CONTEXT-BASED SPATIAL METADATA CLEANING USING QGIS

V. Parmar* and S. Sheoran
Department of computer science & engineering, Indira Gandhi University Meerpur, Rewari, India
*vintiparmar1487@gmail.com

ABSTRACT

In present day world spatial metadata is explosively used in every sphere of human activities but anomalies in the data obstruct the extraction of information with its full potential. Dirty data gives wrong output and hampers the decision making process. Therefore, the data cleaning is an essential requirement for spatial metadata analysis and decision making. In this study a context-based cleaning of population data set of district Gurugram is accomplished using QGIS in conjunction with record linkage technique. Spatial metadata of district Gurugram obtained from state authorities in shape file format is fed to QGIS to get rid of the dirty data. The cleaned data obtained from our framework are analysed and visualized for effective understanding of populace context in densely populated and highly industrialised district.

Keywords: *spatial data, data cleaning, dirty data, data visualization.*

Introduction

Data cleaning is an essential activity for analysis of data and information management [1]. Data cleaning is defined as the process of removing unwanted, useless and corrupt data from data sets by using error detection and correction functions [2]. Spatial data is widely available through internet and used by various organizations and people but spatial data collection is very time consuming and expensive process. The spatial data collected once can be used many times for different applications [3]. Spatial data needs data cleaning for its proper use. Spatial data have attribute data which may contain several extraneous fields, missing values and duplicate values. These type of errors enter during data collection and extraction phase hampers the quality of the data [4]. These errors need to be determined, detected and corrected to prevent its further propagation and make it a fit candidate for further analysis. The quality of data regarding their suitability for particular application is determined through quality parameters mentioned below:

- *Completeness*: Completeness refers to the number of extra committed and omitted objects in the dataset to the reference dataset.
- *Accuracy*: Accuracy refers to the deviation of the recorded values in terms of time and position to the actual ground truth.
- *Consistency*: Consistency is fidelity to the rules of conceptual schema and logical rules of data structure. It is the degree to which data closely stored according to the physical structure of the dataset.

In this paper we have used record linkage similarity measures algorithm to perform data cleaning on attribute data of spatial data using GIS functions in special context of district Gurugram situated in the National Capital Region of India. The clean and error free data is analysed and visualized to understand its quality.

Related Research

Data having some position attribute on the surface of the earth is called as spatial data. Spatial data contain information about specific location of natural phenomenon. The advancement of knowledge and data transmission technologies has changed the way of using spatial data. Further the data containing information about these data are known as spatial metadata. There is enormous increase in usage of geospatial data between people and organisation. Geographic metadata is easily available on Internet and can be downloaded and reused. Many organisations built spatial data infrastructures for handling geospatial data [3]. These dataset can be used for planning and making decisions. Such geospatial data are used in variety of

application including healthcare, weather forecasting, forestry, landfill site selection. The users of geospatial data are utilizing geographic information system for their specific application without knowing the risk of data misuse. Naive users are not good in understanding and handling of geographic information. There is high risk of misusing spatial metadata for unlawful activities in society and ecological derogation. Spatial data is used by various people and organization for myriads of planning and management activities. For proper analysis of spatial data and right use for any specific application, spatial data should be checked and if errors exist in the attribute data they need to be cleaned. Spatial data preprocessing is very time consuming process and must be carried out meticulously. In this research we have proposed a data cleaning model to remove corrupt, incomplete and inconsistent data using GIS functions for getting accurate results. In data cleaning process all unwanted, useless and corrupt data are removed from the datasets [4]. Before performing data cleaning all missing fields, duplicate records and inconsistent data are identified and then data cleaning is performed to enhance data quality features such as accuracy, correctness, completeness and consistency. Spatial data is gathered from various multiple data sources having anomalies and errors and cannot be considered fit for analysis planning and decision making purposes. Reduced data quality leads to false outcomes and finally became reason for improper use of assets and resources [5]. Error free data having salient data quality is always required and to achieve this data cleaning is highly needed. Data cleaning process needs significant Knowledge, experience, time and money for detection and correction of absurd data [1]. Data cleaning process enhances the accuracy, non-duplication, uniqueness, completeness and validity of data [2, 8]. Koshley and Halder (2015) have identified following stages and sources where spatial metadata are prone to errors and petrified[4]:

- *Data collection stage:* Incorrect recording procedures and inaccurate use of equipments is the main reason for allowing entry of errors in the data collection phase.
- *Data input stage:* Errors in digitizing and different forms of data entry leads to spatial errors at this stage.
- *Data storage* stage: *Spatial precision and hardware cause errors in spatial data storage.*
- *Data manipulation stage*: Boundary errors and wrong class intervals are the source of error at data manipulation stage.
- *Data output stage*: Scaling and wrong output device creates error at this stage.

Problem Statement

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 any data cleaning algorithm. Dirty spatial data produce invalid results that hamper the data analysis and data visualization process. In this paper spatial data is cleaned and then analysis and visualization is performed using data visualization algorithm to address the following research questions:

- How the errors enter in the different stages of spatial data management.
- How data cleaning can be performed using spatial technologies.
- How the cleaned data can be visualized using visualization techniques.

Data Set

We have collected the spatial data of population census 2011 in district Gurugram from Society for Geoinformatics and Sustainable Development (SGSD). This spatial data contain large metadata set in its attribute table about rural and urban areas, literacy rate, sex ratio etc. Many of the fields are empty and have duplicate data in the available dataset. So using data cleaning techniques we have cleaned this data in QGIS to analyze population distribution in different regions of the study area. Population of India is growing at a very high pace. Population of Gurugram has witnessed a 73.93% increment in last decade due to urbanization and industrial growth. The high rate of growth in population is the main reason for many diseases, deaths, pollution,

and unemployment [6]. We need to pay high attention to this alarming rate of growth in population, areas of high density population need to be analyzed and reasons for this increment should be identified and checked. The main aim of the paper is to clean the available data using some data cleaning technique for valuable data analysis and visualization using dense pixel display visualization technique [7]. This analysis will help us to find the highly populated areas in the district.

Data Cleaning Techniques

Misinterpretation of results and improper usage of data generates error. Spatial data quality is an essential feature needed for proper usage of data. From user perspective data should be advantageous for use and from data producer point of view data quality should be justified to requirements. Data producer and user should be meticulous about the spatial data quality. Userare using spatial data extensively using GIS tools for numerous application. For good spatial data quality data should be complete, accurate, consistent and unique. To achieve these quality features data cleaning is required. Data cleaning techniques are used to remove dirt from data i.e useless data from the dataset [8]. Data cleaning is performed to remove duplicate entries, null entries, useless attributes, misspelled entries etc. Following are some of the techniques used by researchers to clean the data [9]:

Border detection algorithm: Border detection algorithm is developed by Arturas Mazeika and Michael H.B Ohlen in 2006. It performs cleaning of string data in two steps. In first step cluster is formed near string data by connecting the border and center of hyperspherical and in the second step the cluster string is cleansed by the repeated cluster. This algorithm is simple and yield clean output for string data.

Token based data cleaning: This technique use smart tokens to identify duplicate records and lowers the dependency of data cleaning on match threshold. High precision is achieved using small length smart tokens. This technique is suitable to domain independent data cleaning and can also be used for refreshing of integrated data.

Record linkage similarity measures algorithm: This technique is used to compare two relational tuples for their similarity. In this technique record linkage and approximate join are used to identify whether two tuples are same. Approximate match predicates are used to calculate the degree of similarity between two data tuples using any of the Atomic Similarity Measures, Functions to combine similarity measures, Similarity between linked entities similarity measures based on their characteristics. After quantifying the degree of approximate match the two data tuples are joined using approximate join by relational duplicate elimination and join technique.

In this study of we have proposed data cleaning framework using record linkage algorithm in QGIS to perform data cleaning on metadata of spatial data. We have preferred record linkage algorithm as it uses sql queries and offer easy deployment, greater flexibility and improved performance as compared to the token based algorithm and border detection algorithm.

Proposed Context-Based Framework

The proposed framework for data cleaning is designed to apply on the attribute data of any spatial metadata. This proposed model is applied for data cleaning of spatial data and tested it in QGIS. Different type of errors such as missing values, duplicate values and extraneous fields are removed using below algorithm:

INPUT: Spatial data from different sources having dirty attribute data.
OUTPUT: Clean Data

- *Select and add*
- *Spatial data (vector /raster data) layers in QGIS software.*
- *For each vector/ ratser data layer*
  - *Open attribute table having attribute data of the spatial data.*

- *Identify the required attribute and extraneous attributes.*
- *Perform similarity measure on the attribute table of the added layers.*
- *Perform record linkage of same attributes using join operation.*
- *Perform data query filter operation to remove duplicate and inconsistent values.*
- *Rectify the errors by adequate algorithm for each data instance.*
- *Present to user for analysis/ractification.*
- *Save the attribute table.*

The flowchart of work architecture in the proposed model is given below:

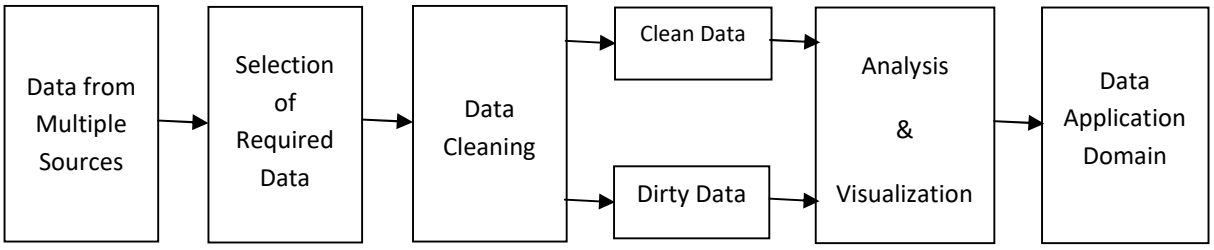


Figure 1: Data Cleaning Flowchart

Context Based Data Cleaning

This section illustrates data cleaning for attribute table of the spatial metadata added in QGIS platform. In this study we have added the vector data of Gurugram district in shapefile format to the QGIS . This spatial data have metadata information contained in attribute table. The two layers added to the QGIS Software are shown in Figure 2. Both layers have their attribute table containing metadata information about the name of tehsils, population, sex ratio, literacy rate etc.

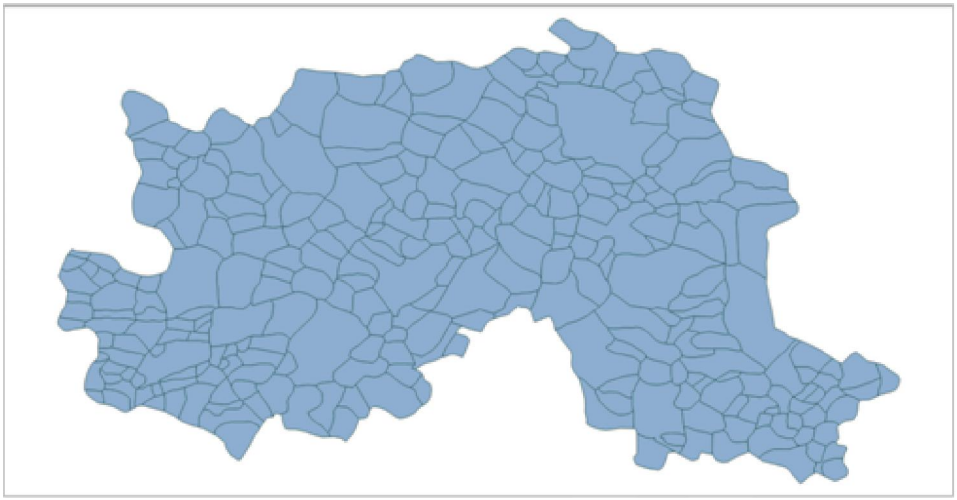
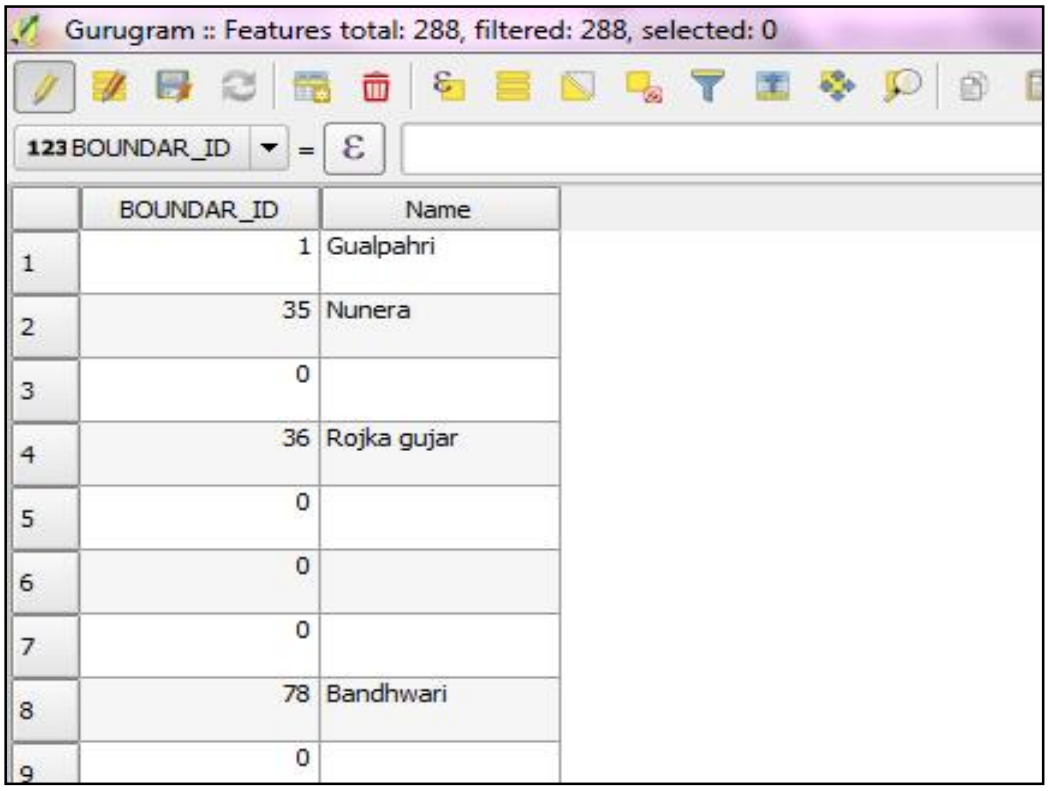


Figure 2: Spatial data of Gurugram district in shapefile format

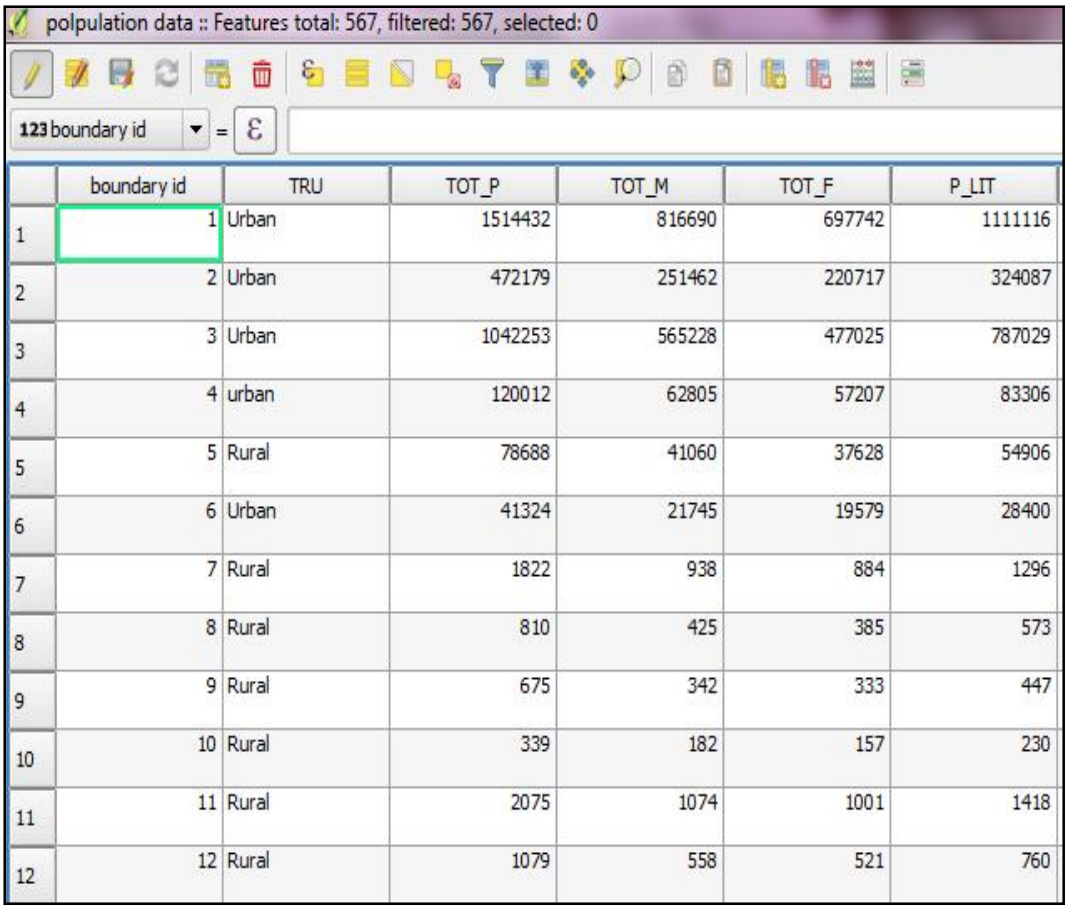
The attribute table in input data contains many extraneous fields and null entries as shown in Figure-3. The attribute table of the layer population data is shown in Figure 4. This table is joined to the Gurugram boundary attribute table after performing similarity measures and record linkage technique and the output is shown in figure 5.


Gurugram :: Features total: 288, filtered: 288, selected: 0

123 BOUNDAR_ID = ε

	BOUNDAR_ID	Name
1	1	Gualpahri
2	35	Nunera
3	0	
4	36	Rojka gujar
5	0	
6	0	
7	0	
8	78	Bandhwari
9	0	

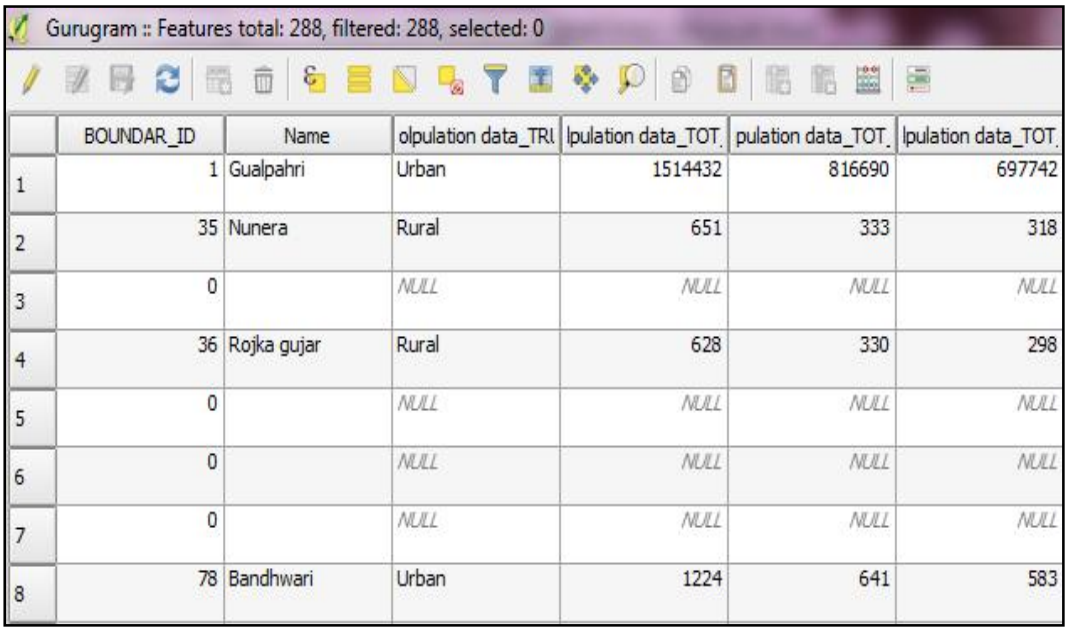
Figure 3: Attribute table of the layer Gurugram Boundary


polpulation data :: Features total: 567, filtered: 567, selected: 0

123 boundary id = ε

	boundary id	TRU	TOT_P	TOT_M	TOT_F	P_LIT
1	1	Urban	1514432	816690	697742	1111116
2	2	Urban	472179	251462	220717	324087
3	3	Urban	1042253	565228	477025	787029
4	4	urban	120012	62805	57207	83306
5	5	Rural	78688	41060	37628	54906
6	6	Urban	41324	21745	19579	28400
7	7	Rural	1822	938	884	1296
8	8	Rural	810	425	385	573
9	9	Rural	675	342	333	447
10	10	Rural	339	182	157	230
11	11	Rural	2075	1074	1001	1418
12	12	Rural	1079	558	521	760

Figure 4: Attribute table of the layer population data

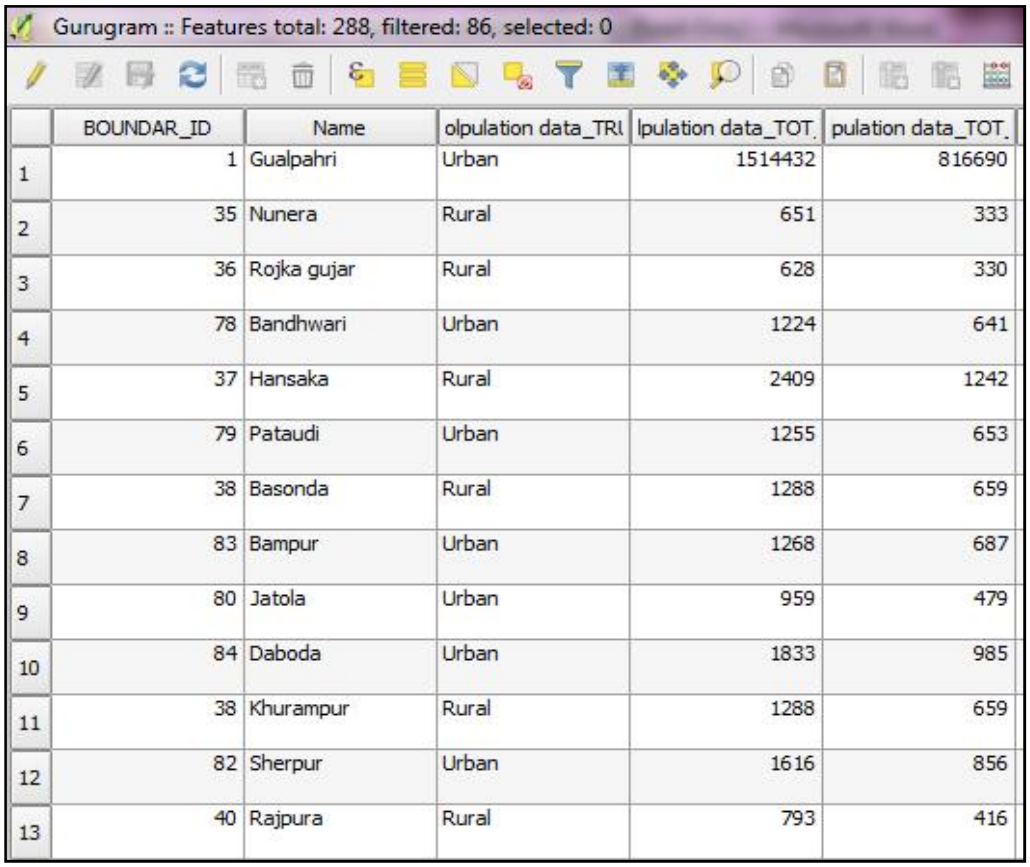


Gurugram :: Features total: 288, filtered: 288, selected: 0

	BOUNDAR_ID	Name	olpulation data_TRl	lpulation data_TOT.	pulation data_TOT.	lpulation data_TOT.
1	1	Gualpahri	Urban	1514432	816690	697742
2	35	Nunera	Rural	651	333	318
3	0		NULL	NULL	NULL	NULL
4	36	Rojka gujar	Rural	628	330	298
5	0		NULL	NULL	NULL	NULL
6	0		NULL	NULL	NULL	NULL
7	0		NULL	NULL	NULL	NULL
8	78	Bandhwari	Urban	1224	641	583

Figure 5: Output after performing Record linkage similarity measures technique on the attribute table of population data and Gurugram boundary

It can be visualized from the figure 3, 4 and 5; various fields are having '0' and null entries which can be cleaned by using the proposed cleaning algorithm. The BOUNDARY _ID having zero entries can be removed by using the query BOUNDARY_ID > '0'in the advance filter section. Figure 6 shows the output of this query. Now we can see that all the values in BOUNDARY_ID is free from zero value.



Gurugram :: Features total: 288, filtered: 86, selected: 0

	BOUNDAR_ID	Name	olpulation data_TRl	lpulation data_TOT.	pulation data_TOT.
1	1	Gualpahri	Urban	1514432	816690
2	35	Nunera	Rural	651	333
3	36	Rojka gujar	Rural	628	330
4	78	Bandhwari	Urban	1224	641
5	37	Hansaka	Rural	2409	1242
6	79	Pataudi	Urban	1255	653
7	38	Basonda	Rural	1288	659
8	83	Bampur	Urban	1268	687
9	80	Jatola	Urban	959	479
10	84	Daboda	Urban	1833	985
11	38	Khurampur	Rural	1288	659
12	82	Sherpur	Urban	1616	856
13	40	Rajpura	Rural	793	416

Figure 6: Attribute table having attributes free from '0' entries and null values

Now we can find the densely populated urban areas by executing another query as shown in Figure7.

Gurugram :: Features total: 288, filtered: 9, selected: 0

	BOUNDAR_ID	Name	olpulation data_TRl	lpulation data_TOT.
1	1	Gualpahri	Urban	1514432
2	90	Hassanpur	Urban	910006
3	60	Dhankot	Urban	20906
4	6	Bhondsi	Urban	41324
5	4	Medawas	urban	120012
6	2	Aklimpur	Urban	472179
7	75	Gurugram City	Urban	872560
8	3	tigra	Urban	1042253
9	4	Dharampur	urban	120012

Figure 7: Urban areas having population greater than 5000

These urban areas can be visualized easily using dense pixel visualization technique as shown in figure 8.

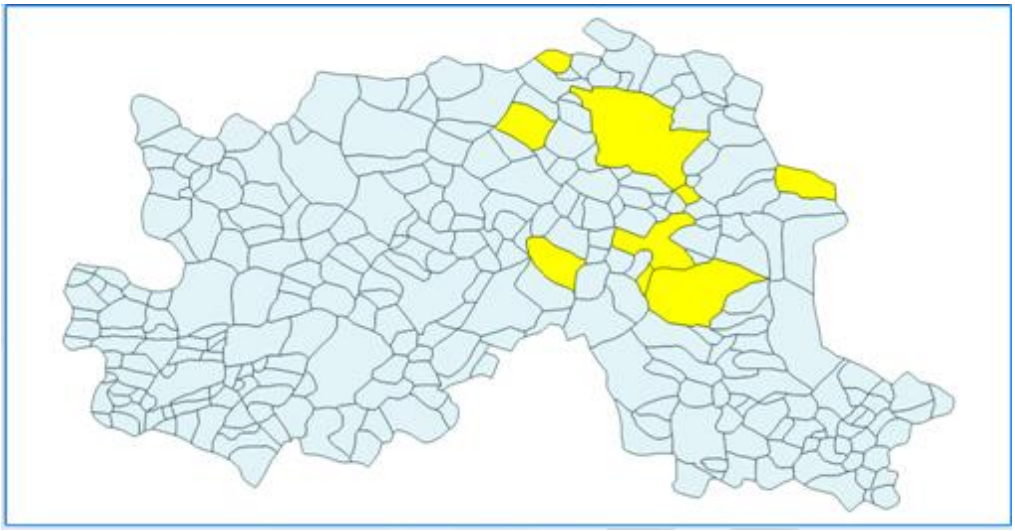


Figure 8: Urban areas having population greater than 5000

This research article performed data cleaning on spatial metadata of district Gurugram through QGIS. The metadata used as input in attribute table was dirty and not fit for the purpose of analysis and decision making as it has many null entries and extraneous fields, So we performed data cleaning and removed the useless data using the data query and GIS functions. The cleaned data is used for analysed and visualized for decision making.

Conclusions and Future Work

The spatial data management is presented by myriads of research paper for different areas such as healthcare, city modeling, remote sensing, image classification, spatial game analytics. In our study we work on cleaning of

the attribute table of the geospatial data to analyse the populated areas in QGIS using dense pixel visualization technique. The cleaned data can be easily analysed and visualized and also can be used for further different applications like finding the landfill site in densely populated area such as district Gurugram located as a satellite city of capital, New Delhi where finding a suitable land for damping waste is a crucial issue. This study has performed three step cleaning using GIS functions but same work can be performed using a single function that can be developed in Phython console with QGIS Software. The use of single function may be fast for performing data cleaning as compared to the three step cleaning done using GIS functions available in QGIS Software. This proposed methodology of data cleaning can be used to develop a prototype for spatial metadata cleaning.

References

1. Prasad, K. H., Faruquie, T. A., Joshi, S., Chaturvedi, S., Subramaniam, L.V., & Mohania M.(2011). Data Cleansing Techniques for Large Enterprise Datasets. Annual SRII Global Conference, San Jose, CA, 135-144.
2. Krishnamoorthy, R., Kumar, S. S., Neelagund, B. (2014). A new approach for data cleaning process. International Conference on Recent Advances and Innovations in Engineering, Jaipur, 1-5.
3. 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.
4. 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.
5. Hamad, K., Quiroga, C. (2016). Geovisualization of Archived ITS Data-Case Studies. IEEE Transactions on Intelligent Transportation Systems, 17(1): 104-112.
6. Kumar, V., Khosla, C. (2018). Data Cleaning-A Thorough Analysis and Survey on Unstructured Data. International Conference on Cloud Computing, Data Science & Engineering (Confluence, Noida, 305-309.
7. Keim, D. A. (2002). Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1): 1-8.
8. Ridzuan, F., Zainon W.M.N.W. (2019). A Review on Data Cleansing Methods for Big Data. Procedia Computer Science, 161: 731-738.
9. Deshmukh, R. R., Wangikar, V. (2011). Data Cleaning: Current Approaches and Issues. IEEE International Conference on Knowledge Engineering, 61-66.