RESEARCH ARTICLE

Clustering of recreational divers by their health conditions in a database of a citizen science project

Tamer Ozyiğit [1], Cuneyt Yavuz [1], Salih Murat Egi [1], Massimo Pieri [2], Costantino Balestra [2], Alessandro Marroni [2]

[1] Galatasaray University Computer Engineering Department, Besiktas, Istanbul, Turkey

[2] DAN Europe Research Division, Roseto, Italy

CORRESPONDING AUTHOR: Tamer Ozyiğit, Ph.D. – *tozyigit@gsu.edu.tr*

ABSTRACT

Divers Alert Network Europe has created a database with a large amount of dive-related data that has been collected since 1993 within the scope of the Diving Safety Laboratory citizen science project. The main objectives of this study are the grouping divers by their health information and revealing significant differences in diving parameters using data mining techniques. Due to the methodology of the project, data cleaning was performed before applying clustering methods in order to eliminate potential mistakes resulting from inaccuracies and missing information. Despite the fact that 63% of the data were lost during the cleaning phase, the remaining 1,169 "clean" diver data enabled meaningful clustering using the "two-step" method. Experienced male divers without any health problems are in Cluster 1. Male and female divers with health problems and high rates of cigarette smoking are in Cluster 2; healthy, overweight divers are in Cluster 3. There are significant differences in terms of dive parameters including pre- and post-dive conditions with respect to each group, such as: exercise level, alcohol consumption, thermal comfort, equipment malfunctions, and maximum depth. The study proves the usefulness of citizen science projects, while data collection methodologies can be improved to decrease potential mistakes resulting from inconsistencies, inaccuracies and missing information. It is hypothesized that if naturally occurring clusters of divers were identified it might be possible to identify risk factors arising from different clusters while merging the database with other dive accident databases in the future.

KEYWORDS: data mining; cluster analysis; scuba diving

1. INTRODUCTION

Citizen science (CS) is becoming a powerful tool in bringing together scientists, decision-makers and the general public. Recent reviews show that the last decades have seen a tremendous increase in CS projects and participation, thanks to technological advances among other elements such as increased environmental concerns. This has helped to make CS outputs reliable contributions to science [1-5]. In spite of a growing number of marine CS initiatives globally, however, the role of CS is much better established in terrestrial environments than in marine contexts [4,5].

Divers Alert Network (DAN) has been collecting dive data since 1993 via an international effort. The aim of this effort is to create a database consisting of a reliable and large number of parameters in order to support dive-related scientific research. Collected by volunteer divers, this database is the first example of dive-related research applications that has spread to the general dive community and is an important milestone in citizen science projects. The database was created to be used for:

1. in-depth epidemiological analysis focusing on habits and risks of the diving community; and
2. investigating additional risk factors correlated with the development of circulating bubbles and decompression sickness.

In 2002, DAN Europe further developed the project by collecting data from the divers' own dive computers in a project named Diving Safety Laboratory (DSL).

The large amount of data collected since the beginning of these projects enables researchers to perform statistical analysis related to diving risk factors, gas bubble formation and decompression illness [6]. In addition, the database constitutes a valuable resource in applying data mining methods for better understanding of the dive community, including different types of divers, their diving behaviors and risk factors.

In this study, the DAN Europe database is preprocessed in order to analyze inconsistencies, inaccurate or missing information, and provide solutions to eliminate the loss of data. The remaining dataset is analyzed using data mining techniques in order to investigate the existence of statistically different diver types according to their demographic information, health condition and medical history and to extract useful information regarding their diving parameters.

2. MATERIALS AND METHODS

2.1 Data

A database of 3,108 divers recorded a total of 50,151 dives. The divers' medical and demographic data was collected using the "DSL Participant Enrollment Form," which included diver health condition and medications used (Appendix 1). This form is completed once by each diver, who also provides address, date of birth, gender, year of dive certification and level of certification. Health condition indicators in the form include past and current presence of: allergy, asthma, back pain, back surgery, cigarette smoking, diabetes, ear/sinus problem, ear/sinus surgery, flu and cold, heart and blood pressure problems, joint/muscle pain, nervous system disorder, peripheral vascular disease, pregnancy, previous decompression illness (DCI), pulmonary problems, and seasickness.

need the text for Appendix 1

Data on dive parameters was collected using the "Daily Dive Log Form," filled by divers before and after dives (Appendix 2). This form reported data about the diver's daily conditions (state of rest, medication, alcohol before, exercise before dive), environmental data (diving platform, diving environment, water temperature, visibility, current), purpose of dive, dive planning method (table or computer), equipment, breathing gas and dive profile (surface interval, maximum depth, and dive time, start and end times). After the dive, divers completed the fields of the form related to problems encountered during the dive, such as equipment malfunctions, workload of the dive and thermal comfort. A series of dives performed in a 48-hour interval are termed "Dive Event."

need the text for Appendix 2

Additionally, the database included digital data collected from dive computers. For this purpose, a specific file format called DAN DL7 was developed in order to allow transfer from personal and dive data from different dive computers in a standard form [7].

2.2 Preprocessing

The data was collected by volunteers and required a cleaning process in order to prevent inconsistencies, as well as inaccurate or missing information. First, divers whose gender was undefined, younger than 7, and older than 70 were excluded from the analysis, as their data is doubtful. In addition to that, divers whose dive activity years (calculated by year of first certification) are greater than 70 years were eliminated.

HOW is their data doubtful? shoulkd we explain or just delete that ?

Height and weight information are used to calculate the body mass index (BMI) of the divers. To obtain accurate BMI data, divers whose height and/or weight information is missing, heavier than 200 kg, taller than 200 cm and shorter than 100 cm were excluded from the analysis as well.

The dive count per dive event was 2.5 dives. The total event count was 20,052 and the average number of dives and events per diver was 16.13 and 6.45, respectively. These statistics show us that some divers in the database participated in multiple dives and dive events during the time the database was first built, since several divers have continued to send profiles over the years.

In order to obtain a correct match between diver and dive-related data, avoid data repetition and eliminate the time effect on variables such as age, dive activity in years, weight and medical history, only the first dives performed by the preprocessed diver data were included in the analysis, so that the number of divers and dives were equalized.

2.3 Data formatting

Originally, DAN Europe stored the DSL data in an MSSQL (Microsoft Structured Query Language) Database. The raw data was transformed into multiple tables arranged by diver, dive, and dive event variables. In the original database, an ID number was assigned to each diver and dive (DiverID and DiveID respectively) as well as each "dive event" (EventID).

There are 13 tables in the database: dive, event, record header, diver identification and demographics, dive header at start, dive profile, dive log details, dive safety report, diver additional anamnesis, additional dive data, additional dive problems and additional dive medications. The database diagram is shown in Figure 1. Because data used for analysis has been separated into these tables, it required aggregation of selected variables into a single table.

Figure 1. Database diagram

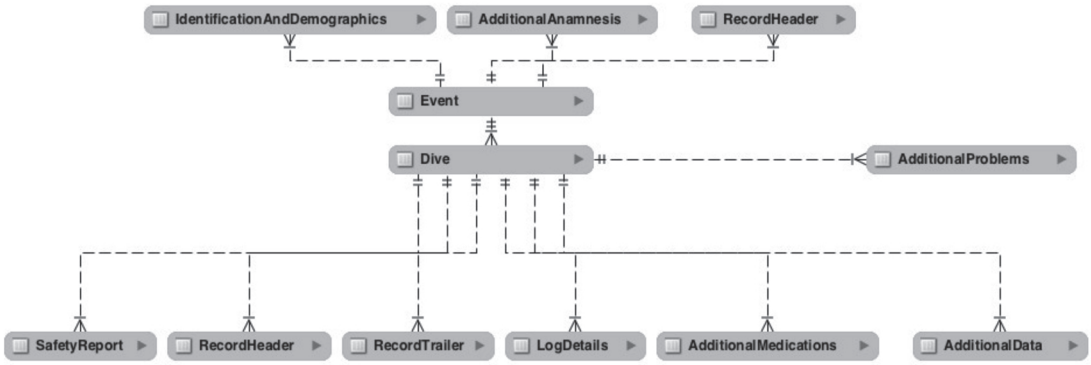


Figure 2. Box plot – age

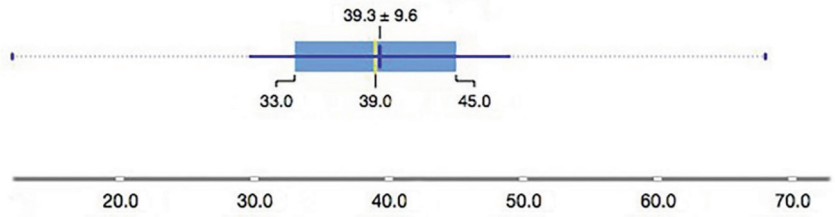
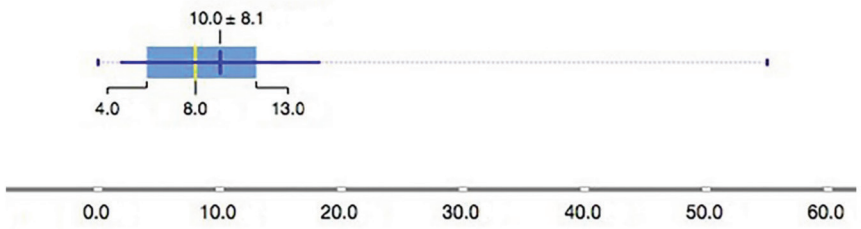


Figure 3. Box plot – dive activity years



In order to keep divers' identities confidential, the relevant data (name, address, email, phone) were removed. Age, dive activity years, height, and weight were numeric variables, while others were binary (0 or 1) variables. Height and weight variables were used to calculate BMI of divers. As two-step clustering – the method we used for grouping the divers – is suitable for categorical data, the numerical variables (age, dive activity years and BMI) were distributed in three categories to be included into analysis by observing the relevant box plots. The box plots for age are shown in Figure 2 and dive activity years are shown in Figure 3.

Observing the box plots above, the optimum intervals of age and dive activity years categories were built by minimum to first quarter, first quarter to third quarter and third quarter to maximum. These categories are shown in Table 1 and Table 2.

Table 1. Age categories

age interval	category
12-30	young
31-45	middle
>45	old

Table 2. Dive activity years categories

dive activity years interval	category
0-5	few
6-13	average
>13	many

Figure 4. Box plot – BMI *(females at top, males at bottom)*

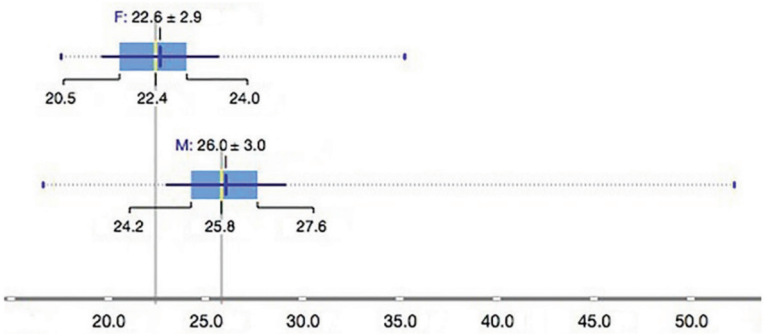


Table 3. BMI categories

interval *(female)*	interval *(male)*	category
0 – 20.58	0 – 24.26	low
20.59 – 24.03	24.27 - 27.68	medium
> 24.03	> 27.68	high

Since weight and height values depend on diver gender, BMI is categorized by gender in Figure 4. BMI categories for male and female divers are shown in Table 3.

2.4 Data processing and analysis tools

The original data tables were transferred to MySQL 5.5, (Oracle Corp. Redwood Shores, California) in order to perform data merging and extract required variables for use in cluster analysis.

For data cleaning, R, a language and environment for statistical computing and graphics developed by John Chambers and colleagues was used [8].

Divers who had similar characteristics were identified and clustered using the two-step analysis using the IBM Statistical Package for the Social Sciences (SPSS, version 20, IBM Corporation, New York).

Two-step clustering is based on forming preliminary groups and reclustering of those groups, then obtaining a cluster tree [9,10]. The advantage of this method is that it can be used with both numerical and categorical data, which makes it suitable for our analysis. The distances between the variable values of the subjects were calculated with log-likelihood distances, and the optimum number of clusters was determined by Schwarz's Bayesian criterion [11].

Chi-square tests on dive-related data on diver clusters and between male and female divers were performed using SPSS as well.

3.1 Results of data cleaning

The citizen science applied by volunteers allows collection of large amounts of data. On the other hand, as the data collection is often not supervised in field, many doubtful, inaccurate and missing data can be found in such databases. As a result of the preprocessing for data cleaning 63% of data was lost: Only 1,169 of 3,108 divers were included into analysis. The filtered divers and exclusion criteria are listed in Table 4.

As seen in Table 4, 54% of diver data is lost in large part due to missing height and weight information. Inclusion of single dives per diver reduced the number of dives from 50,151 to 1,169 thereby inducing a 97% reduction in the number of dives as well.

3.2 Clustering results

The two-step cluster analysis yielded an optimum number of three divers. The clusters with their frequencies in diver-related variables and the chi-square test p-values, which indicate the level of discrimination between clusters, are given in Table 5. The frequencies of distinguishing variables for each cluster are shown in bold.

Table 4. Preprocessed data

exclusion criterion	number excluded
height and/or weight missing	1678
age (<12 and >70)	146
dive activity years (>70)	2
weight (>200)	1
height (>200 and <100)	8
sex (undefined)	4

Based on the results listed in Table 5 we see that the divers are divided into three naturally grouping clusters. Cluster 1 consists of male only, middle-aged divers (mean 39.96) without any health problems.

Table 5. Diver clusters

		-------- CLUSTERS --------			
two-step 3 clusters		**1 count**	**2 count**	**3 count**	**chi-square p-value**
diver sex	male	**464**	309	**185**	<0.001
	female	0	**211**	0	
age category	young	55	118	0	<0.001
	middle	**287**	292	102	
	old	122	110	83	
activity years category	few	0	**259**	35	<0.001
	average	**295**	**177**	**78**	
	many	**169**	84	**72**	
BMI category	low	137	157	0	<0.001
	medium	**327**	252	0	
	high	0	111	**185**	
allergy	present	0	**72**	0	<0.001
asthma	present	0	**11**	0	0.001
back pain	present	0	**16**	1	<0.001
back surgery	present	0	**8**	1	0.021
cigarette smoking	present	0	**130**	0	<0.001
diabetes	present	1	4	0	0.258
sinus problem	present	0	**44**	0	<0.001
heart problem	present	0	**21**	3	<0.001
muscle pain	present	0	**16**	0	<0.001
nervous system disorder	present	1	1	0	0.825
vascular disease	present	0	3	0	0.153
seasickness	present	0	**17**	1	<0.001
peripheral disease	present	0	3	0	0.153
previous DCI	present	2	**9**	1	0.101
pulmonary problems	present	0	**6**	0	0.023
total # divers		464	520	185	

The second cluster is formed by less experienced male and female divers with an average age of 36.94. All female divers are in Cluster 2. A high number of cigarette smokers is one of the particularities of this group. Health problems are seen in higher rates than in the other two clusters. Allergy, ear/sinus problems and seasickness have particularly high rates of occurrence.

Cluster 3 is formed by older experienced male and overweight divers. This group has an average age of 44.40 and average BMI of 29.68, but no health problems were reported.

The chi-square tests show that there are highly significant differences between clusters in most diver-related variables. The exceptions are diabetes, nervous system disorder, vascular disease, peripheral disease and previous DCI, all of which are rarely seen.

The percentages (for categorical variables) and averages (for numerical variables) of dive-related variable for diver clusters and male-female divers are given in Table 6.

3.3 Cluster comparisons

As a general feature of the database, we can say that we sampled a group of divers who were performing relatively safe dive practices regardless of their clusters, with the exception of the use of alcohol: They were well rested before the dive, had a low level of exercise during the dive, and wore good thermal protection.

For the **"alcohol before dive"** category the p-value of the chi-square test for all clusters was 0.072, which showed that there is a weak dependency between clusters and alcohol before the dive. For a deeper analysis of significant differences between groups, pairwise chi-square tests were performed. For Cluster 1 and Cluster 2, the p-value was 0.987. For Cluster 1 and Cluster 3, the p-value was 0.040. The most significant difference was found between Cluster 2 and Cluster 3 with a p-value of 0.043. Compared to other clusters, the older and overweight divers in Cluster 3 had significantly higher rates in the "alcohol before dive" category.

In the **"exercise before dive"** category, the chi-square p-value for all three clusters was 0.002, showing that exercise level before dive differed significantly according to clusters. The most active group before the dive was Cluster 1. The percentage of moderate exercise is high in Cluster 3, a group of healthy but overweight divers. The Cluster 2 group of male and female divers with health problems had the highest percentage of "no exercise." When we performed pairwise comparisons of clusters, the p-value of the chi-square test between Cluster 1 and Cluster 2 was 0.0007; between Cluster 1 and Cluster 3 it was 0.022; and between Cluster 2 and Cluster 3 it was 0.550. The most significant difference was between Cluster 1 and Cluster 2 (healthy versus unhealthy divers).

For **"state of rest before dive,"** the chi-square p-value for all three clusters was 0.012, showing that this variable was firmly dependent on cluster. The p-value of the chi-square test between Clusters 1 and 2 was 0.006; between Clusters 1 and 3 it was 0.23 (which is not significant); and between Clusters 2 and 3 it was 0.196. The biggest difference was between Cluster 1 and Cluster 2. We can say that Cluster 1 differed significantly from Cluster 2 in state of rest before dive. Divers in Cluster 1 stated they much more "rested" than those in Cluster 2.

For **"thermal comfort"** during dive, the chi-square p-value was 0.097, which means thermal comfort during the dive depends weakly on the clusters. The chi-square p-value performed on Cluster 1 and Cluster 2 was 0.149. For Cluster 1 and Cluster 3 the p-value of the chi-square was 0.259, and for Cluster 2 and Cluster 3 it was 0.076. This means that the biggest difference in thermal comfort was between Cluster 2, the group of generally unhealthy male and females, and Cluster 3, a group of older, experienced and overweight divers.

In the **"diving platform"** category, the chi-square p-value was 0.084, demonstrating that different clusters may show a slight tendency to differ in the choice of diving platform. Comparing Clusters 1 and 2, the p-value of the chi-square test was 0.009, showing that these two diver groups were very distinct in their use of diving platform. For Clusters 1 and 3 chi-square p-value was 0.733 (which is not significant). The p-value of the chi-square test between Clusters 2 and 3 was 0.506, indicating independence between these clusters. Divers in Cluster 2 preferred the "small boat" and "charter boat" categories when compared to the other two clusters.

The chi-square p-value for all clusters in the **"workload"** category was 0.113, meaning there were no significant differences between diver groups. The chi-square p-value for Clusters 1 and 2 was 0.147; for Clusters 1 and 3 it was 0.225; and for Clusters 2 and 3 it was 0.213. These results also indicate that there were no differences between any cluster combinations.

In **"breathing gas,"** the p-value of the chi-square test was 0.014, indicating significant differences between clusters. The chi-square p-values of a pairwise test between Cluster 1 and Cluster 2 was 0.001; for Cluster 1 and Cluster

Table 6. Dive-related parameters percentages and averages

dive-related variable		cluster 1	cluster 2	cluster 3
alcohol before dive	yes %	40.52	40.77	49.73
exercise before dive	none %	20.04	31.35	29.73
	light %	69.18	57.88	57.3
	moderate %	9.7	9.42	12.43
	heavy %	1.08	1.35	0.54
state of rest before dive	rested %	93.1	86.92	90.27
	tired %	6.68	12.69	8.65
	exhausted %	0.22	0.38	1.08
min water temp.	°C (mean)	15.8	16.74	16.74
thermal comfort	hot %	0.43	0.77	0.54
	pleasant %	90.73	86.35	92.97
	cold %	8.62	12.12	5.41
	very cold %	0.22	0.77	1.08
platform	beach/shore %	18.53	18.85	18.38
	small boat %	65.73	68.08	66.49
	charter boat %	1.29	3.85	2.7
	live-aboard %	1.51	1.73	1.08
	other %	12.93	7.5	11.35
average max. depth	meters	29.32	27.43	28.82
workload	resting %	40.09	32.31	36.76
	light %	49.78	57.12	55.68
	moderate %	9.05	9.23	5.41
	severe %	0.86	1.15	1.08
	exhausting %	0.22	0.19	1.08
breathing gas	air %	84.48	90.38	87.03
	nitrox %	12.28	8.27	10.81
	trimix %	3.02	0.58	2.16
	other %	0.22	0.77	0
apparatus	scuba open %	89.22	96.54	92.97
	rebreather %	0.86	0.38	0
	other %	9.91	3.08	7.03
equipment malfunction	BC %	0.65	1.35	0.54
	breathing ap. %	0.43	0.96	0.54
	depth gauge %	0	0.19	0
	dive computer %	0	0	0
	face mask %	1.08	2.88	4.32
	fins %	0	0.77	0
	thermal protect. %	0.22	2.31	0
	weight belt %	0	0.96	0
	other %	0.22	0	0.54
any symptom	yes %	10.13	5.77	9.73

3 it was 0.780; and for Cluster 2 and Cluster 3 it was 0.110. Male and female divers in Cluster 2 used air chiefly, while nitrox and trimix practice was higher in other clusters.

For **"apparatus"** we found significant dependency in clusters, with a chi-square p-value less than 0.001. In general, the most common use is open-circuit scuba, whereas healthy male divers in Cluster 1 and older divers in Cluster 3 prefer other types of apparatus to that used by the divers in Cluster 2. The pairwise chi-square tests p-value for Cluster 1 and Cluster 2 was less than 0.001; for Cluster 1 and Cluster 3 it was 0.222; and for Cluster 2 and Cluster 3 it was 0.048. Observing the chi-square p-values we can say that inexperienced divers use scuba, while the use of different types of apparatus increases with dive activity years.

In **"equipment malfunction"** the highest percentages were in Cluster 2, particularly in thermal protection. The chi-square p-value for all clusters was 0.011, which indicated significant differences between clusters. The p-value of the chi-square test between Clusters 1 and 2 was 0.015; between Clusters 1 and 3 it was 0.414; and between Clusters 2 and 3 it was 0.084. The most important difference was between Clusters 1 and 2, the healthy versus unhealthy divers. We can say that Cluster 2 was the most distinct of the three clusters.

In the **"any symptom"** category the Cluster 1 group of healthy and experienced divers had the highest percentage, with a chi-square p-value of 0.030. This indicated high significant dependency of this variable to clusters. The p-value between Clusters 1 and 2 was 0.015; between Clusters 1 and 3 it was 0.993; and between Clusters 2 and 3 it was 0.096.

3.4 Male-female comparisons

All 211 female divers were in Cluster 2, but there were 309 male divers in this cluster as well, making it a mixed cluster. In order to reveal differences between male and female divers, their dive-related data was examined in the same way as diver clusters.

The percentage of **alcohol before dive** for male divers was 44.78%, and for female divers it was 29.86%. The p-value of the chi-square test was less than 0.001. It was evident that male divers consumed more alcohol than females before dive.

None of the female divers performed heavy exercise before dive. The percentage of **"no exercise"** is higher in females, while males have higher percentage of moderate exercise. However, the p-value of the chi-square test was 0.453, indicating no significant difference between males and females in terms of exercise before dive.

For **"state of rest before dive"** the p-value of the chi-square test was 0.085, indicating an insignificant difference between male and female divers in pre-dive state of rest.

For **"diving platform"** the biggest difference was in charter boat platform (1.98% for males and 5.69% for females). The p-value of the chi-square test was 0.036.

There was no significant difference in the **"workload"** category between males and females. Light, severe and exhausting workloads for female divers are higher. The p-value of the chi-square test was 0.165, indicating that workload did not depend on sex.

Breathing gas did not depend significantly on gender, as the p-value of the chi-square test was 0.131. According to Table 6, 95.73% of the female divers used open-circuit scuba, while 6.89% of male divers used other types of apparatus. The p-value of the chi-square test was 0.185, which indicated no significance for dependency of diving apparatus to gender.

Male divers experienced **equipment malfunction** issues in 9.68% of dives, while this rate was 16.67% in females. We can see that the "face mask problems" rate was very high in female divers compared to male divers. The p-value of the chi-square test was 0.002, indicating a significant difference between males and females.

There was no significant difference in symptom occurrence between male and female divers: 7.83% for male divers and 9.48% for female divers. The p-value of the chi-square test was 0.513.

4. DISCUSSION

Science and society – i.e., the general public – have not yet fully embraced CS. This is due to a number of aspects that require better development, from which a number of open issues clearly emerge. These include: timely feedback to participants/support and recognition for participants; validity and contribution to science and policy/decision-making; funding and operational cost management; and legal and ethics aspects [12,13]. Relying on practical support by volunteers, CS initiatives tap into individuals' knowledge, creativity, financial contribution (at least in kind), and hands-on work. Properly designing CS projects and data validation remain challenging processes, as they are lengthy and costly, both in terms of finances and workforce. Where data mining and other quality-check techniques are applied, up to 60%

of CS data may need to be rejected [12,13]. Our study resulted in a similar rejection ratio (65%) but still provided useful scientific information thanks to the amount of data gathered.

We observed that clusters obtained with two-step cluster analysis had significant differences not only for data on divers, but also dive-related data. The resulting groups had meaningful characteristics. One group, Cluster 1, consisted of male divers with no health problems. Cluster 2 consisted of male and female divers who had higher rates than other two clusters in the following characteristics: cigarette smoking, allergy, asthma, back pain, ear/sinus problems and seasickness. We referred to this group as divers who declare health problems. Cluster 3 consisted of older and overweight male divers with many dive activity years.

In dive-related data, the group of unhealthy divers in Cluster 2 differed from others in exercise before dive (highest rate of no exercise), state of rest before dive (highest rate of tired), equipment malfunctions (especially thermal protection problems). The divers in this cluster used mostly scuba for diving apparatus and air for their breathing gas.

The group of older, experienced divers in Cluster 1 differed from other groups in the "alcohol before dive" category (highest rate). Another risk factor for this group was a significantly higher BMI than the other diver groups.

The most important difference in the group of middle-aged male divers in Cluster 1 was their significantly higher rate in the "any symptom" category during and after diving. Although they can be considered as a healthy diver group, the symptom occurrence rate is critical for this group. We can say that the divers in Cluster 1 carry risk because they had the highest rate in the "any symptom" category. Use of apparatus other than scuba was more popular in this group, and they made slightly deeper dives than the other two groups.

The most important risk for Cluster 2 is that they have more health problems and a higher rate of cigarette smoking than the other groups. They also experienced equipment malfunctions more than other divers.

For Cluster 3, the biggest risks were age, being overweight and consuming alcohol before dives. Although these were mostly experienced divers, these factors increase the risk of decompression sickness.

We also observed the percentages and averages of dive-related data by gender. Male and female divers were significantly different in consumption of alcohol before dive, minimum water temperature, and number of equipment malfunctions. Male divers consumed much more alcohol; on the other hand, female divers had more equipment problems.

As our database does not include dive injuries, our study aimed to explore the differences between diver types and their behaviors, but not the consequences of these behaviors as they relate to dive injuries. Beckett and Kordick conducted such a study, which investigated the effects of dive certification, pre-existing medical conditions, diving frequency, alcohol, smoking and use of illicit substances on dive injuries [14]. They did not find any positive association between these variables and dive injuries. The exception was dive certification, which significantly decreases the risk of injury.

LIMITATIONS

Regarding the structure of the database, the biggest problem for analysis was the fact that there were more than one dive and event records for a diver under different dates, and diver data was collected for each dive. This caused repetition in the database. Additionally, as the database has been built over years, the values of some variables have changed over time. For example, a diver may have participated in an event at the age of 30 and in another one at the age of 35. Considering that diver age is also an important variable for clustering, it is necessary to eliminate time effect from variable values. As a robust solution to this problem, only the first dives of divers were included in analysis, which caused loss of a large amount of data. Further analysis can be carried out in observing the data of the "aging" divers who continue to participate in multiple events over the lifetime of the database.

In order to improve the effectiveness of the database, critical data such as weight and height should be collected more carefully for all divers, as these are essential information in assessing diver fitness. It would be beneficial to convert the forms given in Appendix 1 and 2 to a digital platform such as a mobile application, where it would be possible to warn the users about incomplete and inaccurate information. Furthermore, the current database consists mostly of data about health information. For much more comprehensive and effective data mining studies, collecting information on diver socioeconomic status is also important (e.g., marital status, having their own diving equipment, diving outside country).

Additionally, having information about nationality and country of residence of divers would allow comparison of differences between countries

Diving experience of divers is essential data for this study. The variable "dive activity years" indicates how long the diver has been diving but does not provide accurate information about total number of dives. Therefore, variables expressing the number of dives a given diver has experienced in his/her lifetime should be included in the database. Likewise, certificate degree information and diving courses taken by divers also proivde information about diver experience. It is advised to include these variables in future databases structure as parameters.

Physiological data about pre-dive and post-dive status is highly important in terms of dive health research. To be able to carry out detailed studies on the relationship between diver type, dive profile and bubble measurement, collecting these data points should be considered crucial. The paper published by Cialoni, et al. is a recent example of such studies [6]. Additionally, more detailed physiological statistics like body water loss measurements and vascular measurements can be obtained, although they require devices that may be difficult to keep at dive sites.

The characteristics of diver clusters can be compared with similar statistics in the general population – e.g., to analyze whether divers differentiate from the non-diver population in terms of variables like health, smoking, and drug and alcohol use.

Another future research direction is investigating the relationship between diver profile, dive profile, and bubble measurement using multivariate statistic tools such as regression analysis (linear, logistic and polynomial) [15], association rules [16] and decision trees. By doing this, risky diver groups, risky dive profiles and the possible results of combinations of divers, dives and dive profiles can be identified.

■

Ethics statement

All experimental procedures were conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013) and were approved by the Academic Ethical Committee of Brussels (B200-2009-039). All methods and potential risks were explained in detail to the participants. All personal data was handled according to the Italian Law on privacy. Written informed consent was obtained from all the participants.

Acknowledgments

The earlier phase of this study was presented in the 16th Industrial Conference on Data Mining, ICDM 2016, New York, New York, on 13-17 July 2016. The conference paper reported the results of analysis for 874 divers who had a single dive recorded in the database.

Funding

This study is supported by the Green Bubbles Project, which has received funding from the European Union's Horizon 2020 research and Innovation program under the Marie Sklodowska-Curie grant agreement No 643712 and Galatasaray University, Scientific Research Project Commission - Project No. 15.401.001.

Author contributions

TO: Drafting the work, data preparation, method selection, analysis, interpretation of results, discussion, performing final revisions.
SME: Data preparation, method selection, interpretation of results, discussion, revision of text, final approval.
CY: Drafting the work, data preparation and cleaning, performing analysis, writing the draft manuscript.
MP: Drafting the work, providing and cleaning data, interpretation of results, discussion.
AM: Providing data, interpretation of results, discussion, revising of text, approval.
CB: Revising of text, approval, interpretation of results, discussion.

REFERENCES

1. Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J. Citizen science: a developing tool for expanding science knowledge and scientific literacy. BioScience. 2009; 59(11): 977-984.

2. Dickinson JL, Shirk J, Bonter D, Bonney R, Crain RL, Martin J, Phillips T, Purcell K 2012. The current state of citizen science as a tool for ecological research and public engagement. Front Ecol Environ. 2012; 10(6): 291-297.

3. Silvertown J. A new dawn for Citizen science. Trends Ecol Evol. 2009; 24(9): 467-471.

4. Tulloch AI, Possingham HP, Joseph LN, Szabo J, Martin TG. Realising the full potential of citizen science monitoring programs. Biol Cons. 2013; 165:128-138.

5. Thiel M, Penna-Díaz MA, Luna-Jorquera G, Salas S, Sellanes J, Stotz W. Citizen scientists and marine research: volunteer participants, their contributions, and projection for the future. Oceanogr Mar Biol: An Annual Review. 2014; 52: 257-314.

6. Cialoni D, Pieri M, Balestra C, Marroni A. Dive risk factors, gas bubble formation, and decompression illness in recreational scuba diving: analysis of DAN Europe DSL data base. Front Psychol. 2017; 8: 1587. doi: 10.3389/fpsyg.2017.01587.

7. Denoble PJ. Divers Alert Network Project Dive Exploration DL7 Standard. Durham, NC, 2006.

8. Chambers J. Software for data analysis: programming with R. ISBN: 0387759360, 9780387759364, Springer, New York, USA, 2008.

9. Chiu T, Fang D, Chen J, Wang Y, Jeris C. A robust and scalable clustering algorithm for mixed type of attributes in large database environment. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 26-29; Francisco, CA, USA: ACM. 2001; p. 263-268.

10. Chang HL, Yeh TH. motorcyclist accident involvement by age, gender and risky behaviors in Taipei Taiwan. Transpor Res Part F. 2007; 10: 109-122.

11. Schwarz, G. Estimating the dimension of a model. Ann. Statist. 1978; 6(2): 461-464.

12. Newman G, Wiggins A, Crall A, Graham E, Newman S, Crowston K. The future of citizen science: emerging technologies and shifting paradigms. Front Ecol Environ. 2012; 298-304.

13. Vayena E, Tasioulas J. 'We the scientists': a human right to citizen science. Philos Tech. 2015; 28: 479-485.

14. Beckett A, Kordick MR. Risk factors for dive injury: a survey study. Res Sports Med. 2007; 15(3): 201-211.

15. Draper NR, Smith H. Applied Regression Analysis 3rd Edition. John Wiley & Sons, 2014, Canada.

16. Adamo JM. Data mining for association rules and sequential patterns: sequential and parallel algorithms. Springer Science & Business Media, USA, 2012.

17. Rokach L, Maimon O. Data mining with decision trees: theory and applications 2nd Edition. World Scientific, Singapore, 2014.

✦