International Journal of
Geo-Information

Article

Exploiting the Potential of VGI Metadata to Develop A Data-Driven Framework for Predicting User's Proficiency in OpenStreetMap Context

Gangothri Rajaram and KR Manjula *

School of Computing, SASTRA Deemed to Be University, Thanjavur 613401, India; gangothri@sastra.ac.in
* Correspondence: manjula@cse.sastra.edu

Received: 31 July 2019; Accepted: 21 October 2019; Published: 31 October 2019

Abstract: Volunteered geographic information (VGI) encourages citizens to contribute geographic data voluntarily that helps to enhance geospatial databases. VGI's significant limitations are trustworthiness and reliability concerning data quality due to the anonymity of data contributors. We propose a data-driven model to address these issues on OpenStreetMap (OSM), a particular case of VGI in recent times. This research examines the hypothesis of evaluating the proficiency of the contributor to assess the credibility of the data contributed. The proposed framework consists of two phases, namely, an exploratory data analysis phase and a learning phase. The former explores OSM data history to perform feature selection, resulting in "OSM Metadata" summarized using principal component analysis. The latter combines unsupervised and supervised learning through K-means for user-clustering and multi-class logistic regression for user classification. We identified five major classes representing user-proficiency levels based on contribution behavior in this study. We tested the framework with India OSM data history, where 17% of users are key contributors, and 27% are unexperienced local users. The results for classifying new users are satisfactory with 95.5% accuracy. Our conclusions recognize the potential of OSM metadata to illustrate the user's contribution behavior without the knowledge of the user's profile information.

Keywords: OpenStreetMap; VGI; metadata; principal component analysis; K-means; multiclass logistic regression

1. Introduction

Volunteered geographic information (VGI) encourages citizens to contribute geographic data voluntarily that helps to enhance geospatial databases. A variety of VGI establishments took shape over recent years, and the one with massive response over a great extent was OpenStreetMap (OSM). OSM is mainly intended to develop a free editable base map for the whole world, acting as a potential asset to commercial or authoritative geographic data. OSM rapidly increased the volume of geographic information contributed by widely distributed users. Thus, OSM became a robust interest for researchers and practitioners focusing on its usability, benefits, and limitations. The notable change created by OSM among geographic information system (GIS) environments is extending the collection of geodata and the development of cartographic products from specialists, geographic surveyors, and cartographers to neogeographers. OSM data are produced unconventionally with richness and heterogeneity of the information collected and maintained.

The use of VGI data for productive research in the fields of routing [1], three-dimensional (3D) modeling [2], land-cover and land-use analysis [3,4], disaster monitoring [5], urban modeling [6,7], and environmental visualization [8] was widely documented in recent years. Assessing the trust and reliability of diverse user-generated OSM data is equally important to support these investigations. There exist three major approaches to assess the quality of VGI as listed below.

i. The initial approach is performing a comparative analysis of VGI and authoritative datasets of the specific region [9]. The significant drawbacks of this approach are licensing restrictions, limited data availability, and high procurement costs. Also, over the years, there arise discrepancies between data and reality due to the slow update rate of authoritative datasets.
ii. The second approach is designing participatory web-based catalogs [10] that provide a collaborative system for users and experts. Its primary focus is to publish and discover VGI and its metadata to ensure its quality. This approach is entirely user-dependent, and it requires awareness and interest among neogeographers and experts to involve and spend their time, which is unpredictable and cannot be assured all the time.
iii. The final approach is to analyze and aggregate the history and provenance of the VGI data [11] to predict quality through the evolution of the data. This approach overcomes the drawbacks as it functions on the timeliness dimension over the history of data and does not depend on any external data sources or user participation. However, it does not provide accurate measurements; instead, it is a proxy measure of data quality.

The methodologies that tend to analyze OSM quality using the history of data consider various quality measures such as reputation, content quality, credibility, trustworthiness, and local knowledge [12–18]. Many significant pieces of research also use quality indicators for quality assessment solely based on data's history. The iOSMAnalyzer framework [19] used more than 25 methods, as well as indicators within collections such as information on study area, routing and navigation, geocoding, point of interest search, map applications, and user information and behavior, for OSM quality assessment. The three widely used categories of quality indicators are [20] data-centered, socio-economic, and contributor-centric indicators. The data-centered indicators focus on coherence, consistency, and VGI metadata. The socio-economic indicators emphasize population density, social deprivation, economic reality, income, and population age. The contributor-centric indicator considers the contributor's interests, location, behavior, education, profiling, and history of contributions. The VGI metadata under data-centered indicators, in general, represent the date of observation, observation methods, equipment used, versions, feature corrections, and changes. Although it is advised to observe metadata for quality evaluation of data, the metadata content in VGI is not sufficient to describe the data. However, many approaches [21] use the metadata available with the creators during data collection or acquisition for specific applications identifying bias and assessments in VGI. In practice, metadata describing individual activities are not available for VGI, which limits the extension of quality assessment approaches for various applications. Also, substantial crowdsourcing methodologies were proposed to create VGI feature metadata using user's search terms and feature label tags [22].

The contributor-centric indicators pave the way for assessing the quality of VGI creations through its creators. Profiling of contributors and history of contributions are a rich source of information to assess the quality of VGI data. These sources help to estimate the consistency and expertise of volunteers evaluating their performance [23] to predict some concerns about VGI quality. The temporal editing patterns and contributor characteristics help in understanding spatial quality, as demonstrated in Reference [24] using intrinsic quality metrics such as contributor density over time and contributor experience to assess the quality of contributions in OSM. The metadata about the volunteers provide some information supporting VGI quality assessment. In general, they constitute age, address, education, area of interest, and other limited descriptions. These VGI content and volunteer-based metadata are two distinct sources of information that are mutually exclusive. They provide some useful insights but do not involve more details on data credibility. Therefore, it is recommended to construct metadata from user logs to support quality evaluation procedures. The construction of metadata could be of any pattern or combination as per the requirements of the type of quality assessment carried out.

The hypothesis of the proposed research is that "understanding the proficiency level of the contributor/volunteer will aid in evaluating the credibility of their VGI content". The rationale of this paper was to test this hypothesis. *Wikimapia,* a popular open-content collaborative mapping project, has a ranking system (http://wikimapia.org/docs/Community) for users that automatically determines experience levels for encouraging user contribution. It works as the user becomes more active with quality contributions; his experience points increase his level from *regular user* to *advanced user.* This ideology forms the motivation of our research, to identify the proficiency of the user based on their contribution behavior to characterize the resultant OSM data. Extracting meta-information about contributors and their contributions from user logs forms the basis of the assessment technique. In the case of OSM, the default OSM metadata contain a date range, a circular map accuracy standard (CMAS), and a tag value. These metadata are not sufficient to assess the credibility of OSM data. Therefore, we construct tailored metadata using extracted meta-information regarding the element, changeset. and users from OSM history to detect the user's proficiency level and, hence, predict the credibility of the contributed OSM data. By the term "proficiency", we mean the ability, skill, and extensive knowledge of a user in OSM activity. This paper accomplishes the above hypothesis from a machine learning perspective using unsupervised clustering and supervised classification techniques.

In this paper, we firstly perform feature selection on the user contribution history of the specific study region to extract necessary features that constitute user metadata. Then, the selected features are reduced using principal component analysis (PCA) for interpreting useful insights from correlation coefficients for metadata summarization. Furthermore, the application of unsupervised K-means clustering on the reduced user metadata features helps grouping users with similar contribution behavior. Finally, a supervised multi-class logistic regression classifier is designed to classify "new users" who were not involved in the earlier clustering phase and who were not familiar with the geographical study region. The user groups are formed based on their OSM activities, the amount of time they spend in OSM, and the validity of their contributions. The proposed methodology on identifying the proficiency level of OSM users using contribution history provides a valuable foundation for some research questions regarding inferences on VGI data quality, such as the following:

- Contributors of which proficiency levels created a particular dataset of interest?
- What is the state and position of OSM users in the particular geographic area of interest?
- What is the average timeframe for newly registered users to be active mappers?
- What type of awareness and training is needed for OSM users of the particular study area to improve the quality of OSM data?
- How can inclusive participation for the growth of OSM data quality be improved?

The remaining sections are structured as follows: Section 2 illustrates the metadata descriptions that help in model development, as well as methodologies to perform exploratory data analysis and learning for investigating user proficiency level. Section 3 applies these methods to the India OSM data and discusses the results. Section 4 concludes the proposed work and suggests possible future works.

2. Materials and Methods

In this section, we describe our methods to analyze user contribution history of OSM data and classify the users based on contributions to predict their proficiency level from a machine learning perspective. To investigate the user contribution history, we perform feature selection to mine the user logs and summarize the OSM metadata related to element, changeset, and user. We apply unsupervised learning using centroid-based clustering, and those cluster results help us to perform supervised learning through a logistic regression classifier.

2.1. OSM Metadata Description

The OSM history can be downloaded using the Geofabrik website as .osh files containing the temporal evolution of entities such as nodes, ways or relations, changeset, number of contributions, timestamps, user identifiers (IDs), and versions. These files are processed in a python environment using OSM library tools and data handlers to parse OSM data and to make data frames out of the history file. These data frames contain a complete set of attributes that describe OSM data. These attributes relate to the user's query about the number of elements each user creates for nodes, ways, and relations, as well as the modification pattern of each user and tag values of OSM elements. Still, these attributes do not directly picture any information about data quality. In this paper, we perform aggregate functions on the OSM history to extract various features for every user based on the following:

- *Time*: lifespan, inscription days, and active days;
- *Modification*: type of modification, version, and total modifications on node, ways, and relations, as well as number of modifications created, deleted, improved, corrected, and up to date;
- *Contribution*: number of modifications on elements by the same user;
- *Changeset*: starting time, ending time, and duration of changesets by users;
- *Element*: number of modifications in elements yielding number of unique contributions.

These features contain unique information about each user that could potentially describe the user's contribution behavior. The time feature group provides temporal structures that explain how users contribute through time. The modification feature group illustrates and accounts for user's modification type. The contribution feature group defines the intensity of the user contributions. The changeset feature group refers to the user modification strategy. The element feature group lists each element modification by user. Grouping of these features into various combinations yields efficient data descriptions such as *element-based metadata, changeset-based metadata, and user-based metadata*.

1. *Element-based metadata* contain all features describing elements. These data are vast and diverse, providing useful perceptions of data quality or user contributions. Also, processing the element-based metadata requires tremendous computing resources due to their size.
2. *Changeset-based metadata* contain features describing the creation, modification, and deletion of elements, hinting productive changesets that lasts longer. These data help to measure the consistency of the OSM data.
3. *User-based metadata* contain features describing the user's modification timestamp, productivity, OSM editor, and other useful user descriptions. These metadata illustrate the validity of the modifications which, along with the user's temporal features, characterizes the contributions volunteered by different users.

In our study, we focus on the user-based metadata to predict the quality of OSM data through contributors by combining unsupervised and supervised learning.

2.2. Model Development

The development of a data-driven user proficiency prediction framework was carried out in two phases, namely, an exploratory data analysis phase and a learning phase. Exploratory data analysis (EDA) is used to produce metadata summarization by reducing the features in user-based metadata. The learning phase combines unsupervised and supervised learning for clustering and classifying OSM contributors. Figure 1 illustrates the workflow of the proposed model.

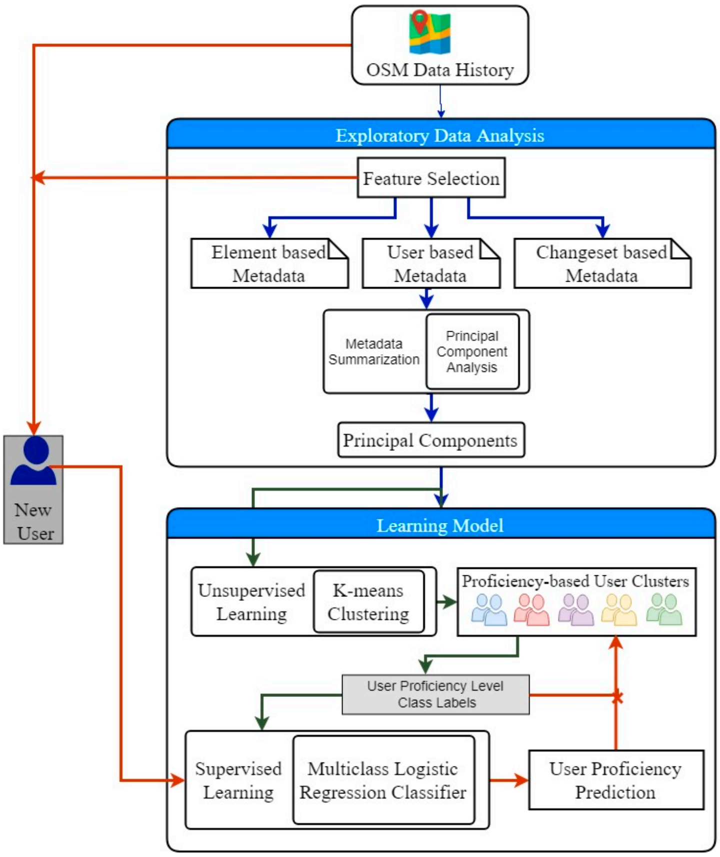


Figure 1. Overview of data-driven framework for user proficiency assessment model.

2.2.1. Exploratory Data Analysis

The user-based metadata need feature normalization as a pre-processing step to express the features into a fixed range (0, 1) to overcome the difference in the feature magnitude. Normalization is carried out in three scenarios:

i. By representing some attributes as a percentage of other attributes, for example, calculating the number of elements created in each type such as node, way, or relation with the total number of modifications in that element type (similar operations are carried out for deletion and improvement).

ii. By calculating the ratios of the number of modifications done in each element type by the total number of modifications and number of changesets edited on a specific editor by all other changesets.

iii. Based on the nature of the variables, for example, the lifespan of the user which cannot exceed the lifespan of OSM.

Finally, the features are scaled to bind them with the same minimum and maximum values for avoiding distortion of metadata using a simple min–max rule.

$$\mathrm{M_x} = \frac{\mathrm{M_{raw}} - \mathrm{Min}}{\mathrm{Max} - \mathrm{Min}}, \tag{1}$$

where Max is the maximum value observed in the user-based metadata feature j, and Min is the minimum value observed in the feature j.

The user-based metadata contain nearly 40 features (Appendix A) comprising attributes describing discrete and temporal values for the volunteer’s contribution to nodes, ways, relations, and changesets.

In this paper, we perform factor analysis on the user-based metadata parsed from the user contribution history. The factor analysis expresses the observed values in the data as functions of several possible causes to identify the importance of the values. We use principal component analysis (PCA) to accomplish factor analysis for identifying patterns in the user-based metadata. PCA quantifies the data by highlighting similarities and differences. In the literature, PCA was used for dimensionality reduction in various applications. However, PCA is also suitable for research problems whose primary concern is to determine the minimum number of factors that will account for maximum variance in the data. The benefits of PCA in this research can be listed as follows:

- There exist only 40 features (Appendix A) in user metadata. Therefore, dimensionality reduction is not necessary. Still, PCA is useful for interpreting user behavior by transforming the original features into a smaller set of linear combinations.
- The interpretation provides uncorrelated components, thereby excluding redundant information on user behavior given by subsets of features in the user-based metadata.
- PCA helps to understand the relationship between the 40 features of the user-based metadata. It offers data interpretation that makes user behavior prediction and clustering more convenient.
- Performing PCA before unsupervised clustering [25] provides effective data reductions, as principle components are the continuous solution for cluster membership indicators in k-means clustering.
- The features are descriptive, containing user information within a set of correlated variables. PCA transforms these variables into principal components, i.e., a small set of values of uncorrelated variables that still contain most of the information.

The application of PCA to the user-based metadata results in metadata summarization, i.e. binding the contribution pattern and behavior of each user within the principal components. These principal components are interpreted to study the characteristics of the user and their contribution pattern. In our methodology, we use PCA to conclude the underlying structure of the user-based metadata. Since PCA calculates the dimensions which maximize variance, it results in compound combinations of features that best describe users.

2.2.2. Combining Unsupervised and Supervised Learning for User Clustering and Classification

Our foremost objective is to identify patterns in user behavior that lead us to the grouping of the users based on their proficiency. At this point, we only observe features but have no established measurements, i.e., we do not have any prior knowledge about the user groups. Hence, it is recommended to perform initial profiling on the user metadata through unsupervised clustering. We concentrate on centroid-based clustering due to its linear space and time complexity [26]. Clustering takes a corpus of observations and divides them into distinctive groups based on similarities. A notable problem with clustering is concerned with the visualization of obtained clusters. A significant solution for this is to preprocess the data using PCA (Appendix B), as discussed in the previous section. As a result, we map the user metadata into the new feature space. Upon applying the clustering algorithm to the user metadata in the feature space, we would be able to distinguish the diverse clusters better. In simple words, PCA helps to reveal clusters.

We aim to segregate user groups with similar behavior traits, but the data in hand are not explicitly labeled. Thus, we tend to use an unsupervised learning technique to perform user grouping. As PCA already reduced our complexity of correlating attributes to identify clusters, we chose the simple K-means clustering for unsupervised learning in user-based metadata. Also, the K-means clustering algorithm is well known to be flexible and straightforward. Our goal is to detect intrinsic properties of user metadata points bound within PCA components that make them fit into the same clusters. It is an iterative procedure refining the dataset to reveal different clusters with the user-defined number of clusters (K) as input parameters. The value of "K" tells the algorithm how many means or centroids is needed to account for "K" clusters. A centroid is a data point that represents the center of the cluster (the mean), and it might not necessarily be a member of the dataset. Before starting the algorithm, the optimal number of centroids or clusters (K) is unknown. The commonly used approach to estimate "K" is to measure the sum of squared errors for a different number of clusters. The "K" value is chosen such that an increase will cause a minimal decrease in the error sum, and a decrease will sharply increase the error sum. This point is known as the "elbow point", which can be used as a visual measure for defining the optimal number of clusters.

The algorithm starts with arbitrarily chosen data points as means of clusters. Iteratively, it re-calculates new means to converge toward final clustering of the user metadata. The algorithm results in final K-means that provide inference about the OSM users. The K-means algorithm (Appendix C) groups the users into various groups relating to the PCA components. Each PCA component describes the characteristics of the user using features related to user contributions in the metadata. The results of K-means provide user clusters with unique values referring to each PCA component. The highest value among the PCA components reflects the user's contribution pattern. This user clustering is carried out without any prior knowledge about the user's profile or skills. The application of the unsupervised learning technique to the contribution history of OSM users provides clear insights into the user metadata.

The combination of PCA and K-means resulted in initial clustering of users into various groups to identify their proficiency level based on contribution history. At this point of experimentation, we partially completed the development of user analysis framework. Unsupervised learning has unique merits for the clustering process, yet supervised learning is the best choice for solving real-world computation processes. In the case of OSM, identification of the user's proficiency seeks instantaneous decision-making. If we apply supervised learning directly, the manual labeling process of observing individual users in search of meaningful patterns on contributions to feed the training phase can be a tedious task. The well-known solution is performing supervised learning after detailing the data using unsupervised learning for having clear insights into user clusters. In simple words, our purpose is matching new users or any OSM user with already existing clusters, thereby combining supervised and unsupervised learning [27] for developing a user analysis-based credibility assessment framework.

The completion of the initial profiling of user metadata using K-means resulted in good understanding by detailing user contributions. Now, we can proceed to building a modeling technique for a specific strategy of classifying new OSM users based on the knowledge gained from unsupervised learning. Beforehand, there was no output variable to guide the user metadata learning process, and we explored the metadata using the K-means unsupervised clustering technique. However, now, our existing user metadata have labels for each user based on the previous study of their contribution history. Hence, we implement a supervised learning algorithm to determine the behavior of a new user entering the study region by training the model using data labels obtained previously. We can label the users based on their contribution and the labels are also user-defined according to the requirements of the problem addressed.

If a problem requires categorization of users based on their local experience, then the principal component (PC) comprising importance to the attribute local_changeset is used. It is a binary classification problem in which the user's classification falls into locally experienced and locally unexperienced (foreign users). Another convincing case is the categorization of users based on their interest toward element type, i.e., assessing a user's high degree of contribution to nodes, ways, or relations, estimating attributes such as way_modification up to date, node_modif-autocorrected, and relation_modification-improved. These problems would result in a multi-class classification problem. Likewise, after the accomplishment of exploratory data analysis and unsupervised learning as discussed above, we can design the supervised learning phase based on the study carried out and the need of the problem in hand.

In this paper, we have clusters of users having similar OSM contribution behavior, which are used to classify the new users. The term "new user" here represents a user who was not involved in the earlier clustering phase (user without labels) and who is not accustomed to the geographical study region. The same person may be an expert or a beginner in OSM mapping in a different region.

For a particular study area, the person might be a foreign user, termed as a "new user" here. When the new user contributes certain OSM elements for this region, we wanted to analyze the OSM ID's user metadata extracted from his previous contributions, thereby classifying the user within one of the previously known clusters. Based on the classified results which label the user's proficiency level, the characteristics of the data they contributed can be estimated.

We chose multiclass (multivariate) logistic regression (MLR) classifier for the accomplishment of the supervised classification of new users. Multiclass logistic regression is a classical algorithm used to predict the categorical relationship between one nominal dependent variable and more independent variables. In this research, the k-means clusters of OSM users are labeled based on their contribution pattern, reflecting their proficiency level. Later, supervised predictive analysis is carried out using MLR with thorough training and testing of user metadata to categorize new users. The training set is used to govern the parameters of the user classification model, and the testing set is used to estimate its performance. The distinctive split between the training and testing sets was 80% and 20%, respectively [28].

The MLR Classification mechanism uses a typical workflow tracked within six stages named inputs, linear model, logits, softmax function, one-hot-encoding, and cross-entropy. All six stages of MLR workflow occur for each observation in the training metadata set.

- The *inputs* to the MLR are the 40 features we have in our user metadata. A key factor to remember while applying MLR is that these features values have to be numerical. If they are not numerical, it is required to convert them using categorical data analysis.
- The *linear model* represents the output of the linear equation that multiplies the set of inputs with the input number of weights. The weight update takes place in the training phase.
- The *logits* are scores obtained as outputs of the linear model. The logits are expected to change with changes in the calculated weights.
- The *softmax function* acts as the heart of MLR. It is a probabilistic function which calculates the probabilities for the given logits. The logits with a higher value get a high probability value, and vice versa.
- The *one-hot-encoding* method represents the target values or categorical attributes into a fixed binary representation. For all input feature sets, the resultant one-hot-encoding matrix is represented using 0 and 1 for the target class. It contains 1 for the target user class for that observation and 0s for others based on the learning assimilated from the training dataset.
- The *cross-entropy* is the distance calculation function that calculates the similarity distance between the probabilities from the softmax function and the one-hot-encoding matrix. The distance is minimum for the correct target user class and maximum for the wrong target user class.

The MLR supervised classification helps in classifying the new OSM users in a particular study region according to features of contribution history. In this research, we combine unsupervised and supervised learning to achieve pattern recognition in user metadata. It helps us to characterize OSM data contributed by a particular OSM user without knowing any personal information about the user.

3. Results and Discussion

This section illustrates the experimental study carried out, along with the results obtained. In this paper, we implemented the proposed approach on OSM data of India for classifying the contributors. Figure 2 shows the boundary extent of the study region under consideration. We could build our own OSM data history for a specific region of interest using the osmium tool (https://github.com/osmcode/osmium-tool). This tool took two input files: the OSM full history dump that contained the entire history of the OSM data sized 56 GB in pbf format (http://planet.openstreetmap.org/planet/full-history/, accessed on 15 July 2019) and a JSON configuration file that described the bounding co-ordinates' extent, output directory path, and output file format. The other two main methods for extracting OSM data are (1) defining a specific area to download the contained XML information from the OSM website using various tools, or (2) downloading freely available data in different formats through the Geofabrik website. This research adopted the latter. The OSM history dump for India contained the entire history of Indian OSM data (https://osm-internal.download.geofabrik.de/asia/india, accessed on 15 July 2019), sized 7 GB in pbf format. The history file contained numerous features describing users, elements, changesets, and modifications. The file included all versions of nodes, ways, and relations that ever existed, including deleted objects and changesets. The history file of India represented data on nearly 27,550 users. The history data represented anonymous edits with no user ID and no username, with a value of 0 for the user ID entity in the pbf file. In our experimentation, we discarded the anonymous edits to make the user analysis more transparent and accurate. At this point, we had a .pbf file for the study region containing every OSM element version through time.

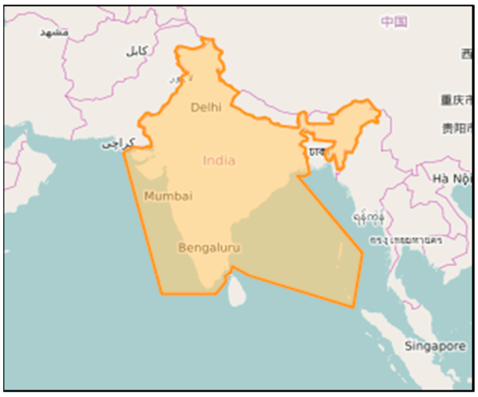


Figure 2. Study area.

3.1. Feature Selection and Metadata Summarization

For ease of access and experimentation, the raw data in the given OSM history underwent a feature selection process to extract useful attributes that constitute OSM metadata to describe the data. The selection of relevant features supports efficient model construction for the user proficiency prediction from OSM history. The features in OSM history were grouped into three categories. The attributes representing element description formed *element-based metadata*, attributes reflecting more information on changeset formed *changeset-based metadata*, and attributes describing behavioral patterns formed *user-based metadata.* In this research, we implemented aggregation operations on these metadata to gather relevant features to know how each user contributed through time, as well as the production of changesets, modification patterns, contribution intensity, and element representations. Some of the aggregation operations are listed below.

- The OSM lifespan of the user could be evaluated using the first contribution date and last contribution date of the user.
- The extraction date and first contribution date provided the number of inscription days.
- The changeset duration (in minutes) could be evaluated using the first date (starting) and last date (ending) of the changeset.
- The mean of the *changeset duration* could be associated with the changeset quantity for each user related to *user-based metadata* attributes. It provided information on how many changesets the user produced in their lifespan, along with its mean duration.
- *The user contribution for each element* could be grouped and counted. It reflected the reliability of the user in their contribution and their interest with unique elements.
- *OSM element versioning* was determined using the *first date* and *last date of the changeset modification.* The maximum and minimum values of the versions were calculated to identify whether the current OSM element was *up to date, and* whether it was *corrected or autocorrected*, to describe the contribution intensity of users.

- All *modifications by the user* were grouped and counted for each element type (node, way, and relation) to identify *total modifications improved, deleted, updated, corrected, and autocorrected* to complete user element representation.

The aggregation operations were user-defined based on the requirement of the research problem addressed. The three metadata categories were built up using these aggregation operations. Among the three metadata categories available, the *user-based metadata* provided more information for user classification and clustering. This phase of exploratory data analysis highly required metadata summarization using PCA. Metadata summarization is a process of identifying and encapsulating useful patterns from given metadata to direct the analytics by emphasizing the connections and variances within the attributes. To choose the number of components for PCA, we followed the rule of thumb, hereby the explained variance proportion was at least 70%. As a result, we selected seven PCA components (Figure 3).

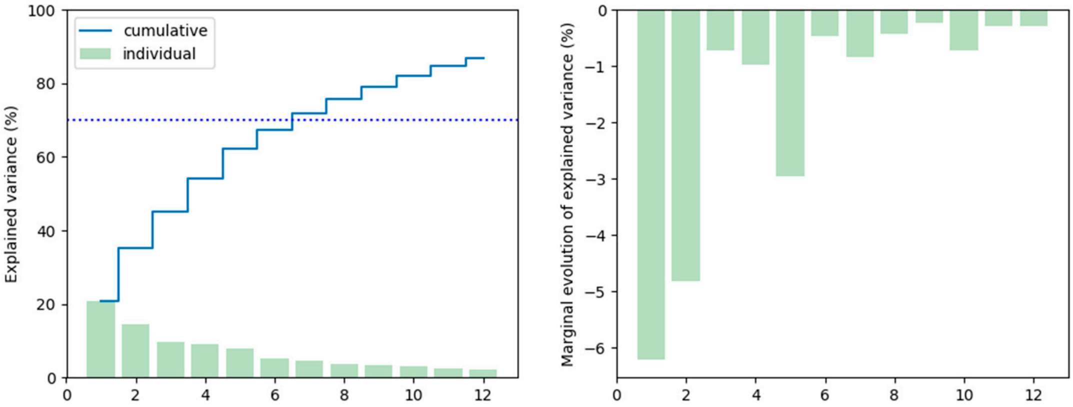


Figure 3. Explained variance for choosing the number of principal components.

The PCA algorithm took the number of components (seven) as the input parameter. After the implementation of PCA on the *user-based metadata,* the information about each user was summarized within the seven user-defined PCA components. These summarization results revealed hidden information that needed further interpretation. The PCA correlation circle in Figure 4 (Figure A1, Appendix D) represents the attributes in a two-dimensional circle with the most explained variance plotted on the horizontal axes and the second most explanatory attributes placed on the vertical axes. Here, if two lines were in the same direction, the attributes were highly correlated; orthogonal lines represented unrelated attributes, and lines that were opposite in direction represented negatively related attributes.

For a better interpretation of results, we plotted them on a heatmap (Figure 5). The user's contribution hidden on metadata attributes was represented within the range −1 (high negative contribution) to +1 (high positive contribution). Each feature (Figure 5) had unique variance values for each principal component. These values determined the combination of features that described the principal components. This observation helped us to draw conclusions about the characteristics perceived in each principal component (Appendix D, Table A2, and Table A3).

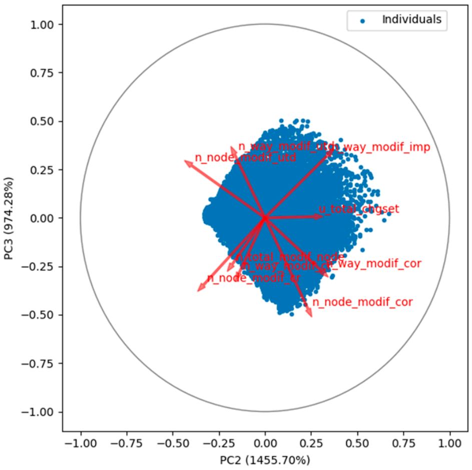
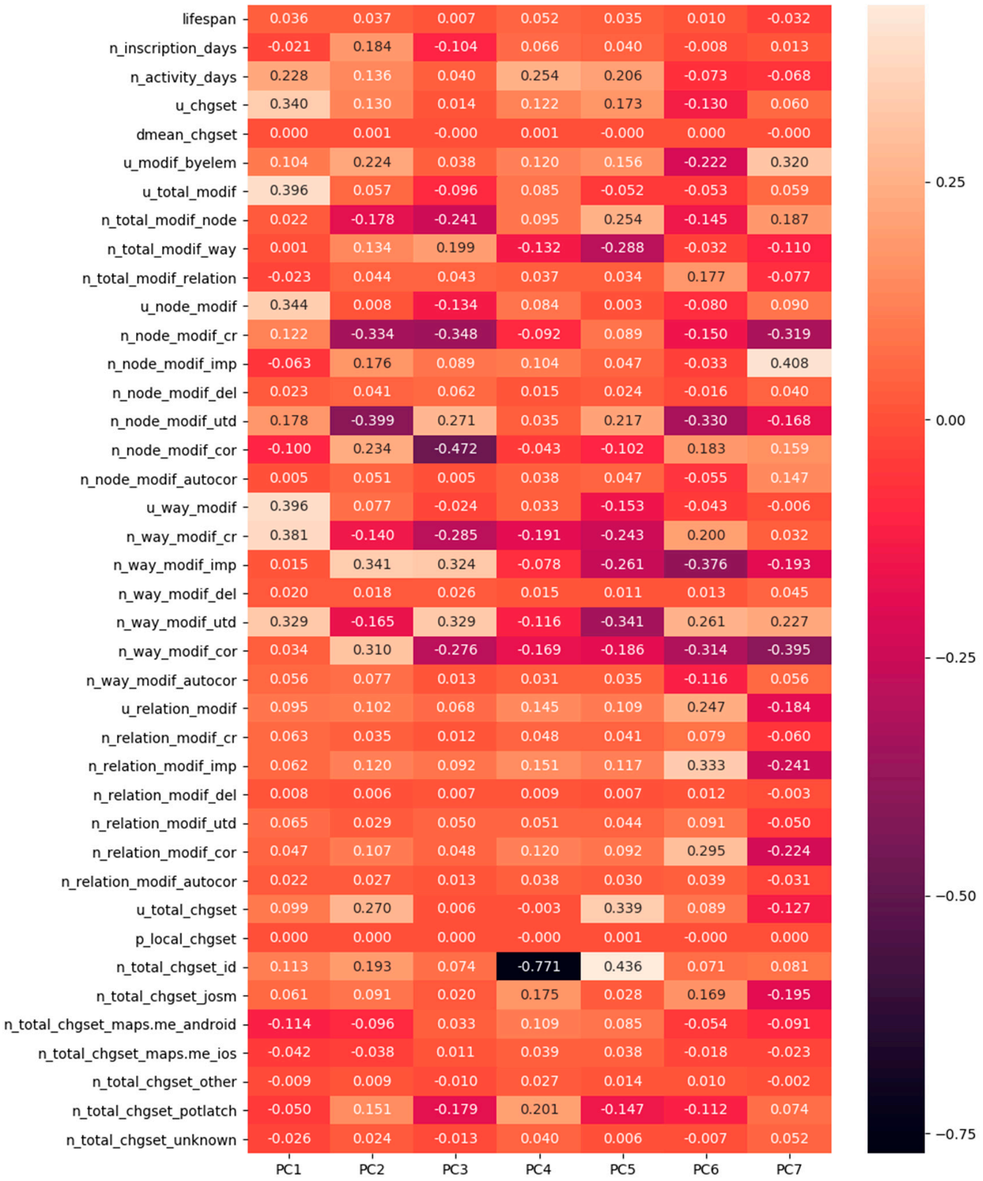


Figure 4. Principal component analysis (PCA) correlation circle.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7
lifespan	0.036	0.037	0.007	0.052	0.035	0.010	-0.032
n_inscription_days	-0.021	0.184	-0.104	0.066	0.040	-0.008	0.013
n_activity_days	0.228	0.136	0.040	0.254	0.206	-0.073	-0.068
u_chgset	0.340	0.130	0.014	0.122	0.173	-0.130	0.060
dmean_chgset	0.000	0.001	-0.000	0.001	-0.000	0.000	-0.000
u_modif_byelem	0.104	0.224	0.038	0.120	0.156	-0.222	0.320
u_total_modif	0.396	0.057	-0.096	0.085	-0.052	-0.053	0.059
n_total_modif_node	0.022	-0.178	-0.241	0.095	0.254	-0.145	0.187
n_total_modif_way	0.001	0.134	0.199	-0.132	-0.288	-0.032	-0.110
n_total_modif_relation	-0.023	0.044	0.043	0.037	0.034	0.177	-0.077
u_node_modif	0.344	0.008	-0.134	0.084	0.003	-0.080	0.090
n_node_modif_cr	0.122	-0.334	-0.348	-0.092	0.089	-0.150	-0.319
n_node_modif_imp	-0.063	0.176	0.089	0.104	0.047	-0.033	0.408
n_node_modif_del	0.023	0.041	0.062	0.015	0.024	-0.016	0.040
n_node_modif_utd	0.178	-0.399	0.271	0.035	0.217	-0.330	-0.168
n_node_modif_cor	-0.100	0.234	-0.472	-0.043	-0.102	0.183	0.159
n_node_modif_autocor	0.005	0.051	0.005	0.038	0.047	-0.055	0.147
u_way_modif	0.396	0.077	-0.024	0.033	-0.153	-0.043	-0.006
n_way_modif_cr	0.381	-0.140	-0.285	-0.191	-0.243	0.200	0.032
n_way_modif_imp	0.015	0.341	0.324	-0.078	-0.261	-0.376	-0.193
n_way_modif_del	0.020	0.018	0.026	0.015	0.011	0.013	0.045
n_way_modif_utd	0.329	-0.165	0.329	-0.116	-0.341	0.261	0.227
n_way_modif_cor	0.034	0.310	-0.276	-0.169	-0.186	-0.314	-0.395
n_way_modif_autocor	0.056	0.077	0.013	0.031	0.035	-0.116	0.056
u_relation_modif	0.095	0.102	0.068	0.145	0.109	0.247	-0.184
n_relation_modif_cr	0.063	0.035	0.012	0.048	0.041	0.079	-0.060
n_relation_modif_imp	0.062	0.120	0.092	0.151	0.117	0.333	-0.241
n_relation_modif_del	0.008	0.006	0.007	0.009	0.007	0.012	-0.003
n_relation_modif_utd	0.065	0.029	0.050	0.051	0.044	0.091	-0.050
n_relation_modif_cor	0.047	0.107	0.048	0.120	0.092	0.295	-0.224
n_relation_modif_autocor	0.022	0.027	0.013	0.038	0.030	0.039	-0.031
u_total_chgset	0.099	0.270	0.006	-0.003	0.339	0.089	-0.127
p_local_chgset	0.000	0.000	0.000	-0.000	0.001	-0.000	0.000
n_total_chgset_id	0.113	0.193	0.074	-0.771	0.436	0.071	0.081
n_total_chgset_josm	0.061	0.091	0.020	0.175	0.028	0.169	-0.195
n_total_chgset_maps.me_android	-0.114	-0.096	0.033	0.109	0.085	-0.054	-0.091
n_total_chgset_maps.me_ios	-0.042	-0.038	0.011	0.039	0.038	-0.018	-0.023
n_total_chgset_other	-0.009	0.009	-0.010	0.027	0.014	0.010	-0.002
n_total_chgset_potlatch	-0.050	0.151	-0.179	0.201	-0.147	-0.112	0.074
n_total_chgset_unknown	-0.026	0.024	-0.013	0.040	0.006	-0.007	0.052

Figure 5. Heatmap of feature contributions on each principal component.

After the detailed inspection of attributes variances within each principal component (PC), we briefly described the seven components as follows:

PC-1, principal component 1 (PC 1), had higher values for the following features: updated changesets, updated total modifications, updated node modifications, updated way modifications, number of way modifications created, and number of way modifications up to date. Furthermore, it had lower values for the following features: inscription days, the number of node modifications corrected, and the number of node modifications improved.

PC1 had a high concentration of node and way modifications with fewer improvements, deletions and corrections. It had a high degree of *way_modifications_up to date*. It had a relatively higher value for *lifespan* with more changesets. PC1 characterized experienced, skillful, and versatile users.

PC-2 had high values for *inscription days.* PC2 was strongly correlated with way and node modifications; it contained contributions by old users (*inscription days*), who were not productive since their inscription. *Potlatch* was the *most used editor*. PC2 provided synonymously high corrected and autocorrected contributions.

PC-3 had a negative value for *inscription days*. It consisted of contributions by recent users (*inscription*). It contained equal contributions to up to date and deleted modifications; it produced a relatively high number of way and relation modifications. PC3 represented quite active users in recent times (*most used editor-android, ios*).

PC-4 had a high value for the *number of activity days.* It corresponded to long-term users with more activity days. The editors included a wide variety right from Potlatch and Josm to Android and IOS. PC4 was impacted by relation modifications representing very productive active users and foreign users (local_changeset).

PC-5 had a high *user changeset* value. Changeset contributions impacted it; often, contributions were autocorrected. PC5 signaled a local, inexperienced, recent user, who was very productive in terms of elements and node modifications.

PC-6 had a marginal value for *lifespan* and negative values for *inscription days* and *activity days*. Relation modifications impacted it; subsequently, other users frequently corrected their modifications. PC6 was a sign of specialization to complex structures and foreign users.

PC-7 had low and negative values for *inscription days, lifespan,* and *activity days*. It indicated novice users and learners with fewer activity days. It had limited contributions to way and node modifications, which were probably equal to the rate of node and way deletions.

3.2. Unsupervised Learning for User Clustering

In the previous section, the results of PCA described the user contribution pattern through the features available in the user-based metadata. Hence, each component constituted specific characteristics depicting the contribution trend. With the use of these components, we learned the contribution characteristics of an individual user. Grouping similar users without any prior labeling of user groups took place through unsupervised clustering. As discussed in Section 2.2.2, we implemented the K-means clustering algorithm on the OSM history dataset to cluster users based on the *user-based metadata*. The input parameter for K-means clustering was the number of clusters of the data. The elbow and silhouette methods are ideal methods for determining the optimized number of clusters. In the elbow method, for a range of values of k, we calculated the sum of squared errors for each value of k. Our goal was to achieve a minimum value for "k" with a low sum of squared errors so as to achieve robust clustering for the dataset. As shown in Figure 6, when we plotted the sum of squared errors on a line chart, we could realize the line as an arm, and the "elbow" on the arm was the best value of k.

Similarly, the silhouette value is a measure of how similar a user is to his cluster compared to other clusters. In Figure 6, for each value of k, the average silhouette is plotted, and the location of the maximum indicates the appropriate number of clusters. As per Figure 6, the optimum number of clusters for our study was five. After implementing the K-means clustering with five as the user-defined number of clusters, the user clusters in the OSM user contribution history were listed out, as shown in Table 1.

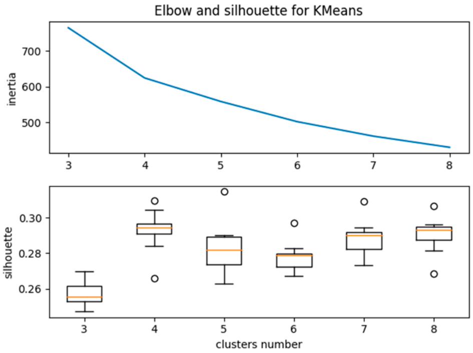


Figure 6. Elbow and silhouette plots for choosing the number of clusters for K-means.

Table 1. K-means clusters with corresponding principal component (PC) values.

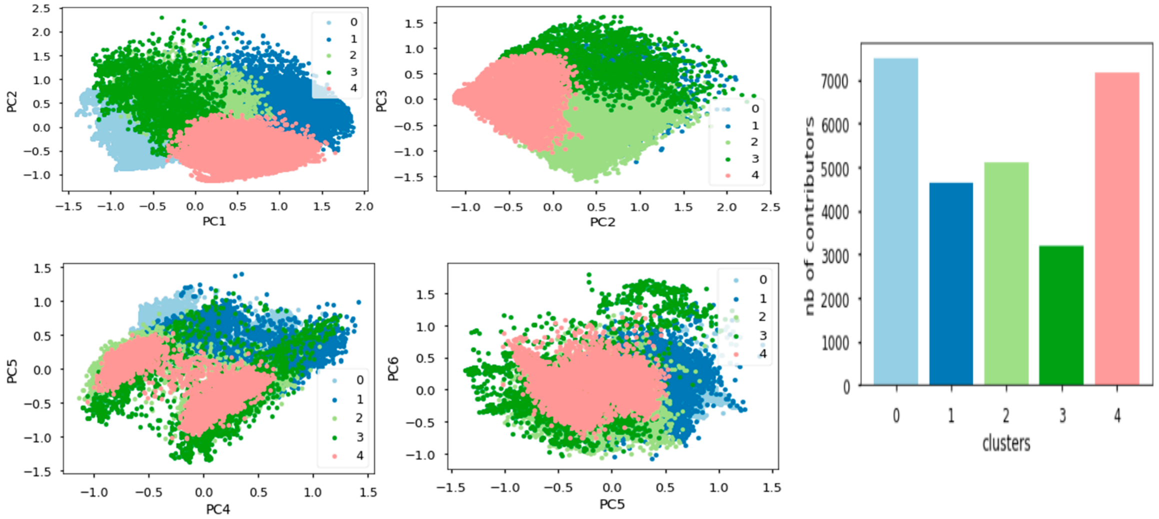
Cluster No.	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total users
C0	−0.807136	−0.367894	−0.091758	0.196653	0.362154	−0.001169	0.078870	7484
C1	0.927565	0.432962	0.125533	0.241770	0.328608	−0.019598	−0.051673	4630
C2	−0.084971	0.507034	−0.505075	−0.252831	−0.159375	−0.181888	−0.120390	5109
C3	−0.557876	0.581777	0.798466	−0.091933	−0.433605	0.147555	−0.134596	3178
C4	0.553027	−0.515710	0.020953	−0.140573	−0.284827	0.078283	0.096693	7153

Each cluster had unique values for all principal components, representing users with comparable contribution patterns. From Table 1, for cluster “0” (C0) the value for PC5 was high; cluster “1” (C1) and cluster “4” (C4) both had high values for PC1, yet, when observing the other PC values, C1 had the second highest value for PC2, whereas C4 had the second highest value for PC7. Cluster “2” (C2) had high values for PC2; cluster “3” (C3) had a high value for PC3. None of the clusters exhibited high values for PC4 and PC6, which means that there existed no long-term very productive active users and no users specialized with complex structures for the study region. Based on these values, the clusters that represented the proficiency level of users were C0 (locally unexperienced contributors), C1 (key contributors), C2 (old one-time contributors), C3 (recent contributors), and C4 (fairly productive contributors) (Table 2).

Table 2. User proficiency level clusters based on contribution behavior. OSM—OpenStreetMap.

Cluster No.	Cluster label	Cluster description	Total users
C0	Naive local contributors	Contributors that are locally unexperienced; proposed mainly changesets	**27%**
C1	Key contributors	Skilled and versatile users; OSM key contributors	**17%**
C2	Old one-time contributors	Old one-time contributors; mainly interested in way and node modifications	**19%**
C3	Recent contributors	Very close to the previous one but a more recent period during which they contributed	**11%**
C4	Fairly productive contributors	Fairly productive users with more corrections	**26%**

Figure 7 depicts the uneven distribution of the contributors within the India OSM history data. The clusters C0 and C4 had the maximum number of users, while C3 had the minimum number of users.



(**a**) (**b**)

Figure 7. (**a**) Individual distribution within clusters, and (**b**) cluster distribution.

To conclude, based on the Indian OSM History data, there were 17% of users revealing a higher level of proficiency to provide continuous valid contributions in OSM. The data provided by these users were characterized to exhibit better credibility, as knowing the way that users contributed gave information about their ability to do so properly. Also, similar to PCA, individual users could be analyzed even after K-means clustering (Table 3).

Table 3. Individual positioning of users within K-means clusters. ID—identifier.

User ID	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Cluster No.
3086	−0.180893	0.522100	−1.273169	−0.218349	−0.570122	0.279113	−0285546	C2
2973	−1.074610	0.205696	−0.791270	0.233733	0.224659	0.454524	0.084166	C0
3744	0.932690	0.558339	−1.026775	0.371771	−0.123918	−0.235809	−0.320904	C1
30047	0.440289	0.534950	−1.285919	0.091572	−0.731935	0.069069	−0.250456	C2
1399	−0.672546	0.310191	−0.847590	0.408293	0.332131	0.331600	0.144158	C0

The users with IDs 3086, 2973, 3744, 30047, and 1399 exhibited a high value for PC2, PC6, PC1, PC2, and PC4, respectively. After the implementation of K-means clustering (Table 3), the users represented a specific pattern with the PC values, placing them into a particular cluster. The users with IDs 3086 and 30047 exhibited similar patterns of PC values when arranged from higher to lower values of PCs; their orders were PC2, PC 6, PC 1, PC 4, PC 7, PC 5, and PC 3 and PC 6, PC 2, PC 1, PC 4, PC 7, PC 5, and PC 3, respectively. Therefore, they fell into the same cluster C2. Also the users with IDs 2973 and 1399 exhibited similar patterns of PC values when arranged from higher to lower values of PCs; their orders were PC 6, PC 4, PC 5, PC 2, PC 7, PC 3, and PC 1 and PC 4, PC 5, PC 6, PC 2, PC 7, PC 1, and PC 3, respectively. Therefore, they fell within the same cluster C0. The user with ID 3744 exhibited the pattern of cluster C1, which was in the order PC1, PC2, PC4, PC5, PC6, PC7, and PC3.

For a better illustration, we presented the manual observation of the user-based metadata of two distinct users with IDs 3086and 3744 (Appendix A). We could conclude that user 3744 had better proficiency than user 3086 based on lifespan, activity days, updated changesets, and updated total modifications. User 3086 contributed only once, while user 3744 made more contributions frequently. Thus, user 3086 was identified as an "old one-time contributor", while user 3744 was a "key contributor".

3.3. Supervised Learning for User Classification

As a result of unsupervised clustering, the various OSM users' cluster labels signifying proficiency level were *naive local contributors, key contributors, old one-time contributors, recent contributors,* and *fairly productive contributors.* Predicting their proficiency level from the evaluation of the contribution pattern in the OSM context helped in the characterization of their contributed data. The unsupervised clustering results could be used to discover user labels that paved the way for supervised learning. As discussed in Section 2.2.2, we established the classification of users using multiclass logistic regression (MLR). The requirement for classification problems in OSM user evaluation is to train the model to predict qualitative targets of contribution behavior for a particular user. For a better understanding of its need, we assumed a specific scenario where a new user arrives at a new geographic location and contributes certain elements to OSM. Before considering the user's contributed data as a source for research and analysis, the question of credibility, data quality, and trust remains unanswered. The K-means clustering provided many insights into contribution behavior and user proficiency level. Based on this study, we developed a supervised learning model which was trained using a previous OSM history dataset with class labels.

Before implementing supervised learning, it was necessary to check class balancing. In our dataset, it is clear from Table 2 that there were impressions on our dataset (India_OSM _History) that were imbalanced. To deal with our imbalanced dataset, we underwent oversampling through synthetic data generation. Our experiments used the synthetic minority oversampling technique (SMOTE). The SMOTE algorithm generates synthetic data of a random set of minority class observations based on feature space similarities from minority samples, rather than data space to shift the bias of classifier learning toward minority classes. This technique can effectively improve the training accuracy of the supervised learning model. Thus, custom-made user-based metadata from OSM history helped to train the MLR learning model. Then, we fed the user metadata with feature values extracted from the OSM history of new users into the trained MLR classifier in the learning process. The classifier automatically classified the data into one of the predetermined contribution behavior labels. To this point, the new user's proficiency level prediction was complete. To evaluate the performance of the classifier, we ran the MLR model using 2000 sample observations from the custom-made user-based metadata. The classifier was evaluated using a confusion matrix (Table 4), and metrics such as accuracy, precision, recall, and specificity were calculated (Table 5).

Table 4. Confusion matrix for classification evaluation. MLR—multiclass logistic regression.

Total Observations ***N* = 2000**		**Predicted**				
		C0	**C1**	**C2**	**C3**	**C4**
Actual	C0	397	1	3	3	1
	C1	2	382	1	2	1
	C2	3	4	365	22	2
	C3	0	2	21	373	2
	C4	1	3	3	4	394
Overall Accuracy of MLR Classifier:						95.5 %

Table 5. Performance metrics for classification evaluation.

User Class	**Evaluation Metric (%)**		
	Precision	**Recall**	**Specificity**
C0	98.51	98.02	99.62
C1	97.45	96.46	99.38
C2	91.02	92.17	97.76
C3	92.33	93.72	98.06
C4	98.5	97.28	99.62
Average	95.56	95.53	98.88

From the metadata of the given 2000 users in the sample, the learning model classified 1911 users correctly within the five predefined contribution behavior user class labels and misclassified 89 users. The diagonal elements of Table 4 represent the *true positive (TP)* values of the MLR classifier confusion matrix. The *true positives* are the total number of users correctly predicted with their actual user class. *True negatives* are the total number of users correctly predicted to not be a part of a user class. Similarly, *false positive and false negative* values occurred when the actual user classes and the predicted classes contradicted each other. Using these values, we calculated the accuracy, precision, recall, and specificity of the model as follows:

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Total\ number\ of\ Predictions}, \tag{2}$$

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{5}$$

The precision value was the ratio of correctly classified users to the total number of appropriately predicted users. High precision indicates that the user label with a particular user class was classified as such. The *recall* value, also called *sensitivity*, of the classification model corresponds to the true positive rate of the considered class. The high values for recall indicated that the model correctly recognized the classes. In our model, we had an average recall value of 95.53% (Table 5), which was slightly lower than the average precision value of 95.56%, which means that the classifier was conservative when classifying the users with an incorrect user class. Also, the average specificity, also known as the true negative rate, was 98.88%, indicating that the MLR model fit the given data well. Also, 95.5% overall accuracy was found to be entirely satisfactory. The values of precision and recall reflected the benefits of balancing the class observations using synthetic data generation. Overall, the results of supervised learning provided a solution to the research hypothesis.

The proposed data-driven model through machine learning methods contributed to the intrinsic quality assessment of VGI. This model is a comprehensive approach that performs analytics over meta-information extracted from OSM history. It does not put forth any quality measures or indicators, as it operates only using the contribution data available in the OSM API. The results of the model demonstrate the various user contribution patterns and user behaviors that help in profiling the user's proficiency level. The outcome of the methodology is consistent with some discussions in the literature; however, different phenomena may arise when used in other geographic regions where diverse OSM users exist. Overall, the combination of unsupervised and supervised learning provides the potential to reveal all possible combinations of features to explore varied contribution patterns and user behaviors for different use cases.

4. Conclusions and Future Directions

According to our study on OSM data in India, the data comprise more contributions from naïve local contributors and fairly productive users. Thus, there is a need for awareness among users in India about the benefits of OSM. The most significant benefit of OSM in a highly populated country like India would be for surveying and humanitarian mapping during emergencies. There is a requirement for more video tutorials in tracing, merging the two ways of routing and addressing. These video tutorials should address the complexity of India's map where streets and avenues look tangled in grids. The tutorials should help the OSM users to handle constant movement in points of interest, road geometry, one-way roads, and building profiles, which are typical in India. The current scenario illustrates 17% proficient users and 30% one-time and recent contributors. It diffuses the reliable estimation of the average timeframe for the transition from registered users to active mappers. Making user communities involving all the five classes with similar neighborhoods would eventually achieve inclusive participation.

This paper presented a comprehensive data-driven approach for exploring the potential of OSM metadata to predict user proficiency based on their contribution behavior. Exploratory data analysis acted as the foundation of the framework upon which we built the unsupervised and supervised learning models. Furthermore, the combination of unsupervised and supervised learning within the model made it more innovative at this early stage of metadata-based analysis in OSM research. The unsupervised learning aided in identifying potential subclasses of users, while supervised learning assisted in building a model for better prediction. By identifying the proficiency level of contributors, the users of OSM can make better-informed decisions based on the characteristics of the data contributed by the user. In addition to Reference [29], we evaluated the contribution behavior beyond changeset data, by structuring all features of OSM history into metadata for assessing the contributors using machine learning techniques. These features distinguish our work from previous OSM quality studies. Our approach describes the use of algorithms to examine user proficiency level in an anonymous OSM context without personalized knowledge about individual users.

This study can be extended to identify whether the quality of OSM users and data is related to the user's familiarity with the area using the classification results and profile information. Although the techniques discussed in this paper are persuasive, subsequently, their combination with quality indicators and reference datasets will improve the interpretation of OSM data quality. In the future, the model can be enhanced using deep learning to enrich the clarity and accuracy of user proficiency assessment. This research is a preliminary study on implementing learning techniques to assess the OSM data credibility through users' contribution behavior in a limited geographic area. The study can be extended to assess users globally using OSM's planet history data, labeling all OSM users accordingly. Global analysis requires high computational resources, more time, and additional aggregations for accurate assessment of users. Global assessment would pave the way for interesting GUI research, allowing the creation of plugins for graphical indications in the user's profile based on their user proficiency level class label, as discussed in this study. In addition to this, using five different colors to represent the five different clusters of users who contributed could help distinguish the OSM elements created by various OSM users. This graphical enhancement would visually help to immediately determine the characteristics of the OSM elements. The proposed model can help to discover and identify specific patterns and important behaviors of users in VGI. This knowledge about the users could be used to notify personalized tutorials and guidelines for them to improve the quality of their contributions, thereby supporting inclusive participation in VGI.

Author Contributions: Gangothri Rajaram conceptualized the idea, designed the experiment, implemented the methodology, and wrote the paper. KR Manjula supervised the research work.

Funding: This research received no external funding.

Acknowledgments: The authors thank SASTRA Deemed to Be University for their financial assistance and infrastructure support. The authors also thank J. Ravishankar, Former Head, and Veeramallu Satya Sahithi, Senior Research Fellow, Land-use and Land-cover Division, National Remote Sensing Center, Indian Space Research Organization for their valuable discussions on OpenStreetMap. The authors thank K. Kannan, Tata

Reality Srinivasa Ramanujan Research chair for Discrete Mathematics, for his advice on learning schemes and mathematical insights into data modeling.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. User metadata features.

Sno.	Features	User ID 3086	User ID 3744
1	Lifespan	0	2061
2	Number of inscription days	4301	4248
3	Number of activity days	1	13
4	Updated changeset	1	21
5	Meantime between changeset	0	0
6	Updated modifications by elements	1	1.006789525
7	Updated total modifications	27	2076
8	Number of total modifications in node	16	1442
9	Number of total modifications in way	11	634
10	Number of total modifications in relation	0	0
11	Updated node modifications	16	1442
12	Number of node modifications created	16	1442
13	Number of node modifications improved	0	0
14	Number of node modifications deleted	0	0
15	Number of node modifications up to date	0	631
16	Number of node modifications corrected	16	811
17	Number of node modifications autocorrected	0	0
18	Updated way modifications	11	634
19	Number of way modifications created	11	615
20	Number of way modifications improved	0	19
21	Number of way modifications deleted	0	0
22	Number of way modifications up to date	0	16
23	Number of way modifications corrected	11	618
24	Number of way modifications autocorrected	0	0
25	Updated relation modifications	0	0
26	Number of relation modifications created	0	0
27	Number of relation modifications improved	0	0
28	Number of relation modifications deleted	0	0
29	Number of relation modifications up to date	0	0
30	Number of relation modifications corrected	0	0
31	Number of relation modifications autocorrected	0	0
32	Updated total changeset	12	22
33	Mean local changeset	0.083333333	0.954545455
34	Number of total changeset in id editor	1	0
35	Number of total changeset in josm editor	0	0
36	Number of total changeset in maps.me android app	0	0
37	Number of total changeset in maps.me ios app	0	0
38	Number of total changeset in other OS	0	0
39	Number of total changeset using potlatch editor	0	1
40	Number of total changeset using unknown editor	11	21

Appendix B

Procedure for Principal Component Analysis in the Proposed Model

PCA was applied to the normalized user-based metadata as follows:

1. The covariance matrix was calculated using variances among all the attributes in the user-based metadata.
2. The eigenvalues and eigenvectors were calculated for the covariance matrix.
3. The number of components for PCA was chosen, for which the explained variance proportion was at least 70% and the eigenvalue was more than 1.
4. The eigenvalues were sorted to understand their significance, and the eigenvector corresponding to the largest eigenvalue was identified as the principal component.
5. A feature vector, which is a matrix of selected eigenvectors, was formed.
6. A matrix of principal components was formed by multiplying the transpose of the feature vector with the transpose of the scaled user-based metadata feature values.

Appendix C

Procedure for K-Means Clustering Algorithm for the Proposed Model

1. K centroids for the user values in the metadata were created randomly based on the predefined value of K. The input for the K-means algorithm was the features depicting users' past contributions contained within the PCA components.
2. K-means allocated every data point in the metadata set to the nearest centroid, minimizing the Euclidean distance between the points to push them (users) into a particular cluster based on distance.
3. The centroids were then recalculated by taking the mean of all data points assigned to that centroid's cluster, thereby reducing the total intra-cluster variance with the previous step. The "means" in the K-means refer to data averages, and the centroids kept moving to their equilibrium position.
4. The algorithm iterated between steps 2 and 3 until the calculated centroids remained the same and the data points stopped switching clusters (i.e., the algorithm converged).

Appendix D

Results of Metadata Summarization using PCA

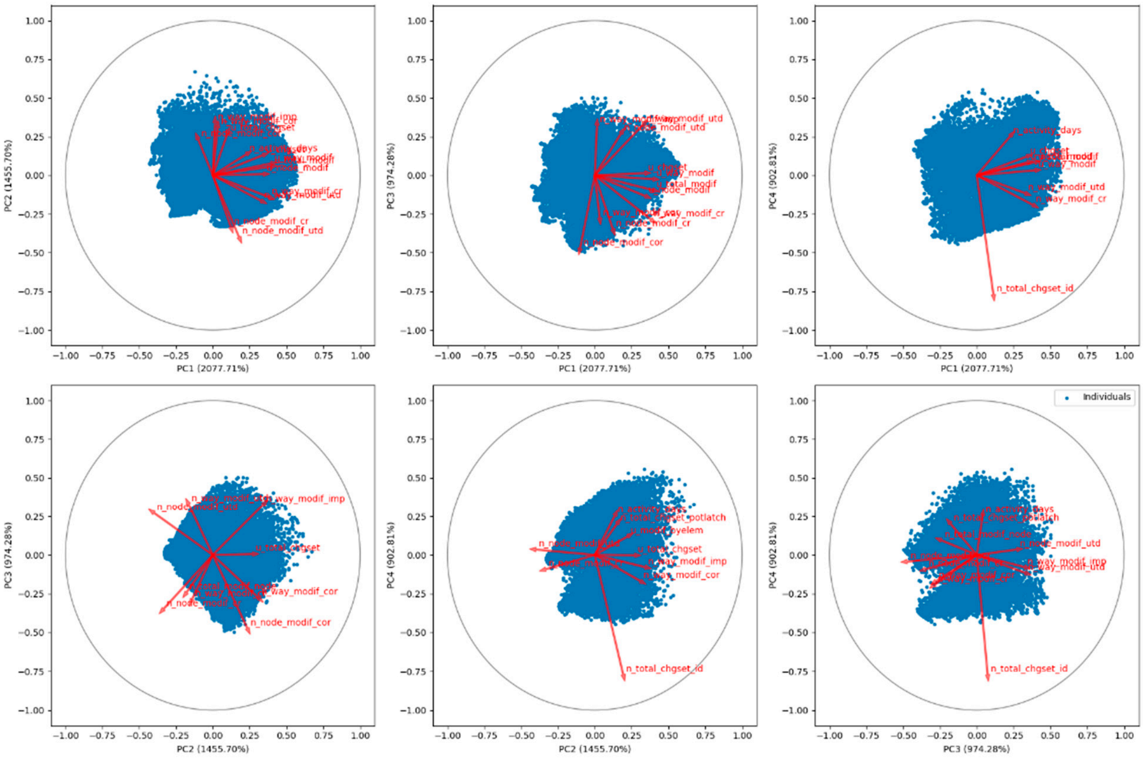


Figure A1. PCA correlation circle.

Table A2. User-based metadata feature contribution for each principal component.

OSM_features	PC1	PC2	PC3	PC4	PC5	PC6	PC7
lifespan	0.035520	0.037039	0.007404	0.051616	0.034914	0.010197	−0.031807
inscription_days	−0.021073	0.183955	−0.103551	0.066396	0.040402	−0.007690	0.012931
activity_days	0.227928	0.136391	0.039658	0.254132	0.206165	−0.072698	−0.067557
Upd_changeset	0.339637	0.130418	0.014266	0.121798	0.172972	−0.129518	0.059517
mean_Changeset	0.000083	0.000724	−0.000375	0.000894	−0.000237	0.000105	−0.000019

Table A3. Individual positioning of users on principal components.

User_id	PC1	PC2	PC3	PC4	PC5	PC6	PC7
3086	−0.180893	0.522100	−1.273169	−0.218349	−0.570122	0.279113	−0.285546
2973	−1.074610	0.205696	−0.791270	0.233733	0.224659	0.454524	0.084166
3744	0.932690	0.558 339	−1.026775	0.371771	−0.123918	−0.235809	−0.320904
30047	0.440289	0.534950	−1.285919	0.091572	−0.731935	0.069069	−0.250456
1399	−0.672546	0.310191	−0.847590	0.408293	0.332131	0.331600	0.144158

The user with ID 3086 exhibited a high value for PC2, which implied that the user was an old user with more way and node modifications. Likewise, the user with ID 2973 exhibited a high value for PC6, implying that this was a foreign user impacted by relation modification. Similarly, the users with IDs 3744, 30047, and 1399 had high values for PC1, PC2, and PC4, respectively, portraying the characteristics of those principle components.

References

1. Syaifudin, Y.W.; Puspitasari, D.; Ariyanto, Y.; Ariyanto, R. The design of road conditions mapping system by utilizing OpenStreetMap spatial data. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Harbin, China, 5–7 July 2019; Volume 523, p. 12045.
2. Boucher, C.; Noyer, J.-C. A General Framework for 3-D Parameters Estimation of Roads Using GPS, OSM and DEM Data. *Sensors* **2017**, *18*, 41. [CrossRef]
3. Luo, N.; Wan, T.; Hao, H.; Lu, Q. Fusing High-Spatial-Resolution Remotely Sensed Imagery and OpenStreetMap Data for Land Cover Classification Over Urban Areas. *Remote. Sens.* **2019**, *11*, 88. [CrossRef]
4. Viana, C.M.; Encalada, L.; Rocha, J. The value of OpenStreetMap Historical Contributions as a Source of Sampling Data for Multi-temporal Land Use/Cover Maps. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 116. [CrossRef]
5. Auer, M.; Eckle, M.; Fendrich, S.; Griesbaum, L.; Kowatsch, F.; Marx, S.; Raifer, M.; Schott, M.; Troilo, R.; Zipf, A. Towards Using the Potential of OpenStreetMap History for Disaster Activation Monitoring. In Proceedings of the 15th ISCRAM Conference, Rochester, NY, USA, 20–23 May 2018; pp. 317–325.
6. Zhang, L.; Pfoser, D. Using OpenStreetMap point-of-interest data to model urban change—A feasibility study. *PLoS ONE* **2019**, *14*, e0212606. [CrossRef] [PubMed]
7. Schiefelbein, J.; Rudnick, J.; Scholl, A.; Remmen, P.; Fuchs, M.; Müller, D. Automated urban energy system modeling and thermal building simulation based on OpenStreetMap data sets. *Build. Environ.* **2019**, *149*, 630–639. [CrossRef]
8. Hadimlioglu, I.A.; King, S.A. City Maker: Reconstruction of Cities from OpenStreetMap Data for Environmental Visualization and Simulations. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 298. [CrossRef]
9. Forghani, M.; Delavar, M.R. A Quality Study of the OpenStreetMap Dataset for Tehran. *ISPRS Int. J. Geo-Inf.* **2014**, *3*, 750–763. [CrossRef]
10. Maué, P. Reputation as tool to ensure validity of VGI. In Proceedings of the VGI Specialist Meeting, Santa Barbara, CA, USA, 13–14 December 2007.
11. Neis, P.; Zipf, A. Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS Int. J. Geo-Inf.* **2012**, *1*, 146–165. [CrossRef]
12. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. [CrossRef]
13. Muttaqien, B.I.; Ostermann, F.O.; Lemmens, R.L.G. Modeling aggregated proficiency level of user contributions to assess the credibility of OpenStreetMap features. *Trans. GIS* **2018**, *22*, 823–841. [CrossRef]
14. Begin, D.; Devillers, R.; Roche, S. Assessing volunteered geographic information (vgi) quality based on contributors' mapping behaviours. *ISPRS—Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2013**, *W1*, 149–154. [CrossRef]
15. Touya, G.; Antoniou, V.; Olteanu-Raimond, A.-M.; Van Damme, M.-D. Assessing Crowdsourced POI Quality: Combining Methods Based on Reference Data, History, and Spatial Relations. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 80. [CrossRef]
16. Rehrl, K.; Gröchenig, S. A Framework for Data-Centric Analysis of Mapping Activity in the Context of Volunteered Geographic Information. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 37. [CrossRef]
17. Yang, A.; Fan, H.; Jing, N.; Sun, Y.; Zipf, A. Temporal Analysis on Contribution Inequality in OpenStreetMap: A Comparative Study for Four Countries. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 5. [CrossRef]
18. Dorn, H.; Törnros, T.; Zipf, A. Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1657–1671. [CrossRef]
19. Barron, C.; Neis, P.; Zipf, A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Trans. GIS* **2014**, *18*, 877–895. [CrossRef]
20. Antoniou, V.; Skopeliti, A. Measures and indicators of VGI quality: An overview. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *II-3/W5*, 345–351. [CrossRef]
21. Esmaeili, R.; Naseri, F.; Esmaili, A. Quality Assessment of Volunteered Geographic Information. *Am. J. Geogr. Inf. Syst.* **2013**, *2*, 19–26.

22. Kalantari, M.; Rajabifard, A.; Olfat, H.; Williamson, I. Geospatial Metadata 2.0—An approach for Volunteered Geographic Information. *Comput. Environ. Urban Syst.* **2014**, *48*, 35–48. [CrossRef]
23. Foody, G.M.; See, L.; Fritz, S.; van der Velde, M.; Perger, C.; Schill, C.; Boyd, D.S.; Comber, A. Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality. *Cartogr. J.* **2015**, *52*, 336–344. [CrossRef]
24. Anderson, J.; Soden, R.; Keegan, B.; Palen, L.; Anderson, K.M. The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters. *Int. J. Hum. Comput. Interact.* **2018**, *34*, 295–310. [CrossRef]
25. Ding, C.; He, X. K-means clustering via principal component analysis. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 29.
26. Solovyov, A.; Lipkin, W.I. Centroid based clustering of high throughput sequencing reads based on n-mer counts. *BMC Bioinform.* **2013**, *14*, 268. [CrossRef] [PubMed]
27. Amershi, S.; Conati, C.C. Combining Unsupervised and Supervised Classification to Build User Models for Exploratory. *JEDM J. Educ. Data Min.* **2009**, *1*, 1–54.
28. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Cambridge, MA, USA, 2016.
29. Yang, A.; Fan, H.; Jing, N. Amateur or Professional: Assessing the Proficiency level of Major Contributors in OpenStreetMap Based on Contributing Behaviors. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 21. [CrossRef]