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

**Abstract:**

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. These fields have huge scope and findings that can be disclosed using data analysis but data quality plays an important role to conclude a strong finding, else it may result in error-prone analysis and predictions.

In the field of earth observatory, the data are generated by various agencies using different tools and techniques. 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. Data quality in GIS is important because accurate and reliable data is essential for making effective decisions. 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. 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 may be due to resolution, visibility or noise.

In this work, we showcase a survey of various works to demonstrate the importance of data quality in raster satelite image data sources.

1. **Motivation**

Currently, a huge amount of satellite data is available from various source varying from low to high resolution with various bands for vegetation and many other applications. But the issue that exists in the current scenario is to evaluate and find the suitable dataset from existing satellites like sentinel 1 to sentinel 7 and landsat 1 to lansat 9. With such a huge data its becomes deficult to identify a useful data for a user defined application with a specific objective. In such case there is a need of a quality metadata and quality check to be attached to the datasets to make filteration and identification of datasets easier for a specific use cases. In this work our aim is to identify exitisting spatial data quality which can be generalized to check for data quality.

1. **Related Work:**

In this field many studies are been performed by various researchers to define the need and how data quality can be defined for earth observation data.

[1] Mauro et.al. 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 on 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 to improve the accuracy of the model.

[2]In this work, the authors has performed a study on the spatial data quality for data from various sources like maps, vector layers and satellite images. The work **showcases a mathematical model to study the data quality accuracy** parameter from various sources and product databases where each product does not fulfil all data quality parameters.

Graphical user interface

Description automatically generated

Figure 1. Spatial data quality in GIS

There exists various type of GIS data type and use cases where different data quality matrix plays an important role. In general, the GIS data can be divided into raster and vector data types, where raster data includes satellite images from various products like MODIS, Landsat, sentinel and many more. On the other hand, vector data are various layers over the map which are generated through the machine like road maps, river maps, location of hospitals and many more location-based information. Both type of data suffer from data Quality issues and resulting 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.

* 1. **Data Quality in Raster data**

Data Quality in satellite images refers to the quality of the image and accuracy of the image in relation to the position and size of the object in the image. Several of the GIS products suffer image quality due to low visibility or resolution and most of the time due to cloud cover in the image. So from the huge data available it is very difficult to identify the useful and correct data for the use case. In order to resolve this data quality plays an important role to define the quality and refine the data using Spatial data quality metrics. Some of the work in the field of image visibility and cloud masking are listed below.

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

Ackerman, S [10] has presented a cloud masking algorithm for (Moderate Resolution Imaging Spectroradiometer) MODIS 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 VCM (visible cloud mask) model. This algorithm is used to classify the various land use 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, can separate clouds and aerosols and cloud shadows.

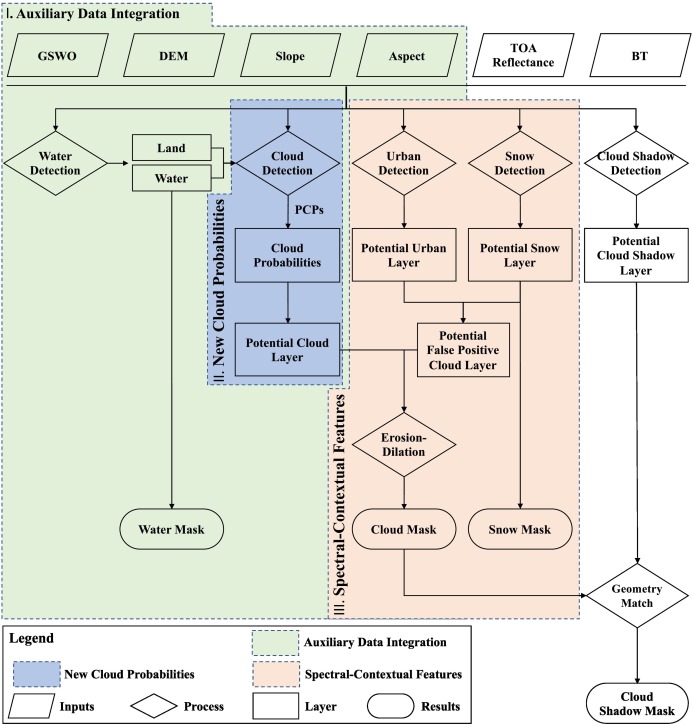
Cesar Aybar et.al. [12] has proposed a deep learning model for cloud detection for Sentinel-2. The model is called CloudSEn12 which is defined to detect cloud, cloud shadow and multi-layer clouds. The model is trained on 49400 image data. The main importance of this model as compared to other models 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

Segal R M. et.al. [13] have proposed and improved 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 bands of 10m. the testing was mostly conducted on images from Fiji island database.

Qiu.S. et.al. [14] in this work has proposed an improved version of 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. Fig X. 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.



Other than above-discussed tools there are various other models which are listed below. This **model** proposes the feasibility and study of various other ML models that can be used for better performance.

Table1: 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-gaofill | deep convolutional autoencode 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 | Model used deeip 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 |

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

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 81.8%. The study area is in Amsterdam and the Ravensburg site.

Barazzetti et.al. [6] studied the accuracy 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 RMSE value of various locations was recorded.

Marangoz, A. M [7] has 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 the lower accuracy in both sentinel and Landsat data with an accuracy of 0.74 and 0.66 correspondingly. 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 high accuracy than lansat-8.

Table

Description automatically generated

Frantz, D.[8] proposed a system called FORCE whihc 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 an image and fuse them to generate a set of 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 & 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 presents that the data suffer from the colour difference. The study also studies the advantages and disadvantages of the various data sources as shown below.

Table

Description automatically generated

* 1. **Accuracy of bands in GIS data**

Albanai et.al .[5] has showcased a model to evaluate the thermal accuracy of Landsat in the band on the sea surface. This study allows checking 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 0.03 over the year. Figure 1 and 2 shows a similar deviation over various seasons.

|  |  |
| --- | --- |
| Chart, bar chart, histogram  Description automatically generated | **Chart, bar chart, histogram  Description automatically generated** |
| Figure1 . Mean-variance in band 10 | Figure2 . Mean-variance in band 11 |

**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] author has demonstrated 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.

Similar work was done [37] to measure the change in land use spread in urban are using GIS where the accuracy of the data has an important role to play. The accuracy of such data needs to evaluate 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] Another standard for data positioning in GIS data [38] is 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.

**Data Completeness**

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 polygon drawn may not provide complete data in such cases the data completeness data quality need to be checked upon.

Similarly other factors the affect data completeness are cloud cover, haze or fog in the atmosphere. As discussed in section 3.1.1 various cloud detection and classification algorithms are been proposed including machine learning models. This allows you 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 landcover, 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 need like land cover or sea cover without processing the data. In such case data completeness allows you to know the data completeness in terms of land cover and sea cover which allows the user know the data quality without computing the data which allows the user to select the high-quality data for analysis.

Table 1. Summary of work on data quality

|  |  |  |
| --- | --- | --- |
| **Data Quality parameter** | **GIS data quality** | **Related Work** |
| Percission | 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] |

**Conclusion:**

The work **showcases the need for data quality in GIS data for various applications**. Many researchers have showcased the need for data quality in GIS applications like land use, forest application and various other studies on climate and farming. But there do not exist any standard for evaluating data quality of GIS data. This raises an issue where selecting a correct dataset that is useful and on the other hand dataset with low data quality may result in low accuracy and even incorrect assumptions. Various European earth observatories reported that data quality of machine-generated GIS data is low quality when tested. **Thus this work aims to identify a generalized data quality benchmark**. The work can identify some of the parameters for SDQ as shown in table 1, where precision, consistency, completeness and accuracy are some of the parameters which should be evaluated for each data before usage. Table 1 also highlights some of the parameters which are clustered under specific data quality assessment. These generalized parameters will be useful for most of GIS data applications.

**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.