# Analysis of User Behavior and Difficulty in Labeling Polygons of a Segmented Image in a Citizen Science Project


MARINALVA DIAS SOARES[1]
RAFAEL SANTOS[2]
NANDAMUDI L. VIJAYKUMAR[2]
LUCIANO VIEIRA DUTRA[1]

INPE - National Institute for Space Research
P. O. Box 515 - 12245-970 - Sao Jose dos Campos - Brazil
[1] (marinalva,dutra)@dpi.inpe.br
[2] (rafael.santos,vijay)@lac.inpe.br



**Abstract.** Citizen science projects attempt to use knowledge, time or computing resources of volunteers to achieve a particular scientific objective. It fosters collaboration between scientists and the general public. Several citizen-science projects are already in execution, some of those successful. In particular, 'wisdom of the crowds' may be used to get a consensual opinion on a particular topic or object (e.g. classification or labeling purposes), by collecting and analyzing opinions of users on those topics or objects. Citizen science projects collect data that is used for its stated purpose (e.g. object classification), but we consider that all the collected data may be used to give interesting insights on the reasoning of the project's volunteers. In this paper we present a citizen-science project (volunteer labeling of imprecisely segmented image regions) and show which information can be inferred, through basic analysis, about individual and collective behavior of volunteers of the project.



**Keywords:** image labeling, citizen science, user behavior.

(Received August 09, 2010 / Accepted December 26, 2010)


## 1 Introduction

Segmentation can be defined as a process that partitions an image into regions (usually polygons), so that elements belonging to each region are similar with respect to some properties [8]. Segmentation itself is just a step towards object identification – the regions obtained through segmentation must be labeled, or identified. Polygons obtained from segmentation must receive a discrete label, usually with semantic information about it. In urban scene applications, for example, one could use labels such as *roofs*, *trees*, *streets*, *pools*, etc. Depending on the application, not all polygons are to be labeled. However, segmentation process may create a huge number of polygons and these polygons may not be properly segmented due to the imperfection inherent of the segmentation algorithms.

Humans interpret scenes using knowledge, experience and visual evidence, but image processing systems for automatic scene identification often cannot use directly the information humans can. Figures 1, 2 and 3 illustrate the process of image segmentation and labeling: a small region in an urban scene is shown in Figure 1; and its segmentation in Figure 2. In Figure 2 there are both oversegmented regions (perceptual objects divided into several regions) and undersegmented regions (regions which contains several different perceptual objects), which are practically unavoidable when using the majority of unsupervised image segmentation algorithms. In Figure 3 there are four regions which were labeled by an expert and painted with dark gray (for ceramic-tiled roofs) and light gray (for trees) over the original segmented image, which also was processed to improve its contrast for publication.

**Figure 1:** Satellite Image.

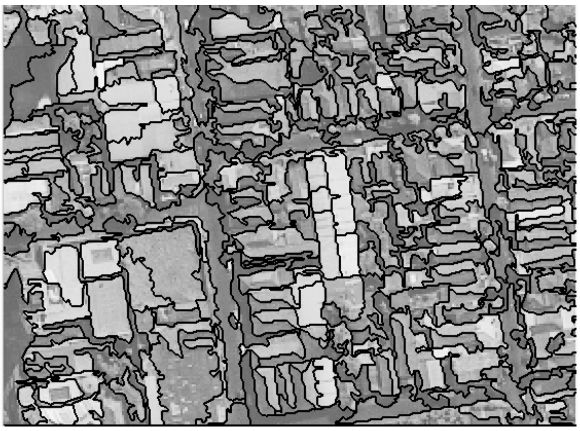

**Figure 2:** Segmentation of the image.

Labels can be assigned by a human expert, with rules, samples or conditions dependent on the object identification task itself. Manual labeling, by specialists, is a repetitive task prone to errors. On the other hand automatic labeling, in order to be successful, must incorporate the knowledge humans use. This is definitely a difficult task, as it is an error-prone process since algorithms cannot reproduce faithfully the knowledge and experience from the users.

A different approach, used in this research, to label objects is to use several different human agents, without the same expertise as the specialist already mentioned. These agents could receive a very brief, superficial training and different polygon labeling tasks to perform. The entire polygon labeling task could be performed by different volunteer users, which could complement each other's opinions, hypothetically leading to good results (since they would use human knowledge and intelligence) without the expensive work of a single expert.

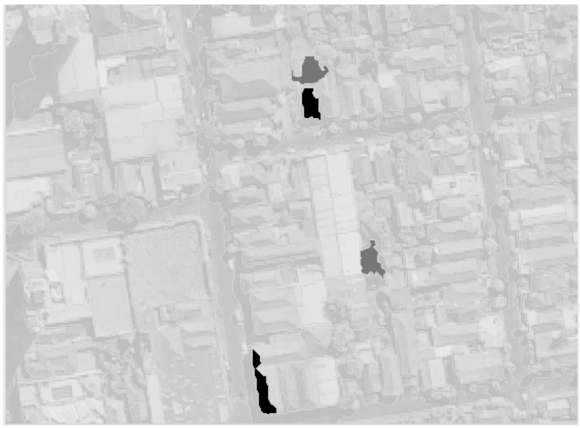

**Figure 3:** Manual labeling of the regions.

The use of communities or networks of citizens who act as participants or observers in some domain of science is often called *citizen science*. Citizen science-based approaches have been used, often with good results, in several different tasks that either could be performed by a lot of work by a human specialist or poorly by an automatic system. Citizen science is more often than not based on volunteer users – the motivation of the participants (often unpaid volunteers) has also been studied [4, 9].

Another term that has been widely used is crowdsourcing [1]. Crowdsourcing is a new and growing innovation tool which is a strategic model to attract an interested, motivated crowd of individuals capable of to solve problems create content and solutions or develop new technologies (using the intelligence and the collective knowledge and volunteers). Crowdsourcing can also be considered a process of distribution of tasks of one to many individuals. These distribution of tasks may have commercial or social purposes. Moreover, the volunteers may be paid for their services or not. On the other hand, the citizen science projects have a scientific basis and traditionally non-commercial, scientists involved in these projects consider the volunteers as learners and volunteers participate in the collection of scientific data or analysis of these data.

There are several projects that aims to promote public engagement with research, as well as with science in general. Some projects differ from each other mainly at the level of user involvement with the project. There are those in which the user participate allowing that their computational resources are used while they are idle (like SETI – Search for Extraterrestrial Intelligence [10], which goal is to detect intelligent life outside the planet

Earth); those that require users' time and work outside the home (like Cornell Lab of Ornithology [3], a nonprofit organization which studies birds and other wildlife) and those that require only that user give a little of their time accessing a site to perform some tasks (like Galaxy Zoo [6], which invites users with Internet connection to classify galaxies accordingly to their shapes and to the Hubble classification scheme).

To use citizen science, it is necessary to create means and methods to present and collect information from collaborators. For the data collected from users to be used, one needs to assess its quality, coherence, relevance, etc. Therefore, the analysis of the collected data may be more important than the data itself – for example, in an object labeling task one may not rely only on most of the opinions from the users, but on the past performance, reliability, inferred knowledge, etc. of those users. Modeling user knowledge is then of major importance when dealing with citizen science.

In this paper we consider a practical application as a case study: labeling of segments or polygons extracted from high-resolution satellite digital images. Figure 4 shows, at a glance, the processes considered for data preprocessing, acquisition and analysis required for the development of this work.

Basically, the data pre-processing (step 1) is the segmentation of a high resolution urban image. The resulting polygons and their statistical attributes are stored in a database; the specialist labels some polygons that are considered to be correct; these labels can be used to assess answers from the non-expert users, effectively allowing the evaluation of the users' ability.

At the second step (data acquisition) data are collected from users to determine which labels are used for the polygons and to model the users' knowledge and behavior. There is a single task for the data acquisition: the creation and deployment of a web-based interface that presents the tasks for the users and collect the users' decisions, which are stored in a database for further analysis. Data collection and analysis are continuous processes – with more data it is possible to perform further, more detailed analysis, and the results of the analysis can serve to identify improvements in the data collection processes. At the third step (data analysis) several different analysis tasks and scenarios can be considered. The analysis presented in this paper is related to users behavior. There are different techniques that are applied to mine web data in order to discover significant patterns about the user behavior.

In this paper we present a citizen-science project (volunteer labeling of imprecisely segmented image regions) and show which information can be inferred, about individual and collective behavior of volunteers of the project. One of the goals of this project is to gain insight into how users behave in front of objects resulting from an imprecise segmentation process. This behavior is analyzed from the point of view of the behavior of an expert in recognizing objects. Such understanding is important to point to some ideas about how to harness the contributions from volunteers to guide the automatic labeling process.

The contributions of this paper include answering the following questions: are users consistent in what they say about a specific object? may the collective opinion question the expert opinion? is the collective opinion reliable enough to direct the process of automatic labeling? is it possible to assess the quality of the labeling initially done by the expert (e.g. by verifying it against the non-expert users)? or is it possible to use the users' knowledge to identify problems on the segmentation process so it can be refined? can patterns and trends between the users be identified (e.g. identification of groups of users who tends to perform better with some labels than others)?; is it possible to evaluate temporal changes on users' performance; can the knowledge of a specific group of users be modeled (e.g. the most precise or reliable) to get information to try and automate the labeling task?

The paper is organized as follows: section 2 describes the experiment in which the analysis in this paper were based; section 3 describes the analysis and results obtained and section 4 describes the future work.

## 2 Experiment

The experiment presented here seeks to obtain more comprehensive understanding of what the volunteer users say about complex and imprecise objects resulting from segmentation process. For this, the experiment presented here uses citizen science to identify objects extracted from a satellite image of Sao Jose dos Campos city in Brazil.

The image size is 900x900 pixels and was segmented with the Spring software [2] with the growing region algorithm in such a way to create an oversegmented image with 2430 polygons. To enable the data collection process described in the previous section, a website was developed in which volunteer users are presented with polygons and a list of options for labeling those polygons. This site is available at `http://www.lac.inpe.br/UrbanZoo`. Figures 5 and 6 show the satellite image with an urban scene chosen for the case study and the site cited above.

Based on the chosen image, we established the following hierarchy of classes (labels) for identification

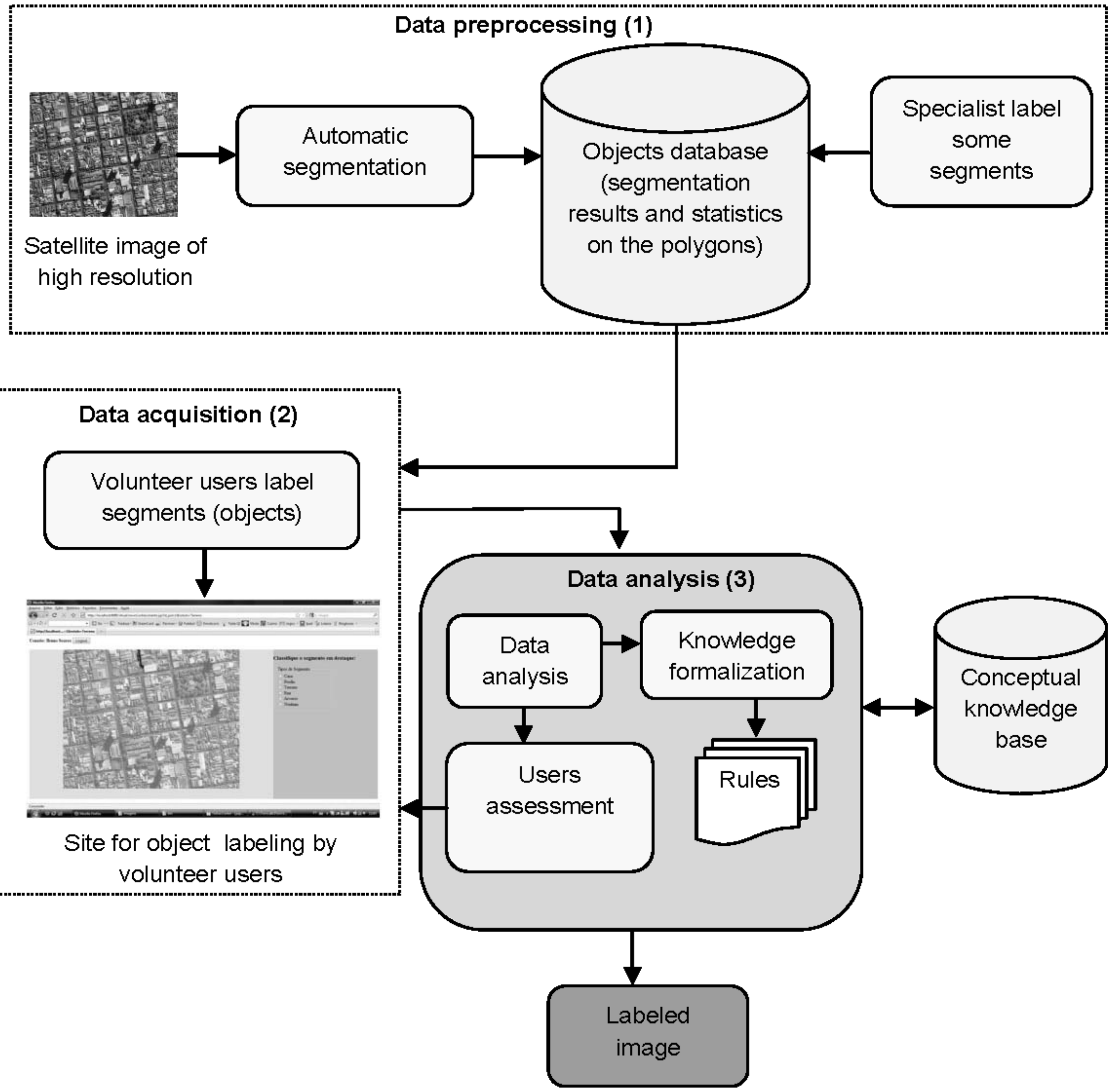


**Figure 4:** Data processing, acquisition and analysis tasks.

of targets of interest: *generic roof*, *ceramic roof*, *tin roof*, *cement asbestos roof*, *tree*, *street*, *swimming pool*, *shadow*, *field* (any kind of vegetation other than trees), *bare soil*, *water* and *mixture* (for when the polygon is composed of different classes of objects). These classes are shown to the user along with the options *none of the above* (in case the user knows the polygon class, but its class is not shown in list of options) and *unknown* (when the user does not know which is the correct class for the polygon).

Users label the polygons by accessing the Urban-Zoo site (which requires registering for identification purposes) and selecting one of the classes for a particular polygon. This is repeated until the user decides to finish his/her interaction with the site. Each interaction (presentation of a polygon and recording of the choice of label by the user) is stored.

This experiment began on April 26, 2010. Polygons were presented as follows: in the first phase of this study, polygons were shown randomly for each user. After the first week of running the site, 3000 polygons with few repetitions (very few objects were labeled more than once) were labeled by 43 users. In the second phase users were presented with a list of 13 polygons


**Figure 6:** Web interface to record users' decisions.

**Figure 5:** Satellite image chosen for the case study.

which were specifically chosen and labeled by an expert user (directly on the database, instead of using the site). The volunteer users were not informed that the polygons were selected in a non-random way. This strategy served two purposes: first, to obtain a significant number of votes from these "known" polygons which class was already known, so a measure of agreement with the expert user's opinion can be calculated; and second, to assess the reliability of each volunteer user (akin to a multiple-choice test).

Two weeks after the new strategy of presenting objects to users, there were 6900 labels with a total of 56 users on the database. Some polygons were labeled up to 35 times and most of the polygons were labeled more than once. In the third phase users were presented with a sequence of 20 polygons, of all the classes. This phase had two goals: first, to ensure that the users label all classes of polygons for further analysis of their ability to recognize polygons of a given class; second, to verify whether non-randomness of the classes influences the user behavior. One week after the beginning of this phase there were 7530 labels with a total of 63 users.

In the fourth phase, users were presented with a sequence of 25 polygons in complex and simple shape. The purpose here was to identify the influence of the shape of the object in the view of the user. Until this moment, there were 69 users and approximately 9500 labeled polygons. During this phase, one domain expert has visited the site and labeled 484 polygons. These polygons together with those that were labeled by another specialist at the beginning of this experiment were used to evaluate the reliability and accuracy level of the user.

## 3 Results and Discussion

In this section, two basic analyses are used to try and characterize the behavior of the users while labeling the polygons.

### 3.1 Characterization of Users According to Their Involvement and Behavior

Is there any relation between the number of interactions of an user with the system (number of labeling tasks) and this user's pace on the collaboration? Can we extract different profiles of users engagement with the project? It is important to characterize the users' involvement with the project so we can use different strategies to present the tasks to the users in order to maximize the return on their volunteer efforts. Other citizen science projects may also use this information to better adapt the task presentation to the users.

Figure 7 shows, for the first 51 days of the experiment , how many labeling each registered user labeled. The X axis represents the dates of April 26 to June 16 and each line on the axis Y represents an user. We can see that there are several "bursts" of collaboration just after the users accessed the site for the first time, which is expected.

Since there are few data available for this analysis (there are relatively few users and less than 60 days of collected data) we can verify the correlation of the frequency of the user and the number of labelings with a simple visualization tool. The data was plotted in a parallel coordinates plot [7], shown in Figure 8. From this plot we can see that there are several *one-time users*, which did almost all of their labeling in a single day, totaling less than one hundred labelings. Most of the users who did a thousand or more labelings did that on a period of more than a week.

In order to characterize the users' pace we devised a simple metric based on how many days it took for the user to label all his/her polygons. For each user we calculated the number of days the user took to label 25%, 50%, 75% and 100% of the polygons presented to that user. Our reasoning is that for few users may be *short-term* volunteers, accessing the site just some days after registration, doing some labeling and then abandoning the project. Users who registered for less than five days at the time the analysis started were not considered for this study.

To see whether the involvement pattern is related to the total number of polygons the user labeled, they were divided into five arbitrary classes accordingly to the total number of tasks they've performed: *c100*, for users who labeled a hundred or less polygons; *c250*, for users who labeled between 100 and 250 polygons; *c500*, for users who labeled between 250 and 500 polygons; *c1000*, for users who labeled between 500 and a thousand polygons and *cplus*, for the users who labeled more than a thousand polygons.

Each line on the plot represents a user; the first four vertical axes represent the number of days to reach each mark (q_25, q_50, q_75 and q_100) and the last vertical axis represents the class for that user. This kind of plot allows the visualization of the progress on the number of collaborations of each user through time, considering the total amount of collaborations. For example, users that use the web site to do a few labellings and don't return later will be represented by flat lines on the bottom of the plot. We can see in Figure 8 that most users that did less than 100 collaborations did them in the initial week.

From the plot on Figure 8 we can see that there is a class of users that are not motivated to access the site and label the polygons more than once. While this is expected in a volunteer citizen science project, one must consider a way to engage these users for more than a single day.

### 3.2 Confusion or Difficulty in Labeling Classes

For this analysis we verify the degree of confusion of the users when labeling certain polygons. More than 400 of the 2430 polygons were labeled by an user with experience in identifiying targets in urban satellite images and are considered as "ground truth" on this experiment. This expert user was exposed to the same interface as the other users, labeling polygons shown in a random order. For each polygon we calculated the entropy of the labels assigned to the polygon by the users. For each polygon $p$ we calculate

$$e_p = -\sum_{c \in C} v_c \log_2(v_c)$$

where $C$ is the set of classes or labels, $c$ is one of those classes and $v_c$ is the percentage of votes for label $c$ in that polygon. This simple metric tends to zero when all votes go to a single label and $log_2(|C|)$ when the votes are equally distributed among the labels – higher values indicate disagreement on the polygons' label.

This metric must be used carefully, since, by definition, a polygon with a single vote will have $e_p$ equal to zero, which indicates total agreement on its label but which is not enough to determine the real label of that polygon. Label assignment entropy on a polygon may indicate a hard-to-label polygon; when we consider the entropy for all polygons we may identify classes that are hard to label.

Since there are few classes and since we considered only polygons which were labeled five or more times and which were labeled at least once by the expert users, this analysis can be performed with a simple query to

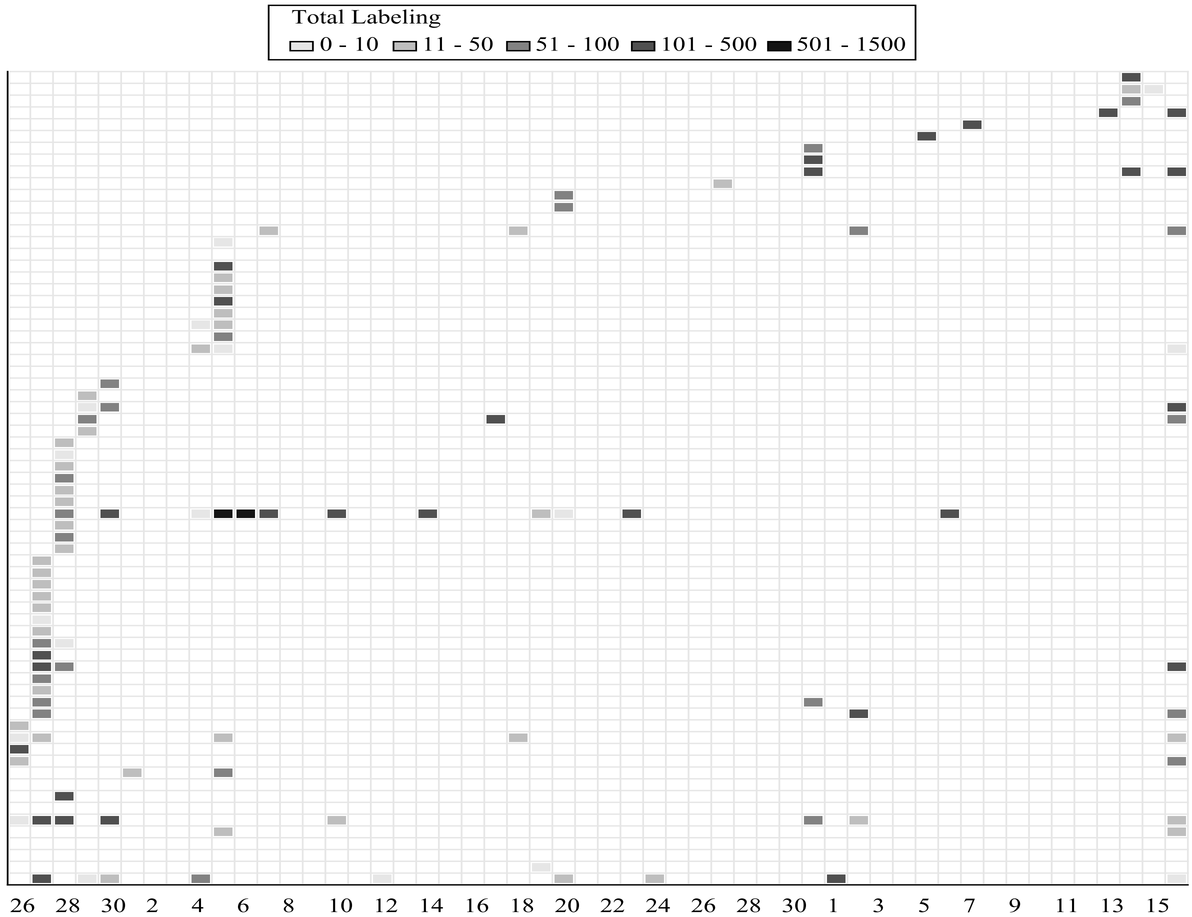


**Figure 7:** User daily interaction.

the collected data and with a simple box plot. First, for each polygon which is both labeled by the expert at least once and labeled by the users at least five times we calculate the entropy based on the users' decisions; and from the entropy values for each class we can create the plot shown in Figure 9.

Figure 9 shows a box plot for the entropy distribution considering all labeled polygons for each class, but only polygons which were labeled at least five times. Box plot shows (in axis Y) minimum, maximum, median values and the first and third quartile for distributions of each class. The axis X shows the classes. We can see that only one class has relatively low median values for its entropy: namely, street.

Other classes, specifically, open field and bare soil present wider spectral on the labeling entropy of their polygons – this is to be expected since there are visual similarities between polygons of those and other classes, and some users manifested (by e-mail) their doubts about the difference between those classes. Yet other classes, namely, unknown, mixture and generic roof (gen. roof) had relatively high entropy values for their polygons since they are, in a way, generic – it is expected some confusion from the users on choosing specific labels for those classes. Considering our expectations, the distribution of entropy values for the class water was a surprise, since it should be easy to identify polygons with this class. One possible reason is that water bodies in the satellite image are parts of a brown-water river, which appearance may confuse the unexperienced user.

Since the collected data is not associated with users' personal data (e.g. names, professions, degrees, etc.) we cannot infer relations between the users' personal or professional characteristics and their performance, which would be very interesting and could open further possibilities for customization of tasks for some users.

Most of the conclusions shown in this section were

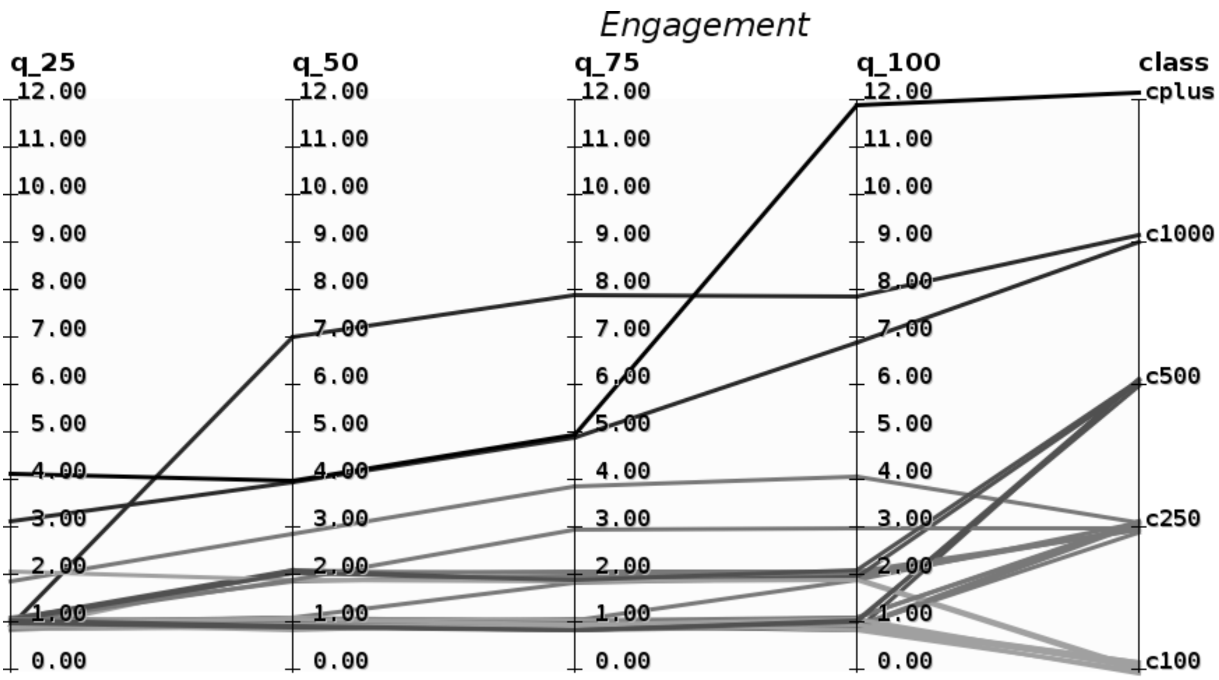


**Figure 8:** Pace of user engagement and volume collaboration classes.

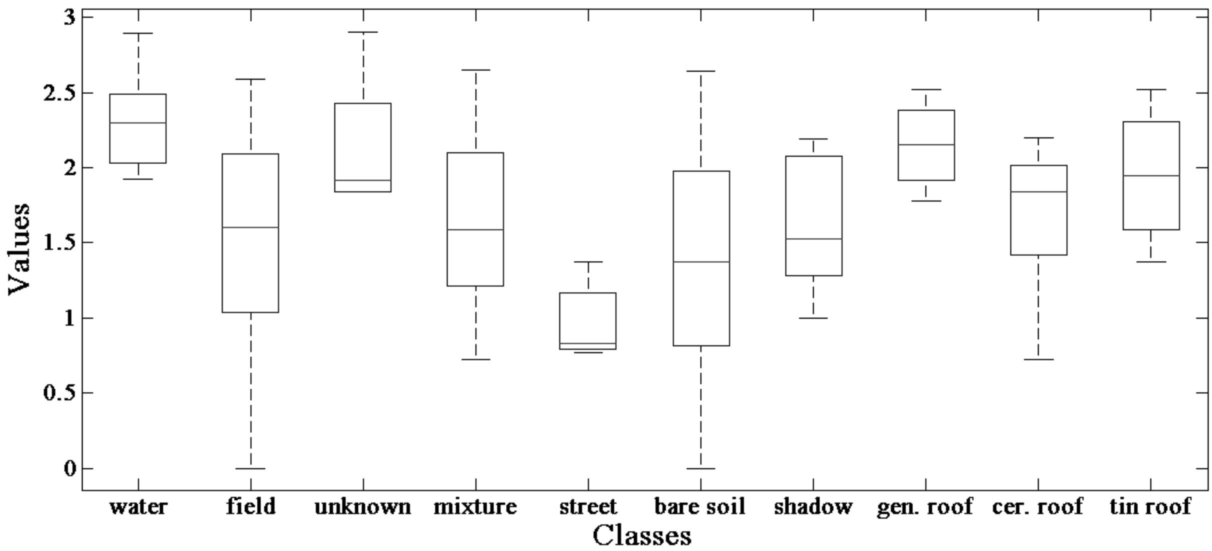


**Figure 9:** Entropy Distribution for the Labeled Polygons.

not obtained through the application of complex or powerful data mining or knowledge discovery algorithms – the questions posed in this section could be answered with simple queries on a transformed set of the original data collected from the users' interactions with the citizen science project site.

## 4 Comments and Future Work

In this paper we presented the general idea of UrbanZoo, our citizen science project in which we used untrained users to label polygons obtained from an imprecisely segmented satellite image. We've shown that besides the users' decisions there is more information that can be used to characterize the users and five insights on the whole process. Two basic analyses, which were done with simple graphical and statistical tools, demonstrated some of these insights.

One interesting analysis is the evaluation of the confusion between labelings, expressed by the entropy calculated for each polygon (see section 3.2). This metric can be considered a measure for "hardness" of labeling which could be compared with spectral and geometric metrics on the polygons [11, 5] to see whether shape complexity is directly related with user confusion on labeling the polygons. Since this is an on-going work, more analyses, using more data and more complex algorithms, will be performed regularly.

## References

[1] Brabham, D. C. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, 14(1):75–90, 2008.

[2] Câmara, G., Souza, R. C. M., Freitas, U. M., and Garrido, J. SPRING: Integrating remote sensing and GIS by object-oriented data modelling. *Computers & Graphics*, 20(3):395–403, 1996.

[3] Cornell Lab of Ornithology. Cornell Lab of Ornithology: Exploring and Conserving Nature. `http://www.birds.cornell.edu`, 2010.

[4] Droedge, S. Just because you paid them does not mean their data are better. `http://www.birds.cornell.edu/citscitoolkit/conference/proceeding-pdfs/`, March 2010.

[5] Frohn, R. C. and Hao, Y. Landscape metric performance in analyzing two decades of deforestation in the amazon basin of rondonia, brazil. *Remote Sensing of Environment*, 100(2):237–251, 2006.

[6] GalaxyZoo. Galaxy Zoo. `http://www.galaxyzoo.org/`, 2010.

[7] Inselberg, A. *Parallel Coordinates – Visual Multidimensional Geometry and Its Applications* . Springer, 2009.

[8] Jung, C. Unsupervised multiscale segmentation of color images. *Pattern Recognition Letters*, 28(4):523–533, 2007.

[9] Raddick, M. J., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., and Vandenberg, J. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. `http://arxiv.org/abs/0909.2925`, March 2010.

[10] SETI@home. What is SETI@home? `http://setiathome.berkeley.edu/`, 2010.

[11] Silva, M. P. S. *Mineração de Padrões de Mudança em Imagens de Sensoriamento Remoto*. PhD thesis, Instituto Nacional de Pesquisas Espaciais, 2006.