oriented database HBase, based on the Hadoop distributed file management system; Bigtable, with powerful backstage support from Google; or Oracle, with the largest number of users, there is no specific solution

for geochemical data. Distributed database centers for geological big data need PB-level data centers to store and analyze complete geochemical data. Consequently, the above database technologies have the following limitations in terms of data storage capacity: (1) the inability to create spatial indices due to lack of spatial extension; (2) difficulty in storing geochemical data based on traditional data structure; and (3) failure to achieve distributed database architecture via sharding of spatial data [24].

Cloud computing technology, NoSQL, and distributed database cluster technology may bring new solutions to overcome these problems for geological big data [25,26]. The establishment of geochemical databases in big data environments aims at innovating data storage structures and spatial index methods to store and analyze data efficiently at minimum cost. Therefore, in this paper, two advanced methods are proposed to solve the disadvantages of large-scale geochemical data storage, especially in geochemical data analysis for geological big data. These two new storage methods provide compact data structure, better performance in storage space, and efficient retrieval speed. This paper proposes two innovative storage methods of geochemical data: one is based on the PostgreSQL hexadecimal stream, and the other improves the GeoJSON storage mode based on MongoDB. This study implements a storage method based on MongoDB in GeoJSON form (SSMG), and a storage method based on PostgreSQL with open location code (OLC) geocoding (SSPOG), in order to achieve efficient retrieval and data compression. To test geochemical data in these methods, we utilized geochemical data and basic geological data from Xinjiang, in shapefile format. Moreover, data compression ratio (DCR) was used to evaluate the storage efficiency of the SSMG method and the SSPOG method. In order to accurately test the performance of the two methods, we simultaneously compared the speed of storage and data compression between the two methods. Finally, conclusions and future directions are discussed.

2. Materials and Methods

2.1. Datasets and Environment

In this research work, geochemical data for Xinjiang, in shapefile format, were selected to test the proposed SSMG and SSPOG storage methods. Xinjiang was selected as the study area. Xinjiang is located in the northwest of China, in the center of Eurasia, covering more than 1.66 million square kilometers, accounting for about 1/6 of China's total territory, and has abundant mineral resources (Figure 1). The establishment of a geochemical database provides data support for the evaluation of mineral resources, groundwater pollution monitoring, and ecological monitoring and evaluation. Geochemical surveys, at home and abroad, along with national geochemical data, have been applied in the process of investigation of mineral resources for decades. Therefore, Xinjiang has abundant mineral resources, which is of great significance in the establishment of a geochemical database.

Shapefile data are often used as a data source for experiments [27]. This experiment was designed to test the performance of the SSMG and SSPOG storage methods using geochemical data. Shapefile is a vector graphics format, which can save the location of spatial elements and related attributes, but this format cannot store the topological information of geographical data. At present, many free programs or commercial programs can read shapefile data. Shapefile can store the location data of spatial features, but cannot store the attribute data of these spatial features in a file at the same time. Therefore, shapefile may also be accompanied by a two-dimensional table file to store attribute information for each spatial feature. A complete ESRI shapefile file consists of a main file (.shp), an index file (.shx), and a table file (.dbf). The main file is composed of a fixed-length header and a variable-length record; it is mainly used to keep spatial feature records. The index file contains a 100-byte header and an 8-byte fixed-length record, recording the location of each spatial feature in the main file. The table file contains the characteristic attributes of each spatial feature in the shapefile file. The corresponding relationship between the table file and the spatial feature record in the main file is established by the index file. Therefore, shapefile data are adopted for the storage of geochemical data. Because the

SSMG method is a storage mechanism based on the MongoDB database, shapefile data inserted into the database become a complete document form. The SSPOG method is based on the PostgreSQL database, which is similar to the form of table file in shapefile, but SSPOG integrates shapefile spatial information into hexadecimal code and stores it in the database.

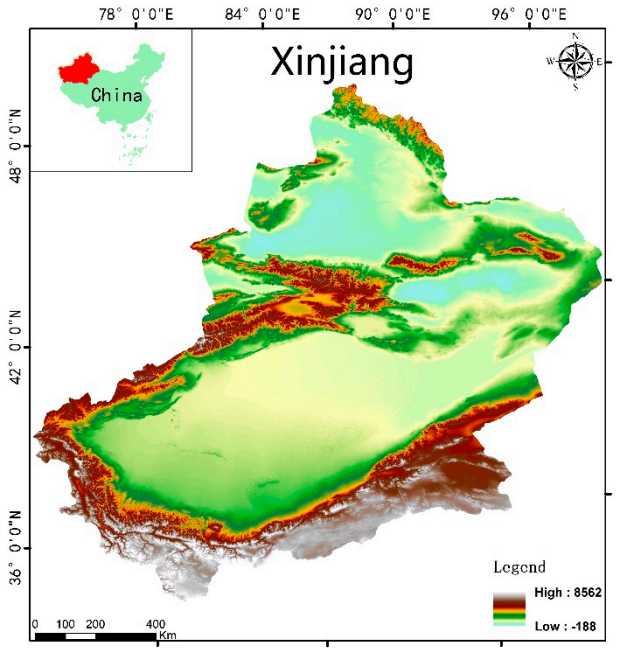


Figure 1. Location map and overview of the study area of Xinjiang (based on map sources: Department of Natural Resources of Xinjiang Uygur Autonomous Regions Xin S (2019) No. 044).

To explain the differences between SSMG and SSPOG, the time consumed by MongoDB and PostgreSQL operations was recorded. Therefore, PostgreSQL and MongoDB were deployed on a single-machine environment, and database visualization software—such as PremiumSoft's Navicat Premium—was deployed to observe the result data. In addition, ArcGIS and QGIS were used to display the result maps, showing the configuration details of each platform (Table 1).

Table 1. Descriptions of testing platforms.

Platform	PostgreSQL	MongoDB
Overview	PostgreSQL runs on a single server.	MongoDB runs on a single server.
Software configuration	PostgreSQL version: 10.0 PostGIS version: 2.4.4 pgAdmin4: 4.4.6	MongoDB version: 4.0.9 MongoDB Compass Community: 1.17
Hardware configuration	CPU: Intel i7-4790 3.4 GHz RAM: 16GB DDR4 3200 MHZ HDD: 1TB 7200 rpm	

2.2. Experimental Design

In our experiment, we tested the SSMG and SSPOG methods with geochemical data in shapefile format. The SSMG method of geochemical data contains two processes—JSONification, and cloud storage—while the SSPOG method of geochemical data contains three processes: geocoding, extension, and data storage. Based on the methodology detailed in Sections 2.3 and 2.4, Python was used to insert geochemical data into the different databases in two ways. In addition, the geochemical data were stored in the

database according to the table structure described in Section 2.5. As shown in Section 3.1, the two storage methods are evaluated by the DCR criterion. Section 3.2 describes the application of geochemical data in the SSMG and SSOG methods. Section 3.3 compares the data storage performance of the two methods through a variety of evaluation criteria and statistical methods.

2.3. SSMG Method

The big data technology group includes three parts: distributed database, parallel computing, and data mining. MongoDB, HBase, Neo4j, and Redis are all popular databases today. MongoDB has the ability to process massive data efficiently [28], supports embedded document objects and array objects [29], and has an automatic sharding mechanism [30]. In addition, MongoDB can provide a high-performance and -availability solution for storing unstructured data. MongoDB stores data in document form. Each document consists of multiple keys and their corresponding values, supports arrays and documents, and can store complex data types. When spatial data are stored in MongoDB, each spatial object is transformed into a JSON object by using the GeoJSON format for spatial data expression, and the spatial and non-spatial attributes of spatial objects are stored in <key,value> mode. Finally, spatial data are serialized into JSON files and stored on disk. GeoJSON defines the following geometric types: Point, LineString, Polygon, MultiPoint, MultiLineString, MultiPolygon, and GeometryCollection. Attributes contain geometric objects and additional information, as well as attribute sets [31]. Compared with the XML data format, GeoJSON supports multiple server-side languages, and is easy to access and extract for the clients, thus reducing the amount of code development on both the server and client sides.

The characteristics of shapefile data stored in GeoJSON are different from relational database storage mechanisms, integrating spatial information and attribute information to ensure consistency [32]. MongoDB was chosen as the container for storing GeoJSON because it is not only a NoSQL distributed database with good performance [33], but also has more advantages in storing document data. In addition, using MongoDB can achieve compatibility with other software. The proposed SSMG method illustrates how to store geochemical data in the form of GeoJSON in the document-type database MongoDB (Figure 2). This method consists of two tiers: JSONification, and cloud storage.

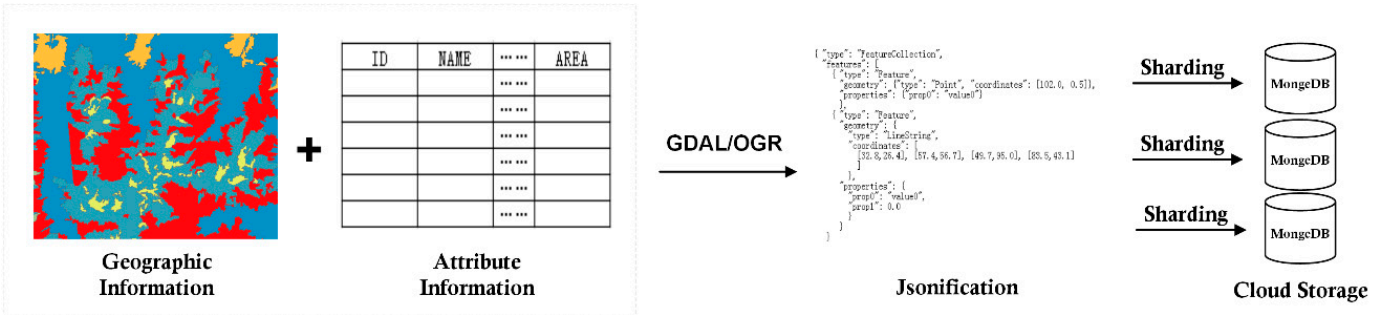


Figure 2. Architecture of the proposed SSMG method.

As a core part of SSMG, the JSONification tier is used to convert geological vector data to GeoJSON format data. The GDAL/OGR spatial database conversion interface is used to process tasks by this tier. The Geospatial Data Abstraction Library (GDAL) is a conversion interface developed by the Open Source Geospatial Foundation (OSGeo) under the Massachusetts Institute of Technology X/MIT license agreement. The OGR Simple Features Library (OGR) is a part of the GDAL, which mainly provides support for vector data, including 84 different types of vector data. The OGR interface treats the shapefile dataset as a whole, and a single shapefile in the dataset as one of the layers. The read driver reads the outer ring clockwise and the inner ring counterclockwise under the polygon specification. If the topological relationship of the shapefile is damaged under the polygon

specification, the configuration option OGR_ORGANIZE_POLYGONS can be reset to complete the analysis of the topological relationship of the original data. The GeoJSON driver supports reading and writing access data in GeoJSON format, as well as the use of GeoJSON for other map service formats, such as GeoServer or CartoWeb. The GeoJSON driver maps five types of element objects—Point, LineString, Polygon, GeometryCollection, and Feature—to new OGRFeature objects. According to the requirements of GeoJSON's specifications, because the members with properties are the characteristics of element objects, every member with properties of OGR objects converted into OGRField type is finally inserted into the corresponding OGRFeature objects. Therefore, the JSONification tier achieves storage of geological vector data in GeoJSON geocoding format.

The cloud storage refers to a distributed database cluster. When more data are stored in the database, a single database cannot meet the storage requirements, nor can it provide acceptable read/write throughput. A distributed database enables the database system to store more data by partitioning the data on multiple other servers. For client users, there is no need to know whether the data are split or not, nor the corresponding server for data sharding. The data sharding task is performed by a route process, which records the storage location of all data and the corresponding relationship between data and shards. The JSONification tier documents the geochemical data, while the cloud storage tier groups the documents into blocks, each consisting of a specified range of keys. The cloud storage tier records the amount of inserted data in each data block, and once the split threshold is reached, the collection of the target database is split. For the client, it simply connects to an ordinary process. In the database service of data request, the location of the target data can be obtained by this process, and the data are collected by the route process and returned to the client. On account of their fast access speed, superior performance, and easy expansion, distributed databases are quite appropriate for geochemical data. Distributed databases provide an easy and fast storage environment for geochemical data.

2.4. SSPOG Method

PostgreSQL is an open-source object-relational database management system, which supports the management of geospatial data. Moreover, some fundamental geometric types have been defined in PostgreSQL. The proposed SSPOG method in this study shows the architecture of SSPOG (Figure 3). The SSPOG method innovatively uses OLC geocoding as the geographic index of vector data, follows a Simple Feature for Structured Query Language (SFS) [34] model to extend geometry objects under Open Geospatial Consortium (OGC) specifications, and stores unstructured geographic data in a spatial database in the form of two-dimensional relational tables. This method consists of three tiers: geocoding, extension, and storage.

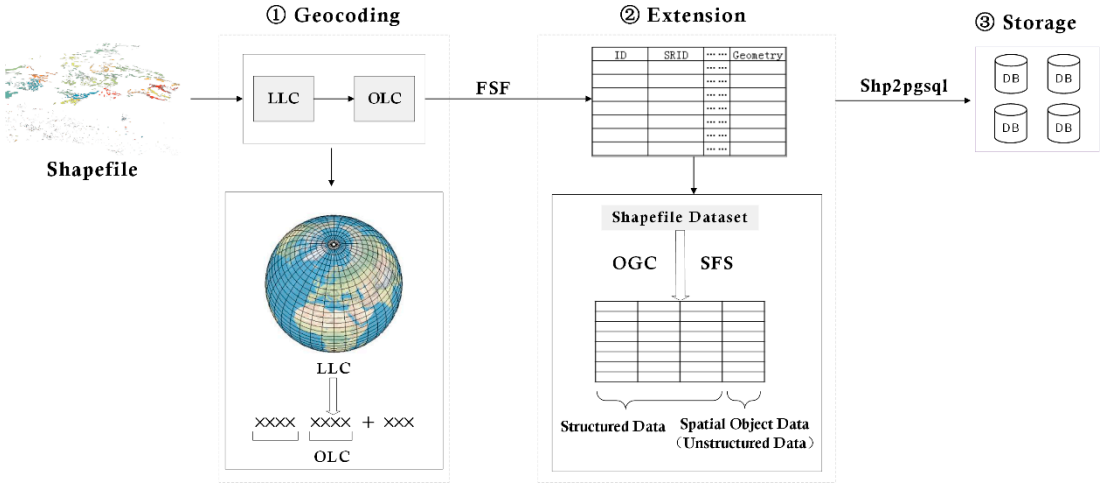


Figure 3. Architecture of the proposed SSPOG method.

The purpose of the geocoding tier is to process the conversion of longitude and latitude of the WGS84 coordinate system to OLC. The input is a large number of longitude and latitude coordinates (LLCs), while the output is a simpler OLC. In the geocoding tier, the conversion interface is transmitted through a dedicated algorithmic reference table supported by the Google Maps spatial engine. The algorithm is authorized to execute under Apache License 2.0. Characters that are not easily confused in more than 30 languages are selected as the OLC code. Meanwhile, each geographic code describes an area consisting of two longitudes and latitudes, as determined by its southwest corner and size. According to the requirement of user request, the geocoding length that meets the accuracy is determined in the geocoding tier. As the geocoding length continues to expand, the target area becomes more precise. When the encoding is extended to 11 characters, the mapping to the Earth's surface can accurately describe the geographical entity with a precision of 3 m. Compared with LLC, OLC coding takes up less space, and is generated by open-source algorithms. OLC coding can identify any part of the Earth, which is an appropriate solution to improve the processing speed and positional identification accuracy of coding.

The extension tier is designed to implement the mapping of geochemical data to geographic entity objects. In order to follow the SFS model specification under OGC, two sets are used to track and report geometries in the database. A collection calls the spatial reference identifier (SRID) to define all known spatial reference systems in the database. The SRID corresponds to a spatial reference system based on a specific ellipse, and can be used for planar or spherical mapping. The extension tier supports the input and output of geological vector data in a variety of formats, including well-known text (WKT), well-known binary (WKB), extended well-known text (EWKT), extended well-known binary (EWKB), and other format types. Among them, the EWKT and EWKB formats are three-dimensional representation formats formally defined by the Structured Query Language (SQL)-Multimedia Part 3 (SQL/MM) specification. According to the request of SFS specification, geochemical data can be fundamentally processed.

The storage tier is the link of executing all types of geochemical data storage. After the model specification of the extension tier, POINT, LINE, POLYGON, POLYGON with a hole, and COLLECTION are used to map geographic entities on the Earth. There are many types of geological data. The client may create geological databases on different topics according to different geological disciplines, including geochemical databases, basic geological databases, and geotectonic databases. Therefore, the storage tier builds different databases according to metadata tables of different topics. Requests for geochemical data are sent through a dedicated job submission interface, which converts the shapefile into spatial databases suitable for insertion into geometric or geographic formats.

Because longitude and latitude require large storage space, and are stored in the form of point features in the database, the efficiency of geochemical data execution is affected. The proposed SSPOG method uses OLC geocoding to accurately describe the common surface elements in geological research with 10–12 characters to meter level, which improves the efficiency of geochemical data, and can quickly and accurately obtain the location information of the target feature. Because the SSPOG method is based on PostgreSQL—a relational database with pluggable type extensions and functional extensions—the spatial and attribute information of geochemical data are therefore used for management in a relational database. Through the extension of geometry objects under the OpenGIS protocol, spatial information is inserted into the database in a hexadecimal system. PostgreSQL distributed function extension technologies—such as Citus, Green Plum, and PL/Proxy—are appropriate choices to support the distributed management of big data technology.

2.5. Design for Data Tables in SSMG and SSPOG to Store Geochemical Data

A dataset is divided into several parts by a relational database, and then stored in the corresponding tables. When the data need to be used, they are spliced together and used. For example, a table describing remote sensing data information is designed according to the third paradigm [35], when different remote sensing data cover a study area. A single

table can be used to store remote sensing images of different time series and read the required data through the association between tables when displaying available remote sensing data. Meanwhile, the geochemical data storage mechanism of SSMG is quite different from this mode. Since this kind of storage unit is a document that supports arrays and nested documents, SSMG can directly describe all attribute information of geochemical data with a documented data structure (Figure 4). Each field in the entity represents a type of information in the SSMG method, and is not a form of table. The association function of a relational database is not necessarily its advantage, but a necessary condition for it to work. In the SSMG method, using its rich document characteristics, it does not require every document to have the same structure, and supports many heterogeneous data scenarios very well. To some extent, association is a pseudo-requirement, which can be avoided by reasonable modeling.

Figure 4. Table designed for geochemical data.

Inheriting the advantages of the geospatial relation–object model, the storage of geospatial set elements conforms to the description and definition of geographic elements by OGC in SSPOG. The structure of the SSPOG method table is mainly divided into two parts: One is a traditional structured attribute column, which meets all the requirements of a traditional relational database paradigm. The other part is the spatial information column, which stores geometric objects in hexadecimal form. Each spatial data record in SSPOG stores a spatial feature, and integrates all tables into a dataset with the same spatial reference system.

1. Geocoding table: Stores the corresponding OLC codes and precise parameters, as well as the converted longitude and latitude coordinates;
2. Geo-Information table: Stores coordinate information of each feature in geochemical data;
3. Age table: Stores sampling time of geochemical data and maximum and minimum ages of strata obtained by field geologists;
4. Tectonic Units table: Stores names of geological tectonic units of different grades in the study area;
5. Zone table: Stores map name and map code of administrative divisions' information and geological maps of the study area;

6. Sample table: Stores the information of the name, description, and type of sampling points when sampling in the field;
7. Geochemistry table: Stores the element content existing in various strata of each sampling point; the first three elements are stored according to the content value.

3. Experiment and Results

3.1. Data Compression Ratio (DCR)

In order to achieve large-scale geochemical data storage, the SSMG and SSPOG methods are used to store unstructured data. There are great differences between the two methods proposed in this research. The former is used to transform the spatial information and attribute information of the shapefile into GeoJSON format and store them in database. The latter is used to extend the spatial information of the shapefile following the OGC protocol, and store it in the database in the form of two-dimensional tables. The increase or decrease in space occupied by data insertion into the database is one of the important evaluation criteria for a data organization mode, and the efficient storage of data is also pursued in the era of big data. Therefore, a new method of evaluating data storage mechanisms—DCR—is proposed in this study. In order to analyze the increase and decrease in the space occupied by two different methods for storing data, firstly, the space occupied by shapefile-encoded experimental data stored on a Windows file system was recorded, which was used as the standard control group for the experiment. Secondly, the experimental data were recorded and stored in different databases using SSMG and SSOG. Thirdly, the amount of space taken up by recording the experimental data in different databases via SSMG and SSPOG was recorded. Finally, the DCR values of different methods were calculated according to (1). The size of DCR represents the efficiency of data storage.

$$R = \frac{D_0 - D_T}{D_0} \tag{1}$$

where R is the DCR of the database, D_T is the space occupied by the experimental group data, and D_0 is the space occupied by the control group data.

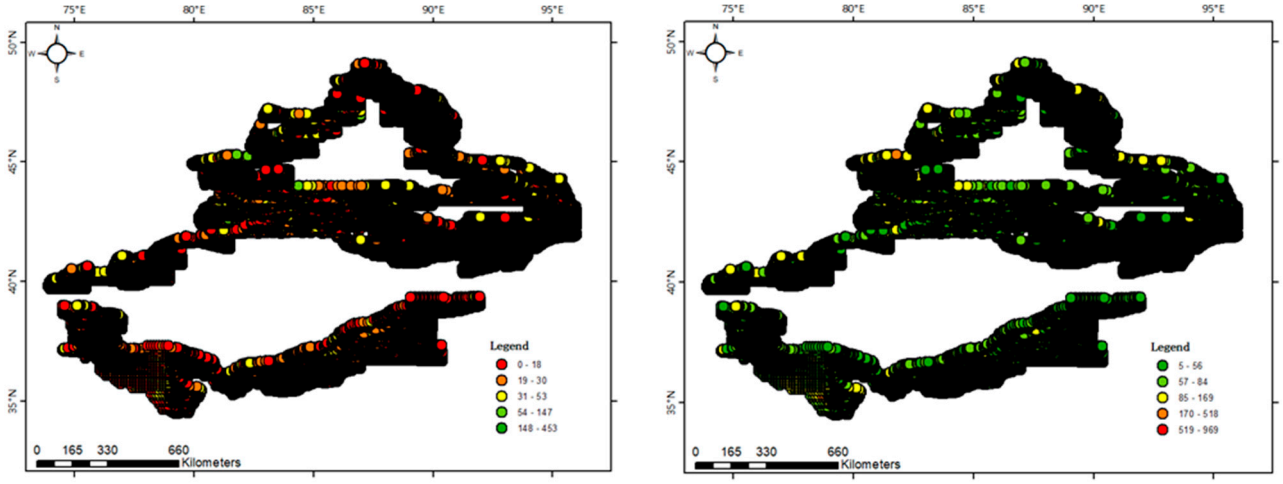
3.2. Geochemical Data Storage and Data Presentation

This study measured the time needed to reconstruct geochemical data into a GeoJSON structure and store it in a two-dimensional table structure. In addition, the time consumed to retrieve data based on the SSMG and SSPOG methods and their corresponding DCR were also measured. The experiment consisted of two steps: storing geochemical data, and mapping them. When using SSMG to store geochemical data, the efficiency of its storage function was evaluated. Three steps were performed in sequence: (1) Clients obtain all the information of geochemical data from the data source by inheriting the GetLayer operation of the GDAL/OGR spatial feature library, and shapefile data are reconstructed into GeoJSON form via the Feature.ExportToJson function. This contains the original data with all the spatial information and attribute information. (2) Clients register data into the MongoDB cluster through the metadata tables already designed in the system to provide data foundation for geological data analysis. (3) At this point, MongoDB divides the documents registered in the database into blocks. When block data reach a threshold, MongoDB divides them into two smaller blocks. Finally, geochemical data are inserted into MongoDB in the form of GeoJSON.

Similarly, when using the SSPOG method to store geochemical data, the efficiency of its storage function was also evaluated. Three steps were performed in sequence: (1) Clients use the DECODE function to encode the shapefile data of the research area, so that each spatial feature can be accurately described by OLC. (2) The SSMG method follows the SFS model specification under OGC to extend shapefile data to geometry objects, describing the spatial information of data in the form of hexadecimal characters. (3) Through the specific model, the structured attribute information and the extended spatial information

are uniformly stored in the two-dimensional table structure, so that clients can analyze spatial data with SQL.

In the process of displaying geochemical data, the geochemical data were retrieved from the database through the application interface accessed by the database, and the data were displayed via the graphical software. Based on the different element content values in geochemical data, the original data were symbolized and displayed, and finally the display results were obtained. The results showed the geochemical element contents of different elements based on shapefile data (Figure 5). Geochemical data contain information about element content in most of the regions. If the kind of data can be used quickly and efficiently, this can provide effective data support for geological big data.



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

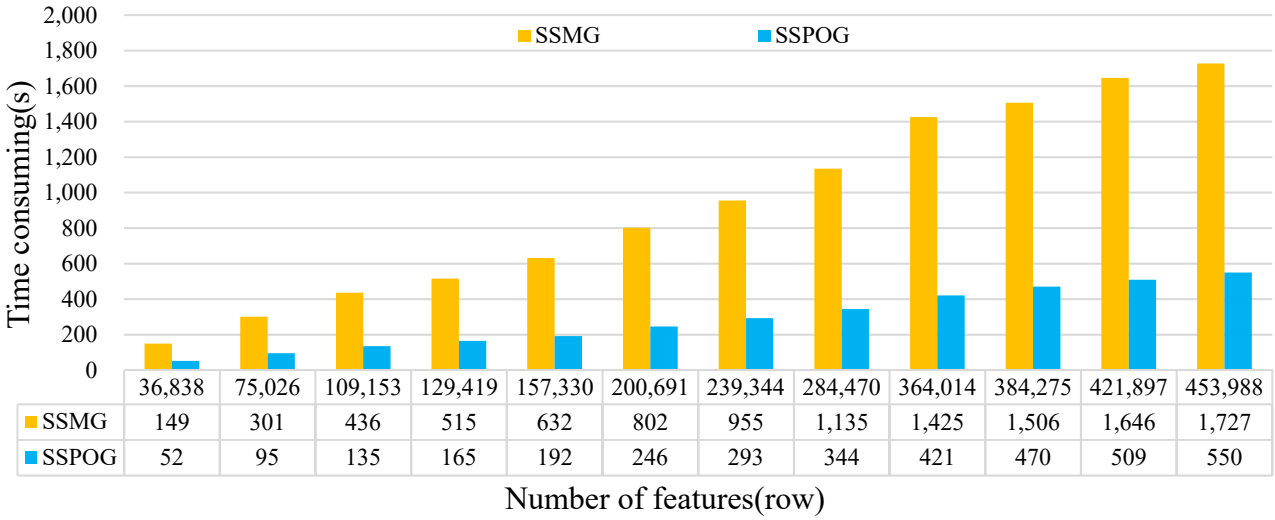
Figure 5. Xinjiang geochemical element display: (**a**) Cu; (**b**) Zn.

3.3. Performance Evaluation

The experiment compared the storage efficiency of SSMG with SSPOG when storing different numbers of features. The SSMG and SSPOG methods are based on open-source servers; the databases of SSMG and SSPOG were MongoDB and PostgreSQL, respectively. Specifically, the experiments of SSMG and SSPOG were carried out in the same hardware environment. Because computer performance would be affected by other processes, the average of three repeated experiments was taken in this experiment. For shapefile data with 129,419, 239,344, and 421,897 features, the time consumed by the SSMG method was approximately 515, 955, and 1646 s, respectively. Meanwhile, the time consumed by the SSPOG method was approximately 165, 293, and 509 s, respectively (Figure 6). When storing 453,988 features, the SSMG method reached approximately 1727 s, while SSPOG reached 550 s. Overall, the SSPOG method was approximately three times more efficient than the SSMG method.

The time consumption growth trend of the SSMG and SSPOG methods was linear with respect to the number of features (Figure 7). The slope of SSMG was approximately 0.0038s/row, while the slope of SSPOG was approximately 0.0012s/row. The SSPOG method is much more efficient than the SSMG method when storing large quantities of geochemical data.

In the same way, this experiment also compared the DCR of the SSMG method with the SSPOG method when storing different numbers of features. For shapefile data with 129,419, 239,344, and 421,897 features, the DRC of SSMG was approximately 22.40%, 22.37%, and 21.43%, respectively, whereas for the SSPOG method it was approximately 53.39%, 53.67%, and 52.07%, respectively (Figure 8). The DRC of SSMG trends to ~22%, while the DRC of the SSPOG method trends to ~53%. Overall, the DRC of SSMG does not reach half that of the SSPOG method.



	36,838	75,026	109,153	129,419	157,330	200,691	239,344	284,470	364,014	384,275	421,897	453,988
SSMG	149	301	436	515	632	802	955	1,135	1,425	1,506	1,646	1,727
SSPOG	52	95	135	165	192	246	293	344	421	470	509	550

Figure 6. Time required for storage of shapefile data using two methods with different numbers of features.

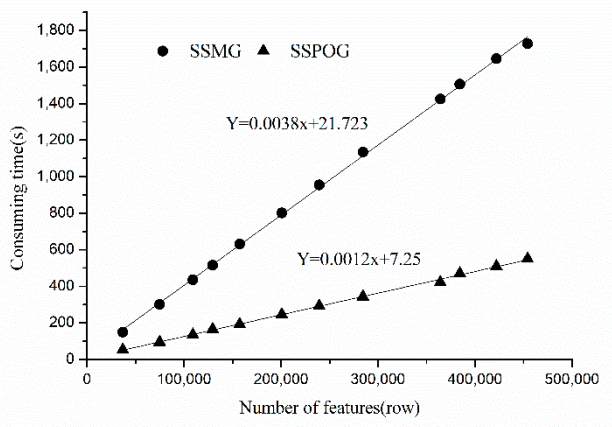
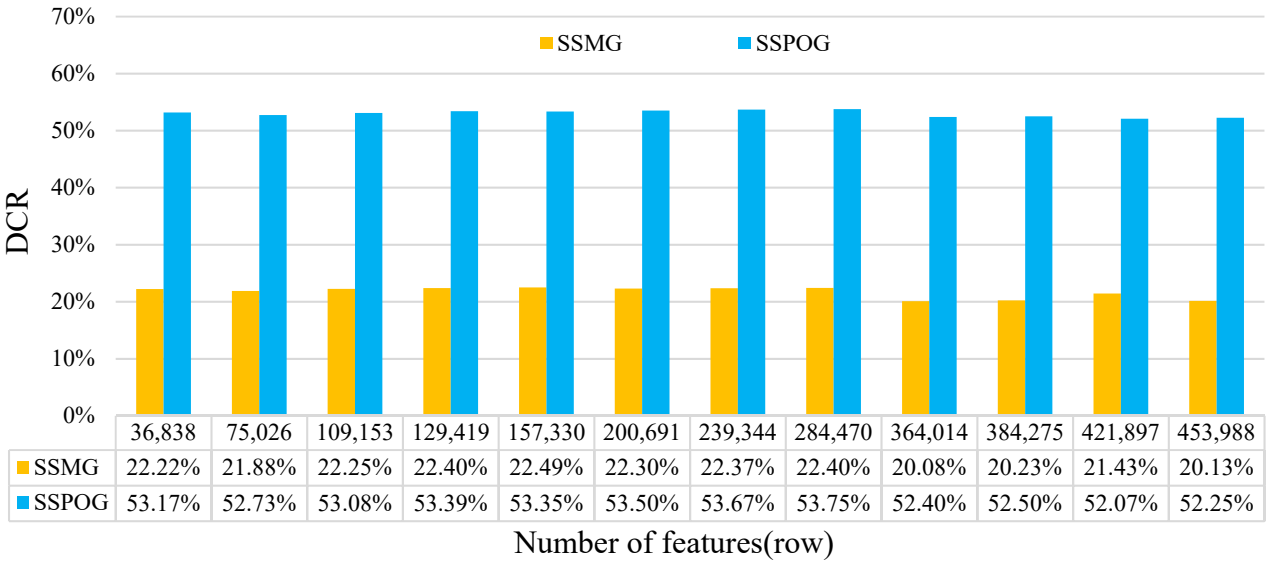


Figure 7. Relationship between the time consumption of geochemical data storage and number of features based on SSMG and SSPOG.

Table 2. Test case of retrieval.

Platform	Test Case	Number of Results
PostgreSQL	SELECT FROM Geodatabse WHERE Ag = 76	1310 rows
MongoDB	db.Geodatabse.find({"properties.Ag": 76})	

In conclusion, the SSPOG method was more efficient when storing different numbers of features. With the number of features increased, the time consumed by SSPOG decreased in comparison with SSMG. Compared with document management systems, the SSMG and SSPOG methods provide new ways to store geochemical data, and support higher storage capacity. Compared with SSMG, SSPOG provides higher and more efficient storage methods (Figures 6 and 8). Meanwhile, using the DCR index, SSPOG provides better compression data capability compared with capacity. However, using different retrieval methods, it is apparent that the SSMG method is better than the SSPOG method in terms of retrieval.



	36,838	75,026	109,153	129,419	157,330	200,691	239,344	284,470	364,014	384,275	421,897	453,988
SSMG	22.22%	21.88%	22.25%	22.40%	22.49%	22.30%	22.37%	22.40%	20.08%	20.23%	21.43%	20.13%
SSPOG	53.17%	52.73%	53.08%	53.39%	53.35%	53.50%	53.67%	53.75%	52.40%	52.50%	52.07%	52.25%

Figure 8. DCR of geochemical data using two methods with different numbers of features. Table 2 shows the performance of testing retrieval under different methods. Dealing with 129,719 features, the time consumed was different with respect to different storage and retrieval methods. Using the collection query method (CQM), the time consumed by the SSMG method was 220 milliseconds. In the same way, the time consumed by the SSPOG method was 2450 milliseconds (Figure 9). Overall, the SSMG method was approximately 10 times faster than the SSPOG method in retrieval.

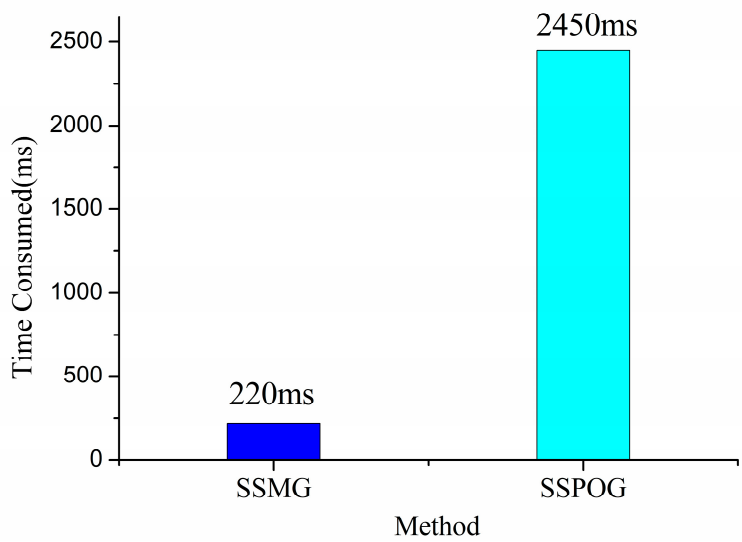


Figure 9. Time consumed for retrieval using SSMG and SSPOG.

4. Discussion

In this experiment, the geochemical data were stored and accessed using the SSMG and SSPOG methods. In the performance evaluation stage, the SSPOG method consumed less time than document methods, such as SSMG. Relational databases are structurally compact and less redundant compared with document databases. The basic structure of shapefile data is to store information in the form of traditional attribute tables. The SSPOG method stores the spatial information of geochemical data as structured data in a relational database after spatial extension. Therefore, the SSPOG method has more advantages than SSMG in terms of saving and compressing data. However, the SSMG method helps to solve the problem of geochemical data storage for retrieval. The document database <key,value> data storage mode eliminates the close relationship between different data in the relational database, and achieves the direct acquisition of target data from the database. Therefore,

the SSMG method performs better in terms of retrieval. The experimental results were compared with one another, and the advantages of SSMG and SSPOG are as follows:

(1) The SSPOG method efficiently stores geochemical data in shapefile format. The SSPOG method can store different types of geographic elements—such as point, polyline, and polygon—in different ways. This storage method enables the same type of data to be invoked to extract multisource data information in geological big data analysis functions. Meanwhile, OLC enables SSPOG to save lots of space and locate target features more accurately, as described in Section 2.2. In terms of storage efficiency and speed, merging two floating-point fields into one character field is an innovation for traditional spatial data storage. As the number of geochemical data increases, so too does the time consumed by SSPOG. Therefore, for the above reasons, the SSPOG method improves the efficiency of storing geochemical data;

(2) The SSMG method innovates the storage form of geochemical data and improves the retrieval efficiency. On account of the increasing accuracy and complexity of geological data description, it is difficult to implement retrieval in large-scale data in an efficient way. The vector format of geochemical data is expressed in the form of <key,value> by SSMG, which breaks through the complex relationships between attributes in relational databases. As mentioned in the conclusion, the storage method is much faster than retrieval in relational database space. Through geochemical data storage in GeoJSON format, this vector data storage method supports a two-dimensional spherical spatial index, and solves the application problem of location-based service (LBS), so it is suitable for large-scale retrieval research. Meanwhile, the clustering technology of MongoDB enables a vector dataset to be segmented and stored on different data nodes, which provides a technological foundation for the distributed analysis and calculation of geochemical data.

Challenges still remain in terms of data storage structure and database organization; more efficient storage methods of geochemical data can be established to achieve geological big data storage. Future work will focus on the following: (1) The OLC unique coding and matching technology of vector features' locations and geometric features can solve the problem of unified coding of elements in geochemical data. Through the uniform coding of geological entities, the matching of geological spatial features can be converted into document format via coding matching, which can improve the matching efficiency of geological data. (2) Storing a large amount of geochemical data in different clusters can make full use of idle computer resources, and improve the data availability and performance of large database retrieval servers. Therefore, database cluster sharding technology will be the focus of our next work.

5. Conclusions and Future Work

This study implemented unstructured spatial data storing methods to improve the storage efficiency of vector data and achieve shapefile data application in the retrieval of geochemical data. Our experiment demonstrated that the SSPOG and SSMG methods achieved creative geochemical data storage and retrieval at a large scale. These two methods showed different performance in storing and retrieving geochemical data. In terms of storage performance, the efficiency of geochemical data storage in SSPOG can be threefold greater than that of SSMG. The SSPOG method showed the advantage of the close data structure of the relational database through spatial extension under OGC standard. In terms of data compression, through the DCR index proposed in this paper, the efficiency of data compression in SSMG was better than that of SSPOG. Meanwhile, the retrieval performance of SSMG was better than that of SSPOG; that is to say, the SSMG method was able to complete real-time geological retrieval tasks with excellent performance when storing geochemical data at a large scale. Because the SSMG method uses a document structure to store geochemical data, it can obtain a looser structure, so it performs better in terms of data compression and retrieval. In fact, 90% of the time consumed in storing geochemical data in SSMG is a process of documentation, which takes only a short time

to insert document data into the database. Therefore, documented vector data have more advantages in optimizing storage space and retrieval.

Compared with the traditional retrieval of geochemically original data, the two geochemical data management models based on big data technology proposed in this paper show effective improvement. It takes less than 1 s to find the target data from 460,000 records, which is an efficiency that cannot be achieved by the traditional geochemically original data management model. On the basis of these management models, the abnormal values in the massive geochemical data can be quickly found and processed. At the same time, the core of geochemical big data analysis is to retrieve the target data from the massive data for processing and analysis, and the methods proposed in this paper can provide efficient technological support. In addition, the SSPOG and SSMG methods have their own advantages and disadvantages in terms of storage and retrieval performance. Under different conditions, different methods can be selected.

At present, the focus of our research is on the improvement of spatial data storage performance and retrieval by range index attributes. In future works, the spatial index will be the focus of our research. In the two methods proposed in this paper, the use of a spatial index can increase the accuracy of data retrieval, and in different application scenarios can also improve the efficiency of data retrieval.

Author Contributions: Framework design, J.W. and K.Z.; methodology, J.W., Y.C., and J.Y.; software, Y.C.; validation, Y.C.; formal analysis, J.W.; investigation, K.Z.; writing—original draft preparation, Y.C.; writing—review and editing, Y.C., J.W., and S.C.; visualization, Y.C.; supervision, P.D.M. and T.V.d.V.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Key R&D Program of China (2018YFC0604001-3), B&R Team of Chinese Academy of Sciences (2017-XBZG-BR-002), National Natural Science Foundation of China (No. U1803117, No. U1803241), and Chinese Academy of Sciences President's International Fellowship Initiative (PIFI Grant No. 2017VCA0002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: We would like to thank the Xinjiang Laboratory of Mineral Resources and Digital Geology of the Chinese Academy of Sciences for guidance and full support.

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

References

1. Zuo, R. Exploring the effects of cell size in geochemical mapping. *J Geochem. Explor.* **2012**, *112*, 357–367. [CrossRef]
2. Staudigel, H.; Albarede, F.; Anderson, D.L.; Derry, L.; McDonough, B.; Shaw, H.F.; White, W.; Zindler, A. Electronic data publication in geochemistry: A plea for "full disclosure". *Geochem. Geophys. Geosyst.* **2013**, *2*, 2001GC000234. [CrossRef]
3. Guo, H.; Wang, L.; Chen, F.; Liang, D. Scientific big data and digital earth. *Sci. Bull.* **2014**, *59*, 5066–5073. [CrossRef]
4. Costa, J.F.; Koppe, J.C. Assessing uncertainty associated with the delineation of geochemical anomalies. *Nat. Resour. Res.* **1999**, *8*, 59–67. [CrossRef]
5. Staudigel, H.; Helly, J.; Koppers, A.A.P.; Shaw, H.F.; McDonough, W.F.; Hofmann, A.W.; Langmuir, C.H.; Charles, H.; Lehnert, K.; Sarbas, B.; et al. Electronic data publication in geochemistry. *Geochem. Geophys. Geosyst.* **2003**, *4*, 8003. [CrossRef]
6. Zuo, R.; Xiong, Y. Big data analytics of identifying geochemical anomalies supported by machine learning methods. *Nat. Resour. Res.* **2017**, *27*, 5–13. [CrossRef]
7. Duscher, K.; Günther, A.; Richts, A.; Clos, P.; Philipp, U.; Struckmeier, W. The GIS layers of the "international hydrogeological map of Europe 1:1,500,000" in a vector format. *Hydrogeol. J.* **2015**, *23*, 1867–1875. [CrossRef]
8. Scarelli, F.M.; Barboza, E.G.; Cantelli, L.; Gabbianelli, G. Surface and subsurface data integration and geological modelling from the Little Ice Age to the present, in the Ravenna coastal plain, northwest Adriatic Sea (Emilia-Romagna, Italy). *Catena* **2017**, *151*, 1–15. [CrossRef]
9. Joseph, M.V.; Dinesh, A.C.; Jayappa, K.S. Quantitative analysis of morphometric parameters of Kali River basin, southern India, using bearing azimuth and drainage (bAd) calculator and GIS. *Environ. Earth Sci.* **2014**, *72*, 2887–2903. [CrossRef]

10. Norder, S.J.; Baumgartner, J.B.; Borges, P.A.V.; Hengl, T.; Kissling, W.D.; van Loon, E.E.; Rijsdijk, K.F. A global spatially explicit database of changes in island palaeo-area and archipelago configuration during the late Quaternary. *Glob. Ecol. Biogeogr.* **2018**, *27*, 500–505. [CrossRef]
11. Hurley, P.T.; Emery, M.R. Locating provisioning ecosystem services in urban forests: Forageable woody species in New York City, USA. *Landscape Urban Plan.* **2018**, *170*, 266–275. [CrossRef]
12. Sullivan, D.G.; Batten, H.L. Little River Experimental Watershed, Tifton, Georgia, United States: A historical geographic data-base of conservation practice implementation. *Water Resour. Res.* **2007**, *43*. [CrossRef]
13. Williams, J.; Labou, S.G. A database of georeferenced nutrient chemistry data for mountain lakes of the Western United States. *Sci. Data.* **2017**, *4*, 170069. [CrossRef]
14. Sánchez-Ruiz, S.; Maselli, F.; Chiesi, M.; Fibbi, L.; Martínez, B.; Campos-Taberner, M.; Gilabert, M.A. Remote sensing and bio-geochemical modeling of forest carbon storage in spain. *Remote Sens.* **2020**, *12*, 1356. [CrossRef]
15. Jarva, J.; Tarvainen, T.; Reinikainen, J.; Eklund, M. TAPIR–Finnish national geochemical baseline database. *Sci. Total Environ.* **2010**, *408*, 4385–4395. [CrossRef] [PubMed]
16. Huang, C.; Shibuya, A. High accuracy geochemical map generation method by a spatial autocorrelation-based mixture interpolation using remote sensing Data. *Remote Sens.* **2020**, *12*, 1991. [CrossRef]
17. Zuo, R.; Carranza, E.J.M.; Wang, J. Spatial analysis and visualization of exploration geochemical data. *Earth Sci. Rev.* **2016**, *158*, 9–18. [CrossRef]
18. Declercq, Y.; Delbecque, N.; De Grave, J.; De Smedt, P.; Finke, P.; Mouazen, A.M.; Nawar, S.; Vandenbergh, D.; Van Meirvenne, M.; Verdoodt, A. A comprehensive study of three different portable XRF scanners to assess the soil geochemistry of an extensive sample dataset. *Remote Sens.* **2019**, *11*, 2490. [CrossRef]
19. Wang, Z.Y.; Zuo, R.G.; Dong, Y.N. Mapping geochemical anomalies through integrating random forest and metric learning methods. *Nat. Resour. Res.* **2019**, *28*, 1285–1298. [CrossRef]
20. Zhu, J.; Wang, X.; Wang, P.; Wu, Z.; Kim, M.J. Integration of BIM and GIS: Geometry from IFC to shapefile using open-source technology. *Automat Constr.* **2019**, *102*, 105–119. [CrossRef]
21. Han, D.; Stroulia, E. HGrid: A Data Model for Large Geospatial Data Sets in HBase. In Proceedings of the IEEE 6th International Conference on Cloud Computing (CLOUD), Santa Clara, CA, USA, 27 June–3 July 2013; pp. 910–917. [CrossRef]
22. Chang, F.; Dean, J.; Ghemawat, S.; Hsieh, W.C.; Wallach, D.A.; Burrows, M.; Chandra, T.; Fikes, A.; Gruber, R.E. Bigtable: A Distributed Storage System for Structured Data. In Proceedings of the 2006 USENIX Symposium on Operating Systems Design and Implementation (OSDI), Seattle, WA, USA, 6–8 November 2006; pp. 205–218. [CrossRef]
23. Zheng, W.; Chengming, L.; Pengda, W.; Jianming, S.; Wei, S. Integrated Storage and Management of Vector and Raster Data Based on Oracle Database. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 639–648. [CrossRef]
24. Cheng, Y.; Zhou, K.; Wang, J.; Yan, J. Big Earth Observation Data Integration in Remote Sensing Based on a Distributed Spatial Framework. *Remote Sens.* **2020**, *12*, 972. [CrossRef]
25. Zhou, L.; Chen, N.; Chen, Z.; Xing, C. ROSCC: An Efficient Remote Sensing Observation- Sharing Method Based on Cloud Computing for Soil Moisture Mapping in Precision Agriculture. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 5588–5598. [CrossRef]
26. Chen, Q.; Liu, G.; Ma, X.; Mariethoz, G.; He, Z.; Tian, Y.; Weng, Z. Local curvature entropy-based 3D terrain representation using a comprehensive Quadtree. *ISPRS J. Photogramm.* **2018**, *139*, 30–45. [CrossRef]
27. Zhang, C.; Fay, D.; Mcgrath, D.; Grennan, E.; Carton, O.T. Use of trans-Gaussian kriging for national soil geochemical mapping in Ireland. *Geochem. Explor. Environ. Anal.* **2008**, *8*, 255–265. [CrossRef]
28. Dai, C.; Ye, Y.; Liu, T.J.; Zheng, J.J. Design of high performance cloud storage platform based on cheap pc clusters using MongoDB and Hadoop. *AMM* **2013**, *380–384*, 2050–2053. [CrossRef]
29. Barnaghi, P.; Wang, W.; Henson, C.; Taylor, K. Semantics for the internet of things. *Int. J. Semant. Web. Inf.* **2012**, *8*, 1–21. [CrossRef]
30. Liu, Y.; Wang, Y.; Jin, Y. Research on the improvement of MongoDB Auto-Sharding in cloud environment. In Proceedings of the 7th International Conference on Computer Science & Education, Melbourne, Australia, 14–17 July 2012; pp. 851–854. [CrossRef]
31. Howard, B. The GeoJSON Format Specification. 2010. Available online: https://geojson.org/geojson-spec.html (accessed on 22 June 2020).
32. Gomes, V.; Queiroz, G.; Ferreira, K. An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sens.* **2020**, *12*, 1253. [CrossRef]
33. Yoon, J.; Jeong, D.; Kang, C.; Lee, S. Forensic investigation framework for the document store NoSQL DBMS: MongoDB as a case study. *Dight Invest.* **2016**, *17*, 53–65. [CrossRef]
34. Solihin, W.; Eastman, C.; Lee, Y.-C.; Yang, D.-H. A simplified relational database schema for transformation of BIM data into a query-efficient and spatially enabled database. *Automat Constr.* **2017**, *84*, 367–383. [CrossRef]
35. Wang, X.S.; Bettini, C.; Brodsky, A.; Jajodia, S. Logical Design for Temporal Databases with Multiple Granularities. *ACM Trans. Databse Syst.* **1997**, *22*, 115–170. [CrossRef]