# Neural network based visualization of collaborations in a citizen science project

Alessandra M. M. Morais[a], Rafael D. C. Santos[a], M Jordan Raddick[b]

[a]National Institute for Space Research, Av dos Astronautas, 1758,
CEP 12227-010, São José dos Campos, Brazil;
[b]The Johns Hopkins University, Baltimore, Maryland, USA

## ABSTRACT

Citizen science projects are those in which volunteers are asked to collaborate in scientific projects, usually by volunteering idle computer time for distributed data processing efforts or by actively labeling or classifying information – shapes of galaxies, whale sounds, historical records are all examples of citizen science projects in which users access a data collecting system to label or classify images and sounds.

In order to be successful, a citizen science project must captivate users and keep them interested on the project and on the science behind it, increasing therefore the time the users spend collaborating with the project. Understanding behavior of citizen scientists and their interaction with the data collection systems may help increase the involvement of the users, categorize them accordingly to different parameters, facilitate their collaboration with the systems, design better user interfaces, and allow better planning and deployment of similar projects and systems.

Users behavior can be actively monitored or derived from their interaction with the data collection systems. Records of the interactions can be analyzed using visualization techniques to identify patterns and outliers. In this paper we present some results on the visualization of more than 80 million interactions of almost 150 thousand users with the Galaxy Zoo I citizen science project. Visualization of the attributes extracted from their behaviors was done with a clustering neural network (the Self-Organizing Map) and a selection of icon- and pixel-based techniques. These techniques allows the visual identification of groups of similar behavior in several different ways.

**Keywords:** Kohonen Self-organizing Maps, Visualization, Citizen Science

## 1. INTRODUCTION

Some scientific projects require the analysis or collection of large data sets, and for some, the analysis or collection itself is a task beyond the capabilities of the team assigned to that project. An approach to make these projects feasible is the delegation of some aspects of it to volunteers[1] – members of the public who participate as assistants in scientific studies.[2] The assistance is usually performed by collecting and organizing data or by labeling and identifying features in already collected data. Such volunteers are most often not paid for their assistance, nor are even necessarily scientists. This approach is called citizen science, and the volunteers are citizen scientists.

Practiced since at least the 1700s,[3] recent citizen science projects use web resources to attract and involve volunteers that will collect, access, review and analyse the data.

Data quality is one of the main issue at these projects. Scientific endeavors require data of high quality, therefore, the fact that citizen science projects intentionally place responsibility for creating data into the hands of non-experts[4] requires that the science team be prepared to scrutinize the data carefully discarding suspect or unreliable data.[2] Mechanisms to enhance the quality and trust of citizen science data can be found at literature,[5–7] as well as, some researches which conclude that data produced by volunteers are as good as data produced by professional scientists.[3,8]

Send correspondence to R. Santos (rafael.santos@inpe.br)

Next-Generation Analyst II, edited by Barbara D. Broome, David L. Hall, James Llinas,
Proc. of SPIE Vol. 9122, 912207 · © 2014 SPIE · CCC code: 0277-786X/14/$18
doi: 10.1117/12.2050183

Researchers that use citizen scientists' collaborations are concerned to find ways to persuade volunteers to get and stay involved with the project until their objective is reached and reducing collaborators' errors.[9]

Visualizing the interaction of citizen scientists through their collaborations with a citizen science project can be used to shed light to some concerns of scientific community engaged with this approach, helping understand how the collaborations are done, understanding the reasons that led volunteers to abandon the project, categorize volunteers accordingly to some parameters with the aim to increase data quality by prioritizing or penalizing some collaborations, and allowing better planning and deployment of similar projects and systems.[9]

In this paper we present an approach to visualize some features of interactions of citizen scientists with a citizen science project using Kohonen Self-organizing Maps. Our approach was applied to a log of activities with more than 80.000.000 collaborations produced by almost 150.000 volunteers of Galaxy Zoo I project,[8, 10] an successful example of engagement success among scientists and volunteers to produce data with quality, which asked volunteers to classify galaxies' images.

This paper is organized as follow: Section 2 will present the Kohonen's Self-Organizing Map and it characteristic which made it a good visualization tool. Section 3 will describe the proposed approach, the data set used and the results, and finally section 4 will present conclusions and directions for future work.

## 2. THE KOHONEN SELF-ORGANIZING MAPS

The Self-Organizing Maps (SOM)[11] is a neural network algorithm based on unsupervised learning. It implements an orderly mapping or projection of high-dimensional data into a regular low-dimensional grid, while preserving the most important topological and metric relationships of the original data.

The SOM consist of $M$ units (neurons) located on a regular low-dimensional grid which represent the map. The grid is usually two-dimensional, particularly when the objective is to use the SOM for data visualization. Each neuronal unit $m$ is associated with a prototype or reference vector $v^m$ of dimension $N$ located at $r_m$ grid position.

The network is trained through of repeatedly presenting each input data vector $x^q$ to the network, finding the reference vector in the SOM that is closest to the input data vector and adjusting this reference vector (called the winning vector or best matching unit, BMU) so it will be even close to the input data vector. Neurons on the SOM that are topologically close to the BMU (limited to an update radius) will also have their reference vectors modified but weighted by the distance to the BMU measured on the grid through a neighborhood function $h$ which size will decrease during the training. The change to the reference vector of the BMU will also be weighted by a learning rate $\eta$, which is usually a value that decreases during the training.

The basic SOM algorithm is shown in Algorithm 1 (adapted from Looney[12]). Prior to the execution of the algorithm, its grid dimensions and lattice type (rectangular, hexagonal or even irregular) must be chosen, as well as a learning rate $\eta$, a function to weight the distance of the neurons to the BMU to determine the size of the update radius and other constraints that may be used to stop the training process.

Two interesting characteristics of this algorithm that are relevant to data visualization are *quantization* and *projection*. Quantization refers to the grouping of similar data vectors in one neuron, while projection refers to the organization of the data into neurons in lower dimensions that preserve the distances (or the order of distances) between the originally high-dimensional data.[13] The SOM algorithm has proved to be especially good at maintain the trustworthiness of the projection, what mean that if two data samples are close to each other in the grid, they are more likely to be close in the original high-dimensional space data.[13]

### 2.1 The Self-Organizing Map as Visualization Tool

Visualization is usually described as the mapping of data to a visual representation. A good visual representation enables the analyzing of several items of information at the same time and maximize data comprehension.[14] Tufte[15] defines graphical excellence as one which shows complex ideas with clarity, precision and efficiency, giving to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space, telling the true and avoiding misinterpretation about data – these principles must be considered when devising visualization tools and techniques.

**Algorithm 1** Basic Self-Organizing Map

```
Require: Grid of reference vectors $v^m$, input data set $x^q$, learning rate $\eta$
Randomize the order of $\{x^q\}$                              ▷ Randomize Data Vectors
q ← 1;
for m = 1 to M do
    for n = 1 to N do
        $v_n^m \leftarrow$ random(0,1)                        ▷ Draw Random Weights
    end for
end for
while train do
    Draw exemplar $x^q$ from the exemplar set;
    for m = 1 to M do
        Compute distance $D_{qm}$                             ▷ Distance between $x^q$ and each $v^m$
        Find $v^{m*}$ with minimum distance $D_{qm}$          ▷ Find BMU
    end for
    Update neurons $m$ mapped by the neighborhood function $h$ with
        $v^m \leftarrow v^m + h\eta(x^q - v^{m*})$            ▷ Reinforce/Reward neurons close to the BMU
    q ← q+1
    if $q > Q$ then
        $q \leftarrow q - Q$                                  ▷ Repeat polling the data vectors if needed
    end if
    Decrease $h$                                              ▷ May be done after $n$ exemplars are presented
    Decrease $\eta$                                           ▷ May be done after $n$ exemplars are presented
end while
```

The challenges and issues of a good mapping are significant when the data to be visually represented is complex. This paper consider as being complex data set those whose combination of dimensionality (attributes) and size (amount of data items which comprise the data set) makes the mapping a hard task to produce a good visual representation through well-established visualization techniques. Examples of these techniques and its limitations are described by Oliveira and Levkowitz.[16]

The SOM algorithm, thanks its characteristic of quantization and projection, have been used as an effective tool to visualization of data.[13, 17–20] Methods for visualizing data using SOM can be grouped in three categories based on the purpose of the visualization:[13] get an idea of the overall data shape and detect possible cluster structure, analyze the prototype vectors and analysis of new data samples for classification and novelty detection purposes. Some of these methods are presented by Vesanto.[13]

## 3. VISUALIZING VOLUNTEERS' INTERACTIONS

Data from interactions between humans and on-line systems (logs) can be considered as complex datasets since they usually contain a large amount of entries and even few attributes for each entry are enough to make its visualization a hard task. In order to visualize the interaction of citizen scientists with the Galaxy Zoo website, we propose the use of the Self-Organizing Map with visual representations of features extracted from the logs. Entries on the log will be tabulated in order to collect data about a study object (e.g. a user or a session) and a visual representation will be created from this data. The data will be processed by a SOM network and the result will be graphically represented by assigning the visual representation of the features as a grid corresponding to the grid of the SOM's neurons.

It is expected that the general appearance of the distribution of the graphical components on the SOM matrix will bring interesting visual clues about the nature of the data. Due to the SOM properties the visual representation obtained will obey some of the principles posed by Tufte,[15] particularly, show the data, induce the viewer to think about the substance rather than about the methodology, avoid distorting what data has to say, present many data in a small space, make large data sets coherent, encourage the eye to compare different pieces of data and reveal the data at several levels of detail.

### 3.1 Data used for visualization – The Galaxy Zoo project

For most of the twentieth century, morphological catalogues of galaxies were compiled by individuals or small teams of astronomers. The modern surveys, like the Sloan Digital Survey (SDSS) which contain data from millions of galaxies, make this approach impractical. Started in June 2007, the Galaxy Zoo I project has the aim of delegate the classification of galaxies to volunteers,[10] becoming a good example of engagement success among scientists and volunteers to produce data with quality.

A website was designed to recruit citizen scientists and collect the data provided by them. Visitors to the site were asked to register and read a brief tutorial, and after that were submitted to a simple multiple-choice test with selected galaxies from the SDSS which were previously classified by astronomers. Visitors who correctly classified 11 or more of the 15 galaxies on the test were allowed to become a citizen scientist.[8]

Through this website the citizen scientists were presented with images showing galaxies and had to answer simple questions about the galaxies' shapes (ellipse, spiral, merge or unknown) and orientation for the elliptical ones (clockwise or counterclockwise). After clicking on the option they believe was the most appropriate, a new galaxy image was automatically displayed and once again the system asked for a new classification. For each classification, the system stored a log of activities with the volunteer identification, the galaxy identification, the timestamps and the classification chosen by the volunteer.

This paper use the log of activities with the volunteers identification anonymized, precluding their identification directly or indirectly. The log cover the period from the Galaxy Zoo site launch in July 8, 2007 until July 7, 2012. During this interval 146,669 volunteers collaborated with the more than 80.000.000 classifications. Some characteristics about the raw data are shown in Tables 1 and 2. Table 1 shows some arbitrary intervals of the number of classifications a volunteer did on the site, showing, for each interval category, the number of volunteers in that interval and the percentage of the total classifications on this interval.

Table 1: Interval of Classifications per Volunteer

| Interval | 1−10 | 10−100 | 100−1,000 | 1000−10,000 | 10,000−100,000 | 100,000−1,000,000 | >1,000,000 |
|---|---|---|---|---|---|---|---|
| Volunteers | 15.880% | 45.350% | 29.740% | 8.315% | 0.688% | 0.026% | 0.001% |
| Classifications | 0.128% | 3.413% | 17.733% | 40.883% | 26.676% | 9.628% | 1.539% |

Table 2 shows arbitrary intervals of number of days the volunteers had recorded interactions with the website and the percentage of total classifications done by users in each interval category.

Table 2: Number of days as Volunteer

| Total of days | 1 | 1-7 | 7-15 | 15-30 | 30-90 | 90-180 | 180-365 | >365 |
|---|---|---|---|---|---|---|---|---|
| Volunteers | 64.161% | 11.309% | 3.563% | 4.377% | 4.436% | 5.146% | 3.704% | 3.303% |
| Classifications | 9.041% | 7.303% | 4.828% | 7.158% | 11.963% | 14.750% | 18.221% | 26.735% |

Since volunteers of the project joined it at different moments of the project's running, in order to avoid bias towards long-time users when extracting features from their collaboration records, we define a threshold of 600 days for collaborations for all volunteers. This means that for each volunteer we considered only the classifications performed during the first 600 consecutive days computed from the day of the first recorded classification. By using this threshold we discarded 3,69% of total classifications, and 4,90% of the total number of volunteers.

### 3.2 Features Selection

The Self-Organizing Map, as described at section 2, needs a collection of input data vectors to its training. This input data can be any numerical data vector, and we explore this feature by selecting different input data vectors, corresponding to features of the different objects being measured, to achieve different visualization results.

Our last study[9] to investigate possible behavior profiles considered 7 attributes extracted from the users' interactions (i.e. galaxy classifications) with the website. The attributes used are based on the following metrics:

- Participation range in days ($p_{u^i}$): the number of days counted from the first day volunteer $u^i$ interacted with the website until the last recorded interaction of $u^i$, so $p_{u^i} \leq 600$;
- Participation count in days ($d_{u^i}$): the number of days for which there were recorded interactions of $u^i$ with the website;
- Maximum classification ($maximum_{u^i}$): the maximum number of classifications done in one single day by volunteer $u^i$;
- Total classifications ($total_{u^i}$): the number of classifications done by $u^i$ during the classification period.

The 7 attributes are computed as follows:

$$a_1 = \frac{p_{u^i}}{600}$$

Attribute $a_1$ is an absolute measure of how much the volunteer $u^i$ was interested in return to contribute to the project. Values close to zero indicates the user joined and left shortly thereafter, while values close to one indicate possible activity during all or almost all the 600 days.

$$a_2 = \frac{d_{u^i}}{p_{u^i}}$$

Attribute $a_2$ is a relative measure of frequency of access to the website by $u^i$. Values close to one means that the volunteer accessed the website every day during his/her collaboration period.

$$a_3 = \frac{d_{u^i}}{600}$$

Attribute $a_3$ complements attributes 1 and 2, measuring the frequency in which $u^i$ accessed the website.

$$a_4 = \frac{maximum_{u^i}}{total_{u^i}}$$

Attribute $a_4$ is a measure of the spread of the classification effort by the days in which $u^i$ contributed. Values close to one indicate volunteers who did the almost all classifications in a single day.

$$a_5 = \frac{total_{u^i}}{average}$$

Attribute $a_5$ indirectly measure how close the volunteer $u^i$ is to the average of classifications done by all the other volunteers.

$$a_6 = d_{u^i}$$

Attribute $a_6$ is the unaltered value of $d_{u^i}$. Reasons to use attributes that are redundant or derivable from others are given below.

$$a_7 = \log_{10}(total_{u^i})$$

Attribute $a_7$ is the log (base 10) of the total number of classifications done by $d_{u^i}$.

Some of those attributes are redundant or may be derived from others, but at this point we want to elicit visual patterns and not determine the best unique subset of attributes that describe volunteers' categories or groups. The visualization technique we've explored allow the use of similar, derived or inversely proportional attributes so visual patterns may emerge.

### 3.3 Experimental Results

As a visualization technique for the SOM results we chose the Parallel Coordinate technique.[21] This technique consists on mapping $k$-dimensional data to a two-dimensional surface by using $k$ equidistant parallel axes where each axis correspond to one dimension (attribute) and are linearly scaled (normalized) according to the minimum and maximum value for that dimension. Each data point is represented as a polygonal line, which intersects each axis at the point which corresponds to the value of the corresponding dimension.[14]

To each cell in the Self-Organizing Map matrix we will associate a Parallel Coordinate icon, which will be drawn using the values associated to the reference vector $v^m$ in a thick gray line. All data vectors that can be assignable to that reference vector (i.e. all the data vectors that would select that neuron as Best Matching Unit when the network is trained) will be drawn in the background in light gray lines. The value over each icon shows the number of data items (i.e. volunteers) which can be considered close (in feature space) to the respective reference vector. The Parallel Coordinates icon will be drawn without titles, scales or any other information, in order to reduce clutter.

Several different subsets of features were used to create and train the SOM and the Parallel Coordinate icons associated to it. The next subsections present some experimental results.

#### 3.3.1 Visualizing Volunteers' Activity

Figure 1 shows the result of the training of a 6 × 6 SOM with the 7 attributes arranged in Parallel Coordinate icons. Each parallel axis corresponds, from left to right, to the order in which they were described in section 3.2.

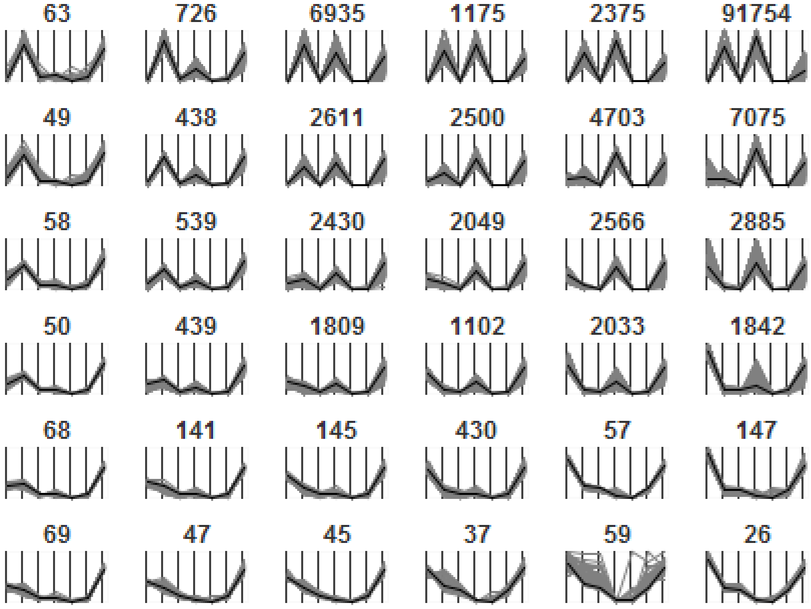


Figure 1: Visualization of volunteers' activity with a SOM network and Parallel Coordinates

The advantages of our approach are the possibility of observation of generic behavior patterns in neighbor SOM cells and concentrations and deviations on generic behavior by the spread of the lines in the Parallel Coordinate icon (when compared with the reference vector corresponding to that neuron on the SOM grid). For example, it is possible to visually identify the following four generic patterns[9] in Figure 1:

- Groups of volunteers identified as "curious": their visual representation have low values for attribute $a_1$, high values for attribute $a_2$ (just because their period is very short, usually one single day) and low values for attribute $a_4$ – these are volunteers who joined the project and done most of the classification in one or few days abandoning the project shortly afterwards. Icons shown near the top right corner of Figure 1 are representative of this category.

- Groups of volunteers that could be labeled "potentials": their visual representation have features of curious but with high values for attribute $a_1$ compared to the values for attribute $a_6$ and medium to high values for attribute $a_4$. This combination of values indicate that they spent long periods without doing any classifications, but come back eventually. Icons shown near the bottom left corner of Figure 1 are representative of this category.
- Volunteers that could be labeled as "dedicated", which visual representation have high values for attribute $a_2$ and predominantly high values for attribute $a_4$. Icons shown near the top left corner of Figure 1 are representative of this category.
- Volunteers that cannot be identified as any of these groups.

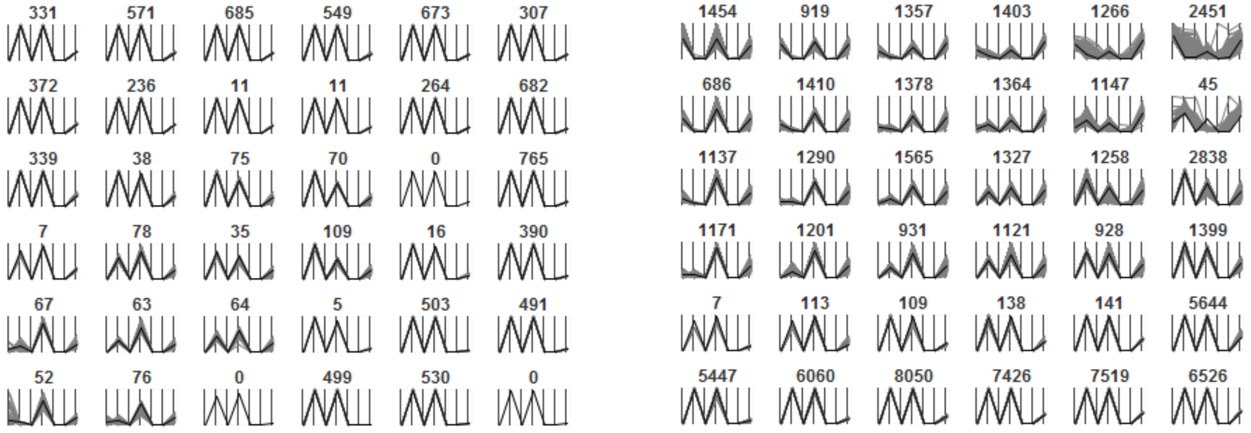
It is also possible to observe some SOM cells in which the reference vector and data associated to that cell are somehow heterogeneous – for example, the cell in line 6, column 5. This occurs because some of the "most different" data vectors (i.e. outliers) get assigned to that cell, choosing it as the Best Matching Unit during the training of the network. It is possible to reduce this heterogeneity by changing the network training parameters and/or its dimensions.

### 3.3.2 Visualizing Volunteers' Activity *and* Accuracy

The previous subsection showed how the Self-Organizing Map and the Parallel Coordinates can be used together to visually classify volunteers' activity in the Galaxy Zoo Project. We would like to investigate whether there are different visual patterns when we consider also the quality of classification done by each user. In order to get an estimate of the quality of classification (accuracy) per user, we used the data from the Galaxy Zoo data release [*], which provides a catalogue of galaxies and their respective classification (spiral, ellipse or uncertain).[10] Basically the classification labels were extracted from many volunteers' classifications, considering only the ones with a certain amount of agreement in order to chose a label for it.

To simplify the analysis we consider only galaxies that were labeled as spiral or ellipse. For each volunteer we compare their classification for each galaxy that was presented to them with the label for that galaxy, calculating a total number of correct classifications, then calculating the percentage of correct classifications per user.

Figure 2 shows two 6×6 SOM grids with Parallel Coordinates icons: on the left side, only data from volunteers who scored 25% or less of correct classifications are shown. On the right side of the Figure 2 only volunteers who scored 75% or more of correct classifications are shown. The attributes used for the Parallel Coordinates icons are the same ones described in Section 3.2. The percentage of correct classifications per user was not used as an attribute for the Parallel Coordinates icons.



(a) Volunteers with 25% or less correct classifications

(b) Volunteers with 75% or more correct classifications

Figure 2: Visualization of volunteers' activity with a SOM network, filtered by percentage of correct classifications

[*] Available at `http://data.galaxyzoo.org/`

On the total we had 8,964 volunteers who did 25% or less correct classifications and 78,226 who classified correctly 75% or more galaxies according to the Galaxy Zoo project – in other words, more users performed well when considering the agreement between their classifications and the labels.

From these Figures it is possible to observe the predominance of volunteers labeled as “curious” – this predominance is also visible in Figure 1, indicating that there is no apparent difference between the different volunteers profiles accordingly to their correct classification score.

### 3.3.3 Visualizing Volunteers’ Activity by Sessions

Sessions are commonly referred as an arbitrary interval of time during which an user interacts with a website. The definition is not precise since we cannot expect an user to have a easily measurable period of time during which he/she is performing activities on the site and nothing else – other activities or the reason for a short pause in the interaction can be caused by several different reasons. Nonetheless we define the interval that defines a session as three minutes – sessions are consider all interactions done by the volunteers without any interruption or inactivity longer than three minutes. All volunteers must have at least one recorded session, and it is expected that some may have several sessions in a day.

From the definition of session, we calculate, for each volunteer, an input data vector containing the number of sessions during his/her collaboration period (600 days or less), the average session length in seconds and the average number of classifications per session. The percentage of correct classifications was also included as an attribute for this example.

Figure 3 shows the result of a 6 × 6 SOM trained with the sessions data, one record per user. The attributes for each axis are in the same order they were described.

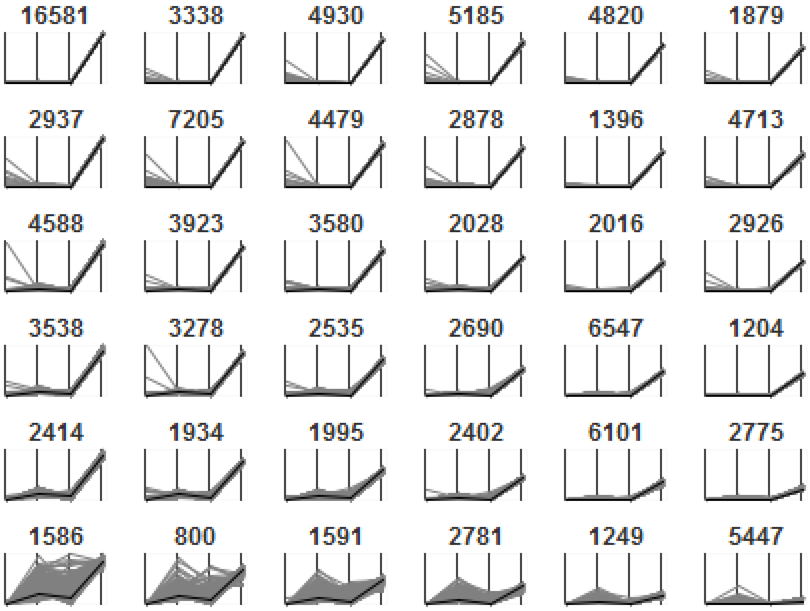


Figure 3: Visualization of volunteers’ sessions with by percentage of correct classifications

In Figure 3 we can also see some patterns on the Parallel Coordinates icons arranged by the SOM: the icons shown near the top left corner of Figure 3 indicates that several volunteers had a few, short sessions, with few collaborations but nonetheless with a good overall correct classification scores. This is the predominant visual pattern: several similar patterns, with smaller classification scores are presented, and spread over the top and right half of the SOM. On the bottom part of the SOM we see lots of volunteers that had lengthy sessions with lots of classifications, also with good overall correct classification scores. The icon on the bottom right of Figure 3 suggest that most of the volunteers with very low correct classifications score had few sessions, with short session lengths and few classifications per session.

## 4. CONCLUSIONS

In this paper we presented an approach to visualize different sets of features as data vectors using a visual representation of multidimensional data using Parallel Coordinates icons, organized in a Self-Organizing Map grid. This approach allows the visualization of significant amounts of data and the visual identification of interesting patterns.

The technique was used for visualization of patterns of interactions between citizen scientists with a citizen science project. The results indicate that it is possible to get some indications on the behaviours of groups of volunteers. More detailed analysis, applied to ongoing and future citizen science projects, may help guide the projects' organizers on selecting subsets of volunteers or deriving labels for the volunteers that describe different aspects of their engagement with the project.

## ACKNOWLEDGMENTS

Alessandra M. M. Morais would likes to thank the Brazilian National Research Council (CNPq) for financial support for the "Evaluation and Implementation of Medium-Sized Scientific Database Management Systems" project (process number 300227-2012/4).

## REFERENCES

[1] Raddick, M. J., Bracey, G., Carney, K., Gyuk, G., Borne, K., Wallin, J., Jacoby, S., and Planetarium, A., "Citizen science: status and research directions for the coming decade," *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey* , 46P (2009).

[2] Cohn, J. P., "Citizen science: Can volunteers do real research?," *BioScience* **58**(3), 192–197 (2008).

[3] Droege, S., "Just because you paid them doesn't mean their data are better," in [*Proceedings, Citizen Science Toolkit Conference. Cornell Laboratory of Ornithology. www. birds. cornell. edu/citscitoolkit/conference/proceeding-pdfs*], (2007).

[4] Prestopnik, N. R. and Crowston, K., "Gaming for (citizen) science: Exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system," in [*e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on*], 28–33, IEEE (2011).

[5] Soares, M. D., *Employing citizen science to label polygons of segmented images*, PhD thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos (2011-06-06 2011).

[6] Alabri, A. and Hunter, J., "Enhancing the quality and trust of citizen science data," in [*e-Science (e-Science), 2010 IEEE Sixth International Conference on*], 81–88, IEEE (2010).

[7] Wiggins, A., Newman, G., Stevenson, R. D., and Crowston, K., "Mechanisms for data quality and validation in citizen science," in [*e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on*], 14–19, IEEE (2011).

[8] Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al., "Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices of the Royal Astronomical Society* **389**(3), 1179–1189 (2008).

[9] Morais, A. M., Raddick, J., and dos Santos, R. D. C., "Visualization and characterization of users in a citizen science project," in [*SPIE Defense, Security, and Sensing*], 87580L–87580L, International Society for Optics and Photonics (2013).

[10] Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., et al., "Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies," *Monthly Notices of the Royal Astronomical Society* **410**(1), 166–178 (2011).

[11] Kohonen, T., [*Self-Organizing Maps*], Springer, 3rd ed. (2001).

[12] Looney, C. G., [*Pattern Recognition Using Neural Networks*], Oxford University Press, 1st ed. (1997).

[13] Vesanto, J., *Data exploration process based on the self-organizing map*, PhD thesis, Helsinki University of Technology (May 2002).

[14] Mazza, R., [*Introduction to information visualization*], Springer (2009).

[15] Tufte, E. R. and Graves-Morris, P., [*The visual display of quantitative information*], vol. 2, Graphics press Cheshire, CT (1983).

[16] De Oliveira, M. C. F. and Levkowitz, H., "From visual data exploration to visual data mining: A survey," *Visualization and Computer Graphics, IEEE Transactions on* **9**(3), 378–394 (2003).

[17] Chantelou, D., Hebrail, G., and Muller, C., "Visualizing 2665 electric power load curves an a single a4 sheet of paper," in [*Intelligent Systems Applications to Power Systems, 1996. Proceedings, ISAP'96., International Conference on*], 126–132, IEEE (1996).

[18] Kohonen, T., "The self-organizing map," *Neurocomputing* **21**(1), 1–6 (1998).

[19] Gorricha, J. and Lobo, V., "Improvements on the visualization of clusters in geo-referenced data using self-organizing maps," *Computers & Geosciences* **43**, 177–186 (2012).

[20] Garcia, J. R. M., Monteiro, A. M. V., and Santos, R. D., "Visual data mining for identification of patterns and outliers in weather stations' data," in [*Intelligent Data Engineering and Automated Learning-IDEAL 2012*], 245–252, Springer (2012).

[21] Inselberg, A., "Visual data mining with parallel coordinates," *Computational Statistics* **13**(1), 47–63 (1998).