GEOWEBCLN: AN INTENSIVE CLEANING ARCHITECTURE FOR GEOSPATIAL METADATA

SAVITA KUMARI SHEORAN , VINTI PARMAR

Department of Computer Science and Engineering, Indira Gandhi University Meerpur, Rewari, India

Manuscript received: September 22, 2021
Revised version: January 10, 2022

SHEORAN S.K., PARMAR V., 2022. *GeoWebCln*: An intensive cleaning architecture for geospatial metadata. *Quaestiones Geographicae* 41(1), Bogucki Wydawnictwo Naukowe, Poznań, pp. 51–62. 16 figs, 1 table.

ABSTRACT: Developments in big data technology, wireless networks, Geographic information system (GIS) technology, and internet growth has increased the volume of data at an exponential rate. Internet users are generating data with every single click. Geospatial metadata is widely used for urban planning, map making, spatial data analysis, and so on. Scientific databases use metadata for computations and query processing. Cleaning of data is required for improving the quality of geospatial metadata for scientific computations and spatial data analysis. In this paper, we have designed a data cleaning tool named as *GeoWebCln* to remove useless data from geospatial metadata in a user-friendly environment using the Python console of QGIS Software.

KEYWORDS: spatial data, spatial metadata, data cleaning, spatial database, GIS

Corresponding author: Vinti Parmar, vintiparmar1487@gmail.com

Introduction

The current enhancement in technology has led to the wide generation and usage of data. Spatial data is also easily available and accessible to different types of users. It is creating opportunities for different types of users as it is freely available and easily accessible. The data are largely used by private and public organisations for planning and data analysis. Spatial data cannot be handled with traditional tools and techniques of data mining, so a new concept of spatial data warehouse and spatial database has evolved. Spatial online analytical processing (SOLAP) tools are used to provide an interactive environment for the representation of spatial data. Spatial metadata is the data that explains spatial data. Spatial metadata was required by mapping organisations for production and management of the dataset. Now advancement in technology and the internet have enhanced the use of geospatial metadata for searching spatial data. Spatial metadata aids users in deciding about the relevant data from large datasets. The data used for map production, urban planning, landcover, and so on. must be free from dirty data, that is useless data from the dataset. However, manual cleaning of data using the geographical information systems (GIS) function is not easy for end users because of the following reasons:

- It requires deep knowledge of the cleaning functions of the software, which is not provided during training.
- Both the cleaned and uncleaned data give the same results in the vector model and lack

© 2022 Author(s)
This is an open access article distributed under the Creative Commons Attribution license

https://doi.org/10.2478/quageo-2022-0004
ISSN 2082-2103, eISSN 2081-6383

visual analysis and do not enhance the confidence of using data for any scientific problem.
- The summary of cleaned data is unavailable and hence cleaning is always required on the same data by repeating all manual cleaning steps as the user is unaware of the deleted fields. The summary holds information about the cleaned fields and succours the future use of data for purposes of analysis.

To handle these issues in data selection and analysis, we have designed a prototype that can clean data in a user-friendly environment by taking input from the user and provide cleaning information that can be saved as CSV files and reused whenever the user needs to know about the quality of data. In this paper, we have designed and developed a novel cleaning prototype called *GeoWebCln* to clean the attribute table of the spatial data of Gurugram district. Finally, analysis and visualisation are performed on the cleaning done by *GeoWebCln*.

Background

Many eminent scholars have worked in this field and their findings are valuable for this research work. Mainly, 22 studies have been found to be very effective and impressive and have provided direction to this research work. Atluri and Chun (2004) state that geospatial data can be accessed by both nongovernmental organisations and government agencies. Spatial data can be accessed by downloading data from geo-portals, sending data on secondary storage devices, and from business partners. Sheoran and Parmar (2020) used the spatial data of Gurugram District for performing multicriteria analysis and decision making. Azri et al. (2014) concluded that metadata is an essential requirement for discovery, assessment, access, understanding and standardisation of geospatial information. Therefore, spatial metadata has become an essential component of any standard data repository. Gaikwad et al. (2014) state that metadata is a vital part of data for describing its relevancy, characteristics, uniqueness, freshness, purpose and interoperability of dataset with the components.

Spatial data quality is affected by the quality of their sources (Jakobsson 2002). Lim (2010) concluded that a reduced data quality leads to unexpected results. Error-free data having salient data quality is a must and to achieve this, data cleaning is highly necessary. When data are integrated from multiple sources, then errors present at a single source also propagates with them and amalgamates at a single place. Eldrandaly et al. (2019) stated that spatial data is gathered from multiple data sources having anomalies and errors and cannot be considered fit for analysis, planning and decision-making purposes. Data cleaning is mandatory for data storage and information management. Zhao et al. (2019) proposed a spatial and temporal compression framework CLEAN to compute and maintain trajectory data. Zylshal (2020) performed visual and statistical analysis for performing topographic correction to reduce reflectance variability in mountainous regions. Bielecka and Burek (2019) compared and analysed research on the quality and uncertainty of spatial data. Data cleaning is performed on a data set using data cleaning functions. Data are first examined to detect errors and then cleaned using data cleaning techniques. Dirty data can be in the form of missing values, duplicate values and extraneous attributes in a spatial dataset. This dirty data spoils the quality and suitability and makes the data unfit for any particular application and scientific problem of data analysis. Errors enter into the dataset at the *data collection, data input, data storage*, *data manipulation, data output stages*. Data cleaning is a time-consuming and expensive process, and requires experience and knowledge. While error-free data only gives accurate results, dirty data can produce wrong results and is risky to use. Data cleaning is performed to remove duplicate entries, null entries, useless attributes, misspelled entries and so on. According to Deshmukh and Wangikar (2011), the following are some of the techniques used by researchers to clean data.

Border detection algorithm

The border detection algorithm was developed by Arturas Mazeika and Michael H.B Ohlen in 2006. It performs cleaning of string data in two steps. In the first step, a cluster is formed near the string data by connecting the border and the centre of the hyper-spherical; in the second step the

cluster string is cleansed by the repeated cluster. This algorithm is simple and yields a clean output for string data.

Token-based data cleaning

This technique uses smart tokens to identify duplicate records and lowers the dependency of data cleaning on the match threshold.

Record linkage similarity measures algorithm: This technique is used to compare two relational tuples for their similarity.

Koshley and Halder (2015) proposed an abstraction-based data cleaning approach that results in instances of abstract domains. Many errors enter in data like typo, measurement and data integration errors that are harmful during decision making and analysis.

Kumar and Khosla (2018) have shown the value of data cleaning by working on the pollution dataset. Dirt, errors and noise present in data hamper the data quality. They performed a survey, analysis and visualisation on dirty unstructured data of air. Li and Chen (2014) stated that the Geospatial Sensor Web performs resource access, query, discovery and resource visualisation. Parmar and Sheoran (2021) performed context-based cleaning on the population dataset using the record linkage technique. Ridzuan and Zainon (2019) state that data cleansing is a time-consuming and complex process, but after data cleansing, the quality of data is enhanced and its verification and validation can be tested. Zou et al. (2018) reviewed the technology and applications of geospatial data for better understanding of large and complex datasets using visualisation platforms.

According to Boella et al. (2019), the Volunteered Geographic Information System collects and distributes user-generated content having geographic components. Keim (2002) classified the visualisation techniques into dense pixel display, iconic display, standard 2D/3D display, and interaction and distortion techniques.

Yoshizumi et al. (2020) defines geospatial analysis as a tool or analysis that was specifically designed for geospatial data or applications. Visualisation techniques have immense potential to communicate geospatial data. Thiyagalingam and Getov (2006) organised and illustrated applications of metadata in a hierarchical fashion. Zhao, Huang (2010) found that quality determination of online analytical processing (OLAP) metadata is difficult due to the structural essential of metadata.

Metadata information quality criteria

Quality is a primary requirement for evaluating the value of any product. Any product that is poor in quality cannot be considered worthy and gets neglected. The data and information quality can be determined by analysing some parameters. Spatial data is complex and requires information that can explain its types and usage. The quality of spatial data is determined by using its metadata. In terms of spatial data, quality refers to completeness, accuracy, and consistency as seen in Figure 1. These spatial data quality elements are consumed by various organisations for different applications. GIS users easily access and edit spatial data using Google earth, google maps, GIS tools and social media sites. This enormous production and use of spatial data create difficulty in maintaining the quality of spatial datasets. Data quality refers to adherence to the excellence of data to meet a given set of objectives. Data quality standards are assessed by mapping agencies and the private sector for producing good results. Data collected from different sources vary in terms of displacements, orientation and resolution.

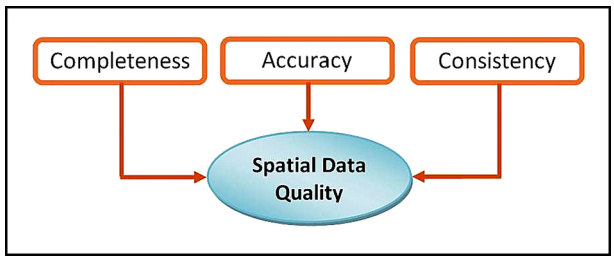


Fig. 1. Data quality parameters.

GIS evolved in the 1960s, and after the 1970s spatial data increased at a high speed due to satellite images. From the 1980s, spatial data quality has become a topic of concern in the GIS community. Acceptance of GIS on a large scale increased the number of digital spatial GIS users in different areas. The fall in prices of computers and easy availability of spatial data in digital

form shifted consumers from analogue to digital data. The management of spatial data in digital form is much easier than managing it in analogue form. Satisfying the quality requirements in spatial data was a big challenge in the geospatial data community. It symbolises the fitness of data for use. Spatial data having quality can be used for any specific application as it is free from errors and produces the right results. The quality requirement is different for different types of users. For data producers, it is related to the inherent nature of the data and for the user it is fitness for use. The presence of errors in spatial data degrades its quality. Errors enter at different stages of data collection, entry and storage in the databases. These errors are identified and detected using error detection procedures to maintain and manage the quality of the data. However, to remove all errors together is difficult in spatial data as it is complex, but proper recognition and reduction of these errors is highly essential to maintain its quality. Depending on the accuracy, completeness and consistency of data, users take decisions in their planning and use.

Completeness

Completeness refers to the presence of all attributes in the dataset to the reference dataset. It is the number of omitted and committed objects. The method of evaluation of completeness determines the quality measure; it can be measured using integer, percentage, Boolean and ratio values. Commission and omission are the sub-elements of completeness.

Consistency

Consistency in spatial data refers to adherence to the syntactic rules used to describe the schema of the database. It can be measured in percentage, ratio, integer and Boolean values.

Accuracy

Accuracy is considered as an important parameter to judge the data quality of spatial data. It is the difference in values in the dataset to the values in the reference dataset.

Spatial data management is presented by a myriad of research papers for different areas such as healthcare, city modelling, remote sensing, image classification and spatial game analytics. Spatial metadata is widely used by different applications without determining the quality of the data. During various stages of data collection, many errors propagate with data. Before application of the data, these errors need to be uncovered and removed using data cleaning algorithm. Dirty spatial data produce invalid results that hamper the data analysis and visualisation processes. Manual cleaning of attribute data of spatial data using GIS functions available in the QGIS software have been performed by Parmar and Sheoran (2021), but the following issues and concerns were found after manual cleaning:

- *Cleaning of data using spatial technologies by non-expert users:* Non-expert users are not aware of available GIS functions that can be used for data cleaning.
- *Assessment of quality of data:* Quality assessment is not possible without any quality information. Users are not provided with any information related to cleaned attributes.
- *Analysis of spatial data:* Spatial data analysis needs complete knowledge about GIS functions. The cleaned spatial data layer needs to store automatically for its proper analysis and visualisation.
- *Security assurance of cleaned* data: Information about deleted attributes is not given and the data if reused lack the information about deleted attributes.
- *Perseverance of quality information:* Cleaned data information should be saved for future purposes.

Cleaning is highly important before making use of data. The data cleaning process involves the following stages:

1. Deep data analysis – detailed analysis is required to know the different types of errors, quality problems, inconsistencies, and anomalies present in the data.
2. Identification of data transformation workflow – data integrated from different sources have different schema, errors, dirtiness and heterogeneity. The degree of heterogeneity and dirtiness in data determines the kind of mapping rules and transformation workflow needed for cleaning.
3. Data testing – verification of transformation workflow and rules is needed to test their accuracy and effectiveness for data cleaning.

4. Transformation – in this phase, transformation steps are executed and query operations are run to clean data.
5. Retreat of data – in this phase dirty data is replaced by clean data to maintain quality at the original source and to prevent doing the same cleaning work again.

Spatial data management is presented by a myriad of research papers for different areas such as healthcare, city modelling, remote sensing, image classification and spatial game analytics. But the untidy data gives wrong results and needs cleaning for proper data analysis and correct decision analysis and cannot be used to solve any of the above scientific problems. Data are massively produced, stored and accessed by millions of people around the globe. Advancement in technology and availability of resources are the main reasons for easy access to spatial data. Spatial data integrated from different sources is full of errors that can reduce the quality of data; hence, these errors need to be removed using any data cleaning method. Data quality can be improved by performing data cleansing on the dataset. Spatial data quality refers to the accuracy, consistency, integrity, and completeness of data.

Cleaning of spatial metadata is not possible by all types of users, but only professional and expert users can perform contextual metadata cleaning. For proper analysis and visualisation of spatial data, it should be free from dirty data. So in this research, we have developed a *GeoWebCln* tool that can clean the dirty metadata by taking input from a user and produce the quality information of cleaned data.

Data source

Vector data of Gurugram District in shapefile format was collected from the Society for Geoinformatics and Sustainable Development (SGSD) and added to QGIS. All the data used in the study are authentic, as they have been received from responsible and authorised agencies and used after verification.

Framework of *GeoWebCln* tool

The framework for the proposed model is depicted in Figure 2. First, collected spatial data is added to the vector layer of the QGIS software. The spatial data layer is investigated for analysis and use. If data contains dirty values and needs cleaning and requires supply of quality information for future use, then the designed *GeoWebCln* tool can be used for cleaning the attribute table. The metadata information of the cleaned layer is given in tabular format and the summary of cleaned attributes is also generated. The quality information about the cleaned attributes and metadata information of the cleaned layer can be exported in the database for future use.

The algorithmic steps followed by the *GeoWebCln* tool for cleaning vector data are given as follows:

Algorithm: *GeoWebCln* Algorithm

INPUT: Spatial data from different sources having dirty attribute data.

OUTPUT: Clean data with summary of cleaning information.

- Select geospatial data (vector data) layers in QGIS software.
- Add geospatial data (vector data) layers in QGIS software.
- Import the *GeoWebCln* tool in the python console of QGIS
- For each vector data layer
  - Select geospatial data (vector data) layers in the *GeoWebCln* tool to clean.
  - Execute the *GeoWebCln* tool using the run tab.
  - (Auto cleaning is done by the *GeoWebCln* tool taking input from users and Perform the below steps)
  - Empty values of the required (area) field are calculated.
  - Null values, duplicate values, 0 values are removed from the attribute table.
  - Extraneous fields are searched and deleted.
  - Cleaning process completed and cleaned layer saved as a new layer.
  - Metadata and summary of cleaned data is generated and displayed.
  - Metadata and a summary can be exported as CSV files for future reference.

The *GeoWebCln* algorithm is followed by the *GeoWebCln* tool in section Visual Analysis Using *GeoWebCln* Tool of this paper for cleaning of attribute table of spatial data added to QGIS. The cleaned layer free from dirt and error can be used for the analysis and visualisation of spatial data. A framework of the *GeoWebCln* Tool is shown in Figure 2.

The following steps are performed in the design of the automated *GeoWebCln tool*:

1. Automatic Cleaning Step.

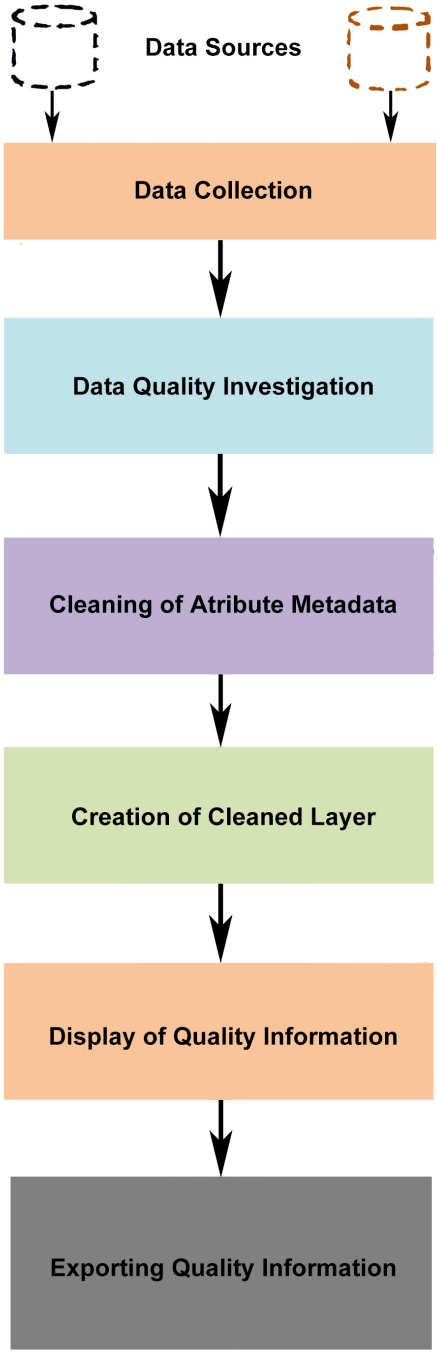


Fig. 2. Framework of the *GeoWebCln* tool.

2. Cleaning Status Message.

Once the user loads spatial data layers into the QGIS, the user can select any layer to start the cleaning process on it. During execution of the code, the program will evaluate various cases in which it will decide whether to remove features/ attributes automatically or ask the user whether it is necessary to remove anything *extraneous*.

The proposed *GeoWebCln* system can be visualised through the function shown in Figure 3.

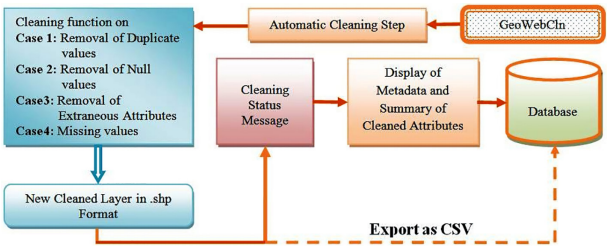


Fig. 3. Data cleaning process using *GeoWebCln* tool.

Automatic cleaning step

Case 1: Removal of duplicate values

Executing the code, the program will select these **features** (blue selection) and it will delete them, as shown in Figure 4, because:

a) They have duplicated the 0 BOUNDAR_ID value;
b) They have NULL Name value.

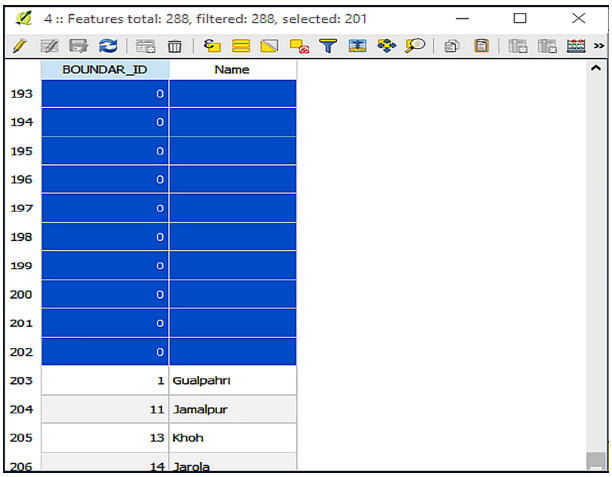


Fig. 4. Duplicate values.

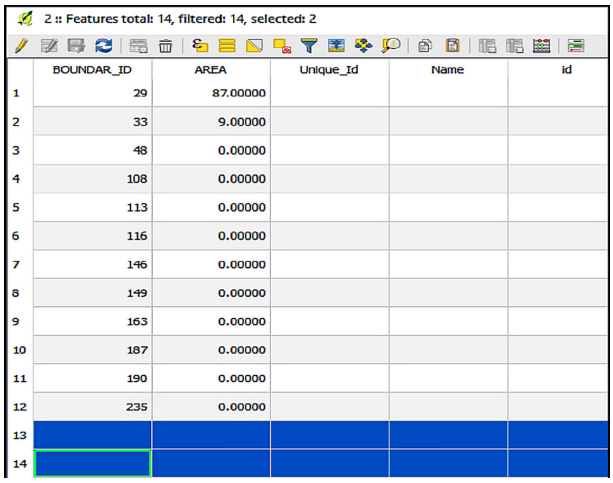
Case 2: Removal of null values

Executing the code, the program will select these features (blue selection) and it will delete them, as shown in Figure 5, because:

a) They have the NULL BOUNDAR_ID value;
b) They have the NULL AREA value.

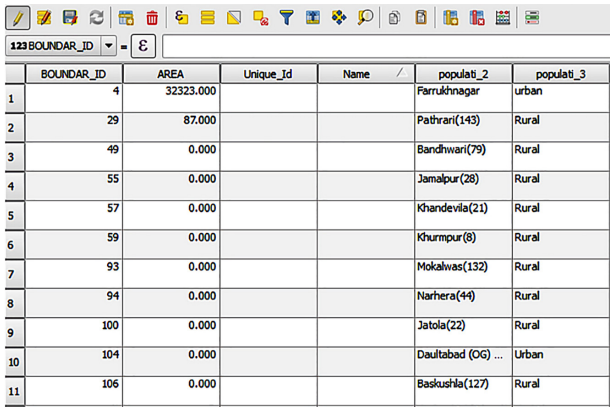
Case 3: Removal of extraneous attributes

Executing the code, the program will select and delete the attributes (Unique_Id, Name) as shown in Figure 6, because they are empty.



	BOUNDAR_ID	AREA	Unique_Id	Name	id
1	29	87.00000			
2	33	9.00000			
3	48	0.00000			
4	108	0.00000			
5	113	0.00000			
6	116	0.00000			
7	146	0.00000			
8	149	0.00000			
9	163	0.00000			
10	187	0.00000			
11	190	0.00000			
12	235	0.00000			
13					
14					

Fig. 5. Null values.

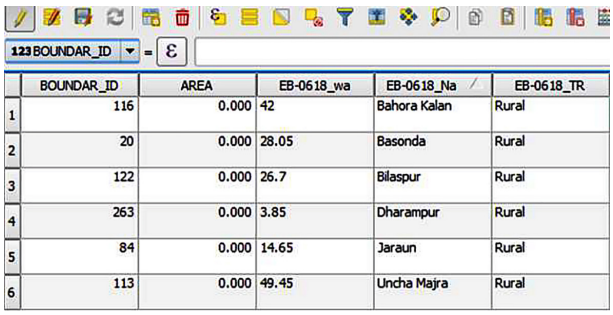


	BOUNDAR_ID	AREA	Unique_Id	Name	populati_2	populati_3
1	4	32323.000			Farrukhnagar	urban
2	29	87.000			Pathrari(143)	Rural
3	49	0.000			Bandhwari(79)	Rural
4	55	0.000			Jamalpur(28)	Rural
5	57	0.000			Khandevila(21)	Rural
6	59	0.000			Khurmpur(8)	Rural
7	93	0.000			Mokalwas(132)	Rural
8	94	0.000			Narhera(44)	Rural
9	100	0.000			Jatola(22)	Rural
10	104	0.000			Daultabad (OG) ...	Urban
11	106	0.000			Baskushla(127)	Rural

Fig. 6. Extraneous attributes.

Case 4: Handling missing values in an important attribute

Executing the code, the program will select AREA and it will compute and fill the value of Area as it is a necessary attribute field and cannot be deleted as shown in Figure 7.



	BOUNDAR_ID	AREA	EB-0618_wa	EB-0618_Na	EB-0618_TR
1	116	0.000	42	Bahora Kalan	Rural
2	20	0.000	28.05	Basonda	Rural
3	122	0.000	26.7	Bilaspur	Rural
4	263	0.000	3.85	Dharampur	Rural
5	84	0.000	14.65	Jaraun	Rural
6	113	0.000	49.45	Uncha Majra	Rural

Fig. 7. Missing values in important attribute.

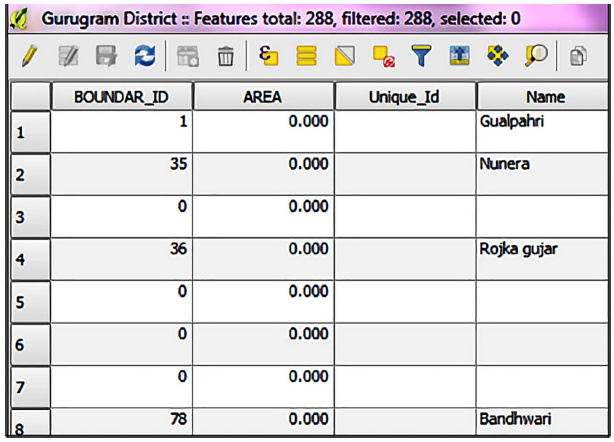
After cleaning the current layer, the *GeoWebCln* tool will show metadata summary information containing percentages corresponding to the amount of deleted and remaining data.

Visual analysis using *GeoWebCln* tool

Analysis of *GeoWebCln* tool is done by adding spatial data to the vector layer of QGIS and executing *GeoWebCln* in Python console of QGIS. Following are steps of *GeoWebCln* analysis:

(A) Adding vector data to QGIS

Spatial data of Gurugram District in shapefile format is added to the vector layer of QGIS platform and the attribute table is analysed as shown in Figure 8.



	BOUNDAR_ID	AREA	Unique_Id	Name
1	1	0.000		Gualpahri
2	35	0.000		Nunera
3	0	0.000		
4	36	0.000		Rojka gujar
5	0	0.000		
6	0	0.000		
7	0	0.000		
8	78	0.000		Bandhwari

Fig. 8. Attribute table of spatial data.

(B) Importing automated *GeoWebCln* in QGIS

Automated *GeoWebCln* is imported to the python console in QGIS and execution of layer cleaning will start to run with the given command as shown in Figure 9.

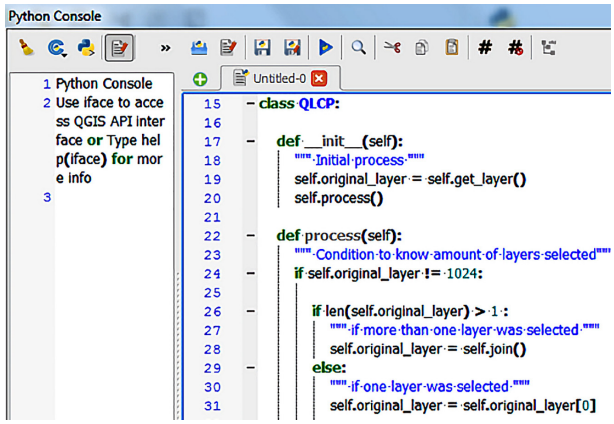


Fig. 9. Geospatial metadata cleaning using *GeoWebCln* tool.

(C) Accepting input from user

The user is asked for selecting deletion of duplicate attribute values as shown in Figure 10.

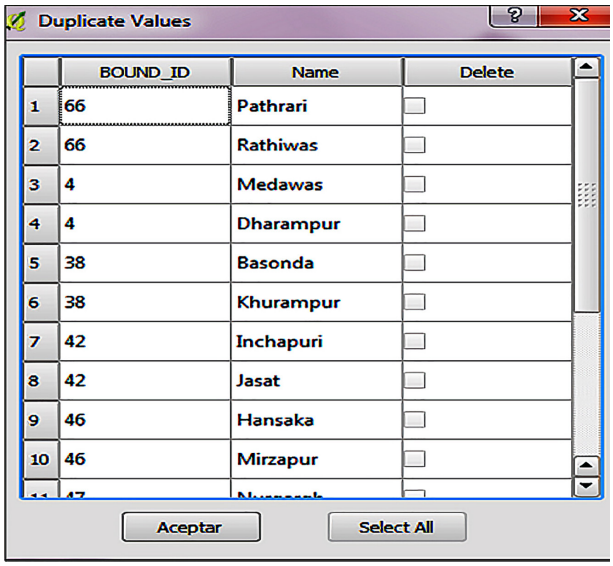


Fig. 10. User interaction with *GeoWebCln* tool.

(D) Cleaned layer saved as new layer

The *GeoWebCln* tool after accepting input from the user cleans the vector data layer and saves it as a different layer as shown in Figure 11.

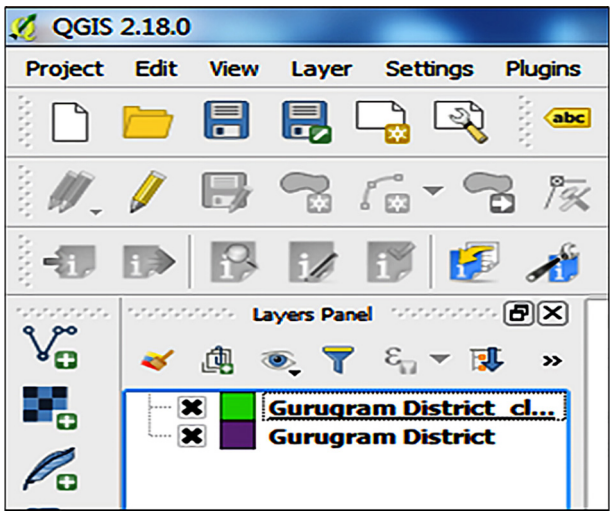


Fig. 11. Cleaned data saved as new layer.

(E) Display of metadata and summary of cleaned layer

The quality information of the cleaned layer is displayed as a summary in tabular form along with the metadata information of the new layer as shown in Figure 12.

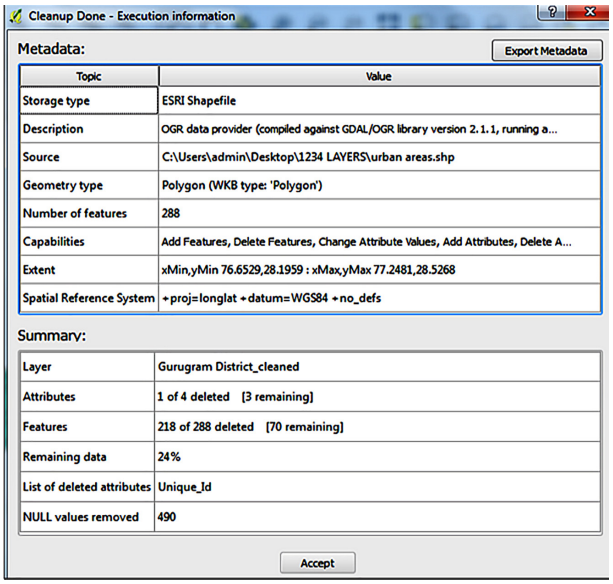
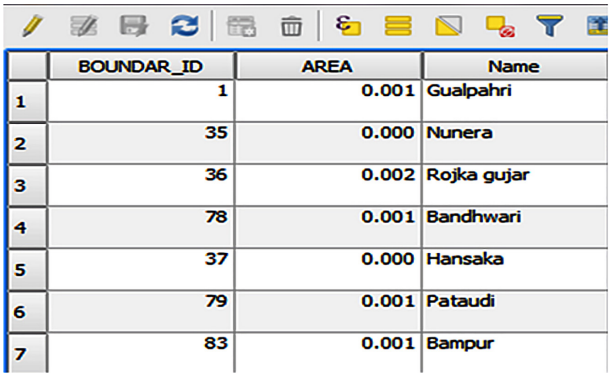


Fig. 12. Quality information of geospatial data.

After execution of the *GeoWebCln* tool, information related to cleaning of the program is also displayed to judge the quality of clean data. Now data are free from null values, 0 values, duplicate values, extraneous attributes and missing values. This *GeoWebCln* tool has removed the dirty data and produces clean data as a separate layer for analysis and visualisation. The attribute table shown in Figure 13 can be checked for anomalies. The data are free from errors and dirt. The summary information generated by the *GeoWebCln* tool can be exported as CSV files for future use.



	BOUNDAR_ID	AREA	Name
1	1	0.001	Gualpahri
2	35	0.000	Nunera
3	36	0.002	Rojka gujar
4	78	0.001	Bandhwari
5	37	0.000	Hansaka
6	79	0.001	Pataudi
7	83	0.001	Bampur

Fig. 13. Attribute table of new layer.

(F) Cleaned layer can be visualised

Spatial data free from errors resulting after execution of the *GeoWebCln* tool can be visualised as a different layer in green, named Gurugram cleaned, which can be used for further analysis,

visualisation and decision making. The vector layer in purple named Gurugram contains dirty data as shown in Figure 14. This visualisation of cleaned and uncleaned data helps the user to understand the regions which are error-free and can be considered for performing any geospatial analysis.

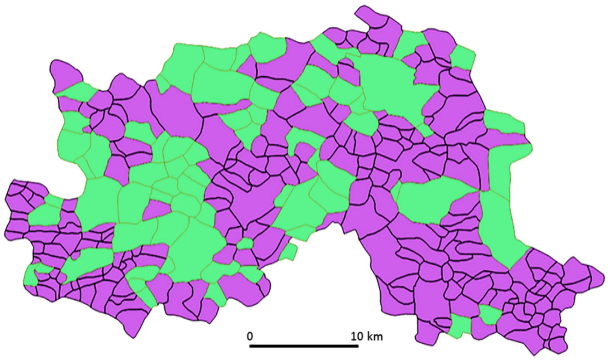


Fig. 14. Visualisation of cleaned and uncleaned data

(G) Metadata and summary information saved as CSV files.

Metadata of spatial data gives information about the storage type, description, source, geometry, extent and spatial reference of the generated layer, as shown in Figure 15. The summary shows the name of the cleaned layer, number of deleted attributes, number of deleted features, remaining cleaned data, list of deleted attributes and the number of null values removed.

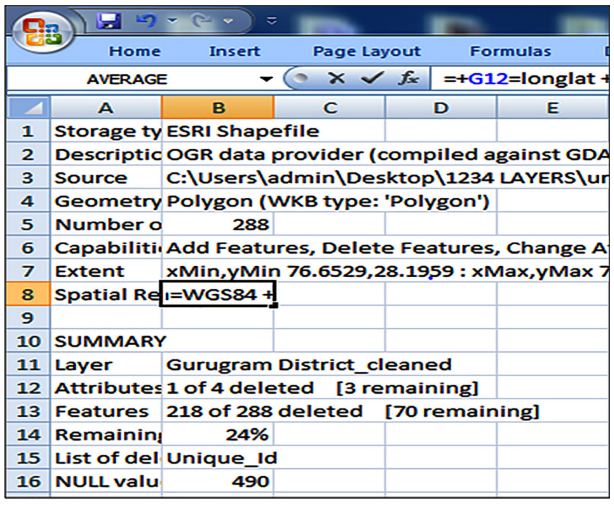


Fig. 15. Metadata and summary of cleaned layer.

Comparative analysis of *GeoWebCln* tool

This section presents the metadata cleaning ability of state of art and *GeoWebCln* tool developed in this research by comparing the cleaning performed using GIS functions done by Parmar and Sheoran (2021) and cleaning is done using *GeoWebCln*. The *GeoWebCln* tool was designed to perform cleaning of the added layer; second, to display the cleaning information to judge the quality of the clean layer and finally, to store the metadata and cleaning summary as a CSV file for future use.

Spatial data of Gurugram District in shapefile format taken from Society for Geoinformatics and Sustainable Development (SGSD) was analysed in QGIS to determine the quality of the data stored in the attribute table. The attribute table of the data was full of errors and their removal requires the use of various GIS functions available in QGIS. Cleaning of attribute data was done using GIS functions by Parmar and Sheoran (2021), but manual cleaning is a tedious, cumbersome, time-consuming and lengthy process. So the *GeoWebCln* tool was used for performing cleaning of contextual metadata of spatial data in a single click by taking input from the user. The cleaning performed by *GeoWebCln* removed all null values, 0 values, duplicate values, missing values and extraneous attributes, and hence produced clean data of good quality. Quality information about the cleaned layer is also provided to the user after execution of the program. QGIS functions are not capable to provide cleaning information about the cleaned attributes. The testing data that we have used is the vector data and its attribute table data of Gurugram District taken from SGSD. The reference data is the cleaned data retrieved after manual cleaning done by Parmar and Sheoran (2021) and the competitive procedure is done by analysing the cleaning output produced by *GeoWebCln* and the cleaning output generated using GIS functions available in QGIS software. Below is the comparison report of the metadata cleaning achieved using QGIS functions and cleaning performed by *GeoWebCln* (Table 1).

The output produced after performing cleaning operation using available GIS functions in the attribute table of QGIS software is analysed. The

Table 1. Comparative analysis of *GeoWebCln* tool.

Cleaning using QGIS functions	Cleaning using *GeoWebCln*
The user must have prior knowledge of GIS cleaning functions and its steps. QGIS is a vast software having various functions. New users are not aware of these functions and need a tutorial before performing the cleaning process.	Users can perform cleaning using a single function with a single click in QGIS. There is no need to analyse the dirty data. A user just needs to import the cleaning function in the Python console of QGIS and click on the run tab. The vector layer will be cleaned.
It is suitable for trained GIS users. The cleaning needs expertise in QGIS and cannot be handled by novice users.	It is suitable for all types of users.
It is a time-consuming process as it requires operation and analysis of various GIS functions such as JOIN, DELETE and SQL Query in Advance Filter Expression.	It is a very fast cleaning process. There is no need for any GIS function and query execution.
It is not an interactive approach as no input is asked from the user. The user is not aware of the work performed by the GIS functions.	It is interactive and user-friendly as input is asked from a user before the removal of duplicate values.
It is less reliable as cleaning performance depends on the skills of the user. If the user chooses wrong functions, then cleaning is not done properly.	It is reliable as cleaning is performed by the *GeoWebCln* tool itself without depending on the skills of the user.
Incapable to provide cleaning information of attributes. The summary of cleaned data is not available.	Provide cleaning information of the attributes as shown in Figure 12.
The cleaned layer cannot be automatically saved.	The cleaned layer is saved as a new layer automatically after cleaning.
Metadata information of spatial data cannot be stored for future use.	Metadata information of cleaned data is exported as CSV files and can be used for comparison and analysis.
Data quality parameters such as completeness, consistency and accuracy cannot be perceived by the users after cleaning as no cleaning information is provided.	Users can easily judge the quality parameters after analysing summary information. This summary information helps users to understand the completeness as several deleted attribute information is given; accuracy of the data is provided by the remaining data in the above summary, i.e. 24% is the remaining data which is completely accurate; consistency information is also provided as 490 null values are removed that were not according to the domain values.
Output as cleaned vector layer is not distinguishable from the dirty layer.	Output as the cleaned vector layer is apparent to the dirty layer as shown in Figure 14. Spatial data in green is cleaned data and is free from errors.

cleaned values cannot be visualised apparently from dirty values as both are stored in the same layer and the user cannot judge whether the data are cleaned or not until the attribute table is analysed.

The output produced by the *GeoWebCln* tool can be visualised in more interactive ways, as shown in Figure 14. The data in green contain the correct value and are free from errors. The data in purple are dirty. The user can easily analyse and visualise the data produced by the cleaning function. This tool is user-friendly, more interactive and provides quality information. The cleaned data can be used to create, update and publish maps online using *qgis2web plugin* in the QGIS software.

GeoWebCln removed 490 null values and 218 extraneous features, whereas cleaning performed by using the GIS function of QGIS software removed only 202 null values and 202 extraneous features. So performance analysis of cleaning done by *GeoWebCln* can be understood by using the below mathematical equations and graph in Figure 16 using Eqs. (1) and (2) as

$$P_a = \left(\frac{G_o - G_i}{T_a}\right) \times 100 \tag{1}$$

where:

- G_o is the number of features cleaned by *GeoWebCln*,
- G_i is the number of features cleaned by GIS function,
- T_a is the total number of features feed,
- P_a is the percentage difference of attribute cleaning efficiency.

$$P_n = \left(\frac{N_v - N_i}{T_n}\right) \times 100 \tag{2}$$

where:
- N_v is the number of null values cleaned by *GeoWebCln*,
- N_i is the number of null values cleaned by GIS function,
- T_n is the total number of values feed,
- P_n is the percentage difference of null value cleaning efficiency.

From Figure 14, the values of G_o = 218, G_i = 202 and T_a = 288. Therefore, using Eq. (1), the calculated value of P_a = 5.56%. Similarly, the values of N_v = 490, N_i = 202 and T_n = 1,152. Therefore, using Eq. (2), the calculated value of P_n = 25.00%. From these calculations, it is evident that performance of the *GeoWebCln* tool is better than the state-of-the-art GIS Function by 5.56% and 25.0% to clean extraneous features and remove null values, respectively. Therefore, *GeoWebCln* proved to be highly efficient in cleaning geospatial data.

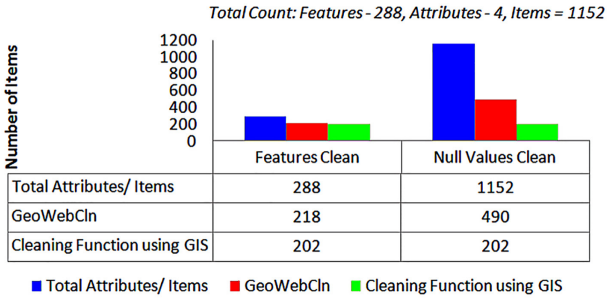


Fig. 16. Performance analysis of *GeoWebCln* tool.

This tool is very helpful for data analysis of spatial data, as cleaned data leads to right analysis while wrong data gives a wrong output; moreover, wrong analysis of data is a big problem and is to be taken seriously. *GeoWebCln* addresses the scientific problem of data analysis in spatial data by working on the cleaning of contextual metadata of attribute table and can be expressed mathematically as below.

The objective is to clean the contextual metadata given in the attribute table of spatial data added to the QGIS software for the right data analysis of spatial data. If M represents the dataset of all attribute tables

$$M = \{M_1, M_2, M_{3\ldots}..M_N\}.$$

Metadata data set would be taken from data sources $\mathbf{D_1}$, $\mathbf{D_2}$, $\mathbf{D_3}$ ..... $\mathbf{D_N}$ and added to the attribute table which contains dirty data. Once the *GeoWebCln* program is run in the Python console of QGIS the required cleaned metadata can be represented as

$$CM = \{M_1, M_2, M_3\ldots\ldots M_i: I \leq N\}.$$

Thus,

$$CM \subseteq M.$$

The cleaned data set CM can be used for data analysis as it is free from missing values, null values, extraneous attributes and so on.

Conclusion

In this paper, an intensive metadata cleaning tool, *GeoWebCln*, is developed and tested on the spatial data of Gurugram District. The attribute table of the vector data was analysed to check the data quality; there were various missing values, null values, extraneous attributes and duplicate values in the attribute table. *GeoWebCln* can clean the data by taking a single input from the user and provide cleaned data as a separate layer. Also, the produced metadata information and summary of cleaned attributes that can be saved as CSV files for future analysis and use. The performance of the tool developed in this research is compared with the GIS function-based cleaning generally used for spatial metadata. The test result on QGIS reveals that our model outperforms the state-of-art by 5.56% and 25.0% to clean extraneous features and remove null values, respectively. It shows that this research succeeds in achieving fast cleaning, analysis, visualisation and storing of quality information.

Acknowledgments

The authors express thanks to Khuspal Dahiya, President, Society for Geoinformatics and Sustainable Development (SGSD), for generously providing the spatial data of the road network, boundary map and census data 2011, and to Rakesh Sheoran for his invaluable support in reviewing this research paper. They also acknowledge the official websites of administration and other anonymous sources whose information and data proved helpful for performing the present study. They are grateful to the reviewers

for providing constructive comments that helped in giving this manuscript its present form.

Author's contributions

SKS analysed the data and work, critically reviewed the contents of the paper, supervised the research and provided the academic material for review. VP conducted the review, collected data, carried out implementation and concluded the study. All the authors approved of the final version.

References

Atluri V., Chun S.A., 2004. An authorization model for geospatial data. *IEEE Transactions on Dependable and Secure Computing* 1(4): 238–254. DOI 10.1109/TDSC.2004.32.

Azri S., Ujang U., Rahman A.A., Anton F., Mioc D., 2014. Spatial access method for urban geospatial database management: An efficient approach of 3D vector data clustering technique. *Ninth International Conference on Digital Information Management (ICDIM 2014). IEEE Conference Publications:* 92–97. DOI 10.1109/ICDIM.2014.6991400.

Bielecka E., Burek E., 2019. Spatial data quality and uncertainty publication patterns and trends by bibliometric analysis. *Open Geosciences* 11(1): 219–235. DOI 10.1515/geo-2019-0018.

Boella G., Calafiore A., Grassi E., Rapp A., Sanasi L., Schifanella C., 2019. FirstLife: Combining social networking and VGI to create an urban coordination and collaboration platform. *IEEE Access* 7: 63230–63246. DOI 10.1109/ACCESS.2019.2916578.

Deshmukh R.R., Wangikar V., 2011. Data cleaning: Current approaches and issues. *IEEE International Conference on Knowledge Engineering*: 61–66.

Eldrandaly K.A., Abdel-Basset M., Shawky L.A., 2019. Internet of spatial things: A new reference model with insight analysis. *IEEE Access* 7: 19653–19669. DOI 10.1109/ACCESS.2019.2897012.

Gaikwad D.B., Wanjari Y.W., Kale K.V., 2014. Disaster management by integration of web services with geospatial data mining. *Annual IEEE India Conference (INDICON)*: 1–6. DOI 10.1109/INDICON.2014.7030685.

Jakobsson A., 2002. *Data quality and quality management –Examples of quality evaluation procedures and quality management in European National Mapping Agencies. Spatial Data Quality.* Taylor & Francis, London: 216–229.

Keim D.A., 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8(1): 1–8. DOI 10.1109/2945.981847.

Koshley D.K., Halder R., 2015. Data cleaning: An abstraction-based approach. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Kochi: 713–719. DOI 10.1109/ICACCI.2015.7275695.

Kumar V., Khosla C., 2018. Data cleaning-A thorough analysis and survey on unstructured data. *8th International Conference on Cloud Computin. Data Science & Engineering (Confluence)*. Noida: 305–309. DOI 10.1109/CONFLUENCE.2018.8442950.

Li J., Chen N., 2014. Geospatial sensor web resource management system for Smart CITY: Design and implementation. *14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*: 819–827. DOI 10.1109/CCGrid.2014.70.

Lim S., 2010. Cleansing noisy city names in spatial data mining. *International Conference on Information Science and Applications*: 1–8. DOI 10.1109/ICISA.2010.5480390.

Parmar V., Sheoran S., 2021. Context-based spatial metadata cleaning using QGIS. *Vidyabharati International Interdisciplinary Research Journal* 12(1): 55–62.

Ridzuan F., Zainon W.M., 2019. A review on data cleansing methods for big data. *Procedia Computer Science* 161: 731–738. DOI 10.1016/j.procs.2019.11.177.

Sheoran S.K., Parmar V., 2020. Identification of alternative landfill site using QGIS in a densely populated metropolitan area. *Quaestiones Geographicae* 39(3): 47–56. DOI 10.2478/quageo-2020-0022.

Thiyagalingam J., Getov V., 2006. A metadata extracting tool for software components in grid applications. *IEEE John Vincent Atanasoff International Symposium on Modern Computing (JVA'06)*: 189–196. DOI 10.1109/JVA.2006.3.

Yoshizumi A., Coffer M.M., Collins E.L., Gaines M.D., Gao X., Jones K., McGregor I.R., McQuillan K.A., Perin V., Tomkins L.M., Worm T., Tateosian L., 2020. A review of geospatial content in IEEE visualization publications. *IEEE Visualization Conference (VIS)*: 51–55. DOI 10.1109/VIS47514.2020.00017.

Zhao P., Zhao Q., Zhang C., Su G., Zhang Q., Rao W., 2019. CLEAN: Frequent pattern-based trajectory spatial-temporal compression on road networks. *IEEE International Conference on Mobile Data Management*: 605–610. DOI 10.1109/MDM.2019.00127.

Zhao X., Huang Z., 2010. A quality evaluation approach for OLAP metadata of multidimensional OLAP data. *IEEE International Conference on Information Management and Engineering*: 357–361. DOI 10.1109/ICIME.2010.5477583.

Zou T., Li W., Liu P., Su X., Huang H., Han Y., Guo X., 2018. An overview of geospatial information visualization. *IEEE International Conference on Progress in Informatics and Computing (PIC)*. Suzhou. China: 250–254. DOI 10.1109/PIC.2018.8706332.

Zylshal Z., 2020. Topographic correction of LAPAN-A3/LAPAN-IPB multispectral image: A comparison of five different algorithms. *Quaestiones Geographicae* 39(3): 33–45. DOI 10.2478/quageo-2020-0021.