ABNORMAL CROWDSOURCED DATA DETECTION USING REMOTE SENSING IMAGE FEATURES

YU Guang[1,2,3], ZHOU Xiaoguang[1,2,3]*, HOU Dongyang[1,2,3], WEI Dongsheng[1,2,3]

[1] School of Earth Science and Information Physics, Central South University, Changsha 410083, China;

[2] Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring（Central South University), Ministry of Education

[3] Key Laboratory of Nonferrous Resources and Geological Hazard Exploration (Hunan Province)

Commission IV, WG IV/4

KEY WORDS: feature space; crowdsourced vector; abnormal data detection; OpenStreetMap; remote sensing image.

ABSTRACT:

Quality is the key issue for judging the usability of crowdsourcing geographic data. While due to the un-professional of volunteers and the phenomenon of malicious labeling, there are many abnormal or poor quality objects in crowdsourced data. Based on this observation, an abnormal crowdsourced data detection method is proposed in this paper based on image features. This approach includes three main steps. 1) the crowdsourced vector data is used to segment the corresponding remote sensing imagery to get image objects with a priori information (e.g., shape and category) from vector data and spectral information from the images. Then, the sampling method is designed considering the spatial distribution and topographic properties of the objects, and the initial samples are obtained, although some samples are abnormal object or poor quality. 2) A feature contribution index (FCI) is defined based on information gain to select the optimal features, a feature space outlier index (FSOI) is presented to automatically identify outlier samples and changed objects. The initial samples are refined by an iteration procedure. After the iteration, the optimal features can be determined, and the refined samples with categories can be obtained; the imagery feature space is established using the optimal features for each category. 3) The abnormal objects are identified with the refined samples by calculating the FSOI values of image objects. In order to valid the effectiveness, an abnormal crowdsourced data detection prototype is developed using Visual Studio 2013 and C # programming, the above algorithms and methods are implemented and verified using water and vegetation categories as example, the OSM (OpenStreetMap) and corresponding imagery data of Changsha city as experiment data. The angular second moment (ASM), contrast, inverse difference moment (IDM), mean, variance, difference entropy, and normalized difference green index (NDGI) of vegetation, and the IDM, difference entropy and correlation and maximum band value of water are used to detect abnormal data after the selection of image optimal feature. Experimental results show that abnormal water and vegetation data in OSM can be effectively detected in this method, and the missed detection rate of the vegetation and water are all near to zero, and the positive detection rate reach 90.4% and 83.8%, respectively.

1. INTRODUCTION

Crowdsourced data are a voluntary geographic information platform, aiming to create and provide free geographic data for the world. With the development of Volunteered geographic information (VGI), several crowdsourced data platforms have sprung up, e.g., Wikimapia, OpenStreetMap (OSM), The Global Learning and Observations to Benefit the Environment (GLOBE) Program and Google Earth, etc (O'Reilly, 2007; Goodchild, 2007; Goodchild, 2009). These platforms have collected a wealth of geospatial data through active or passive means (Vatsavai & Chandola, 2016; Linda et al., 2016). However, the existing crowdsourced data platform generally lacks effective quality control measures (Heipke & Christian, 2010), due to the un-professional of volunteers, there is much poor quality objects in crowdsourced data; furthermore there is even a phenomenon of malicious labelling. As a result, the quality of crowdsourced data has become a key issue restricting its wide application. For example, On April 24, 2015, a "Google Maps fan" produced a set of spatial objects using Google Map Maker. These objects formed a figure of “Android robot indecent Apple logo”. It has become a typical event for the problem of quality control of crowdsourced data.

In order to solve the problem of quality control for crowdsourced geographic data, at first, many scholars have tried to know the quality problem by evaluating the quality directly using high-precision professional vector data as reference data, from the following aspects, i.e., location accuracy, completeness, logical consistency, topic accuracy, time accuracy, the latest situation and availability (Haklay & Mordechai,2010; Mohammad et al.,2014; Zhou et al.,2014; Zhou,2018; Lyimo et al.,2020). However, due to the difficulty and high cost of obtaining high-precision professional vector data in practical applications, it is suitable for only local-area crowdsourced data quality evaluation. Therefore, some researchers explored methods of using crowdsourced data acquisition processes (historical records) and contributors' reputations to evaluate and control the quality of crowdsourced data (Nasiri et al., 2018; Almendros-Jiménez & Becerra-Terón, 2018). Keßler et al. (2011) evaluated the quality of OSM data target objects by tracking the history of volunteer operations (confirmation, correction, and rollback) of the target object. Barron et al. (2014) analyzed the factors affecting the quality of crowdsourced geographic data, focusing on historical records, and proposed a comprehensive evaluation framework for OSM data quality including 25 methods and indicators. Zhao et al. (2016) proposed a model for calculating the credibility of volunteers based on the similarity of versions. Jacobs et al. (2020) used OSM data (including historical and volunteer data) in the Gatineau area of Ottawa as the research data, using unsupervised machine learning methods to expose many experienced

contributors, and based on this to evaluate the quality of OSM data. However, these methods often consider only the object's shape, size, topology errors, and the credibility of users, etc. they cannot distinguish the object's category attribute errors and the objectivity. Moreover, for the data contributed by new users, since there are neither historical data nor editing and modification by the others, the existing methods cannot evaluate the credibility of the fresh users and the data quality contributed by them, so it is impossible to judge whether they are usable.

In order to solve the problem of "Android robot indecent Apple logo", the author have tried to overlapped the objects and the corresponding images, and it is found that though the set of objects with regular shapes, and has no topological errors, it is obvious that this set of objects is seriously inconsistent with the images. Since the spatial object is a true reflection of the real geographic spatial phenomenon, fidelity is the most important characteristic of the spatial object. However, the authenticity of such a set of objects is doubtful. Therefore, how can we check the authenticity?

Remote sensing image is the snapshot of the real world with authenticity, and remote sensing classification using the different features of the images has always been an important research region use of (Adesina & Mavomi , 2014; Sofina & Ehlers , 2017). Crowdsourced data are often generated by users who are independently edited referenced by remote sensing images. Therefore, introducing remote sensing images is a reasonable way to detect the abnormal crowdsourced data. Based on this observation, this paper proposed a crowdsourced vector data anomaly detection method based on remote sensing image features. In this method, the remote sensing image is segmented by crowdsourced vectors to obtain image objects with clear boundaries, locations and a priori category attribute. By comparing the labeled category with the image features of the same category of objects, abnormal data with incorrect category can be identified.

The remainder of this paper is organized such that Section 2 outlines the proposed abnormal crowdsourced data detection method, including the sampling design, sample refinement by the iteration of remote sensing feature selection and outlier sample elimination, and the abnormal objects detection. Experimental results and discussion are presented in Section 3 and Section 4, respectively. Finally, the main conclusions are drawn in Section 5.

2. MODEL AND ALGORITHM

In this paper, we use remote sensing images as the ground truth, and assured that most crowdsourced data are true and credible. Based on this assumption, a method for detecting anomalies in crowdsourced vector data using the characteristics of remote sensing images is constructed. This method first uses pre-processed crowdsourced vectors to perform mask segmentation on remote sensing images to obtain image objects with clear boundaries, locations, and a priori category attribute. Then, regular grid random sampling is used to select image objects with prior category information as initial samples. Image object samples are classified and optimized according to the feature contribution to construct a feature space. Then, the local reachable density is used to define the abnormality index, and the abnormal objects in the initial sample are found and eliminated by selecting the appropriate threshold to obtain the preferred sample. Finally, abnormal data in the crowdsourced vector are found by combining the optimal sample and the abnormal index. The key models and methods in this paper include the automatic extraction of test samples, the construction of the remote sensing image feature space, and development of an anomaly detection method based on local reachable density and its implementation algorithm. The proposed abnormal crowdsourcing data detection framework is shown in Figure 1.

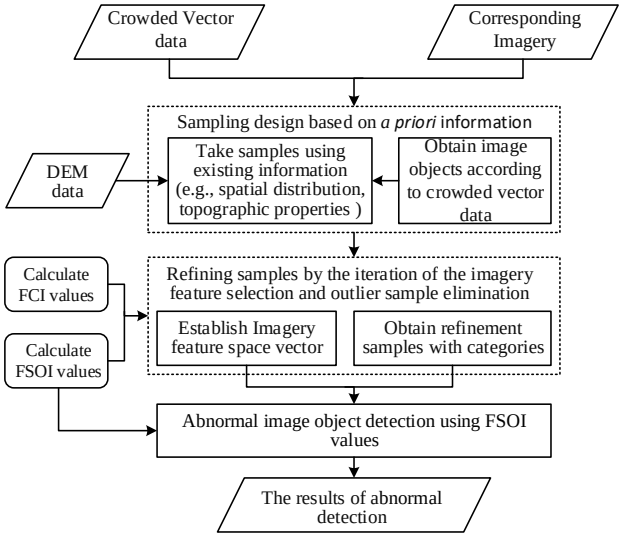


Fig 1. The framework of the proposed abnormal crowdsourced data detection

2.1 Automatic Extraction of Test Samples

In the process of detecting crowdsourced data anomalies, the quality of samples directly determines the accuracy. The sample must be random and global. If the image objects of the same priori category of sample contain multiple categories, it will directly affect the optimization of the feature parameters, which in turn will affect the accuracy of the results of abnormal data detection. In fact, though there are many incorrectly labeled abnormal objects in crowdsourced vector data, but the proportion is still relatively small compared with correctly labeled. Sample selection mainly adopts stratified random sampling and cluster sampling methods. The main considerations include the reasonable distribution of sample spatial heterogeneity, the accuracy of spatial locations, and the accuracy of sample attributes (Scepan,1999; Zhao et al.,2014; Chen et al.,2016). Since the crowdsourced data come from volunteers, their regularity of distribution is based on the volunteers' preferences and interests. During sampling, the randomness and uncertainty of the data distribution should be fully considered, and due to the complexity of the data quality, it has to be pre-processed to achieve better sampling results. Therefore, this article mainly uses a regular grid method to randomly select samples. First, the sampling area is divided into $r * c$ regular grids. The size of the grid can be established according to the size of the sampling area. The number of selected samples is given in the total number of objects of the corresponding category. The number of samples drawn by each grid Obj_{cou} can be expressed by the formula (1):

$$Obj_{cou} = \frac{Grid_{r*c}}{Sum_{obj}} * Int_{obj} \quad (1)$$

In formula (1), Sum_{obj} represents the total number of objects in the sampling layer, $Grid_{r*c}$ represents the number of objects contained in the $r * c$th grid, and Int_{obj} represents the desired samples taken.

2.2 Feature Space Construction of Feature Category

In optical remote sensing images, the spectral features and non-spectral features of the same categories have high similarities, and there are greater differences between different categories. Spectral features can express the rules of electromagnetic wave reflection of surface objects, and different surface objects can be

distinguished by calculating the pixel value of each band (Lv et al.,2018). However, "similar spectra" affect the spectral feature classification accuracy. Using texture features for change detection of remote sensing image can significantly improve the accuracy of change detection (Ulaby et al.,1986;Du et al.,2014). Therefore, Wei Dongsheng and Yang Wentao (2020) weighted 14 texture feature parameters based on the gray-level co-occurrence matrix (GLCM). The texture feature parameters for the greater contribution of each category are selected for detecting damaged buildings after an earthquake. However, because the texture features of the same category in high-resolution images are also complex and changeable, using only 14 texture features cannot fully reflect the differences between categories. Vegetation indices are the simplest and most effective measure of the vegetation condition on the surface, among which the normalized difference greenness index (NDGI) is commonly used (Meyer & Neto,2008). It selects the brightness values of the red band and the green band which can better distinguish the spectral features of vegetation from other categories. In addition, the values of each band of each object also have large differences in different categories. Therefore, this paper introduces the NDGI, band average, maximum, minimum and 14 texture features to construct the optimal feature space. Among them, the calculation formula of the NDGI, I_{NDGI} is as formula (2):

$$I_{NDGI} = \frac{I_G - I_R}{I_G + I_R} \tag{2}$$

In formula (2), I_G and I_R represent the red band and the green band value of the image object, respectively.

According to the calculation method of the texture feature contribution index (Wei Dongsheng, Yang Wentao,2020), f_i $(i = 1,2, \dots, i)$ are feature parameters, all of which are normalized, and p_i^j $(i = 1,2,3, \dots, i)$ is the feature contribution index corresponding to the prior category of the *i*-th feature parameter of image object *j*. The calculation method is shown in formula (3). The feature space of image object *j* is $Fsp(ob_j)$, and the feature space $Fsp(ob_j)$ based on information gain can be described by formula (4):

$$p_i^j = \frac{GainRat(C_j, f_i) * 100\%}{max_C \left\{ \frac{GainRat(C_j, f_i)}{max_f \left(GainRat(C_j, f_i) \right)} \right\}} \tag{3}$$

$$Fsp(ob_j) = \left[p_1^j f1, p_2^j f2, p_3^j f3, \dots, p_i^j fi \right]^T \tag{4}$$

In formulas (3) and (4), $GainRat(C_j, f_i)$ is the information gain rate of the *i*-th feature parameter f_i to the image object *j* corresponding to the prior category C_j. Since each feature has a different contribution to each category, the larger the contribution index is, the higher the recognition of the category in the construction of the optimal feature space, the higher the recognition of the feature. Therefore, according to the size of the contribution index, the feature parameters are divided into high contribution (60%-100%), medium contribution (40%-60%) and low contribution (0%-40%) groups. In general, when the contribution of a feature is between 0 and 40%, it is considered that the recognition of the feature as part of the category is extremely low and should be excluded; when it is 40%-60%, it is considered that the feature is not recognizable as part of the category. When the contribution of a feature is between 60% and 100%, the feature is considered highly recognizable as part of the category and should be selected.

2.3 Detection Method Based on the Local Reachable Density

General anomalies are data that deviate from most of the data in a data set (Hawkins,1980). In this article, abnormal data can be defined as the user's incorrect labeling of crowdsourced data, which mean the wrong data for attribute tagging for surface category. After constructing the optimal feature space, using the idea of density-based anomaly detection of crowdsourced data. The reachable local density $LRD(ob_j)$ represents the reciprocal of the average reachable distance from all objects in the *k*-th neighborhood of object ob_j to object ob_j, as shown in formulas (5) and (6).

$$LRD(ob_j) = \frac{N_k(ob_j)}{\sum_{i \in N_k(ob_j)} Rdis_k(ob_j, ob_i)} \tag{5}$$

$$Rdis_k(ob_j, ob_i) = \max\left(k - dis(ob_i), d(ob_j, ob_i)\right) \tag{6}$$

If an object is relatively distant from other objects in the feature space, then its local reachability density is small, and the probability that this object belongs to the same category as the object in its designated *k* neighborhood is small, and vice versa. $N_k(ob_j)$ represents the *k*-th neighborhood of the object ob_j. There should be at least *k* objects in the *k*-th neighborhood of the object ob_j. $Rdis_k(ob_j, ob_i)$ represents the *k*-th reachable distance between objects ob_i and ob_j,which means the maximum value of the distance between the *k*-th closest point to ob_i and the distance from object ob_j to ob_i. The Euclidean distance between two targets is adopted as distance measure. Suppose the anomaly index of the image object ob_j is $Fsoi$.

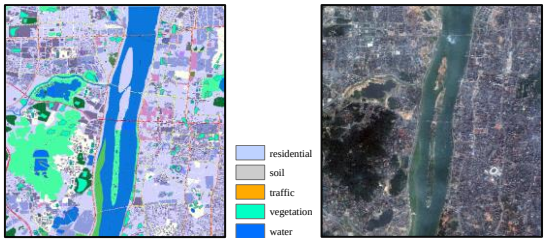
$$Fsoi(ob_j) = 1 - \frac{LRD(ob_j) * 100\%}{\max\limits_{i \in D} LRD(ob_j)} \tag{7}$$

D represents all objects in a category. The value of $Fsoi(ob_j)$ is between 0 and 100%, and its value reflects the probability that objects ob_j belongs to the category. The larger the value is, the smaller the probability that the sample object belongs to the prior category. If the value is greater than a certain threshold, the object can be considered abnormal data with incorrect attribute categories. Otherwise, the object is considered to be data with the correct attributes.

3. EXPERIMENT ANALYSIS

3.1 Experimental Data

This article takes OSM data of Changsha city China (it is the capital Hunan province, and it is a big city in the middle-south China; the coordinate range is 28.15934°N-28.24013°N, 112.91696°E-112.99370°E) as the experimental data. The Bing image contains 7424*8704 pixels with red, green, and blue bands selected for the anomaly detection experiment, and has a spatial resolution of 1.2 m. As showed in Figure 2, the OSM platform provides Bing images by default for volunteers to contribute. The all data are provided by OSM platform. Due to the attribute information of the OSM data is represented by the key and values in the tag, and the type of object cannot be determined directly. Therefore, this article uses a rule-based OSM data model conversion method to make the crowdsourced vector data have a relatively standard initial category (Zhou Xiaoguang et al., 2015).



(a) OSM vector data from 2020 (b) Bing remote sensing from 2018

Fig 2. Experiment data for anomaly detection

3.2 Results and Analysis

In order to achieve the full automation of the anomaly detection process, this paper uses C# language, ArcGIS Engine, etc. to develop a prototype system for crowdsourced data anomaly detection based on image features. The prototype has includes data loading, image cropping, grid sampling, TFCI calculation and crowdsourced data anomaly detection, etc. function. Its main interface is shown in Figure 3.

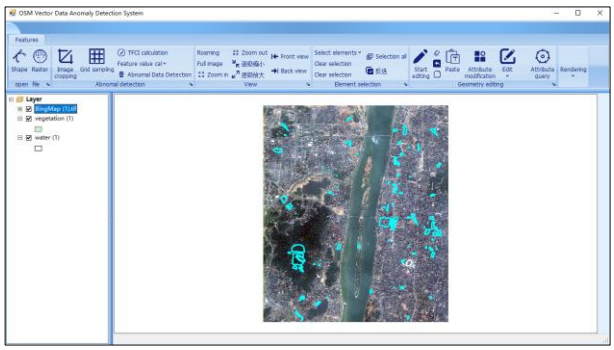
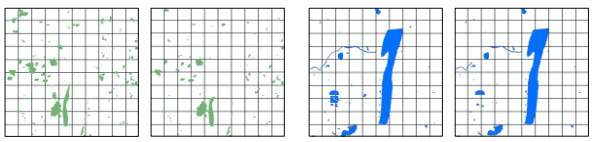


Fig 3. The interface of abnormal crowdsourced data detection prototype system

Since the experimental area is located in an urban fringe area, vegetation and water objects are relatively abundant in the crowdsourced data, this experiment uses vegetation and water as example to test the above method, and extracts the vegetation and water layers from the 2020 OSM crowdsourced vector data, and the sample exaction size of the field is 100 m×100 m. Then, the sampling layer is used to segment the 2018 image data to obtain the image objects. According to the distribution characteristics of the vegetation and water system image objects, a random sampling method is used to sample each grid. Figure 4 shows the results of the sampling layers and sample layout.



(a) (b) (c) (d)

Fig 4. Schematic diagram of the sample layout. (a),(b) respectively represent the sampling layer and sampling result of vegetation;(c),(d) respectively represent the sampling layer and sampling result of water

The test area in this paper contains a total of 325 vegetation objects, of which 278 are correctly labeled and 47 are incorrectly labeled. In addition, there are 272 water objects; 241 water objects are correctly labeled, and 31 water objects are incorrectly labeled. First, 207 vegetation sample objects are randomly selected from 325 vegetation objects through the grid, of which 180 samples are correctly labeled and 27 are incorrectly labeled; 135 water sample objects are selected from 272 water system objects, of which 120 samples are correctly labeled and 15 samples are incorrectly labeled.

In the anomaly detection of remote sensing images, the feature space vector of the sample image object is constructed first, and the optimal feature parameters of the two categories are determined according to the initial category attributes of the image object and the contribution of each feature, as shown in Figure 5, where the vegetation selection is the angular second moment (ASM), contrast, inverse difference moment (IDM), mean, variance, difference entropy, and NDGI constitute the optimal feature space vector, and the water selection includes four feature parameters, the IDM, difference entropy and correlation and maximum band value (BandMax), to constitute the optimal feature space vector.

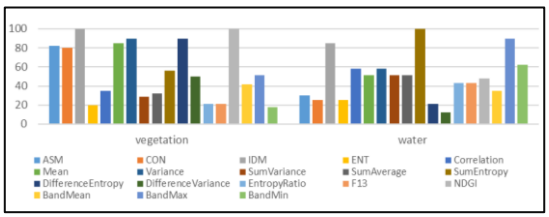


Fig 5. Feature contribution indices

When calculating the local reachable density, the value of k will affect the calculation result of the abnormality index. Tables 1 shows the sample abnormality inspection accuracy when the abnormality threshold is set to 60% or more under different k values. Among them, the abnormality threshold (AT) represents the FSOI values, the positive detection rate (PDR) represents the total number of correct detections divided by the total number of abnormalities detected, and the missed detection rate (MDR) represents the total number of true abnormalities minus the total number of correct detections and then divided by the total number of true abnormalities.

k	AT	Vegetation		Water	
		PDR/%	MDR/%	PDR/%	MDR/%
20	90	66.7	63	100	60
	80	48.4	44.4	66.7	33.3
	70	56.3	0	55.6	0
	60	51.9	0	46.9	0
30	90	85.7	33.3	100	60
	80	69	25.9	69.2	40
	70	60	0	60	0
	60	54	0	51.7	0
50	90	100	81.5	33.3	86.7
	80	81	37	50	53.3
	70	65.9	0	60.9	6.7
	60	50.9	0	48.4	0
75	90	100	88.9	100	86.7
	80	93.3	48.1	55.6	66.7
	70	53.3	40.7	52.6	33.3
	60	47.6	25.9	58.3	6.7
100	90	100	92.6	100	93.3
	80	100	70.4	100	80
	70	47.6	63	72.7	46.7
	60	50	44.4	68.4	13.3

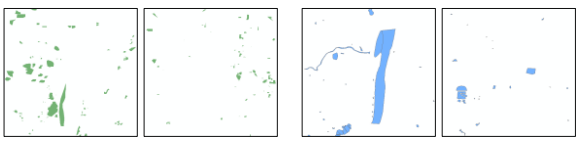
Tab 1. Analysis of the vegetation sample anomaly detection accuracy

Table 1 shows that when the abnormality threshold is 70% and *k* is 50, the MDR of vegetation is 0, and the highest PDR is 65.9%. The highest PDR is 60% when the abnormality threshold is 70% and the *k* value is 30, the MDR of water is 0. In the process of detecting abnormal data, the values of *k* and the abnormality threshold will affect the results and statistical accuracy. When the value of *k* is large, the number of samples with high abnormality will decrease, and the MDR will be correspondingly higher, which will lead to failure of abnormality rejection. When the value of *k* is in a certain interval (*k* is 30-50 in Table 1), the number of samples with a high degree of abnormality is relatively large, and the MDR is low. When the abnormality threshold is larger, the PDR will increase, but the MDR will increase. Therefore, in the calculated result, an MDR of zero can be achieved by setting an appropriate *k* value and a lower abnormality threshold, and then all samples with correct attribute categories, that is, preferred samples, can be obtained. It can be seen from the results in the table that when *k* is between 1/6 and 1/4 of the total number of samples and the abnormality threshold is 60%, it can basically ensure that all abnormal samples are eliminated, that is, an MDR of 0 can be achieved.

Therefore, in the process of removing abnormal samples, the *k* value of vegetation is set to 50, the *k* value of the water is set to 30, and the abnormality threshold is set to 70%. In the preferred sample library obtained after the abnormal samples are eliminated, there are 166 vegetation samples and 95 water samples. Although some of the normal samples are inevitably removed in the abnormality removal process, it will not affect the detection of the overall crowdsourced vector data because in this method, it is necessary only to ensure that the selected samples after removal are all correct data. A new feature space is built by optimizing the samples. When k is 1/4 of the total number of objects and the abnormality threshold is set to 70%, the abnormality value of each crowdsourced vector data sample of each category in the experimental area is calculated to obtain the accuracy of the experimental results (Table 2). The DA represents the detected abnormalities, the CD represents the Correct detections, and the TA represents the true abnormalities. Table 3 shows that under the selected parameters, the MDRs of the vegetation and water are all zero, and the PDRs reach 90.4% and 83.8%, respectively. Compared with the experiment that does not use the preferred sample library to directly perform anomaly detection, the accuracy is much improved.

Category	DA	CD	TA	PDR/%	MDR/%
Vegetation	52	47	47	90.4	0
Water	37	31	31	83.8	0

Tab 2. Accuracy of crowdsourced vector data anomaly detection



（a1） （a2） （b1） （b2）

Fig 6. the results of OSM abnormal data detection. (a1), (b1) represent the normal OSM vector data of vegetation and water, (a2), (b2) represent the OSM abnormal vector data of vegetation and water, respectively.

Figure 6 indicates the abnormal detection results of the water and vegetation categories in the experimental area. For each crowdsourced vector object, the corresponding abnormal value is automatically calculated. The data with an abnormal value greater than 70 will be highlighted and superimposed on the image data. As shown in Figure 7.b, there are obvious data of suspecting malicious annotations. When the crowdsourcing user provides this part of the vector data, its category attribute is marked as a water system. Depending on the remote sensing image visual judgment, the area where the data are located is actually a vegetation coverage area. Through the proposed detection method, similar annotation errors can be effectively and automatically detected, thereby providing users with more accurate crowdsourced vector data.

(a) Anomalous vegetation data (b) Anomalous water data

Fig 7. Abnormal data detection results

4. DISCUSSION

As we all know, data quality is the principal basis for determining its availability. Due to the non-professional of the contributors, OSM data category attribute poses a huge challenge for its use widely. Thus, this paper proposes an abnormal OSM data detection method based on the optimal feature space of each category to the reference image and the local reachable density. Although the OSM abnormal data detection in this article is mainly proposed for OSM data, this method also can be used for the other crowdsourced vector data that are mainly edited by using reference images. Experiments have proved the feasibility of the OSM abnormal data detection method based on image features proposed in this paper. However, the method proposed in this article still has some shortcomings, and the following aspects still need to be further studied and discussed:

(1) In the process of OSM abnormal data detection, the quality of the image will also bring uncertain factors. For example, the image texture characteristics of some buildings are very complicated, and it is not suitable for feature optimization; in many remote sensing images, buildings and tall trees have large shadows on the ground due to the sunlight angle. In addition, since the image is more or less affected by clouds, shadows and aerosols when projected on the ground, etc., it will also make it impossible to calculate the corresponding preferred features of the image objects segmented by the vector.

(a) Clouds (b) the shadow of building (c)the shadow of high tree

Fig 8. the problem of remote sensing

(2) In this article, the value of "*k*" means *k*-th neighborhood of object ob_j to object ob_j, it is important for the abnormal samples

distinguish from OSM vector data. In this paper, we make *k equals* 1/5 to 1/4 of the total number of samples. 1/5 to 1/4 are empirical values obtained through experimental analysis, although the optimal parameters can be determined, it is still necessary to carry out a lot of experiments and manual parameter tuning. How to adaptively select the value of *k* based on the number of samples and the type of features is still an open issue to be researched.
(3) The preferred feature space in this article is for the image object of the OSM data preferred sample. The reference image has a significant role in promoting the selection of features. Therefore, whether the migration of the preferred sample can be guaranteed is a question that needs to be discussed in this article.

5. CONCLUSIONS

A crowdsourced vector data anomaly detection method is proposed in this paper based on remote sensing image features. In this method, the crowdsourced vector data is used to segment the corresponding remote sensing imagery to get image objects with a priori information (e.g., shape and category) from vector data and spectral information from the images. Then, the sampling method is used to obtain the initial samples, although some samples are abnormal object or in poor quality. A feature contribution index (FCI) is defined based on information gain to select the optimal features, a feature space outlier index (FSOI) is used to automatically distinguish the outlier samples and changed objects. The initial samples are refined by an iteration procedure. After the iteration, the optimal features can be determined, and the refined samples with categories can be obtained; the imagery feature space is established using the optimal features for each category. At last, the abnormal objects are identified with the refined samples by calculating the FSOI values of image objects.
An abnormal crowdsourced data detection prototype system is implemented, ASM, contrast, IDM, mean, variance, difference entropy, and NDGI of vegetation; and the IDM, difference entropy and correlation and maximum band value of water are used to detect abnormal data after the selection of image optimal feature. Experimental results show that abnormal water and vegetation data in OSM can be effectively detected using this method, and the missed detection rate of the vegetation and water are all near to zero, and the positive detection rate reach 90.4% and 83.8%, respectively. Though the OSM anomaly data detection in this paper is mainly aimed at OSM data, this method may also be used for the other crowdsourcing vector data which was contributed mainly based on reference images.

REFERENCES

Adesina, G.O., and Mavomi, I., 2014. "Landuse and Landcover Change Detection of Jebba Lake Basin Nigeria: Remote Sensing and GIS Approach." *Journal of Environment & Earth Science,4* (5),119-127.

Almendros-Jiménez, J.M.,Becerra-Terón, A.,2018. Analyzing th e Tagging Quality of the Spanish OpenStreetMap. *ISPRS Int. J. Geo-Inf*, *7*(8),323. https://doi.org/10.3390/ijgi7080323.

Barron, C. , Neis, P. , & Zipf, A. , 2014. A comprehensive fram ework for intrinsic openstreetmap quality analysis. *Transactions in GIS*, *18*(6),877-895.

Du, P.J., Samat, A. , Gamba, P. , et al., 2014. Polarimetric sar image classification by boosted multiple-kernel extreme learnin g machines with polarimetric and spatial features. *International Journal of Remote Sensing*, *35*(23-24), 7978-7990.

Chen, F. , Chen, J. , Wu, H. , Hou, D. Y. , et al., 2016. A lands cape shape index-based sampling approach for land cover accur acy assessment. *Science China Earth Sciences*, *59*(12), 2263-22 74.

Goodchild, M. F. , 2007. Citizens as sensors: the world of volun teered geography. *Geojournal*, *69*(4), 211-221.

Goodchild, M. F. , 2009. Neogeography and the nature of geogr aphic expertise. *Journal of Location Based Services*, *3*(2), 82-96.

Haklay, M. , 2010. How good is volunteered geographical infor mation? a comparative study of openstreetmap and ordnance sur vey datasets. *Environment & Planning B Planning & Design*, *93* (4), 3-11.

Hawkins, D. M. 1980. Identification of outliers. Biometrics, 37 (4), 860.

Heipke, C., 2010. Crowdsourcing geospatial data. *Isprs Journal of Photogrammetry & Remote Sensing*, *65*(6), 550-557.

Jacobs, K., Mitchell, S. W.,2020. Openstreetmap quality assess ment using unsupervised machine learning methods. *Transactio ns in GIS*,*24*(5),1280-1298.

Keßler, C., Trame, J., Kauppinen, T., 2011. Provenance and Tru st in Volunteered Geographic Information: The Case of OpenStr eetMap.

Linda, S., Peter, M., Giles, F., et al., 2016. Crowdsourcing, citiz en science or volunteered geographic information? the current st ate of crowdsourced geographic information. *International Jour nal of Geo-Information*, 5(5).

Lv, P., Zhong, Y., Zhao, J., et al., 2018. Unsupervised change d etection based on hybrid conditional random field model for hig h spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(7),4002-4015.

Lyimo, N. N., Shao, Z., Ally, A. M., et al.,2020. A fuzzy logic-b ased approach for modelling uncertainty in open geospatial data on landfill suitability analysis. *International Journal of Geo-Info rmation*, 9(12),737.

Meyer, G. E. , Neto, J. C. ,2008. Verification of color vegetation indices for automated crop imaging applications. *Computers an d Electronics in Agriculture*, 63(2), 282-293.

Mohammad, F., Mahmoud, D .,2014."A Quality Study of the O penStreetMap Dataset for Tehran." *Isprs International Journal o f Geo Information 3*(2),750-763.

Nasiri, A., Abbaspour, R.A., Chehreghan, A., et al.,2018. Impro ving the Quality of Citizen Contributed Geodata through Their Historical Contributions: The Case of the Road Network in Ope nStreetMap. *ISPRS Int. J. Geo-Inf. 7*(7), 253. https://doi.org/10. 3390/ijgi7070253.

O'Reilly, T., 2007. "What Is Web 2.0: Design Patterns and Busi ness Models for the Next Generation of Software." *Communicat ions and Strategies*,*65*,17-37.

Scepan, J. ,1999. Thematic validation of high-resolution global l and-cover data sets. *Photogrammetric Engineering & Remote Se nsing*, *65*(9), 1051-1060.

Sofina, N. , Ehlers, M.,2017. Building change detection using hi gh resolution remotely sensed data and gis. *IEEE Journal of Sel ected Topics in Applied Earth Observations & Remote Sensing*, *9*(8), 3430-3438.

Ulaby, F. T., Kouyate, F., Brisco, B., et al.,1986. Textural Infor mation in SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, *24*(2), 235-245.

Vatsavai, R., and Chandola, V.,2016 “Guest editorial: big spatia l data.” *GeoInformatica ,20*,797-799.

Wei, D.S., Yang, W.T.,2020. Detecting damaged buildings usin g a texture feature contribution index from post-earthquake rem ote sensing images. *Remote Sensing Letters*, *11*(2), 127-136.

Zhao, Y.J., Zhou, X.G., Li, G.Q., et al.,2016. A Spatio-Tempora l VGI Model Considering Trust-Related Information. *ISPRS Int ernational Journal of Geo-Information*, *5*(2), 10.

Zhou, P., Huang, W., and Jiang,J.,2014. "Validation analysis of OpenStreetMap data in some areas of China.".*The International Archives of Photogrammetry, Remote Sensing and Spatial Infor mation Sciences*, *40*(4): 383.

Zhou, Qi.,2018. "Exploring the relationship between density and completeness of urban building data in OpenStreetMap for qual ity estimation." *International Journal of Geographical Informati on Science*, *32*(2), 257-281.

Zhou, X.G. , Zeng, L. , Jiang, Y., et al.,2015. Dynamically inte grating osm data into a borderland database. *ISPRS Internationa l Journal of Geo-Information*, *4*(3), 1707-1728.

Zhao, Y.Y., Gong,P. ,Yu,L.,et al., 2014. “Towards a common va lidation sample set for global land-cover mapping.” *Internationa l Journal of Remote Sensing*, *35*(13), 4795 - 4814.