Thakur, J Remote Sens GIS 2019, 8:2

Journal of Remote Sensing & GIS

Research Article Open Access

Cadastral Boundary Extraction and Image Classification Using OBIA and Machine Learning for National Land Records Modernization Programme in India

Thakur V[1*], Doja MN[1], Ahmad T[1], Rawat R[2]

[1]Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Milia Islamia University, New Delihi, India

[2]National Centre of Geoinformatics, National e-Governance Division, Ministry of Electronics and Information Technology,New Delhi, India

[*]**Corresponding author:** Vinay Thakur,Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Milia Islamia University, New Delhi, India, Tel: +9868227040; E-mail: roopalirawat111@gmail.com

Received date: July 08, 2019; **Accepted date:** July 26, 2019; **Published date:** August 07, 2019

Abstract

The present work is based on the dynamic approach for the extraction of cadastral boundaries and image classification using machine learning algorithms. The efforts are focused on easing the map digitization process in the country. The Large Scale Mean Shift Segmentation algorithm was used for the delineation of cadastral boundaries from two different types of study regions taken up for study, based on their landforms-hills and plains. The quality of segmentation was measured by AssesSeg software. Models using classifiers-Random Forest and Support Vector Machines were trained and their efficiency was tested on multiple images. The behavior of models was observed based on the landforms. The error matrices were generated based on the reference data. We tested these models as demonstrator for updating old maps through image analysis and on the basis of their performance, considered the potential of using them to update land records data in the country. This research shows the possibility of adapting the supervised machine learning methods for the extraction and classification of geographical features using satellite imagery.

Keywords: Feature extraction; Support vector machines; Random forest; Image segmentation; Machine learning; Geographic information system; Image classification

Introduction

The National Land Records Modernization Programme was launched by the Government of India in 2008. During the initial stages, the data digitization was started in the various blocks and then the data was merged to district level. Land record computerization is one of the earliest e-governance projects initiated in India and a majority of states have reached a stage wherein they could capture mutation details as a part of automation process and provide Record of Rights (RoR) to a common man in service mode [1]. One of the basic steps towards the computerization process is introducing the image of the cadastral region into a digital format. Adding the area of the parcel as an attribute is very useful for the owners and the authorities [2]. The cadastre maps or imageries are scanned and digitized using specific software and every feature on the map is brought into the editable environment. A parcel is the unit of land that has to be defined on cadastral map by indicating its boundary and assigning it a unique identity. With the advancement in image processing techniques in past few years [3], it has now become possible to extract meaningful information from the satellite imagery. Remote sensing images cover wide areas and contain vital information regarding the land. To extract this meaningful information from High Resolution Images, several researchers have worked on OBIA techniques in which the image objects are analyzed, instead, of pixels. Regarding the previous work in the similar area [4], extracted the cadastral information from World view images using mean-shift segmentation application in QGIS. Their results were quiet satisfactory. They analyzed that parcel boundaries in flat and non-vegetated areas were extracted more accurately compared to other areas. In one study [5], a methodology was proposed combining super pixel and supervised classification for the delineation of agriculture parcels automatically using machine learning approach. Edge detection technique and object based classification was used [6] for extraction of land information automatically from high resolution imagery. They developed a complete new algorithm for this work. Multi-Scale Object-Specific Segmentation (MOSS) was introduced [7] for automatic delineation of segments at multiple scales from high resolution scenes.

Segmentation

For the extraction of information from any imagery, segmentation is a pre-processing step. The process divides the image into regions or objects of homogeneous pixel values [8-10]. In image processing, segmentation is the process of grouping the pixels of same intensity. The process simplifies the image into meaningful segments or objects. The uses of segmentation process are image simplification, image classification, image compression, edge detection, Object-Based Image Analysis (OBIA), feature extraction and object recognition. The most effective segmentation algorithms are obtained by carefully customizing combination of components. The parameters of these components are tuned for the characteristics of the image modality used as input and the features of the objects to be segmented. Image segmentation algorithms are categorized as-(1) point based or pixel based (2) edge-based and (3) region-based. In point based segmentation, the pixels of the image are separated into different segments using threshold method. In edge-based segmentation, the edges in the image are tracked and linked into contours to represent the boundaries of image objects whereas in Region-based segmentation, which we have implemented, the image is divided into regions or objects of homogeneous pixel values. Segmentation can affect image classification and efficiency. It is not easy to extract desired categories of objects with controlled quality from a very large

image and none of the method has been reached a level to be considered as operational [11]. Trial and error is still a standard approach for achieving proper segmentation of the objects of interest. Knowledge of the analyst about imagery and segmentation experience also counts towards the successful results [12]. Some previous work done so far included–region based segmentation [12] in which large agriculture fields were extracted from the Landsat scenes in an area in Germany using detection models. Various segmentation methods were implemented [6] using object based analysis to extract the land information automatically.

Mean Shift segmentation was introduced by Fukunaga and Hostetler [13], is a non-parametric iterative algorithm and considers feature space as an empirical probability density function. Clustering is the most important application of this algorithm. While processing the image using mean shift implemented in the Orfeo Toolbox, three parameters are to be assigned; A Spatial radius of the neighborhood-Sr The range radius-Rr and Minimum region size-Mr.

The adjacent pixels of the smoothed images [14], whose range distance is below the range parameter, and whose spatial distance is below the spatial parameter, are grouped together. In one of the work [5] split and merge algorithm was used to characterize the aerial images based on their grey levels and later merges the features that have the same textural similarities. The process here involved is done tile-wise where the segments whose size in pixels is below the minimum region size parameter are deleted (zero labeled) and the segments which are not useful are merged with the adjacent region based on their radiometric values. This is controlled by the minimum object size parameter. In the final step of the processing, the segmented image is changed into a relatively smooth vector file. The Mean Shift approach can be understood through mathematical approach provided by [15].

Classification model using machine learning

Two powerful machine learning classifiers SVM and Random Forest, implemented in Orfeo Toolbox were trained for classification of land use and land cover classes. Many researchers have previously described the performance of both of these algorithms. SVM method was used [7] to delineate river boundaries from satellite images. An OTB pixel based SVM classification method was developed [2] for the satellite image analysis and feature extraction for their urban disaster risk assessment project. In two other studies, [16,17] machine learning algorithms along with maximum likelihood classification were utilized for the classification and interpretation of different types of crops. A study was conducted of an area in United Kingdom with seven land cover classes, comparing the results obtained from random forest classifier and support vector machines [18]. In his study, it was found that RF classifier performed equally well to SVMs in terms of classification accuracy and training time and that RF required less number of user defined parameters compared to SVMs. In the study derived thematic maps were derived using RF and SVM classifiers and computed their accuracies [19].

SVM classifier is a supervised non-parametric statistical learning method which separates the given set of labeled training data with a hyper plane which finds the maximum distance between two classes. SVMs are able to handle problems where classes are not linearly separable. It transforms the data using a kernel function which breaks the query into a series of binary classification problems using a one-against-one approach so that binary classifiers are trained. Here, the Gaussian Radial Basis Function (RBF) kernel was used. Random Forest classifier consists of a number of trees, where the leaf nodes of each tree are labeled by estimates of the posterior distribution of the classes over the image. Each internal node contains a test that best splits the space of data to be classified. An image is classified by sending it down every tree and aggregating the reached leaf distributions.

Materials and Methods

Study area

Data sets for two study areas-'A', plain region of Nalanda District, Bihar (85°59′ 11″ W, 25°03′ 65″ S) mostly characterized by small agricultural parcels with full canopy covered crops and village settlements which comprised approx. 21 sqkm of the land and 'B' the hilly region of Mandi District, Himachal Pradesh (31°42′ 38″ N76°55′ 53″ E) comprising total 14 sqkm of area. The Quick Bird satellite images (0.6 m Panchromatic and 2.4 m Multi-spectral spatial resolution) were used acquired on 4 December 2015 (Figures 1 and Figure 2).

Process flow

Small subsets from study areas were tested against different segmentation parameters to avoid the long processing time. The resultant output is a labeled vector image where each segment is assigned a label. This dataset was used to generate machine learning models. Different sets of parameters of Sr, Rr, Mr combinations were tested before finalizing the following set-For area 'A'-2,10,15 and for area 'B'-5,15,10 (Figure 3).

Datasets and training of classifier

Instead of using the entire image to learn the model, random ROIs (Region of Interest) were created from the training images based on the visual description of the features. The ROIs were generated for following five classes each for 'Train data' and 'Validation data': (a) Agriculture (b) Open/Fallow (c) Built-up (d) Shadow (e)

Figure 1: Himachal Pradesh and Bihar (Displayed in dark tone).

Vegetation in order to train each classifier we selected the above classes. Two training sets for each class, consisting of 120-120 points each, were selected interactively by displaying the raw image on the computer screen and selecting a 1213 × 1277 homogeneous area. The

classifiers were trained using the training and reflectance data for all the bands.

The image statistics were computed and combined with the ROIs. The classifiers SVM and Random Forest were trained based on labeled geometries. As we performed object based classification, the data used for training is the vector data that was generated as the result segmentation process. The segments were used as minimum classification units instead of pixels. Linear kernel is used by default in the application. The resultant classified images are provided as Figure 4 for plain area and Figure 5 for hilly area.

Quality assessment of segmented boundaries

The results of segmentation were qualitatively measured by a command line tool 'AssesSeg'[20]. This tool is incorporated with modified version of the supervised discrepancy measure 'Euclidean Distance 2'. The ED2 evaluates the segmentation by calculating the metric which measures the geometric differences between the generated image objects and reference data. The segmentation quality is computed by:

- Potential Segmentation Error (PSE)
- Number of Segments Ratio (NSR)

Following is the mathematical explanation of the tool as defined [20].

1. $ED^2 = \sqrt{(PSE)^2 + (NSR)^2}$
2. $PSE = \frac{\sum Si|r_k|}{\sum|r_k|}$
3. $NSR = \frac{|m-v|}{m}$
4. $PSE_{NEW} = \frac{\sum|S_i - r_k| + [n \times \max(|S_i - r_k|)]}{\sum|r_k|}$
5. $NSR_{new} = \frac{|m-n-[n \times v_{\max}]|}{m-n}$

Figure 2: Process flow of the work.

$\max(|S_i - r_k|)$-Maximum under segmented area with respect to single RO$\vartheta_{\max}$-Maximum number of corresponding segment found for one single RO$\sum|r_k|$-Total area of ROs (m-n)

The PSE is the measurement of ratio between total area of under segments and the Reference Objects (ROs) whereas NSR is the measure of arithmetic discrepancy between the number of ROs (m) and the reference segments (v). If ED2 value is small, it indicates a good segmentation quality whereas higher value is the indication of poor segmentation quality.

Results

Visual interpretation

In an image, the edges are formed where there is a sudden change in the intensity of pixels [21]. The reference data we have used here is that of land records. In our analysis, we observed that the mean-shift approach is still appropriate as the extracted boundaries are identifiable, especially in the plain study area, where most of the boundaries are matching the reference data but in case of hilly study region, the algorithm is not able to generate satisfactory results. This could be due to the complex nature of the terrain where the boundaries of the parcels are difficult to locate in the image (Figure 6).

Extracted boundaries and assessment using AssesSeg

The referenced cadastral boundaries were overlaid on the extracted boundaries (after merging and smoothing) and visual analysis was done. Different parameter combination were tried and tested until the satisfactory outcome is generated. It was difficult to identify cadastral boundaries on the subset image of Himachal whereas features in the plain region of Bihar, the segmented features quiet matched the referenced geometries. The quality of resultant segments was tested against the referenced Segments using

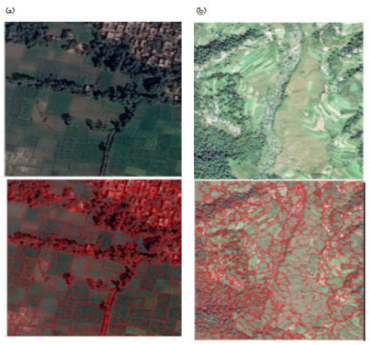

Figure 3: (a) Study area A: Raw image subset (Bihar) and mean shift segmented image and (b) Study area B: Raw image subset (Himachal) and meanshift segmented image.

Figure 4: Study Area A (L-R)-Raw Image, SVM classified, RF classified.

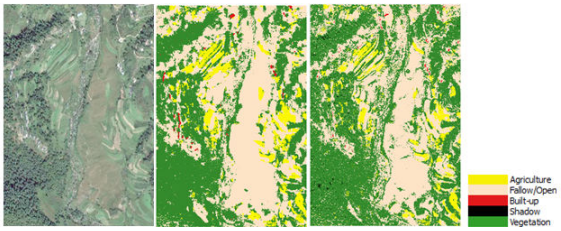

Figure 5: Study Area B (L-R)RawImage, SVM classified, RF classified.

AssesSeg tool (Figure 7) [20]. The graphical representation is depicted. It can be seen that in both the images that with the decrease in the number reference objects, the Euclidean Distance Value is increased. The ED evaluates the segmentation quality. The high ED value states the poor segmentation quality (Tables 1 and Table2).

Discussion

Accuracy assessment

For the evaluation of land cover classification, accuracy assessment of the classified maps was an important step in the analysis. Object-based accuracy assessment was performed on the classified outputs and their corresponding reference data. The relationship between both the datasets was compared using a confusion matrix approach.

Three accuracy measures, the Overall Accuracy (OA), User's Accuracy (UA), Producer's Accuracy (PA) were used to access the accuracy of models, where;

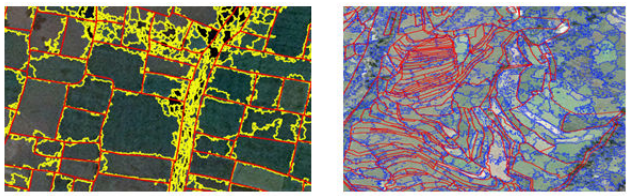

Figure 6: (a) (Left) Study area A: Referenced boundaries (Red) Extracted boundaries (Yellow) (b) (Right) Study area B: Referenced boundaries (Red) Extracted boundaries (Blue).

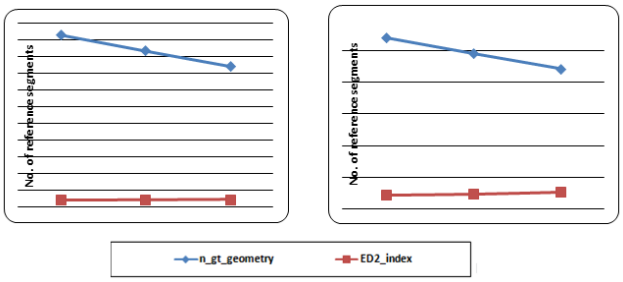

Figure 7: (a) (Left) Study area A: Bihar; ED measurement (extracted segments vs. referenced segments). (b) Study area B: Himachal; ED measurement (extracted segments vs. referenced segments).

- PA-pixels that belonged to a class but failed to get classified by the model,
- UA-pixels that belonged to a class but labeled wrongly to another and
- OA-Total classification accuracy

The model was implemented on remaining images of both the study areas and results were evaluated. The accuracy assessments of the classifications generated are shown in Tables 3-7 which summarizes the overall accuracies of both the classifiers on different types of datasets.

Assessment of classification results for plain area

Table 3 shows OA for SVM classifier for plain area is 86.66%. Shadow class has a higher PA (98.40) followed by Built-up and other categories. The PA of Fallow class has the lowest value (44.36%).

Table 4 displays the OA for RF classifier for plain area which is 69.17%. Here also Shadow class has the highest PA (98.62) and class 'Fallow' has the lowest PA of 23.72%.

Assessment of classification results for hilly area

Table 5 shows OA for SVM classifier for hilly area is 93.43%. Vegetation class has a higher PA (97.89%) followed by Fallow and other categories. The PA of Built-up class has the lowest value (79.63%). Values were not found for Shadow class as classifier did not observed any area falling under this class category.

Tables 6 and 7 display the OA for RF classifier that is 87.57%. Here also Vegetation class has the highest PA (93.26%) and Shadow class has the lowest PA of 52.74%.

From the above tables, result was concluded that SVM model predicted features more accurately in both the images. In plain area, Shadow class has the highest PA for both the classifiers whereas in contrast, the PA for the Fallow class are lower in both the classifiers due to most shadow pixels are classified as Built-up pixels.

Segmentation Parameters	No. of reference segments	Euclidean Distance
2_10_15	515	22.30
2_10_15	467	22.85

2_10_15	421	24.24

Table 1: ED values with respect to number of references of Bihar region.

Segmentation Parameters	No. of reference segments	Euclidean Distance
5_15_10	271	21.54
5_15_10	246	23.08
5_15_10	222	26.14

Table 2: ED values with respect to number of references for Himachal region.

Error matrix	Reference Data							
Classification	Agriculture	Fallow/Open	Built-up	Shadow	Vegetation	Total	PA[%]	UA[%]
Agriculture	78885	0	0	18	803	79706	95.26	98.96
Fallow/Open	2916	9419	133	9	16	12493	44.36	75.39
Built-up	10	11811	6661	0	240	18722	97.92	35.57
Shadow	6	0	0	10537	28	10571	98.40	99.67
Vegetation	992	0	8	144	5880	7024	84.39	83.71
Total	82809	21230	6802	10708	6967	128516	OA[%]= 86.66	

Table 3: Error matrix- SVM Model for plain area.

Error matrix	Reference Data							
Classification	Agriculture	Fallow/Open	Built-up	Shadow	Vegetation	Total	PA[%]	UA[%]
Agriculture	61296	31	0	25	488	61840	74.02	99.12
Fallow/Open	4003	5036	363	0	117	9519	23.72	52.90
Built-up	1600	16163	6438	0	755	24956	94.64	25.79
Shadow	12	0	0	10561	33	10606	98.62	99.57
Vegetation	15898	0	1	122	5574	21595	80.00	25.81
Total	82809	21230	6802	10708	6967	128516	OA[%]= 69.17	

Table 4: Matrix - RF Model for plain area.

Error matrix	Reference Data							
Classification	Agriculture	Fallow/Open	Built-up	Shadow	Vegetation	Total	PA[%]	UA[%]
Agriculture	3987	0	0	0	14	4001	84.84	99.65
Fallow/Open	166	9697	106	0	140	10109	95.41	95.92
Built-up	0	30	649	0	35	714	79.63	90.89
Shadow	0	0	0	0	0	0	0.0	nan
Vegetation	546	436	60	91	8776	9909	97.89	88.56
Total	4699	10163	815	91	8965	24733	OA[%]=93.43	

Table 5: Error matrix - SVM Model for Hilly area.

Error matrix	Reference Data							
Classification	Agriculture	Fallow/Open	Built-up	Shadow	Vegetation	Total	PA[%]	UA[%]

Agriculture	3443	2	0	0	20	3465	73.27	99.36
Fallow/Open	560	9370	345	0	491	10766	92.19	87.03
Built-up	0	85	438	0	0	523	53.74	83.74
Shadow	0	0	0	48	93	141	52.74	34.04
Vegetation	696	706	32	43	8361	9838	93.26	84.98
Total	4699	10163	815	91	8965	24733	OA[%]=87.57	

Table 6: Error matrix - RF Model for Hilly area.

Subset Areas	SVM	RF
Bihar	86	69
Himachal	93	87

Table 7: Accuracy percentages of models based on classified images.

This occurred due to the similar spectral reflectance values of Fallow and Built-up land. In hilly area, the Shadow class has the lowest PA as most of the shadow areas are misclassified under vegetation class due to the dominancy of the vegetative areas in some regions which suppressed the spectral reflectance of the shadow and other classes.

Conclusion and Recommendations

This study demonstrates the possibility of extraction of parcels through mean shift segmentation approach. In our work, only those boundaries of the parcels which are clearly visible are detected well. Most of the extracted segments in case of plain region show close proximity with the reference features. Although, the extracted segments are not matching the edges of the boundaries completely and are non-linear in shape. In future work, we would focus on linear conversion of the extracted boundaries as well.

There is a lot of over-segmentation in case of hilly region. This aroused due to the incompatibility of the image with the processing parameters. The complexities of the images also affect the performance of the developed system and other factors like tuning parameters, mixed pixels and lack of reference data may lead to inappropriate results.

But overall it was observed that if machine is trained well, it could yield promising outcomes and the digitization and classification task could be performed with much ease with minimizing effort on human operator.

Recommendations for future work

- Testing other segmentation algorithms and analysis of parameters for delineation of different cadastral features.
- Processing of image using block processing method so that large images could be used at once.
- Developing a standalone tool for the subdivisions or mutation of the parcels for updating the land records.
- Developing a multipurpose cadastral system for the management and maximum utilization of computerized land records.

References

1. Venkatesh DS, Pradhan S, Omlata TV, Misra DC (2014) Land records data integration strategies-certain methodological solutions. J Land Rural Studies 2: 131-143.
2. Zou Z, Lin X (2013) Geoinformatics production for urban disasters risk reduction: A zero cost solution. Conference paper, International Congress on Modelling and Simulation.
3. Michel J, Grizonnet M, Jaen A, Harasse S, Hermitte L, et al (2012) Open tools and methods for large scale segmentation of Very High Resolution satellite images. Open Source Geospatial Res Education Symposium 179-184.
4. Wassie YA, Koeva MN, Bennett RM, Lemmen CHJ (2018) A procedure for semi-automated cadastral boundary feature extraction from high-resolution satellite imagery. J Spatial Sci 63: 75-92.
5. García PA, Gonzalo MC, Lillo SM (2017) A machine learning approach for agricultural parcel delineation through agglomerative segmentation. Intern J Remote Sens 38: 1809-1819.
6. Suresh M, Jain K (2013) Colorimetrically resolution enhancement method for satellite imagery to improve land use. 14th Esri India user conference.
7. Geoffrey A (2005) An automated object-based approach for the multiscale image segmentation of forest scenes. Intern J Applied Earth Observ Geoinfo 7: 339-359.
8. Coleman GB, Andrews HC (1979) Image Segmentation by Clustering. Proceedings of the IEEE 67: 773-785.
9. Jamet O, Dissard O, Airault S (1995) Building extraction from stereo pairs of aerial images: Accuracy and productivity constraint of a topographic production line. In: Gruen A, Kuebler O, Agouris P (eds) Automatic extraction of man-made objects from aerial and space images, Monte Verità (Proceedings of the Centro Stefano Franscini Ascona).
10. Ma WY, Manjunath BS (1997) Edgeflow: A technique for boundary detection and image segmentation. Proceedings of IEEE conference on computer vision and pattern recognition 744-749.
11. Zevenbergen J, Augustinus C, Antonio D, Bennett R (2013) Pro-poor land administration: Principles for recording the land rights of the underrepresented. Land Use Policy 31: 595-604.
12. Mueller M, Segl K, Kaufmann H (2014) Extracting characteristic segments in high-resolution panchromatic imagery as basic information for object-driven image analysis. Canadian J Remote Sens 29: 453-457.
13. Fukunaga K, Hostetler LD (1999) The estimation of the gradient of a density function with applications in pattern recognition. IEEE Trans Info Theory 21: 32-40.
14. Liu W, Duan Y, Shao K, Zhang L (2007) Image smoothing based on the mean shift algorithm. IEEE Intern Conf Control and Automation 1349-1353.
15. Meer P, Comaniciu D (2002) Mean Shift: A Robust Approach toward Feature Space Analysis. IEEE transactions on pattern analysis and machine intelligence. 25: 281-288.
16. Lee RY, Ou DY, Shiu YS, Lei TC (2015) Comparisons of using Random Forest and Maximum Likelihood Classifiers with Worldview-2 imagery

for classifying Crop Types. Proceedings of the 36th Asian Conference Remote Sensing Foster, Resilient Growth Asia, Quezon City, Philippines.

17. Saini R, Ghosh SK (2018) Crop classification on single date Sentinel-2 imagery using Random Forest and Support Vector Machine. The intern arch photogram. Remote Sens and Spatial Infor Sci 425: 683-688.
18. Pal M (2005) Random forest classifier for remote sensing classification. Internl J Remote Sens 26: 1217-1222.
19. Adam E, Mutanga O, Odindi J, Abdel-Rahman EM (2014) Land-use/cover classification in a heterogeneous coastal landscape using Rapid Eye imagery: Evaluating the performance of random forest and support vector machines classifiers. Intern J Remote Sens 35: 3440-3458.
20. Novelli A, Aguilar MA, Aguilar FJ, Nemmaoui A, Tarantino E (2017) AssesSeg-A command line tool to quantify image segmentation quality: A test carried out in Southern Spain from satellite imagery. Remote Sens 9: 40.
21. Christophe E, Inglada J, Giros A (2008) Orfeo toolbox: A complete solution for mapping from high resolution satellite images. Intern Archives Photogra. Remote Sens Spatial. Info Sci 8: 37.