**SPATIAL DATA QUALITY IN GIS DATA: A REVIEW**

Punit Gupta1 and Gavin McArdle2

1,2 University College Dublin, Dublin, Ireland

punit.gupta@ucd.ie

**Abstract:**

Understanding spatial data quality is important in Geographic Information Science (GIS) applications. Spatial data are used in a variety of critical areas, including urban planning, environmental management, emergency response, and natural resource management where the accuracy and precision of spatial data can have a significant impact on decision-making, especially when used with predictive analysis. A review of the importance of spatial data quality in GIS data is necessary to understand the factors that affect the quality of spatial data and strategies used to interrogate and maintain spatial data quality. While there is no standard definition for spatial data quality, typically the term refers to the accuracy, completeness, consistency, and currency of the data. One of the key factors that affect spatial data quality is the data acquisition phase. The accuracy of spatial data can be compromised due to errors introduced during data collection, such as measurement errors or errors in processing raw data. Therefore, it is essential to ensure that data collection procedures are well-designed and accurately executed to minimise such errors. In this work, a review of various applications of spatial data quality in GIS is carried out. The goal is to provide a generalized SDQ (Spatial Data Quality) benchmark to reduce errors in spatial data across various domains.

1. **Introduction:**

Data quality plays an important role in any form of data analysis and predictive analysis. Over the years big data environments like cloud computing, geographic information (satellite images and other earth observatory data) and healthcare have attracted researchers as data were seen as the new oil. These fields have a huge scope but data quality plays an important role in concluding a strong finding, or else it may result in error-prone analysis and predictions.

In the fields of Geographic Information Science and earth observation, data are generated by various agencies using different tools and techniques [1-5]. This can result in an error or incomplete data. Such incomplete data or low-quality data used for analysis may result in low accuracy or even misleading results. In general, data quality is important because accurate and reliable data is essential for making effective decisions [4,5]. Poor data quality can lead to incorrect conclusions and poor decision-making. In GIS, data quality refers to the degree to which the data meets the requirements for its intended use. This includes factors such as accuracy, precision, completeness, and consistency [2]. To ensure data quality in GIS, it is important to use high-quality data sources, properly maintain and manage the data, and regularly verify and validate the data to ensure it is accurate and up to date. Additionally, proper documentation and metadata are essential for understanding the quality of the data and for ensuring that it is being used correctly.

GIS data primarily consists of raster and vector data types. Both types of data sources and databases suffer from different types of data quality issues and can be assessed with different metrics. In the raster data type, the database mostly suffers from the satellite image quality and the quality of data in the image source due to resolution, visibility, or noise. On the other hand, vector data suffers from missing values, null value errors, data replication and value out-of-range errors, among other data quality issues.

Currently, a huge amount of satellite data is available from various sources varying from low to high resolution with various bands for vegetation and many other applications like ocean data, precipitation time series data, soil temperature data, and object accuracy in the vector layer. But the issue that exists in the current scenario is to evaluate and find a suitable dataset from existing satellite data (Sentinel 1 -7 and Landsat 1-9) and other GIS vector data like time series data, Census and other surveys. With such a huge volume of data, it becomes difficult to identify useful data for a user-defined application with a specific objective. Even when suitable data sources are located, the data can have errors or be of low data quality [42]. In such cases, there is a need for quality metadata and quality checks to be attached to the datasets to make filtration and identification of datasets easier for specific use cases.

In this work, a survey of various works, to demonstrate the importance of data quality in Geographic Information System (GIS) data, is discussed. The data can be raster satellite image data sources for applications like cloud cover detection, ocean data and object detection or vector data which includes the geo-sensor data readings and man-made data from surveys and field data. Vector data also includes time series data. In this work, we aim to identify existing spatial data quality measures which can be generalized to check for data quality before the data is used for an application which may result in error-prone output due to low data quality. The data quality metrics may vary from application to application as discussed in Section 2. So this work aims to identify common SDQ parameters for multiple applications.

The remainder of the paper is organized into three sections. In section 2 an extensive survey of work done in the field of spatial data quality for GIS is discussed. Finally, the conclusion section discusses the outcome and implications of this work.

1. **Survey**

In the field of GIS, many studies are being performed by various researchers to define the need to define data quality for earth observation data.

There exist various types of GIS data types and use cases where different data quality metrics play an important role. In general, GIS data can be divided into raster and vector data types as shown in Figure 1, where raster data includes satellite images from various products like MODIS, Landsat and Sentinel among others. On the other hand, vector data are typically user-generated data layers added to a map which are generated through field surveys to get data like road maps, river maps, location of hospitals and many more location-based information. This also includes data from various GIS surveys and time series data. Both types of data suffer from data quality issues and can result in poor results and analysis. In this section, we introduce various quality indexes in raster and vector with some of the related work in that domain. Figure 1 shows two types of GIS data. The first is raster data (e.g. satellite imagery) and the second is vector data which is numerical data that can be moisture, pressure, humidity, sea salt content, etc., and many more user-recorded or user-generated data from surveys.

Graphical user interface, text, application, chat or text message

Description automatically generated

**Figure 1. Types of GIS data**

Spatial data quality in general for GIS data can be evaluated under four different categories for both raster and vector data types which are as follows and shown in figure 2 [2]:

1. **Precision**
2. **Consistency**
3. **Completeness**
4. **Accuracy**

The four SDQ (Spatial Data Quality) parameters can be used to evaluate data quality in GIS. For raster data precision, completeness and accuracy are the main parameters on the other hand for vector data precision, consistency and completeness play an important role.

Diagram

Description automatically generated

**Figure 2. Classification of Spatial Data Quality (SDQ).**

Graphical user interface

Description automatically generated

**Figure 3. Spatial Data Quality (SDQ) in GIS Raster Data.**

Graphical user interface, application

Description automatically generated

**Figure 4. Spatial Data Quality in GIS Vector Data**

Figure 3 and Figure 4 show the SDQ parameters for raster and vector data respectively. The next sections describe these SDQ metrics in more detail.

* 1. **Precision**

For raster data precession is evaluated as the image accuracy and the metadata quality which includes bands and other data like depth and number of bands [2]. Data Quality in satellite images refers to the quality of the image and precision of the image concerning the position and size of the object in the image. Several of the GIS products suffer image quality due to low visibility or low image resolution.

* + 1. **Precision of bands in GIS data**

Albanai et.al . [5] showcased a model to evaluate the thermal accuracy of Landsat in the band on the sea surface. This study checks the computational accuracy of satellite images with live data as compared to the vector data available from sea beakers. The work uses bands 10 and 11 from Lansat-8 and compares the accuracy which comes out to be a deviation in accuracy with a mean standard deviation of 0.03 over the year. Figure 5 and Figure 6 show a similar deviation over various seasons for bands 10 and 11. The work showcased a deviation in raster data when it was compared with real data from sea beakers.

|  |  |
| --- | --- |
| Chart, bar chart, histogram  Description automatically generated | **Chart, bar chart, histogram  Description automatically generated** |
| **Figure 5. Mean-variance in band 10 [5]** | **Figure 6. Mean-variance in band 11 [5]** |

* 1. **Completeness**

Completeness [43] is defined as the accuracy of the data in the raster image which can be cloud coverage, or land cover accuracy whereas in vector data it is defined as the percentage of missing data or null values. Where data quality is defined as the precision of detecting clouds in an image with cloud shadow and further classification.

* + 1. **Cloud cover and masking**

Cloud cover in satellite imagery plays a key role in determining the completeness of the area of interest. In this section, we review the state of the art in cloud cover detection and removal to improve completeness in GIS data. Cloud detection and cloud shadow contributed to completeness in SDQ. This review plays an important role because there exist many cloud detection models and cloud classification models but a comparison of each for a specific dataset is required. This review also contributes to the review of various cloud classification and removal models for Sentinel and Landsat data.

Ackerman, S [10] has presented a cloud masking algorithm for the MODIS (Moderate Resolution Imaging Spectroradiometer) database. The algorithm uses MODIS and LIDAR data from the Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) Program Southern Great Plains (SGP) site in Lamont. The algorithm is trained to find the cloud mask in the image with high accuracy. It uses 3 years of MODIS data.

Kopp, T [11] has proposed a (Visible Infrared Imager Radiometer Suite) VRIIS model for detecting cloud masks. This model used the VCM (visible cloud mask) model. This algorithm is used to classify the various land uses like cloud, land, soil, water, coastal & snow. This is a product of the Joint Polar Satellite System program, the algorithm is defined for the MODIS database. The model can define multi-layered clouds and can separate clouds aerosols and cloud shadows.

Cesar Aybar et.al. [12] have proposed a deep-learning model for cloud detection for Sentinel-2. The model is called CloudSEn12 which is defined to detect cloud, cloud shadow and multi-layer clouds. The model is trained on 49400 image data. The main importance of this model as compared to other models is it can differentiate between thick and thin models. The work is also compared with other existing models like Fmask, Sen2Cor and UNetMob. The figure below shows the performance of CloudSEN-12 with various other existing models for cloud and cloud shadow classification.

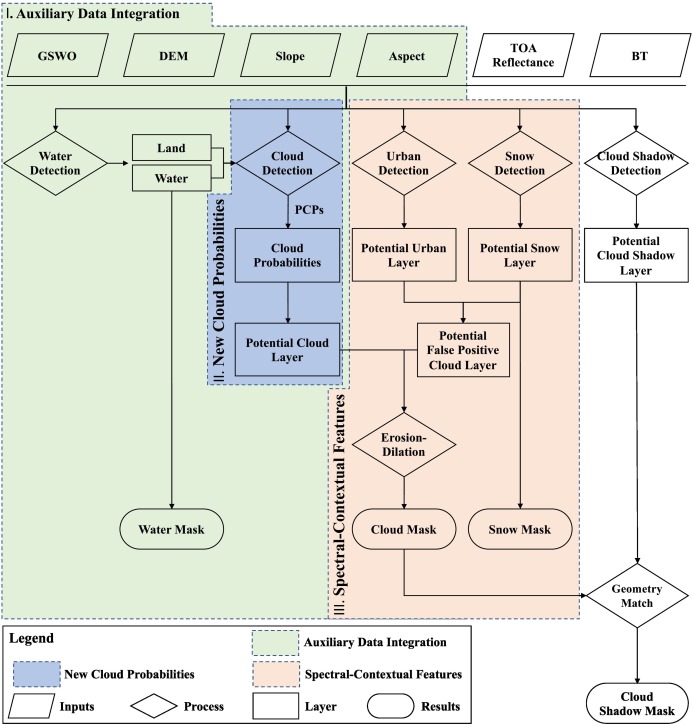
Chart, histogram

Description automatically generated

**Figure 7. Performance of various cloud detection models[12]**

Segal R M. et.al. [13] have proposed and improved the S-2 cloud mask algorithm using the CNN model. The work provides better accuracy for cloud detection compared to the original S-2 cloud mask. The work uses sentinel-2 data for testing and training the model, with 13 spectral bands and a resolution of 10m. The testing was mostly conducted on images from the Fiji Island database.

Qiu.S. et.al. [14] in this work has proposed an improved version of the FMASK algorithm for Lansat4, Landsat 8 and sentinel-2 images. This is one of the tools which allows cloud masking for multiple datasets available with high accuracy. This work demonstrates FMASK 4.0, a version of the algorithm integrated with separate models for cloud masking over land and water to maintain high accuracy. Figure 8. shows the working of FMASK 4.0 where various auxiliary data are integrated for training purposes and detection of cloud, cloud shadow, urban detection and snow detection.



**Figure 8 Cloud shadow detection and mask generation [14]**

Additionally, there are various other machine learning models for cloud mask generation which are presented in Table 1. The FMASK machine learning model proposes the feasibility and study of various other ML models that can be used for better performance in terms of accuracy of cloud detection.

Table 1: cloud detection and masking techniques

|  |  |  |
| --- | --- | --- |
| **Reference** | **Model** | **Model used** |
| 15 | SEN12MS-CR-TS | SEN12MS-CR-TS |
| 16 | SECloud Mask | spectral-temporal classifiers |
| 17 | Fmask | fusion of Images and Auxilary data |
| 18 | dsen2-cr | Deep residual neural network |
| 19 | DEcloud | Deep learning model |
| 20 | Luojia1-Cloud-Detection | Threshold-based cloud detection |
| 21 | Deep-gap fill | deep convolutional autoencoder for cloud detection and gap filling |
| 22 | CloudFCN | Full CNN |
| 23 | Ukiscsmask | convolution neural network |
| 24 | STGAN | Cloud Removal using Spatiotemporal Generative Models |
| 25 | Cloud-Net | fully convolutional network (FCN) based cloud detection |
| 26 | CloudMattingGAN | GAN |
| 27 | ES-CCGAN | haze removal using cycle generative adversarial network |
| 28 | Cdnet | basic CNN with low dataset and low accuracy |
| 29 | GLNET | CNN-based cloud and non-cloudy classification |
| 30 | CDNetV2 | CNN-based model cloud detection and removal |
| 31 | AISD | Deep learning model for shadow detection |
| 32 | Cloud-GAN | Deep learning GAN model |
| 33 | Mec-GAN | https://github.com/andrzejmizera/MEcGANs |
| 34 | CloudXNet | https://github.com/shyamfec/CloudXNet |
| 35 | SEnSEl | https://github.com/aliFrancis/SEnSeI |

In [15] presents a new remote sensing dataset aimed at cloud removal in multitemporal images. The authors start by highlighting the importance of remote sensing data in various applications, including land use and land cover classification, crop yield estimation, and urban planning. However, the presence of clouds in the images can significantly affect the accuracy of these applications. To address this issue, the authors introduce SEN12MS-CR-TS, a new data set that includes multimodal and multitemporal remote sensing data with and without clouds.

In another work a model was proposed to remove the noise from the images and new pixels were generated using a geometric median. The authors in [16] propose an API named SECloud Mask to regenerate pixels and fill the noise in the image with high-quality pixels where noise can be cloud and cloud shadow.

FMask[15] is a tool kit and algorithm aimed at identifying clouds, cloud shadows and snow in satellite images. The toolkit was released in 2015 and has been improved over the period with the latest release of FMask 4.0. The tool is made for Landsat 4-8 and sentinel 2 satellite images. The model uses Haze Optimized Transformation (HOT) for the prediction of clouds and snow in images. The tool is used to define the Normalized Difference Snow Index (NDSI) and Normalized Difference Cloud Index (NDCI).

In this generation of artificial intelligence, various works are being proposed using deep learning and neural networks. Various trained machine learning models are produced using deep learning, artificial neural networks, CNN, RNN and many more. In [18] a similar work is presented for cloud detection and removal from sentinel-2 images using deep neural networks. The work showcases the collection of huge satellite data and the training of the data for cloud detection using deep RNN which is a neural network with a large number of hidden layers and neurons. The work is useful for detecting and removing clouds from images and regenerating the removed pixels using an optical representation of near-land structure.

Another work using a deep CNN-based machine learning model [21] resulted in a tool called Deep-gapfill. The tool is an image gap filling model using a deep convolutional neural network which is trained for filling the pixels in radar images. This work is just a demonstration since it is not trained with a huge dataset. Another research using CNN (Convolutional neural network) for the identification of cloud in satellite GIS data is presented in [22]. CloudFCN [22] is a CNN-based detection machine learning model for any raster images. The model identifies thick clusters of cloud and their shadow over the area. The model is trained with Landsat and sentinel images for training purposes. The work uses RGB band images for training purposes. The work is compared with SVM, PCA and single-pixel neural networks (NNs) [39,40,41]

Similar work for cloud detection using a fully convolutional network [23, 25] is proposed and used in tools named Cloud-Net and Ukiscsmask. Ukiscsmask is trained using Landsat OLI dataset over a U-Net CNN model for cloud detection the work is an extension of existing work where this model extends the cloud classification to five classes (“shadow”, “cloud”, “water”, “land” and “snow/ice”). Where before this only 3 classes existed (Cloud, land, and no cloud).

On the other hand, Cloud-Net [25] is a trained machine-learning model using CNN for cloud detection in Landsat 8 data. The model is very specific due to its training data restrictions. The work is compared with the existing FMask model for the accuracy of cloud detection. The proposed cloud-Net model proved to provide better accuracy in terms of detection of cloud in Landsat 8 data.

Some of the similar proposed ML-based toolkits for cloud and cloud shadow detection are Cdnet and GLNET [27, 28, 29]. These are some simple CNN-based models for cloud detection and classification into thick and thin clouds. Cloud shadow detection using deep learning is shown in AISD [31] where deep Deeply Supervised convolutional neural network for Shadow Detection (DSSDNet) is used to improve the cloud shadow detection raster Landsat data. In [32] a Distortion Coding Network method is proposed for cloud detection. In [33] another cloud detection algorithm is proposed using GAN which is an unsupervised model with higher accuracy than any other model but needs huge data for training. Similar work using machine learning is proposed in [34, 35] for cloud detection for various satellite datasets. Since the accuracy in GIS models depends on the quantity of datasets trained and the variety of datasets, so new developments are taking place to make the model more accurate.

After cloud detection and removal, the empty pixels need to be filled/generated. For this, mathematical models [16, 21] are proposed. In some of the newer research, Machine learning models and deep learning models are used to improve the accuracy and quality of the pixels. In [26] author has proposed a Generative adversarial network to use a deep neural network to generate similar pixels for replacing cloud pixels.

This data quality refers to the amount of useful data out of the whole data set. In the case of earth observatory data where various platform provides satellite images based on AOI (Area of Interest) in such cases, a polygon drawn may not provide complete data in such cases the data completeness quality needs to be checked.

Similarly, other factors that impact data completeness are cloud cover, haze or fog in the atmosphere. As discussed above various cloud detection and classification algorithms have been proposed including machine learning models. This allows users to know the useful or visible data that can be used for analysis. Similarly, classification algorithms allow you to know the degree of the visible area, partially visible or cloud-covered area.

Data completeness plays an important role in various applications like land cover, forest cover and sea or water bodies. In these specific GIS applications users are interested in knowing the quality of data in terms of useful data for their needs like land cover or sea cover without processing the data. In such cases data completeness allows you to know the data completeness in terms of land cover and sea cover which allows the user to know the data quality without computing the data which allows the user to select the high-quality data for analysis.

* 1. **Accuracy**

In this section, a review of existing work for accuracy in vector data is presented. Accuracy in GIS is the degree to which information on a map matches real-world values. It is an issue that pertains both to the quality of the data collected and the number of errors contained in a dataset or a map. The study covers the review of various types of accuracy in vector data based on data like soil data [1], atmospheric pressure [2], income [42] and many more. Where accuracy can be defined as the

In [1] data quality for watershed data which is a time-series data is discussed. Mauro et.al. [1] presented a study on the importance of data quality in watershed streamflow and sediment data analysis. The work showcases the study of fine sediment yield in the Goodwin Creek watershed of 21.3 km. The work is a study of the effect of various spatial data, and geomorphology on land use and land cover maps. The work uses various existing models like Soil and Water Assessment Tool (SWAT) and AVSWAT to study the performance. The result shows that GIS data has a significant effect on the models to predict the streamflow and sediment data analysis where the data quality plays an important role in improving the accuracy of the model.

In [42] a study on SDQ for American Community Survey Data 2013 is been performed. This study showcased the data quality errors in the American census data in various parameters like age and income where discrepancy in these parameters for some counties was very high using mean and median as data quality parameters.

* + 1. **Accuracy of the object in GIS data**

This section showcases the work done in improving and evaluating the accuracy of object detection algorithms like tree detection, roof detection, and ship detection among other downstream tasks.

Zhan, Q [4] has showcased a study on accuracy in object identification and placement in **vector maps**. The work showcases the study on the error and changes in accuracy in object detection to find the exact object like streets, buildings, trees and many more. The author has given a model to match the vector data which is a combination of lines and points which allows finding the changes like missing objects or errors in the data. On comparison of different data, the accuracy was found to be 81.8%. The study area is in Amsterdam and the Ravensburg site.

Barazzetti et.al. [6] studied the accuracy using RMSE ( Root-Mean-Square Error) of the **images** between sentinel 2 and Landsat-8 images where the comparison of the image registered at 10 m and 15 m are taken into consideration. The work also studies the accuracy of various bands B1-B11 using RMS (root-mean-square error). The study showcases that error in various reference bands 4(10m), 5(20m) and 9(60m) where RMS error was recorded in each image which varies from 0.19-0.55. This can also be used to define the correctness of the data. The study was conducted for images of various countries where the RMSE value for each country was evaluated and where a variation in the RMSE value of various locations was recorded.

Marangoz, A. M [7] studied the accuracy of land use classification between Sentinel-2 and Landsat-8 images. The work aims to first define the land use classification using Sentinel images and compare the accuracy using RGB and NIR bands. In the second phase, the same process is done with Landsat images to find the land use and classification in the image. The work has showcased lower accuracy in both sentinel and Landsat data with an accuracy of 0.74 and 0.66 correspondingly for RGB and NIR bands. The work also studies the accuracy of object-based classification where the accuracy of the sentinel and Landsat was recorded to be 80.7% and 73.4%. This showcases that for land use and object-based classification, sentinel images have higher accuracy than Landsat-8.

**Table 2. RMS pixel quality of various Bands [7]**

Table

Description automatically generated

Frantz, D. [8] proposed a system called FORCE which is a tool to generate images with high accuracy for land use that combines the images from sentinel, Landsat, NANA and ESA. The tool is designed to take multiple images and fuse them into one to generate a single image and bands which has high-accuracy data. FORCE is a data fusion tool to improve the spatial resolution of land surface images using Landsat and Sentinel ARD.

In [9] Kocaman. S et.al. have studied the image quality and geometric quality of Landsat 7 and 8 where various issues were highlighted in the global database at zoom levels and in the histogram which was further improvised by histogram and other techniques. The work highlights that the data suffer from the colour difference. The study also studies the advantages and disadvantages of the various data sources as shown in Table 3.

**Table 3**. Advantages and disadvantages of various GIS products [9]

Table

Description automatically generated

* + 1. **Structural Accuracy in GIS data**

In this section, some of the work of structural data quality in GIS data and its role is showcased. In [36] the authors demonstrate the use of GIS data to measure the accuracy of a bridge deformation. This refers to the evaluation of degradation of data accuracy which allows you to evaluate any error in a structure like bridges, buildings and high-rise structures. This work uses a ground-based radar system to collect the structural data and then further comparison and evaluation The work was able to evaluate the accuracy of the deformation in a bridge.

Similar work was done by the authors in [37] to measure the change in land use spread in urban areas using GIS where the accuracy of the data has an important role to play. The accuracy of such data needs to be evaluated to measure the consistency in the data collected and the data showcased. This work uses thematic accuracy to evaluate the correctness of the data. In another work [38] a standard for data positioning in GIS data [38] is the National Standard for Spatial Data Accuracy (NSSDA) which is used in the US for positional accuracy of data in GIS data using a normal distribution. Where the normal distribution defines the spread of position error at a specific location.

**Summary**

Table 4 shows a summary of the work where the SDQ benchmark can be defined as precision, consistency, completeness and accuracy for any GIS data which can be raster or vector. This benchmark SDQ will allow users to evaluate data, which can be raster or vector. This will allow users to select appropriate data before moving on to further analysis. These SDQs will also user to select data for analysis based on accuracy, completeness and precision this will allow a user to get the required data in the application domain may be land cover, ocean, forest cover or forest fire analysis but on the other hand if the data has low completeness in that case user has large amount of data but less useful data for its application.

**Table 4. Summary of Work on Spatial Data Quality**

|  |  |  |
| --- | --- | --- |
| **Data Quality parameter** | **GIS data quality** | **Related Work** |
| Precision | Image Resolution,  Quality of Bands,  Number of Bands | [3,5-7,13] |
| Consistency | Logical consistency | [1-2] |
| Completeness | Useful Land data, Useful Sea data,  Useful Forest data | [10-33] |
| Accuracy | Structural Accuracy,  Accuracy of bands,  Accuracy of the object,  Spatial accuracy,  Temporal accuracy,  Thematic accuracy | [4-9] [36-38] |

1. **Conclusion**

The work showcases the need for data quality in many GIS applications. Many researchers have showcased the need for data quality in GIS data for land use, flood mapping, marine applications, forest application and various other studies on climate and farming to improve the accuracy of downstream applications . However, there is no single standard for evaluating the quality of GIS data for a specific task. This raises an issue when selecting the correct dataset that is useful and on the other hand, a dataset with low data quality may result in low accuracy and even incorrect assumptions. Various European earth observatories reported that the data quality of machine-generated GIS data is low quality when tested [3-5]. This work aims to identify generalized data quality benchmark parameters to evaluate data quality in raster and vector data using SDQ. The paper has identified the metrics for SDQ as shown in Table 4, where precision, consistency, completeness, and accuracy are some of the parameters that should be evaluated for each data before usage. Table 4 also highlights some of the parameters that are clustered under specific data quality assessment. These generalized parameters will be useful for most GIS data applications. In future work, these SDQ parameters will be used to evaluate the data quality of raster data and assess its usefulness for a given use case.

**References**

[1] M. di Luzio, J. G. Arnold, and R. Srinivasan, “Effect of GIS data quality on small watershed stream flow and sediment simulations,” *Hydrol Process*, vol. 19, no. 3, pp. 629–650, Feb. 2005, doi: 10.1002/hyp.5612.

[2] S. Ying, Y. Lei, and J. Zhanming, “Evaluating spatial data quality in GIS database,” *2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007*, pp. 5962–5965, 2007, doi: 10.1109/WICOM.2007.1463.

[3] Trigila, A., Iadanza, C., & Spizzichino, D. (2010). Quality assessment of the Italian Landslide Inventory using GIS processing. *Landslides*, *7*(4), 455-470.

[4] Zhan, Q., Molenaar, M., Tempfli, K., & Shi, W. (2005). Quality assessment for geo‐spatial objects derived from remotely sensed data. *International Journal of Remote Sensing*, *26*(14), 2953-2974.

[5] Albanai, J. A., & Abdelfatah, S. A. (2022). Accuracy assessment for Landsat 8 thermal bands in measuring sea surface temperature over Kuwait and North West Arabian Gulf. *Kuwait Journal of Science*, *49*(1).

[6] Barazzetti, L., Cuca, B., & Previtali, M. (2016, August). Evaluation of registration accuracy between Sentinel-2 and Landsat 8. In *Fourth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2016)* (Vol. 9688, pp. 71-79). SPIE.

[7] Marangoz, A. M., Sekertekin, A., & Akçin, H. (2017). Analysis of land use land cover classification results derived from sentinel-2 image. *Proceedings of the 17th International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*, 25-32.

[8] Frantz, D. (2019). FORCE—Landsat+ Sentinel-2 analysis ready data and beyond. *Remote Sensing*, *11*(9), 1124.

[9] Kocaman, S., Debaecker, V., Bas, S., Saunier, S., Garcia, K., & Just, D. (2020). Investigations on the Global Image Datasets for the Absolute Geometric Quality Assessment of MSG SEVIRI Imagery. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 1339-1346. Doi: 10.5194/isprs-archives-XLIII-B3-2020-1339-2020

[10] Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., & McGill, M. (2008). Cloud detection with MODIS. Part II: validation. Journal of Atmospheric and Oceanic Technology, 25(7), 1073-1086.

[11] Kopp, T. J., Thomas, W., Heidinger, A. K., Botambekov, D., Frey, R. A., Hutchison, K. D., ... & Reed, B. (2014). The VIIRS Cloud Mask: Progress in the first year of S‐NPP toward a common cloud detection scheme. *Journal of Geophysical Research: Atmospheres*, *119*(5), 2441-2456.

[12] Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Yali, R., ... & Gómez-Chova, L. (2022). CloudSEN12-a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2.

[13] Segal-Rozenhaimer, M., Li, A., Das, K., & Chirayath, V. (2020). Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). *Remote Sensing of Environment*, *237*, 111446.

[14] Qiu, S., Zhu, Z., & He, B. (2019). Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment*, *231*, 111205.

[15] Ebel, P., Xu, Y., Schmitt, M., & Zhu, X. X. (2022). SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-14.

[16] Roberts, D., Mueller, N., McIntyre, A. (2017). High-dimensional pixel composites from Earth observation time series. IEEE Transactions on Geoscience and Remote Sensing, PP, 99. pp. 1--11.

[17] Zhu, Z., Wang, S., & Woodcock, C. E. (2015). Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote sensing of Environment*, *159*, 269-277.

[18] Meraner, A., Ebel, P., Zhu, X. X., & Schmitt, M. (2020). Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, *166*, 333-346.

[19] Cresson, R., Narçon, N., Gaetano, R., Dupuis, A., Tanguy, Y., May, S., & Commandre, B. (2022). Comparison of convolutional neural networks for cloudy optical images reconstruction from single or multitemporal joint SAR and optical images. *arXiv preprint arXiv:2204.00424*.

[20] Ou, J., Liu, X., Liu, P., & Liu, X. (2019). Evaluation of Luojia 1-01 nighttime light imagery for impervious surface detection: A comparison with NPP-VIIRS nighttime light data. *International Journal of Applied Earth Observation and Geoinformation*, *81*, 1-12.

[21] Cresson, R., Ienco, D., Gaetano, R., Ose, K., & Minh, D. H. T. (2019, July). Optical image gap filling using deep convolutional autoencoder from optical and radar images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 218-221). IEEE.

[22] Francis, A., Sidiropoulos, P., & Muller, J. P. (2019). CloudFCN: Accurate and robust cloud detection for satellite imagery with deep learning. *Remote Sensing*, *11*(19), 2312.

[23] Wieland, M.; Li, Y.; Martinis, S. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. Remote Sensing of Environment, 2019, 230, 1-12.

[24] Uzkent, B. U., Sarukkai, V. S., Jain, A. J., & Ermon, S. E. (2019). Cloud removal in satellite images using spatiotemporal generative networks.

[25] Mohajerani, S., & Saeedi, P. (2019, July). Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 1029-1032). IEEE.

[26] Zou, Z., Li, W., Shi, T., Shi, Z., & Ye, J. Generative Adversarial Training for Weakly Supervised Cloud Matting Supplementary Material.

[27] Hu, A., Xie, Z., Xu, Y., Xie, M., Wu, L., & Qiu, Q. (2020). Unsupervised haze removal for high-resolution optical remote-sensing images based on improved generative adversarial networks. *Remote Sensing*, *12*(24), 4162.

[28] Yang, J., Guo, J., Yue, H., Liu, Z., Hu, H., & Li, K. (2019). CDnet: CNN-based cloud detection for remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(8), 6195-6211.

[29] Sun, H., Lin, Y., Zou, Q., Song, S., Fang, J., & Yu, H. (2021). Convolutional neural networks based remote sensing scene classification under clear and cloudy environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 713-720).

[30] Guo, J., Yang, J., Yue, H., Tan, H., Hou, C., & Li, K. (2020). CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(1), 700-713.

[31] Luo, S., Li, H., & Shen, H. (2020). Deeply supervised convolutional neural network for shadow detection based on a novel aerial shadow imagery dataset. *ISPRS Journal of Photogrammetry and Remote Sensing*, *167*, 443-457.

[32] Zhou, J., Luo, X., Rong, W., & Xu, H. (2022). Cloud Removal for Optical Remote Sensing Imagery Using Distortion Coding Network Combined with Compound Loss Functions. *Remote Sensing*, *14*(14), 3452.

[33] Hasan, C., Horne, R., Mauw, S., & Mizera, A. (2022). Cloud removal from satellite imagery using multispectral edge-filtered conditional generative adversarial networks. *International Journal of Remote Sensing*, *43*(5), 1881-1893.

[34] Kanu, S., Khoja, R., Lal, S., Raghavendra, B. S., & Asha, C. S. (2020). CloudX-net: A robust encoder-decoder architecture for cloud detection from satellite remote sensing images. *Remote Sensing Applications: Society and Environment*, *20*, 100417.

[35] Crisler, M., Essig, R., Estrada, J., Fernandez, G., Tiffenberg, J., Haro, M. S., ... & Sensei Collaboration. (2018). SENSEI: first direct-detection constraints on sub-GeV dark matter from a surface run. *Physical review letters*, *121*(6), 061803.

[36] Erdélyi, J., Kopáčik, A., & Kyrinovič, P. (2020). Spatial data analysis for deformation monitoring of bridge structures. *Applied Sciences*, *10*(23), 8731.

[37] Herold, M., Scepan, J., & Clarke, K. C. (2002). The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. *Environment and Planning A*, *34*(8), 1443-1458.

[38] Zandbergen, P. A. (2008). Positional accuracy of spatial data: Non‐normal distributions and a critique of the national standard for spatial data accuracy. *Transactions in GIS*, *12*(1), 103-130.

[39] Li, P., Dong, L., Xiao, H., & Xu, M. (2015). A cloud image detection method based on SVM vector machine. *Neurocomputing*, *169*, 34-42.

[40] Ahmad, A., & Quegan, S. (2012). Cloud masking for remotely sensed data using spectral and principal components analysis. *Engineering, Technology & Applied Science Research*, *2*(3), 221-225.

[41] Hughes, M. J., & Hayes, D. J. (2014). Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. *Remote Sensing*, *6*(6), 4907-4926.

[42] Wong, D. W., & Sun, M. (2013). Handling data quality information of survey data in GIS: A case of using the American Community Survey data. *Spatial demography*, *1*, 3-16.

[43] Wang, S., Zhou, Q., & Tian, Y. (2020). Understanding completeness and diversity patterns of OSM-based land-use and land-cover dataset in China. ISPRS International Journal of Geo-Information, 9(9), 531.