IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 9, 2019, accepted June 10, 2019, date of publication June 17, 2019, date of current version July 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923533

Towards Organization Management Using Exploratory Screening and Big Data Tests: A Case Study of the Spanish Red Cross

MARGARITA RODRÍGUEZ-IBÁÑEZ[1], SERGIO MUÑOZ-ROMERO[1,2,3], CRISTINA SOGUERO-RUIZ[2], FRANCISCO-JAVIER GIMENO-BLANES[4,5,6], AND JOSÉ LUIS ROJO-ÁLVAREZ[1,2,3], (Senior Member, IEEE)

[1]Department of Business Economics, Universidad Rey Juan Carlos, Madrid, Spain
[2]Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain
[3]Center for Computational Simulation, Universidad Politécnica de Madrid, Madrid
[4]Department of Communications Engineering, Universidad Miguel Hernández, Elche, Spain
[5]Center for Computational Simulation, Universidad Politécnica de Madrid, Madrid, Spain
[6]Central Office, Cruz Roja Española, Madrid, Spain

Corresponding author: José-Luis Rojo-Álvarez (joseluis.rojo@urjc.es)

This work was partly supported by FINALE Project under Grant TEC2016-75161-C2-1-R, by KERMES Network under Grant TEC2016-81900-REDT/AEI from Spanish Government, by DTS 17/00158 grant by Institute of Health Carlos III (Spain), and by Cruz Roja Española (Spanish Red Cross).

ABSTRACT With the emergence of information and communication technologies, a large amount of data has turned available for the organizations, which creates expectations on their value and content for management purposes. However, the exploratory analysis of available organizational data based on emerging Big Data technologies are still developing in terms of operative tools for solid and interpretable data description. In this work, we addressed the exploratory analysis of organization databases at early stages where little quantitative information is available about their efficiency. Categorical and metric single-variable tests are proposed and formalized in order to provide a mass criterion to identify regions in forms with clusters of significant variables. Bootstrap resampling techniques are used to provide nonparametric criteria in order to establish easy-to-use statistical tests, so that single-variable tests are represented each on a visual and quantitative statistical plot, whereas all the variables in a given form are jointly visualized in the so-called chromosome plots. More detailed profile plots offer deep comparison knowledge for categorical variables across the organization physical and functional structures, while histogram plots for numerical variables incorporate the statistical significance of the variables under study for preselected Pareto groups. Performance grouping is addressed by identifying two or three groups according to some representative empirical distribution of some convenient grouping feature. The method is applied to perform a Big-Data exploratory analysis on the follow-up forms of Spanish Red Cross, based on the number of interventions and on a by-record basis. Results showed that a simple one-variable blind-knowledge exploratory Big-Data analysis, as the one developed in this paper, offers unbiased comparative graphical and numerical information that characterize organizational dynamics in terms of applied resources, available capacities, and productivity. In particular, the graphical and numerical outputs of the present analysis proved to be a valid tool to isolate the underlying overloaded or under-performing resources in complex organizations. As a consequence, the proposed method allows a systematic and principled way for efficiency analysis in complex organizations, which combined with organizational internal knowledge could leverage and validate efficient decision-making.

INDEX TERMS Big Data, machine learning, organization management, organization efficiency, prediction model.

I. INTRODUCTION

The Red Cross is an international organization whose activities focus on preventing and mitigating suffering through humanitarian action to support the victims of armed conflicts and natural disasters, as well as on preventive actions to promote social welfare and quality of life [1]. According to its 2017 Annual Report, Red Cross in Spain (CRE, from **Cruz Roja Española** in Spanish) had over 200,000 volunteers throughout the country, organized in local centers.

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

Their offered services were divided into specific action plans for social intervention, focused on care and employment for vulnerable groups, youth, environment, healthcare, and care for people in need of relief and in emergencies. In each action plan, CRE collects information on all the interventions undertaken in its local centers [2]. This information is usually stored in a repository managed by CRE, and the stored data include the center, the action plan, the type and subtype of intervention, the responsible staff, sociological data on the person or group for whom the action was designed, and data related to the intervention performance, including starting and ending dates and the use of resources. These data have been recorded over the years and they represent today a source of information that can be exploited to draw conclusions and help CRE to improve its organizational processes and to optimize its resource efficiency.

In a more general context, we can say today that Big-Data technologies have emerged and developed rapidly in recent years [3], and they are showing the potential to extract relevant information from large amounts of data in a number of fields [4]. The review from several state-of-art related works in the next section indicates that this is a growing field, whereas still a principled approach is to be developed. In 2013, the American Red Cross already had underlined the opportunities in the application of Big-Data technologies and techniques to assess the needs of humanitarian aid efforts [5].

Accordingly, in the present study we analyzed how to exploit organization databases in the early stages of their analysis towards an efficient implementation of Big-Data systems [5]. In these early stages, when little quantitative information is available about their efficiency, it is recommended to screen the database through simple, yet statistically-founded methods. This has at least three main advantages. First, data quality and database management are paid special attention by the owner organizations after any systematic data screening, as this works as a reminder bringing again into scene the expected relevance of the accumulated data [6]. Second, univariate descriptions in large databases are highly informative by themselves when accumulating in large numbers of recordings. And finally, the detailed knowledge about the single-variable nature and its screening yields a solid basement for subsequent higher-level multivariate analysis.

Therefore, categorical and metric single-variable tests are proposed and formalized in this work to provide a mass criterion that can identify regions in organization forms with clusters of relevant variables. Following previously developed methodology and experience [7], [8], nonparametric bootstrap resampling is used to establish the statistical tests, each single-variable test is represented on a visual statistical plot, and all the variables in each form are jointly visualized in the so-called *chromosome representation*. The method is used to perform a Big-Data exploratory analysis on the follow-up forms of CRE, based on the number of interventions and from a by-record basis.

The paper is organized as follows. In Section I, *Background*, a contextual reference summarize some relevant works in the field. In Section II, *Theoretical Proposal*, the basis of the proposed screening analysis are presented, including the use of proportion and mean difference tests, and their application for performance analysis in the absence of specific quality measurements for the intervention recordings. In Section III, *Organization Description and Database*, the NGO structure and the assembled database are described. Two result sections are presented in the paper. Section IV, *Results Obtained on CRE Big Data*, includes a detailed and methodological description of the experiments and the structure of the outputs, including their graphical interpretation. Section V specifically shows and describes the final results on the database through the mentioned methodology on the pre-defined view-model of the Humanitarian Organization. Finally, Section VI, *Discussion and Conclusion*, draws the principal conclusions and presents a basic discussion.

II. BACKGROUND

A growing amount of academical and applied systems have been proposed in order to optimize the organization management from very different viewpoints, strategies, and problem statements based on their currently available large databases [4], [9]. Emergency organizations have already started to scrutinize their available activity recordings in order to optimize their efficiency in several dimensions [10]. A system has been delivered that uses Big-Data-based statistical information to analyze the emergency rescue services offered by hospitals [11], and it provides the responsible managers with a way to identify shortcomings in the performance of emergency rescue operations and to reveal opportunities for optimization. A theory has been stated on the needs of emergency response organizations in order to develop their collaborative capacity, based on two large-impact emergency events, namely, the 2004 Asian Tsunami and the 2008 Wenchuan Earthquake, with data provided by the Taiwan Red Cross organization [12]. This study sheds light on the field of disaster management, while authors acknowledge and point out a number of factors than Red Cross should examine before generalizing these results when confronting a disaster situation in different natural disasters, different medical organizations (structure, number of doctors), different country infrastructures and population, or technological developments, among many others [12].

The practical application of knowledge discovery in databases technology to engineer project management has been scrutinized, ranging from management difficulties to optimization of engineering project management and to project progress control [13]. A system has been described which uses Big-Data based statistics to analyze questionnaire data and to evaluate the formation of the project management system, including the administrative system, its organization, and its process control [14]. A documentation infrastructure has been proposed to improve project management, project execution, and team communication, consisting of a documentation model and a supporting environment to capture, store, and retrieve knowledge from organization databases [15]. The application of data mining techniques has been also applied to create a risk management system that can help to make the right decisions faster and more accurately [16]. Specifically in the education sector, a study was carried out based on data mining of statistical data for effective educational project management, better educational policies, and improvement in educational learning strategies [17].

In the healthcare environment, we can find many different, yet concurrent approaches to exploit Data Science and Big-Data analytics in order to improve the organization efficiency from the available electronic and health recordings. For example, an intensive care unit readmission prediction model has been described [18] to admit patients into intensive care units based on an automatic learning technique, in which the variables available in the inhospital health record are used in real time. A decision support tool has been introduced based on a data mining system for simulation and optimization in an emergency department in Kuwait [19], with real data that were gathered for 24 working hours and allowed to reduce hospital expenses and provide better treatment to patients. A complete framework has been presented [20] based on a simulation model and a practical methodology for identifying bottlenecks in the interface between the emergency department and the hospital ward. The model considered the relationship among patient urgency, treatment, availability, and the occurrence of waits for treatment.

Other fields are also starting to analyze the scope of the Big-Data exploitation of their organization forms. In the criminology field, several systems have been highlighted that used Big-Data-based crime statistics to build crime prediction strategies, which could help to reduce the amount of crimes [21]. In the textile industry, the varying effectiveness of data-mining techniques has been described [22], showing that classification techniques are of greater interest than clustering techniques to solve problems in this area, whereas some more steps are still needed to build efficient predictive systems. New techniques have also been proposed which can be used with such datasets and subsequently applied to discrete variable multi-objective problems related to production systems [23]. In the electricity industry, a data mining statistical tool has been created that can download, parse, and store data from electricity-market operator websites, to ensure that stored data are constantly updated and reliable [24].

The method expanded in the present work was partially used in [8] to analyze data from 10 years in the telecommunication network of the Spanish high-power electric network, and it made evident non-trivial statistical shapes in the data distribution reflecting the implicit maintenance behaviour, while highlighting the significantly relevant features from the usual maintenance reporting forms.

To sum up, the need of processing raw data and exploring valuable and potentially useful information obtained from them has arisen in many areas of science [25], [26], medicine [27], engineering [28], marketing [29], and pharmaceutical [30], among many others [7]. Today information technology applications analyze data and convert them to valuable knowledge in an efficient way. However, few systematic analysis are still available to deal with heterogeneous and large databases made by forms.

III. THEORETICAL PROPOSAL

A database of intervention recordings in an organization for society attention can be seen as a multidimensional data structure, often corresponding to a SQL structured type. Sometimes a map-reduce based solution can be fruitful in order to extract useful information from queries on the original database in an embarrassingly parallel way [31], which can be helpful in most of the possible data arrangements. If we denote the set of interventions in a database as $\{C^i, i = 1, \cdots, I\}$, the data structure for each intervention can be seen as given by a concatenation of J features, denoted as $\{F_j, j = 1, \cdots J\}$. Often we will have S form sheets from different views of the multidimensional cube, denoted as $\{F^s, s = 1, \cdots, S\}$.

On the other hand, our database often will have form fields with different types. Let us assume that each feature can belong to one type in a set of different possible data types. Two of the most usual feature types are categorical and metric (denoted here as $\mathfrak{C}$ and $\mathfrak{M}$, respectively). A hybrid-vector notation can be established to work with similar statistical descriptions and vector descriptions, which is presented here to support the univariate statistical description of the form features in the database. Similar notations have been established and used before in Big-Data applications for event detection in telecontrol electric networks and for tourism management [7], [8]. If we can previously identify two different groups of interventions, we will be able to generate similar statistical tests and we can then use a vector representation supporting the detection of statistical differences in features from a form.

On the one hand, be F_j a metric variable, then denoted by M_j. Its probability density function (*pdf*) [32] is denoted as $f_{M_j}(M_j)$. Sometimes we can use some convenient criterion to establish two groups in the interventions, denoted as G_1 and G_2. The conditional distributions for this variable fulfill

$$f_{M_j}(M_j) = P(G_1)f_{M_j}(M_j|G_1) + P(G_2)f_{M_j}(M_j|G_2) \tag{1}$$

where $P(G_1), P(G_2)$ are the *a priori* probabilities of the interventions in each group. Each conditional density has its own distribution parameters, and without loss of generality, we denote their conditional mean, deviation, and *pdf* shape (for example, presence of multimodality of heavy tails) as follows,

$$m_j^{G_1}, \quad m_j^{G_2} \tag{2}$$

$$\sigma_j^{G_1}, \quad \sigma_j^{G_2} \tag{3}$$

$$f_{M_j}(M_j|G_1), \quad f_{M_j}(M_j|G_2) \tag{4}$$

We can define their group differences and use them as statistic measurements, i.e.,

$$\Delta m_j = m_j^{G_1} - m_j^{G_2} \tag{5}$$

$$\Delta \sigma_j = \sigma_j^{G_1} - \sigma_j^{G_2} \tag{6}$$

$$\Delta f_{M_j} = f_{M_j}(M_j|G_1) - f_{M_j}(M_j|G_2) \tag{7}$$

so that they can be readily used to make statistical tests to detect significant differences in the two intervention groups.

On the other hand, be F_j a variable with $F_j.type = \mathfrak{C}$, then denoted by C_j. This variable can have a set of possible values or categories among a discrete set, which is $F_j.value = \{v_k^j, k = 1, \cdots, K_j\}$, and where K_j is the number of possible categories for variable C_j. The probability mass density (*pmf*) of that categorical variable is given by $P(v_k^j)$, which is the proportion of this category in a set of observed interventions. Its two-group conditional probabilities are

$$P(v_k^j|G_1), \quad P(v_k^j|G_2) \tag{8}$$

We can define a convenient statistic with their *pmf* differences,

$$\Delta P(v_k^j) = P(v_k^j|G_1) - P(v_k^j|G_2) \tag{9}$$

and its category differences can be grouped in a feature vector,

$$\boldsymbol{\Delta p}_j = [\Delta P(v_1^j), \cdots, \Delta P(v_{K_j}^j)]^T \tag{10}$$

This test can be used to analyze date-type features, in terms of week day, month day, month, and year. In order to obtain efficiently these tests, a bootstrap resampling procedure can be applied over all the computed counts for each feature and category at hand, as described for example in [8].

The above elements give a detailed and vast amount of statistical information about the database, still its visualization is complex because of the variety of data types in the different forms. Further advantage can be taken from the statistical notation and methods previously established, as follows. For the s^{th} form, let us assume that we have J_M metric variables, and the statistical tests for the j^{th} feature is in a row vector,

$$\Delta \boldsymbol{r}_j = [\Delta m_j, \Delta s_j] \tag{11}$$

We can define the chromosome representation of this form as the concatenation of vectors with the statistical difference markers for all the metric and all the categorical features, as follows:

$$\Delta \boldsymbol{c} = \left[\bigcup_{j=1}^{J_M} \Delta \boldsymbol{r}_j, \bigcup_{j=1}^{J_C} \Delta \boldsymbol{p}_j \right] \tag{12}$$

where $\bigcup$ denotes the row vector concatenation operator. This chromosome representation gives an overview of the relevance of each feature in each form sheet, which can be reinforced by using graphical joint representation of the bootstrap significance tests previously calculated. The *pdf* shape in metric features is not included in the chromosome, as previous works show the convenience of analyzing them separately.

In early stages of a Big-Data screening on a database, some *a priori* criterion can be missing in order to establish two groups. We inspired here in two possible criteria to establish them in these circumstances, namely, the Pareto principle and the Six-Sigma principle. In short, the Pareto principle establishes that in some processes, roughly the 30% of the causes are responsible for roughly the 70% of the effects, and vice-versa. This can be related with the acceptable assumption of an underlying exponential distribution in this process. Let us assume that one of the fields in the intervention of the database is an agent of the intervention in wide sense, for instance, a working division or its manager, a geographical location, an end-user, or some other suitable one. We denote this set of agents as C_g, which corresponds to the g^{th} categorical variable, and then its *pmf* is $P_g(v_k^g)$. We can now define a threshold $K - 0$, in such a way that

$$\sum_{k=1}^{K_0} P_g(v_k^g) \leq e^{-1} \tag{13}$$

and then define the groups as follows,

$$G_1 = \{v_k^g / k \leq K_0\} \text{ and } G_2 = \{v_k^g / k > K_0\} \tag{14}$$

This way, G_1 (G_2) represents the set of the categories in the organization which have a larger (smaller) number of interventions, and the performance criterion is this number of interventions itself. This should be obviously taken with caution as a performance criterion, because factors such as cost or quality are not necessarily associated to the number, but they still can represent a starting point. A more refined approach can be given inspired by the Six-Sigma principles, which work with the two tails of statistical distributions. We can establish two thresholds, K_1 and K_2, with $K_1 < K_2$, and three groups,

$$G_a = \{v_k^g / k \leq K_1\}, \tag{15}$$

$$G_b = \{v_k^g / K_1 < k \leq K_2\} \tag{16}$$

$$G_c = \{v_k^g / k > K_2\} \tag{17}$$

In this case, G_b can be considered as the usual performance in terms of number of interventions, whereas G_a (G_c) should be considered the top-performers (the low-performers) with respect to the intermediate group, and pairs of comparisons can be done to identify the relevant features characterizing the top- (the low-) performers with respect to the intermediate by making $G_1 = G_a$ and $G_2 = G_b$ ($G_1 = G_c$ and $G_2 = G_b$). This nomenclature should again be taken with caution when driving conclusions, but it will useful here to follow the analysis.

IV. ORGANIZATION DESCRIPTION AND DATABASE

CRE is organized on the basis of a pyramidal structure where the activity unit is found in Local Assemblies, so-called *locations* hereafter. The structure presents several consolidation levels in terms of management and coordination activities,

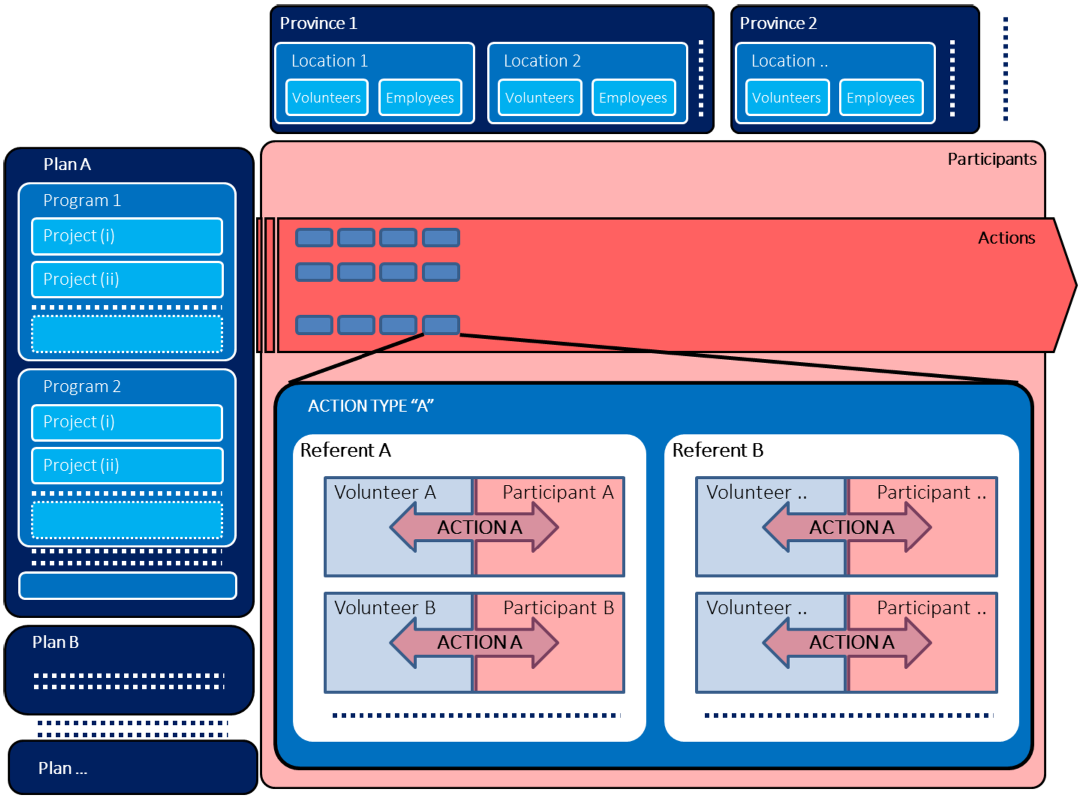


FIGURE 1. Functional and operational organization of spanish red cross. This picture shows the CRE structure including actions as well as intertwine human resources perspective of volunteers and employees.

namely, Provincial offices, Autonomous offices, and Central Office. Provincial Offices act as a support and management body for the Local Assemblies, among other functions. Above the Provincial Offices, and with a similar structure to them, are the Autonomous Offices. And finally, the Central Office at national level consolidates the activity developed by the entire organization in all the national territory.

A. ACTION PLAN AND MANAGEMENT MODEL

If the words organization and management are directly related to the operational planning of the activity to be developed, then we can say that the Red Cross approaches it by following the Action Plan Model. This model obeys to a strict process and structure, starting with the Mission and the Vision of the organization, and right after establishing a four-year Strategic Plan. This Strategic Plan includes all general and specific objectives to be developed over the term, as well as key initiatives and priorities. Taking this information into account, the derived key performance indicators (KPI) cross the boundaries of the planning phase and become the basic elements for permanent monitoring, as they are meant to be the main variables that match the effective activity. These variables are the most relevant ones as they will allow the management to evaluate the impact of strategies and the efficiency of organization structures over the years. Accordingly, these variables are the focus of this work for their fundamental managerial value.

Under the perspective of the real efforts of volunteers and professionals, the humanitarian institution built a conceptual tree-structure, where the atomic unit is the *Action*. This *Action* stands for the effective activity that is supposed to be performed by the volunteers or by the professional staff. *Actions* are then framed into *Projects*, and *Projects* are turned into *Programs* when aggregated. Finally, *Programs* are consolidated into *Plans* or *Schemes* (see Fig. 1). It can be pointed that, since *Actions* are the atom of the real activity, this information undertakes a relevant part of the database, so that the analysis of such structured information should be scrutinized in detail. This way of structuring the activity is parallel to the way the organization is internally shaped, giving birth to a Business Unit, or more precisely speaking in this case, to *Activity Units*. For organization accountability throughout this tree, it becomes a good way of exploring the reality as it very much relates to physically separated groups of people, as well as to processes and eventually also to assets within the organization.

In order to offer a clear picture of the actions and how they spread over the effective activity, we can illustrate that in the same way that in physics a limited number of different types of atoms pack together to give birth to all the existing materials, in the Red Cross we have less than two tens of different types of actions that are combined to create all the existing real activities carried out by volunteers and professionals. These basic actions are in some cases, but not always, structured by *Type* and by *Subtype* for a better consolidation and a more comprehensive representation. All these activities are present throughout a cross-sectional organization view in all the departments and in all the activity units.

Following the above mentioned view, the reader could consider that an organizational model based on departments and actions is far from the modern conceptualization of efficient organizations. New organizations are more shaped or based on processes, or even at present times, in departments according to knowledge units. Therefore, it could be erroneously concluded that the organization and activity has not evolved over the years to accommodate to the very today. However, the organization has taken all the necessary steps to improve its efficiency and it is currently operating in a process-management model. Still, the extensive databases of an organization with over 150 years, which serves to more of 5 million people only in Spain, requires the existence of a solid ans steady support system for its activity. This represents an almost-deeply hardwired-shaped database that allows a real-time and non-stop operation in an always changing reality. Looking almost a paradox, the organization makes it possible through this rock-solid underlying data structure, which has been maintained over the years with a steady structural model that allows its analysis, supervision, and improvement, based on common structures over the years. Maintaining such a database structure is not against developing activity processes, but instead it serves to keep working fluid a real-knowledge based organization.

B. VOLUNTEERS, REFERENTS, AND PARTICIPANTS

Red Cross is an organization of volunteers, where the activity is carried out essentially by them. But the special characteristics and profile of the volunteers themselves (such as not 24 × 7 availability), or some times the lack of specific professional skills, force the organization to also incorporate professionals under its payroll. In this scenario, it is necessary to characterize the activity carried out by each of these groups to better organize the activity and also to ensure the institutional essence of remaining a volunteer organization, and not just an organization with some volunteers.

The real and basic activity or actions are expected to be done by *Volunteers*, whereas coordination and special activity could then be performed by hired professionals or by volunteers. This needed coordination is performed by a special profile of members, so-called the *Referents*, and regardless they are volunteers or professionals under a payroll, they assume the responsibility over a group of people receiving help and over the volunteers working with them.

Finally, in order have the full picture of the organization and of the people involved, the objective pursued by the Red Cross is to serve individual people who require assistance from the organization. These people served or helped by the Red Cross are referred to as *Participants* in this study, being the axis on which all the activity rotates and the ultimate goal for the entire organization (see Fig. 1).

C. DATABASE DESCRIPTION

The huge database of the organization advises the realization of a first study bounded in terms of the number of records for an efficient management and modelling, in such a way that the results obtained for the restricted analysis in this study can be later extended and refined. Therefore, the present study considered the records of the activities of years 2014 to 2017 of the humanitarian organization in two Spanish provinces selected as representative of the organization. A total of 57 local assemblies or locations are part of the two provinces, and the location variable is one of the key elements of analysis in this study. The complex structure of the organization and the activity makes it difficult to find a single view of the database providing the necessary information and framework for the proposed analysis, so that the information was structured in 4 tables according to the work perspective, which repeat (or not) some variables for better by-part analysis.

The first table was related to the *Referents*, and it contained the structured data around the referring field. It included the activity developed by each of the 504 referents of these two provinces. In this case, the table contained the following categorical variables: province, location, program plan, project, action type, and action sub-type. The table also incorporated the numerical variables year, number of participants, and number of interventions.

The second table showed the activity from the point of view of the *Participants*. In addition to all the aforementioned information, the table conveyed all the categorical or numerical demographic information, such as sex, collective, postal code, language, type of home, or family situation, among many others. This table contained about 250,000 records.

The third table presented the view of the *Person who performs the action*, and in addition to the categorical and numerical information collected in the Referent table, it incorporated singularly relevant information in this regard. This is the case of categorical variables such as the day of the week in which the activity is carried out, whether the person performing the activity is volunteer or professional, or the date and time for the activity is carried out, among others.

The last table conveyed the activity carried out according to the activity nature itself, and it is called the *Action* table. The actions were collected here in a consolidated manner by activities, also consolidating the participants' view, but not from the referent or volunteer view. This table picked up again the variables of the referring table, accumulating in the same way that in the activity table those categorical variables of the date, time, day of the week for the activity, or the voluntary character or not of the person who performed the action.

V. RESULTS OBTAINED ON CRE BIG DATA

In this section we present the results of the implemented Big-Data model on the previously described database. For this purpose, the specific notation and proportion model are exhibited. Then, the observed organizational and productive models are classified, for a more suitable mapping

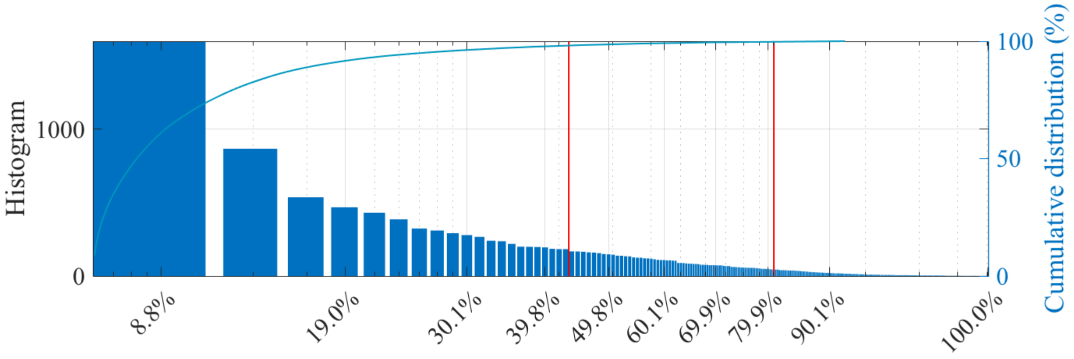


FIGURE 2. Histogram of the realizations (bar graph), cumulative of realizations (line graph), together with separation of the groups described in the text (vertical lines). Interventions are weighted in terms of the number of users each one reaches in the society.

of organizational reality, both in the categorical and in the numerical variables. After that, the results themselves are summarized in the later subsections.

A. NOMENCLATURE AND DATA MODEL DESCRIPTION

The data model used for the analysis was analyzed with a three-group structure, according to the histogram of interventions. Rather than the interventions themselves, each intervention was weighted according to the number of users it affected. The three groups were established for $K_1 = 0.40$ and $K_2 = 0.80$. A an example, if we consider for instance the categorical variable *Activities* and the table under study is *Referents*, then the three groups are structured in such a way that, once the database is sorted decreasingly: (a) The first group incorporates all the first set of Referents and Activities adding up to the 40% of the total activities weighted by the number of users; (b) The second group incorporates the sequentially registered Referents that inject up to the total the next 40% of the weighted activities; And (c) the third group will be considered as the remaining up to the 100% of the weighted activities registered. Hence, in this case we can state that the first group, from now on identified as G_1, includes the top-performers in terms of activities developed that from an aggregated point of view consolidate the 40% of the total effort. As shown in Fig.2, G_1 includes the Referents producing from about 60,000 up to 15,000 activities-user, G_2 includes the Referents raising from 15,000 to 9,000 activities-user, and G_3 includes all the remaining ones.

In general terms and from an organizational and business perspective, we could consider the G_a group as the *Top Performers*, G_b as the *Standard*, and G_c as the set of *Underperformers*. It is important to point out that this nomenclature should not be understood in the business sense in this example, nor in others, due to several reasons:

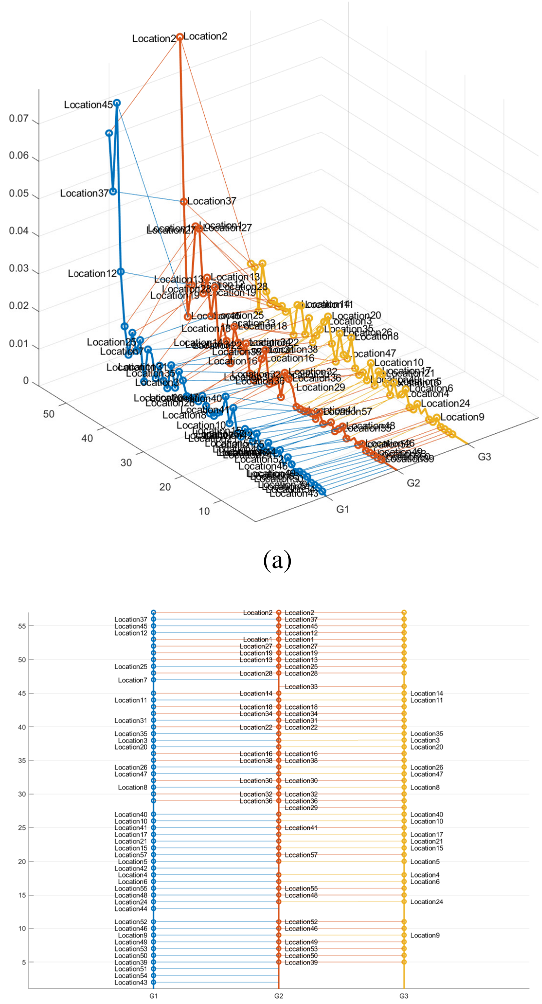
- *Diversity of consolidated activities.* Present cross-sectional work incorporates, without distinction, activities such as the delivery of food to vulnerable people, which might be considered as easily extensive in terms of function and with low intensity in terms of efforts. Caution has to be taken when comparing this to other actions, such as the employment training and re-qualification programs, as these often involve several months and thousands of hours of effort but they are considered as one single action or activity.
- *Diversity of Referent Profiles.* This humanitarian organization of volunteers also has professionals who sometimes concentrate, at least formally, in the referencing function as they are always present. But in other cases, especially in smaller locations, this profile is covered exclusively by volunteer staff, as no professional is available for such function.
- *Volunteer Availability.* By the same token, in cases where the Referent function is covered by Volunteers, there is no fixed number of hours each of them can devote, therefore, this role might sometimes be distributed among a variable number of the Referents.

Therefore, the characterization under these denominations should be understood as a way of conceptualizing in order to argue, discuss, and elaborate on the different models, but not to define any punitive or rewarding character among these three groups. In the same way that a worker who spends 40 hours a week cannot be rewarded for having a large number of activities under his reference function, it is not intended to underestimate the activity of volunteers who use their time and devotion to attend and help the most vulnerable people. On the contrary, this classification should help us to understand, and if it was case also to improve, the organizational efficiency and the use of resources in order to facilitate the best working environment while maximizing the activity.

For a better resolution, we often used logarithmic units when representing metric variables, and for mathematical consistency and illustrative purposes, the null values were replaced by 0.1 prior to this operation, so that they could be readily observed in the numeric variable distribution panel. As mentioned before, categories and subcategories of variables were treated equally in this study as they have equivalent meaning with different aggregation level.

B. REPRESENTATION OF CATEGORICAL VARIABLES

For the representation of the results with categorical variables we used 3-dimensional (3D) plots where the three mentioned profiles and their statistical comparative



(a)

(b)

FIGURE 3. **Example of profile plots for Referents in terms of location categories. (a) Example of the *Profile-Plot* for the three representative profiles of this categorical variable. The representation of Referent table by locations for both provinces are here combined. (b) Overhead or 2D *Profile-Plot* for the same example categorical variable. This view yields a clear perspective of the different models, as the name of the locations appears in the predominant corresponding side. The text before the G_1 vertical line reflects the predominance of G_1 vs G_2, whereas the text to the left of the G_2 vertical line shows the prevalence of G_2 vs G_1. A similar reading of the situation can be done if the text is located on the right of G_2 (of G_3), which express the predominance of G_2 over G_3 (of G_3 over G_2) in that category.**

behaviour are shown. In this representation, the normalized and weighted number of times that the categorical variable appears are presented on the $z-$axis. In the $x-$axis, the G_a, G_b, and G_c groups are represented, whereas on the $y-$axis the comparative variable is represented, sorted from the highest to the lowest prevalence of its categories. This representation shows the predominance over groups for each of the categorical variables. In order to provide a better understanding and a comprehensive scenario of the representation, several examples are presented next. Figure 3 shows an example of this 3D representation (in Panel (a)), and the same example but from an overhead or zenith perspective (Panel (b)). In order to identify this type of plot among others included in this work, we will refer to it as the *Profile-Plot*, and it represents either the 2D or the 3D plot of the three Representative Profiles simultaneously in one single figure. Different data patterns can be scrutinized with this representation, which are summarized next.

Dissociated Pattern. In the particular case of a selected categorical variable, such as the *Location* variable, high values of G_1 vs G_2 in a given table (such as activities) indicate that the Top-Performers group is predominant in this location compared to those considered as Standard. Similarly, a G_3 with significantly higher values than G_2 indicate in that same location that the group of Under-Performers are predominant vs the standard. Consequently, if both circumstances are simultaneously present, it could be interpreted as the existence of a dissociated reality of two clearly separated groups of referents in terms of their productivity. This reality must be considered relevant for the analysis because it supposes the existence of clearly separated needs between both groups, therefore, depending on the volume of people in these groups they could demand the need to establish special policies, both to address this reality or to correct it. In any case, this kind of pattern is considered more than relevant form an organizational perspective.

Normalized Pattern. Following a similar analysis, in the case of a categorical variable in which the G_2 is clearly superior to its precedent (G_1) and to its subsequent (G_3), this could indicate a broadly normalized reality in the environment of the activity, with values close to the standard in that location.

Overloaded Pattern. In another possible organizational pattern for the case of dichotomous variables, we can find the result of having a much higher proportion of Top-Performers (G_1) compared to the rest of the groups. This would indicate in our example for a specific *Location* a strong concentration of activities in some few Referents. It is necessary to note that this reality is not an isolated analysis, but instead it is the result of statistical comparative analysis with all the locations, meaning that this kind of result will be statistically relevant and it should be carefully analyzed after. We point out again that the analysis does not indicate an operational malfunctioning, but rather a differential situation with respect to the statistically consolidated and normalized data.

Participatory Pattern. The existence of a high predominance of the Under-Performers in front of the other two groups will illustrate the statistical presence of a very high group of G_3 participants who share the carried-out activities. Note that it is precisely at this point that the aforementioned descriptive nomenclature becomes patent, since the existence of a high number of Under-Performers reveals positive aspects such as the existence of many of them. This fact is something that the humanitarian organization pursues, since the impulse of many volunteers in the various activities brings balance and soundness to the model. This being true, it could also incorporate negative aspects to the contrary, if tailored tools are not in place to manage it, such as communications training. But as we were informed by the organization, this is a more than desirable challenge as it should be the fundamental role of professionals in the payroll of volunteer-based organizations.

To end-up with these descriptive patterns, just recall that this classification is not intended to be a qualifier or disqualifier of any organization, but rather it aims to give a descriptive representation of special or singular situations present in that organization, which should allow a credited and subsequent analysis for a proper management and the proposal of principled initiatives for improvement.

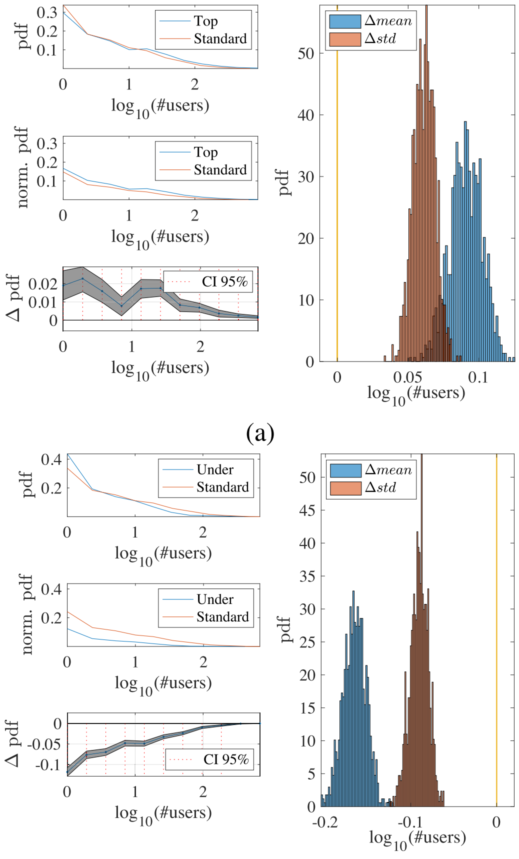
C. REPRESENTATION OF NUMERICAL VARIABLES

As stated in the theory section, the numerical variables are represented in terms of their mean and standard deviation of the statistical differences of their group values. In this setting, a representation with logarithmic transformation of such a difference can characterize the variable under study visually and quantitatively. Accordingly, and for example the *Top-Performers* or G_1 vs the *Standard* or G_2, if the logarithmic histogram (i.e., its 95% confidence interval) overlaps the zero value, this indicates that there is no significant difference between the number of elements of that variable that participate in one and another group. On the other hand, if the histogram of the mean difference does not overlap zero, it denotes a preponderance in one of the groups.

Figure 4 shows an example of this representation corresponding to the number of participants, as a numerical variable, and the statistical distribution of the decimal logarithm of the difference between the number of volunteers between G_1 and G_2 (up) and between G_3 and G_2 (down). In the first case the distributions of the differences in mean and in standard deviation do not overlap zero, indicating that significant differences are statistically detected, and more, that both statistics (mean and standard deviation) are larger in G_2 group. The upper panel shows positive significant differences in the mean and in the standard deviation, showing that the Top-Performers present a higher number of participants than the *Standard* group. We can state that the selected numeric variable is statistically predominant or not in the defined group when the analysis of the differences between the two groups does not overlap the zero, and the histogram occupy the consistent side. Note finally that the complete histograms of the numerical variable, as well as its estimated difference band (in grey) are represented in order to give a comprehensive description of the distribution in the two groups under comparison, beyond only the mean and the standard deviation.

D. CHROMOSOME PLOTS

In addition, a wider view of the relevant information in the database could be exhibited through a bit more complex but interesting plot, jointly including all the categorical and all the numerical variables from a given data table. This plot represents a database perspective, as it displays all of the variables sequentially in terms of their differences between two groups. The compared groups are usually G_1 or G_3 (Top-performers or Low-performers) compared vs G_2 (Standard).



(a)

(b)

FIGURE 4. Example of statistical tests for numerical variables. In each panel, the left column represents the two distributions for the variable in each group (normalized to unit area and normalized with respect to the a priori probabilities of the groups, on the top and medium panel), as well as the difference between these distributions and their confidence bands in gray (bottom panel). Dashed vertical red lines denote the band region where significant differences are found in the group-conditional *pdf* profiles. An example is shown for the number of participants in top-performer vs standard groups (a) and in low-performers vs standard groups (b).

An example of this representation can be seen in Fig. 5 for both provinces (A and B) included in this analysis. What is interesting about this view is that this plot can be uniquely linked to each province or location, and it reflects the profile of the activity performed under the statistical perspective, allowing to establish similarities and differences among different provinces as a door-key shape.

VI. RESULTS ON THE DIFFERENT VIEWS

Based on the aforementioned statistical analysis tools, a number of circumstances can be scrutinized on the different views or tables and the set of variables represented in each of them. Below are the most relevant results for these views.

A. REFERENT VIEW

The *Referent View* table consolidates all the actions by Referent. This view combines into a single registry all the activities

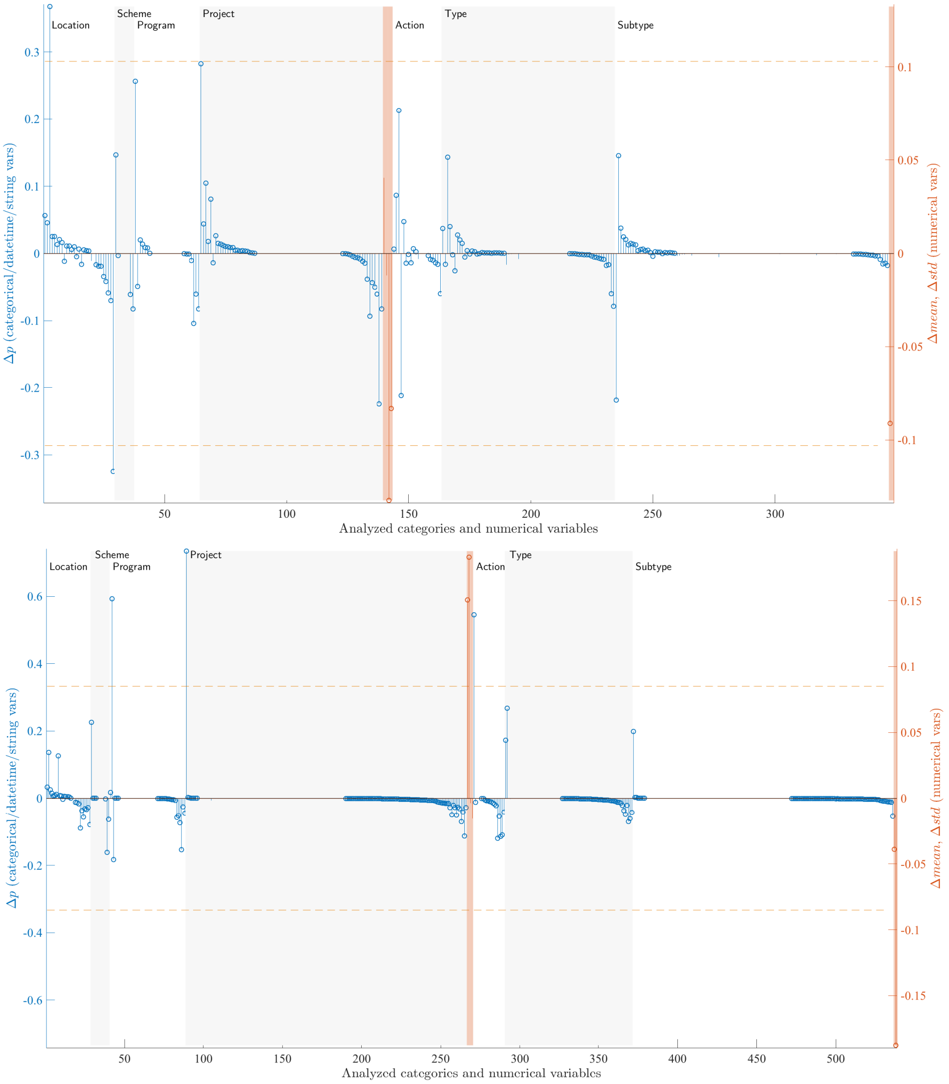


FIGURE 5. Examples of chromosome plots. (a) Representation of the chromosome plot on referent view for province A (top) and for province B (down). The comparisons of groups are depicted for all the aforementioned elements in terms of their probability differences (in blue), namely, for categorical variables location, scheme, program, project, action, type, and subtype. On an additional right vertical axis, numerical variables (in orange) are characterized in terms of their difference of means and standard deviations. Circles indicate significant differences are obtained in those variables or categories. Gray and white shadowed regions indicate the different variables to be distinguished from their neighbors.

carried out under the umbrella of a single Referent within a specific Program and Project. In other words, all those activities with the very same Action, Type, and Subtype, under the responsibility of a given Referent, are stored as one single record. According to this view, the main results were as follows. Larger locations in terms of activity often showed a Dissociated Pattern profile, with the exception of the largest one in each province. This last one corresponded to the always singular situations of the capital of that province, which concentrates a high number of employees reshaping the standard local behaviour and showed a Normalized Pattern. On the other hand, in those locations with lower intensity of the activity, the Overloaded Pattern was highly predominating, both in the Top-Performers and in the Standard groups, as it can be seen in Fig. 3 (b).

Regarding the analysis of the categorical variable *Plan* (or Scheme, as displayed in Fig. 6, left), this variable has also a great reflection in its subcategories *Program* and *Project* (corresponding with subcategorizations of each plan). The Normalized Pattern predominates in a number of plans, with some singular exception which presents a Dissociated Pattern. This variable also highlights the existence of the Participatory Pattern especially in the most active programs. When this singularity arose during conversations with the humanitarian institution, it was accepted as the special reality of this one. Accordingly, the results showed the good performance of the developed method, recognizing this outlier as a special well-known situation. In the case of the categorical variable *Programs*, it is necessary to note the existence of two highly active programs, Programs 12 and 3, whose patterns

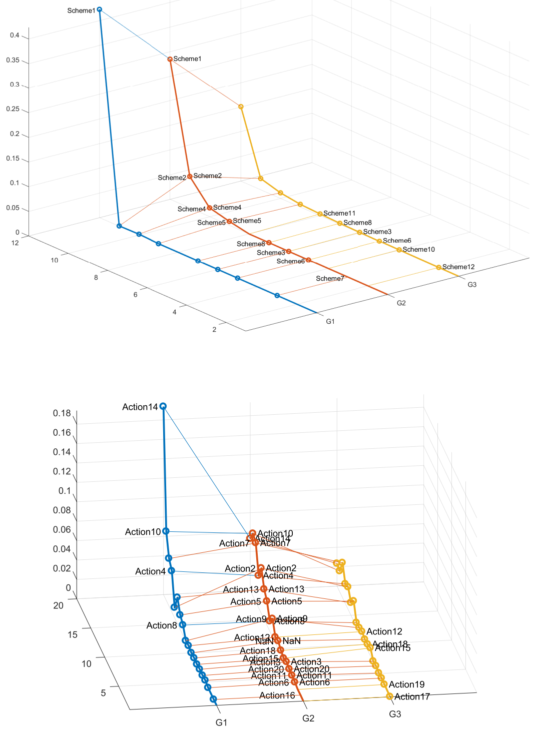


FIGURE 6. 3D representation of referents-view over variable Schemes (up), and over variable action type (down).

stand out both for their breadth and intensity, and for their overload in the Top-10 group of performers. Regarding the variable *Actions* (see Fig. 6, right), the Normalized Pattern stands out, except for the Monitoring activities, which are strongly linked to telephone activities and present an Overloaded Pattern. Regarding numerical variables, we can point out the behavior of Uniform Patterns both for the case of the volunteers and for the participants, where the preponderance of the Standard group stands out in front of the *Top-Performers* and of the *Under-Performers*.

An example of chromosome plots of G_1 vs G_2 for the Referent View can be seen in Fig. 5 for provinces A and B. As earlier mentioned, this plot offers an interesting perspective as it gives a holistic representation of the organization activity in the area of analysis. In particular for our example, we can readily observe the following points. Information can be extracted by the length of the abscissa axis, as it represents the number of different items that are available for each analyzed variable. It is clear in this setting that province B incorporates a much wider range of activities, going up to 550, whereas province B barely reaches 350, which is almost halve of it. A deeper analysis can be seen in the profile for every type of present variable, and how the visual profile identified each province and the activity on an almost unique way. In order just to explore the potential of this perspective for the humanitarian organization, we can see that for the variable Project both provinces have a similar shape in those projects occupying the last positions, which corresponds to projects with a lower number of Referents. For this segment, the Standardized Pattern G_2 is reproduced. But if we look at the projects with larger number of referents, a much larger predominance of G_1 (Top-performers) is present in province A. It is also significant than in province B there is one single project with unparalleled predominance of Top-performers. This situation will probably require a review to understand if it is justified for any reason. According to members of the organization, in this particular case it corresponds to a project that is very specific to the province, as this one contributes not only to the province but also to other regions. It is also interesting at this point to indicate that, depending on the specific sorting of the abscissa axis, the perspective and comparison within each variable could offer additional outlooks.

B. PARTICIPANT VIEW

The Participant View included the absolute complete database, and it incorporated up to about a quarter of million records (more than 256,000). This table incorporates each action carried out with a participant in a separate record. Therefore, it is the most complete view that will allow to consolidate all the information available for the analysis. In this case, group G_1 corresponded with the accumulated 25% of the actions, group G_2 to 63%, and group G_3 to the remaining ones. These ranges correspond in this case with performance in terms of activities per participant, from 23 to 7 in G_1, from 6 to 4 in G_2, and from 3 to 1 in G_3.

From the *Location* point of view, the most predominant pattern was the Normalized one, although in the case of locations with less activity, the Participatory Pattern was more predominant. This reality suggests that in the locations where the number of actions per user is larger, and this happens in the intermediate group, which mostly corresponds to a closer number of helps received by the users. However, in the locations with lower number of actions, a lower number of helps or activities per user is also predominant.

From the point of view of the categorical variable *Country of Origin*, it is relevant to point out that the most repeated variable was the local country, and it follows a Normalized Pattern. Apart from this one, the next most frequent pattern was the Participatory with some few exceptions of Dissociated Pattern. It can be interpreted in this setting that local participants, in addition to being a majority, receive predominantly intermediate assistance and attention, while other nationalities present a more dissociated reality, so that some of them receive large attention in terms of number of helps, and on the contrary less than locals.

Regarding the variable *Educational Level*, it should be noticed that the pattern was mostly Standardized, with the exception of Graduates of Arts and Crafts that followed an Overloaded Pattern, and in the case of those who have a School Certificate, who presented a Participatory Model. The most predominant language was the local, which followed a Normalized Pattern, compared to the second one, which presented a Dissociated Pattern, whereas the rest of them followed a more Participatory Pattern.

Regarding variables *Professional Activity* and *Labor Situation*, it is relevant to tell that the number of participants not setting it were a wide majority, and also their pattern was Participatory. It is also relevant that, in statistical terms, those participants who did not set or complete this variable were getting less attention from the Humanitarian Entity. When this singularity was confronted with the institution, it was suggested to be consistent with the idea that in the case of not detailed participant evaluation being developed, it is not possible to define a tailored solution to attend the participant needs, which will guide the help that has been developed by the Humanitarian Organization.

C. VOLUNTEERS AND ENVIRONMENT VIEWS

Focusing on the *Volunteers* table, and attending to the existing data distribution, a new structure of G_1 to G_3 was generated. Top-Performers were herein those ones that consolidated up to 50% of the activity, whereas the Standard were those ones adding the subsequent set up to 75%. Regarding this view, it is not possible to find homogeneous patterns along the *Locations* where the activity takes place, showing Overloaded, Participatory, Dissociated, and to a lesser extent, Normalized Patterns.

Under the perspective of *Plan or Scheme* in Fig. 6, two schemes concentrated a significant part of the volunteers. Those two schemes showed an Overloaded Pattern, although the second in relevance exhibited a slight trend to the Dissociated Pattern. As for variables *Programs* and *Projects*, the Overloaded Pattern standed out, with a better participation of the Dissociated Pattern in the second place. As for the analysis carried out by the *Days of the Week*, it turned out that the existence of Overloaded Pattern was widespread. From an interpretation standpoint, we can note that Top-Performers are volunteers every day of the week without exception, in other words, the variable coding the day of the week when the action is performed is not of any interest from either statistical or technical analysis viewpoints.

Regarding the *Environment View*, everything previously expressed was basically reproduced, as there were no new variables to introduce in the analysis, although some of the effects indicated above were more clearly appreciated herein. For instance, this was the case of variable *Location* in the three predominant patterns. The Overloaded Pattern was observed in the two most active locations, followed by a Normalized Pattern in the larger ones, and by a Participatory Pattern in the small locations.

VII. DISCUSSION AND CONCLUSION

Decision making is always a real challenge in organizations, and in the specific case of non-governmental and non-profit organizations, their efficiency is directly transferred to enlarge their social impact and goes to the people in more need. Sophisticated and complex organizational state-of-the-art initiatives are often more related to new economy organization or to very large organization than to the so-called third sector. It is not that common for the social sector to cope with state-of-art technology and research in any field, and especially in the management side, which can limit the set of available tools. We might think that the reason for this lack of research activity in the NPO-NGO management models is not related to the lack of interest from their side, but rather it is probably associated to other facts. Examples of them could be disuses in the difficulty to allocate budget and effort on such activities, which may look externally far from their core business. This is specially relevant in this kind of organization due to the fact that these organizations are strongly scrutinized by their financing bodies (public administrations and/or the society in general), so that rather might not see it as justifiable or the associated expenses could be rejected from direct funders.

However, in general terms and according to the literature, the early stage planning and the critical report analysis have a relevant impact on the continuous improvement for almost any organization. These techniques always rely on the analysis of tracked records and data, normally related to a specific activity or economic model, and not so often they incorporate a dynamic view of the organization, or a multiple perspective for a wide range of simultaneous activities or profiles of people participating. Big-Data analytics is one of the concepts that supported our analysis in this work. Several reviews and tutorials can be found in this setting (as seen in Section II) which not only highlight the possibilities of these techniques in the area, but also their applications in other fields.

In this paper, we proposed that the acquisition of knowledge and evidence of the multidimensional reality of a complex organization as CRE can facilitate the decision-making processes as well as the implementation of organizational protocols, even when based on a single-variate data model. Our results demonstrate that the statistical analysis based on Big Data can open a new and wider perspective of the organizational dynamics, offering an opportunity to identify efficiency gaps and spaces where the lack of resources can be limiting the activity. Additionally, the wide spread organization and the vast creativity of the volunteers of the humanitarian organization attached to those in need and the lack of resources, could arise other efficient (yet difficult to imagine from a corporate position) models that require to be deeply analyzed. The strategy proposed in this paper opens a new perspective to organizational analysis and management. The method was applied to a database provided by CRE, in order to obtain exploratory conclusions regarding the organizational model and its use of resources and human capital. The database that was analyzed gathered the number of interventions on a by-record basis (according to the province, location, program plan, project, type and sub-type of action, year, and number of participants). The results of the simple and principled analysis provided relevant information and evidence that can help management in decision making while reducing uncertainty. The results presented herein reveal that the principled combination of raw bootstrap, map-reduce, mean and standard deviation tests, density tests, and proportion tests, can extract valuable knowledge for decision makers. The CRE managers agreed that it is important to have this information for support prior to making any changes in the actual organization. This is particularly significant, since the methodology presented here takes a descriptive rather than a causal approach, which could be the focus of subsequent efforts. Categorical and metric single-variable tests on the database were proposed and formalized, and in addition, nonparametric bootstrap resampling was used as a wide-use nonparametric criterion to establish the statistical cut-off tests.

As the key conclusion of this work, we would like to call attention on the global result that a statistical approach based on exploratory screening and Big Data tests can show relevant information on the structure of a certain organization, and it could become a powerful tool for management in the organizational field. We are currently witnessing the proliferation of new and disruptive organizational models, some of them really successful (such as Google) and some other not yet validated (such as ING Agile-Organization), where the traditional quantitative and structural descriptions may not be that appropriate to characterize. These new more flexible and dynamic organizational approaches incorporate a relevant number of additional elements not strictly measurable or not strictly static enough to be described and analyzed. The singularity of CRE, and specially the fact that a very relevant part of the activity is developed by a floating work-force, the volunteers, and the unstable demand from the equivalent to customers (the participants), make together this organization a good test bed for these new organization models. CRE is forced to maintain a flexible and dynamic organization, not because it wanted to be the state-of-the-art in organization, but because of its special situation. We understand that this makes this organization a realization similar to other fashioned companies, and as a consequence, we would like to argue in that regard that the Big Data statistical modeling described in this paper may open new opportunities for a continuous evaluation of efficiency and asset allocation in new dynamic and flexible organizations. Hence, we consider that Big-Data models as the one described here, once analyzed and carefully implemented in a certain company, could provide outcoming features to be considered as valid inputs for management report, balance scorecards, and organizational modeling.

In terms of limitations, and although a large database was used in this work, it should be considered that just two provinces were included in this analysis, and so a much larger effort including the 52 national provinces could be addressed in the future. Additional studies are possibly including the regional perspective or further classification of the provinces according to certain criteria. Another limitation of this work could be seen from the categorical variables that were not weighted by any activity or results, but directly analyzed as they had been recorded in the systems, and so supplementary information could be extracted from a deeper analysis considering other relevant ratios suggested from the humanitarian organization. Also this work is based on a univariate approach, but a principled multivariate approach could also boost the results and knowledge in an organizational approach.

REFERENCES

[1] (2019). *Web Site Corporativa de Cruz Roja Española*. [Online]. Available: https://www.cruzroja.es/principal/web/cruz-roja/nuestra-historia

[2] J. Senent, (2017). *Memoria 2017 Cruz Roja Española, Spanish Red Cross, Intern.* [Online]. Available: https://bit.ly/2QsDWwq

[3] I. Taleb, M. Serhani, and R. Dssouli, "Big data quality: A survey," in *Proc. IEEE Int. Congr. Big Data*, Jul. 2018, pp. 166–173.

[4] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[5] A. Monaghan and M. Lycett, "Big data and humanitarian supply networks: Can Big Data give voice to the voiceless?" in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, Oct. 2013, pp. 432–437.

[6] S.-G. Lee, "Challenges and opportunities in information quality," in *Proc. IEEE 9th Int. Conf. E-Commerce Technol., IEEE 4th Int. Conf. Enterprise Comput., E-Commerce E-Services*, Jul. 2007, p. 481.

[7] P. Talón-Ballestero, L. González-Serrano, C. Soguero-Ruiz, S. Muñoz-Romero, and J. L. Rojo-Álvarez, "Using big data from Customer Relationship Management information systems to determine the client profile in the hotel sector," *J. Tourism Manage.*, vol. 68, pp. 187–197, Oct. 2018.

[8] J. R. Feijoo-Martínez, S. Muñoz-Romero, C. Soguero-Ruiz, M. Castro-Fernández, and J. L. Rojo-Álvarez, "Event analysis on power communication networks with big data for maintenance forms," *IEEE Access*, vol. 6, pp. 72263–72274, 2018.

[9] J. Philip and J. Hancock, "The influence of big data on talent decisions," *Acad. Manage. Global Proc.*, p. 73, Jun. 2018.

[10] D. Vanni, G. Palasciano, P. Vanni, S. Vanni, and E. Guerin, "Medical doctors and the foundation of the International Red Cross," *Internal Emergency Med.*, vol. 13, no. 2, pp. 301–305, Mar. 2018.

[11] P. Sefrin, A. Haendlmeyer, and W. Kast, "Performance of the emergency service results of a nationwide analysis of the german red cross in 2014," *Notarzt*, vol. 31, no. 4, pp. S34–S48, Aug. 2015.

[12] L. Y.-H. Allen, "Organizational collaborative capacities in disaster management: Evidence from the taiwan red cross organization," *Asian J. Social Sci.*, vol. 39, no. 4, pp. 446–468, Jan. 2011.

[13] Y. Jia, "Research on practical application of big data mining technology in engineering project management," in *Proc. Int. Conf. Edu. Technol. Econ. Manage.*, 2015, pp. 53–60.

[14] W. Yan-fang, J. Peng-tao, and D. Xiao-ning, "Application data mining techniques for status quo of project management system," in *Proc. Int. Conf. Eng. Bus. Manage.*, vols. 1–8, Mar. 2010.

[15] K. Becker and C. Ghedini, "A documentation infrastructure for the management of data mining projects," *Inf. Softw. Technol.*, vol. 47, no. 2, pp. 95–111, Feb. 2005.

[16] X. Deng, Q. Li, D. Li, and E. Zhang, "Application of data mining in risk management of construction projects," *Appl. Mech. Mater.*, vol. 513, pp. 2165–2169, Dec. 2004.

[17] J. Villanueva, V. Rodríguez, F. Ortega, and C. González, "Data mining applied to the improvement of project management," Citeseer, Rijeka, Croatia, Aug. 2012. doi: 10.5772/46830.

[18] J. C. Rojas, K. A. Carey, D. P. Edelson, L. R. Venable, M. D. Howell, and M. M. Churpek, "Predicting intensive care unit readmission with machine learning using electronic health record data," *Ann. Amer. Thoracic Soc.*, vol. 15, no. 7, pp. 846–853, Jul. 2018.

[19] M. A. Ahmed and T. M. Alkhamis, "Simulation optimization for an emergency department healthcare unit in Kuwait," *Eur. J. Oper. Res.*, vol. 198, no. 3, pp. 936–942, Nov. 2009.

[20] R. Ceglowski, L. Churilov, and J. Wasserthiel, "Combining data mining and discrete event simulation for a value-added view of a hospital emergency department," *J. Oper. Res. Soc.*, vol. 58, no. 2, pp. 246–254, Feb. 2007.

[21] S. Prabakaran and S. Mitra, "Survey of analysis of crime detection techniques using data mining and machine learning," *J. Phys. Conf. Ser.*, vol. 1000, Apr. 2018, Art. no. 012046.

[22] P. Yildirim, D. Birant, and T. Alpyildiz, "Data mining and machine learning in textile industry," *Wiley Interdiscipl. Reviews-Data Mining Knowl. Discovery*, vol. 8, no. 1, p. e1228, Feb. 2018.

[23] S. Bandaru, A. Ng, and K. Deb, "Data mining methods for knowledge discovery in multi-objective optimization: Part B—New developments and applications," *Expert Syst. Appl.*, vol. 70, pp. 119–138, Mar. 2017.

[24] I. Pereira, T. Sousa, I. Praça, A. Freitas, T. Pinto, Z. Vale, and H. Morais, "Data extraction tool to analyse, transform and store real data from electricity markets," in *Distributed Computing and Artificial Intelligence*. Berlin, Germany: Springer, 2014, pp. 387–395.

[25] K. Yang, Q. Yu, S. Leng, B. Fan, and F. Wu, "Data and energy integrated communication networks for wireless big data," *IEEE Access*, vol. 4, pp. 713–723, 2016.

[26] J. Feijoo, J. Álvarez, J. López, J. Romera, and M. Cárdenes, "Design of a physical and logical secure ip telecommunication network architecture in the spanish TSO red eléctrica de españa," in *Conseil International Grands Reseaux Electric*. Paris, France: CIGRE, Aug. 2010.

[27] P. Genevès, T. Calmant, N. Layaida, M. Lepelley, S. Artemova, and J.-L. Bosson, "Scalable machine learning for predicting at-risk profiles upon hospital admission," *Big Data Res.*, vol. 12, pp. 23–34, Jul. 2018.

[28] W. Alves, D. Martins, U. Bezerra, and A. Klautau, "A hybrid approach for big data outlier detection from electric power SCADA system," *IEEE Latin Amer. Trans.*, vol. 15, no. 1, pp. 57–64, Jan. 2017.

[29] G. Krishnam, "Marketers, big data and intuition—Implications for strategy and decision-making," *Acad. Manage. Global Proc.*, p. 101, Jan. 2018.

[30] K. Grishikashvili and Giacomo Carli, "Big data and organizational capabilities: A grounded study in a pharmaceutical company," *Acad. Manage. Global Proc.*, p. 151, Jun. 2018.

[31] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[32] C. Grinstead and J. Snell, *Introduction to Probability*. Providence, RI, USA: American Mathematical Society, 2012.

MARGARITA RODRÍGUEZ-IBÁÑEZ received the Ph.D. in communication and marketing, in 2008, from the Rey Juan Carlos University of Madrid, Spain. Since 2008, she has been a Full Professor with the Department of Business in different Spanish universities with the main teaching areas of Marketing and Communication. She is currently a Postdoctoral Researcher with Biomedical Engineering and Data Science Group, Rey Juan Carlos University. Her current research interests include new technologies and its impact on organizations and society, data science, marketing, and machine learning.

SERGIO MUÑOZ-ROMERO received the B.Sc. degree in telecommunication engineering and the Ph.D. degree in machine learning from the Universidad Carlos III de Madrid. He has led pioneering projects, where the machine learning knowledge was successfully used to solve real Big Data problems. He is currently a Researcher with the Universidad Rey Juan Carlos. Since 2015, he has been the Head of Data Science and Big Data with Persei vivarium. His current research interests include machine learning algorithms and statistical learning theory, and their applications to Big Data.

CRISTINA SOGUERO-RUIZ got the Ph.D. degree in machine learning with applications in healthcare, in 2015, with the Joint Doctoral Program in Multimedia and Communications in conjunction with University Rey Juan Carlos and University Carlos III. She was supported by FPU Spanish Research and Teaching Fellowship (granted in 2012). She won the Orange Foundation Best Ph.D. Thesis Award by the Spanish Official College of Telecommunication Engineering. She has published several papers in JCR journals and international conference communications. She has participated in several research projects (with public and private fundings) related to healthcare data-driven machine learning systems. Her current research interests include machine learning, data science, and statistical learning theory.

FRANCISCO-JAVIER GIMENO-BLANES received the Telecommunications Engineer degree, in 1995, Diploma of Advanced Studies in taxes and business administration and the Ph.D. in communications technologies. He is currently National Vice President of the Spanish Red Cross, President of the Red Cross with the Valencian Community and University Professor with the Miguel Hernández University. With 25 years of professional experience, Javier Gimeno has devoted almost half of it in industry, both in Spain and internationally, holding various management positions as the Director of Strategic Planning for a major cable television operator, a Manager for Strategic Planning with Telefónica DataCorp, a member of the Telefónica's Chairman's Office, and a Corporate Development Manager with the Fuertes Group. In the academic field, he held management positions as Deputy Vice Chancellor and Deputy Director with the School of Engineering. He has been over 13 years of experience as a Teacher and a Researcher with the university.

JOSÉ LUIS ROJO-ÁLVAREZ received the B.Sc. and the Ph.D. degrees in telecommunication engineering from the University of Vigo and University Politécnica de Madrid, in 1996 and 2000, respectively. He is a Professor with the Department of Signal Theory and Communications in University Rey Carlos (Spain). His research interests include statistical learning methods for signal and image processing, arrhythmia mechanisms, robust signal processing methods for cardiac repolarization, and Doppler image post-processing. He has coauthored more than 120 international papers and has contributed to more than 160 conference proceedings.