Computer Science and Information Systems 19(3):1093–1114 https://doi.org/10.2298/CSIS210830030P

An Innovative Quality Lane Change Evaluation Scheme based on Reliable Crowd-ratings

Konstantinos Psaraftis, Theodoros Anagnostopoulos, and Klimis Ntalianis

Department of Business Administration, Division of Information Systems and Decision Making
University of West Attica,
250 Thivon & P. Ralli, Egaleo 12241, Greece
kostaspsaraftis@hotmail.com
theodoros.anagnostopoulos@uniwa.gr
kntal@teiath.gr

Abstract. Intelligent Transportation Systems (ITSs) and their applications are attracting significant attention in research and industry. ITSs make use of various sensing and communication technologies to assist transportation authorities and vehicle drivers in making informative decisions and provide leisure and safe driving experience. Data collection and dispersion are of utmost importance for the proper operation of ITSs applications. Numerous standards, architectures and communication protocols have been anticipated for ITSs applications. In recent years, crowdsourcing methods have shown to provide important benefits to ITSs, where ubiquitous citizens, acting as mobile human sensors, help respond to signals and providing real-time information. In this paper, the problem of mitigating crowdsourced data bias and malicious activity is addressed, when no auxiliary information is available at the individual level, as a prerequisite for achieving better quality data output. To achieve this goal, an innovative algorithm is designed and tested on a crowdsourcing database of lane change evaluations. A three-month crowdsourcing campaign is performed with 70 participants, resulting in a large number of lane changes evaluations. The proposed algorithm can negate the noisy ground-truth of crowdsourced data and improve the overall quality.

Keywords: crowdsourcing, intelligent transportation systems, subjective ratings, lane change evaluation, bias reduction, malicious activities, fuzzy logic.

1. Introduction

Road accidents constitute a major social problem in modern societies. Approximately 1.35 million people die every year on the roads worldwide, and another 20 to 50 million sustain non-fatal injuries as a result of road traffic accidents [1]. It is estimated that lane-change crashes account for 4 to 10 % of all crashes. These injuries and fatalities have an immeasurable impact on the families affected, whose lives are often changed irrevocably by these tragedies, and on the communities in which these people lived and worked [2]. In the meantime, careful and lawful drivers are not rewarded for their responsible driving. According to [3] in the Belonitor project in Denmark by rewarding the participants' good driving behaviour, the percentage of kilometres they travelled within the speed limit increased from 68% to 86%, and the number of kilometres driven a safe distance from the car in front rose from 58% to 77%. However, as soon as the feedback and reward system ended, most drivers returned to their old habits. Most participants acknowledged that the combination of feedback and reward had a strong positive effect. For driving evaluation reports in general, there are three main sources of data that researchers are utilizing. The first source of data collection is by sensors [4] during the act of driving vehicles. The second main category is from the CAN bus data [5], which basically records everything on the running state of the car, such as car speed, steering wheel, gps location, brake and other. The third category includes video data [6], which often derive from a mobile phone or a dashcam. To this end, the first research question is addressed: Is it possible to perform a reliable driver's evaluation out of video data by utilizing crowdsourcing solutions?

Crowdsourcing, in general terms, is the act of taking a job traditionally performed by a designated employee and outsourcing it to an undefined, generally large group of people (a "Crowd") in the form of an open call [7]. Technically speaking, Crowdsourcing is a distributed problem-solving and production model. In such a model, initially, the problems are formulated in a format that can be understood easily by technical and non-technical people. This model is used in many applications. For instance [8], proposed an approach to develop a crowd sensing framework to allow an easier cooperation between the citizens and the authorities by collecting information on crimes and suspects through an e-participatory infrastructure. Specifically, in transportation it has emerged as a novel mechanism for accomplishing temporal and spatial critical tasks with the collective intelligence of individuals and organizations.

In ITSs, Xiao Wang [9] performed a quantitative analysis of related research topics and categorized seven kinds of main crowdsourcing based ITSs services: Crowdsourced geospatial data collection, which is contributed by non-expert end-users for altruistic reasons, which both fully utilize end-users' significant local expertise and provide better data and temporal coverage. Urban traffic planning and management, which focuses on bus arrival time prediction, common trajectory pattern identification, shortest-path computing, optimal route planning, customized deployment of cycle length and signal transition time of traffic light systems, travel information recommendation, etc. Green transportation, which aims to reduce fuel consumption and carbon emission, and to provide a highly efficient trip mode for both public and private transports using crowdsourcing based mobile applications. Social navigation, which leverages public online information with users' social network resources, providing real time exploration in novel and strange environments. Road condition monitoring and assessment, which enables people to effectively take part in solving time-spatial critical traffic tasks without generating the additional financial burden on the transportation agencies with the help of social media sites like Facebook, Twitter, YouTube, and Flickr. Smart parking, which is a long-standing problem in ITSs, because searching for street parking and navigating to it in a crowded urban area impose great societal and environmental challenges. Traffic network construction and communication, where employees' ubiquitous roadside units and vehicular ad hoc networks integrate the capabilities of new generation wireless networks and provide infrastructural support of the inter-vehicle, vehicle-to-roadside and inter-roadside communications in hybrid vehicular ad hoc networks.

In this research, a crowdsourced geospatial data collection solution for peer-to-peer lane change vehicle evaluations is proposed. To this end, the second research question is addressed: How can workers' bias and malicious responses be reliably mitigated? It is proved that data acquired from tasks that comprise a subjective component (e.g. ratings, opinion detection, sentiment analysis) is potentially affected by the inherent bias of crowd workers who contribute to the tasks [10]. In addition, workers may not take tasks seriously. Gadiraju [11] in his research analysed the malicious behaviour in the crowd and defined five categories of untrustworthy workers. Ineligible Workers (IW), who do not comply to the prior stated pre-requisites, e.g. 'Please attempt this task only after you successfully complete 3 tasks. Fast Deceivers (FD), who give random answers in order to finish a task as fast as possible, e.g., entering random numerical values. Rule Breakers (RB), who do not provide the required quality of the answer, e.g., giving 1 keyword, when the task requires at least 3 keywords. Smart Deceivers (SD), who conform to the rules but give semantically wrong answers. Finally, Gold Standard Preys (GSP), who follow instructions and provide valid responses but are caught with providing different answers on repeated test questions during the evaluation. In the context of crowdsourcing, subjective tasks with numerical responses, three of the mentioned categories are applied, namely FD, SD and GSP. Consequently, an algorithm to negate this effect is proposed as a worker might provide biased and/or malicious feedback.

The rest of the paper is structured as follows: Section 2 contains related work. In Section 3, the proposed crowdsourcing framework is described and explained. Section 4 describes the 3-month crowdsourcing campaign conducted and discusses the experimental results. Last, section 5 concludes this paper by pointing out some future research directions.

The three major contributions of this paper are summarized below:

- An innovative algorithm is designed to negate the effect of bias and malicious activities with regards to subjective crowdsourcing environments.
- A crowdsourcing peer-to-peer evaluation framework is proposed, which mainly focuses on lane-change driving acts.
- A large-scale crowdsourcing campaign is carried out with regards to lane-change evaluations and results demonstrate the effectiveness and overall data quality improvement.

2. Literature Review

One of the main and most challenging issues that still exist in crowdsourcing applications, especially in subjective studies where no ground truth exists, is ensuring the reliability of workers' ratings. For that purpose, in case of crowdsourced datasets that record auxiliary information from participants (such as gender, age, income or education level), the work in [12] proposed to apply quasi-randomization techniques in which pseudo-inclusion probabilities are estimated based on covariates available for samples and non-sample units. In other approaches such as in [13, 14], in order to reduce sample bias and adjust the non-probability samples to the target population distributions, pseudo-sampling weights are estimated that are predictive of the outcome of interest and/or the probability of selection. Wang et al. [15] propose a multilevel regression and post-stratification (MRP) method, which is an extension of the hierarchical regression modelling. The work in [16] focuses on debiasing crowdsourcing answers to estimate the average innate opinion of the social crowd with a small number of samples and depends on the social dependency among workers. Other common techniques used to correct worker bias are Bayesian Additive Regression Trees (BART), Inverse Probability Bootstrapping [17], the Least Absolute Shrinkage and Selection Operator, LASSO [18] and the Propensity Score Adjustment [19]. However, these approaches record large samples of highly relevant variables. In this research, minimal information is available at the individual level because it is common that workers in volunteered geographic information applications do not provide auxiliary individual information apart from the measure of interest and the geographical information.

Researchers have also studied worker bias estimation techniques in crowdsourcing platforms. The works in [20, 21] study the data annotation bias, when data items are presented as batches to be judged by workers simultaneously and propose models to characterize the annotating behavior on data batches. However, they focus on binary answers and their goal is to properly categorize each data item instead of estimating worker bias and malicious activity. In [21], the authors show that crowdsourcing workers have both bias and variance and propose an approach to recover the true quantity values for crowdsourcing tasks with an unsupervised probabilistic model to jointly assess task difficulties. In [22] the proposed scheme aims to solve the above problem by building and using probabilistic graphical models for jointly modeling task features, workers' biases, worker contributions and ground truth answers of tasks, so that task-dependent bias can be corrected. In order to achieve effective models, the aforementioned approaches need a large number of worker responses and additionally, they do not consider workers’ malicious activities at all. On the contrary, the proposed algorithm performs well and improves data output even when the number of worker responses is low. In addition, it detects and negates the effect of malicious activities.

Quality control in the data output is also approached with task assignment techniques. For example, [23] investigates the accuracy of workers by evaluating their performance on the completed tasks and predicts which tasks the workers are well acquainted with; [24] propose a framework comprising an inference model and an online task assigner. They prove that inserting a gold standard question helps estimate the worker accuracy and supports blocking of poor workers; [25] estimate the workers’ accuracy according to their previous performance and the core quality-sensitive model is able to control the processing latency; [26] developed a quality-sensitive answering model, which guides the crowdsourcing engine to process and monitor the human tasks. The model achieves reliable results by providing an estimated accuracy for each generated result based on the human workers' historical performances. Different from the previous approaches, in this research, only non-auxiliary subjective tasks are considered. Moreover, while these works focus on selecting the reliable workers to perform the tasks, they do not consider workers’ bias or malicious activity combined in their responses and they do not focus on estimating quantitative values, such as evaluations. Authors in [27] aim to reliably identify crowdsourced events by selecting a small subset of human sensors to perform tasks. Similarly, to the algorithm proposed in this research, they exploit linear regression to estimate worker bias in each task and attempt to eliminate it, the moment workers provide their ratings. However, in their setting, they assume that, although answers might be subjective, all are considered as truthful. Hence, they do not take malicious activities into consideration at all.

Researchers in the field have acknowledged the importance and need for techniques to deal with inattentive workers, scammers, incompetent and malicious workers. Authors in [11] analyzed the prevalent malicious activity on crowdsourcing platforms and studied the behavior exhibited by trustworthy and untrustworthy workers. Eickhoff et al. [28] aimed to identify measures that one can take in order to make crowdsourced tasks resilient to fraudulent attempts. The authors concluded that understanding worker behavior better is pivotal for reliability metrics. Difallah et al. [29] reviewed existing techniques used to detect malicious workers and spammers and described the limitations of these techniques. In another relevant work by Gadiraju et al. [11], the authors proposed to design and plan micro-tasks such that they are less attractive for cheaters. In order to do so, the authors evaluated factors such as the type of micro-task, the interface used, the composition of the crowd and the size of the micro-task. All previous research, however, does not take into consideration workers' bias in the datasets.

To the best of the authors' knowledge, no study has reported a similar algorithm to evaluate workers' performance and the use of crowdsourcing techniques for lane-change vehicles' evaluation. A novel algorithm is presented to mitigate bias and malicious activity on workers' ratings and improve overall quality of data output.These instructions and the corresponding MS Word document template are based on the corresponding Springer instructions and MS Word document template for preparing camera ready papers to be published in the Springer series Lecture Notes in Computer Science.

The preparation of manuscripts which are to be reproduced by photo-offset requires special care. Papers submitted in a technically unsuitable form will be returned for retyping, or canceled if the volume cannot otherwise be finished on time.

3. Framework Overview

3.1. System overview

A first high-level overview of the proposed architecture is presented in Figure 1. In general, the requester will post tasks to a crowdsourcing platform as an input. These tasks will be given to a pool of workers for evaluation. The workers' submitted responses will be the raw data for the system. Raw data will be processed through the recommended algorithm to negate crowdsourced data bias and malicious activities. Finally, the proposed algorithm will output the data with improved overall quality.

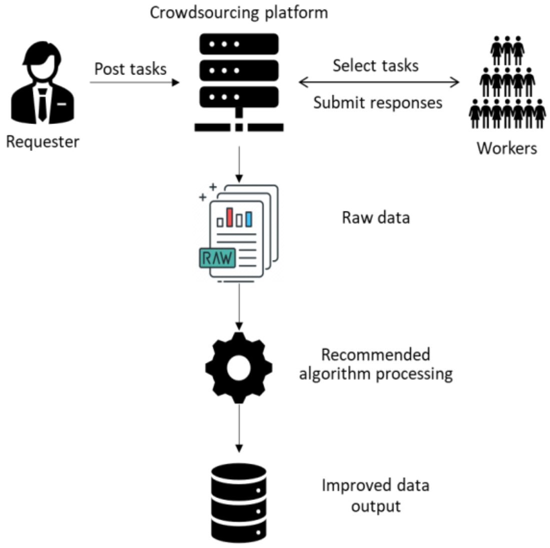


Fig. 1. System Overview

3.2. Problem formulation

Given a set of workers denoted as $w \in W$ that take part in the system and a set of events denoted as $e \in E$, the goal is data acquisition regarding an evaluable event. During this process, it is very important to acquire as many reliable ratings R for as many events as possible. More specifically, the goal is the goal is to populate a database of tuples of the form $T = < w, e, r >$, where *w* is a worker, *e* is the event and *r* is the rating provided by w for an event e. Given that subjective events that are evaluable may contain bias and/or malicious activity, the framework of data acquisition has the following two secondary goals. First goal is that worker w should not be biased. Therefore, bias, denoted as $bias_w$, which indicates the likelihood of providing ratings above or lower than the average rating must be calculated. Second goal is that worker should not be deceitful. The level of maliciousness, if any, is denoted as $cheat_w$ and must be measured. Both goals, calculating $bias_w$ and $cheat_w$ are important for the realization of the primary goal, acquiring reliable ratings for events. In the next section, the framework that realizes the above goals to achieve the primary goal of acquiring reliable ratings for as many evaluable events as possible from a crowd of workers is described.

3.3. Framework

In a nutshell, the framework works as follows. Initially, a crowdsourcing system with a well of events to be rated into E itemset $e_1, \dots, e_n$ is formulated. Each event e is associated with the following attributes: $< id_e, lat_e, long_e >$. Attribute id_e is the event's unique identifier in the framework, and $lat_e, long_e$ corresponds to the current location in terms of latitude, longitude of the event. The total set of possible ratings R is:

$$\begin{Bmatrix} 1\ (Dangerous), \\ 2\ (Needs\ improvement), \\ 3\ (Neither\ good\ nor\ bad), \\ 4\ (Very\ good), \\ 5\ (Excellent) \end{Bmatrix}$$

Ratings are mapped to values between 1 and 5 to stay compatible with the 5-star rating paradigm proposed by [30] and used by most recommender systems. In the system, a worker w has the following attributes: $< id_w, bias_w, cheat_w, prev_w[] >$ where id_w is the worker's unique identifier in the system, $bias_w$ represents bias based on expertise and skills when estimating the evaluation of an event, $cheat_w$ correspond to the worker's maliciousness score, and $prev_w[]$ is used to store information about the evaluations completed by the worker w. The calculated value for the event's rating is denoted as val_e^W and computed based on input from all workers in the set W. The crowdsourcing answer of worker w to the event e is denoted as $a_{w.e}$. In the rest of the section, the proposed algorithm for mitigating bias and malicious activity is described.

3.4. Worker Bias

Similar to the intuition of authors in [27], worker ratings are considered having a bias which is defined as a linear function of their answers (*x*-axis) with respect to the difference of their ratings from the average rating when all workers in *W* are considered (*y*-axis). Each worker $w \in W$, who is requested to evaluate an event, is assumed to provide an answer $a_{w.e}$ with a bias $b(a_{w,e})$ and thus the estimated debiased response is:

$$val_{w.e} = \frac{\sum_{w \in W} a_{w,e}}{|W|} + b(a_{w,e}) \#(1)$$

The bias $b(a_{w,e})$ is defined as a linear function of the worker's response. Thus, it is possible to estimate the difference from the average value for each worker's response $a_{w.e}$. Consequently, linear regression is exploited, to adjust the worker's bias estimation whenever the worker provides a response. Linear regression is a useful tool in many applications to model the relationship between a scalar response and one or more explanatory variables. For this framework, the matrix points are defined from the worker's response $a_{w.e}$ and its difference from the average value of the event. Thus, the response $a_{w.e}$ provided for each event's evaluation and the respective difference from the average rating is recorded and the ratio among these two dimensions is defined. This is computed easily using simple linear regression that produces a linear function:

$$b\left(a_{w,e}\right) = \mu * a_{w.e} + v \ \ \#(2)$$

where μ is the slope and v is the intercept of the line, which are calculated from the linear regression. This way, an estimation for the difference of each worker's rating compared to the average rating from all workers by computing $b\left(a_{w,e}\right)$ for each rating $a_{w.e}$ is made.

3.5. Worker maliciousness

In this section, the quality of each worker is described by estimating the worker's weight in a simplified setting. To estimate each worker's maliciousness, a fuzzy logic controller is utilized.

Fuzzy logic has proven itself as a promising mathematical approach for addressing subjectivity, ambiguity, imprecision, and uncertainty of linguistic expressions [31]. A fuzzy logic inference system may contain many inputs and outputs and allows the implementation of the rules described in a natural language. An explanatory diagram is shown in Figure 2. The input consists of numerical signals and are called 'crisp', which are later translated into the fuzzy sets through the fuzzification process. A fuzzy set is a pair consisting of linguistic variables. Different membership functions are used to perform the fuzzification process. Often, triangular, or trapezoid functions are used to keep the computational cost low. After the transformation into linguistic variables, the inference rules could be applied. After that process, a so-called defuzzification is needed to generate a sharp output value.

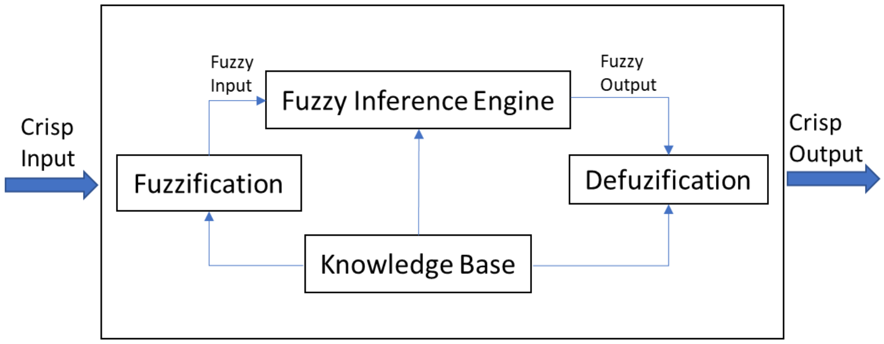


Fig. 2. Fuzzy logic inference diagram

Fuzzy logic controllers are used in many applications. For instance [32], developed a system with the internet of things (IoT) concept for making right decision according to the situation for monitoring and determining fire confidence and reduce the number of rules by doing so sensor activities also reduced and extend battery lifetime as well as improve efficiency whereas [33] developed an efficient and intelligent IoT communication system to ensure data security with network consistency.

The proposed fuzzy logic maliciousness evaluator has a ‘2 input – 1 output’ structure. First input is based on a Euclidean distance similarity measure between the vector distance of the estimated debiased worker’s evaluation to the estimated average debiased evaluation. Second input is the relative normalized number of times the worker’s estimated vote is above the estimated debiased average evaluation of the event. More formally, the fuzzy logic controller will solve the following problem:

Given two sets of numbers between 0 and 1, where 1 is excellent, that respectively represents the Distance Score and the Over-Under Score of a crowdsourcing worker, what maliciousness weight should be assigned?

In the following paragraphs, the calculation process of the controller’s input is described.

For the first input, namely *Distance Score*: let w_i be a random worker, *W* the whole set of workers and *E* events. Also, let $< val_{w_i,e_1},\ val_{w_i,e_2}, \dots , val_{w_i,e_E} >$ be the vector of debiased ratings the random worker provided. Similarly, let $< \frac{\sum_{w \in W} val_{w,e_1}}{|W|}, \frac{\sum_{w \in W} val_{w,e_2}}{|W|}, \dots , \frac{\sum_{w \in W} val_{w,e_E}}{|W|} >$ be the vector of average debiased evaluations. Based on the Euclidean-distance similarity vector similarity, the *Distance Score* is computed as follows:

$$Distance\ score\ (w_i) = \frac{1}{1 + \sqrt{\sum_{j=1}^{E} \left(val_{w_i,j} - \frac{\sum_{w \in W} val_{w,e_j}}{|W|} \right)^2}} \#(3)$$

This result in a value in the range of [0,1]. The higher the score, the better the results for the worker, indicating that no malicious activity is detected. Of course, any other suitable similarity measure can be seamlessly used instead, depending on the context for which the framework is used.

For the second input, namely *Over-Under score*: Once again, w_i is assumed to be a random worker, *W* the whole set of workers and *E* is the total number of events. For each crowdsourcing event, the number of times the worker’s estimated vote is above the estimated debiased average evaluation of the event, $N_{w_i}(voted\ over)$ and the number of times less $N_{w_i}(voted\ under)$ is tracked. Consequently, the *Over-Under score* is calculated as follows:

$$Over - Under\ score\ (w_i) = \frac{\left| N_{w_i}(voted\ over) - N_{w_i}(voted\ under) \right|}{E} \#(4)$$

In a similar manner as with the *Distance Score*, this result in a value in the range of [0,1]. With this score alone, it is not safe to make assumptions about the worker. Combining *Distance* with the *Over-Under score*, fuzzy sets can be created. The set of rules for the fuzzy logic system are shown in Table 1. To train the fuzzy inference system, two datasets of evaluations were formulated. For the first dataset, a local police department of Athens is contacted where three traffic enforcement officers (experts) assisted by performing lane-change event evaluations. Specifically, officers performed

132 lane-change evaluations and for each event, the averaged of their response was calculated to minimize any possible bias. Their evaluations are all considered truthful and, for the framework, the ground truth of these events. For the second dataset, for the same set of events, the average of the evaluations made by the 70 workers from the crowdsourcing campaign was calculated. Finally, based on the two sets of evaluations, the fuzzy logic controller's fuzzy rules and membership functions were manually adjusted by trial and error until the root mean square of the framework was minimal.

Table 1. Fuzzy logic evaluation rule-base

Worker weight		Over-Under score		
		Low	Average	Excellent
Distance score	Low	Low	Below Average	Average
	Average	Below Average	Average	Above Average
	Excellent	Average	Above Average	Excellent

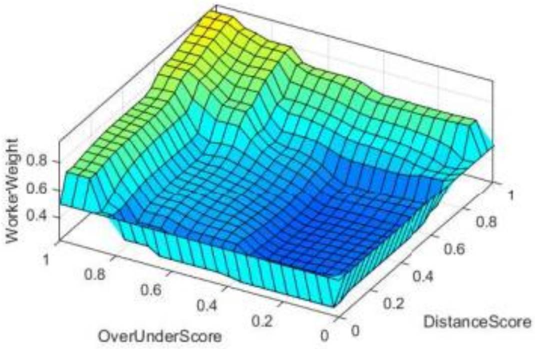


Fig. 3. Worker weight fuzzy evaluator 3D surface

The output variable represents the maliciousness of a worker and is denoted as: $cheat_w \in [0,1]$. The three-dimensional surface of the designed fuzzy logic inference mechanism is displayed in Figure 3 and the corresponding fuzzy logic rules in Figure 4. Variable $cheat_w$ provides an estimation on how malicious a worker may be in the ratings provided. The closest to the upper bound, 1, the better the score for the worker. Hence, to better improve the data quality, the top-K workers are selected with the highest score achieved. The next step after acquiring the fuzzy logic's output and selecting the top-K workers is to assign each worker a weight which is calculated as shown in equation (5) where K is the selected set of workers with the best score.

$$weight_w = \frac{cheat_w}{\sum_{w \in K} cheat_w} \#(5)$$

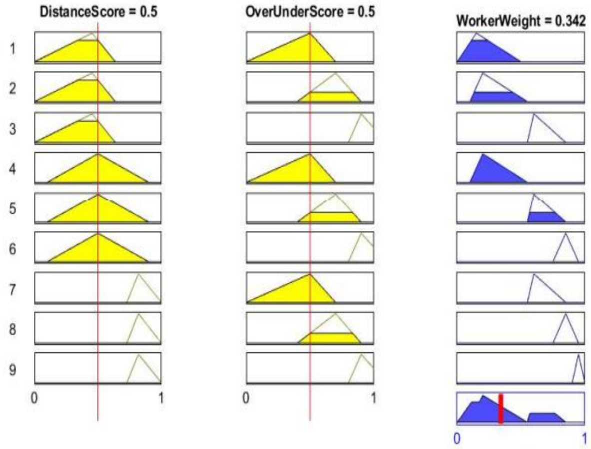


Fig. 4. Worker weight fuzzy evaluator rules

3.6. Event output computation

Finally, the last step is to determine the output of the event. The output of the event is computed using the following equation:

$$val_e^K = \sum\nolimits_{w \in K} \left(\left(a_{w,e} - b\left(a_{w,e}\right) \right) \left(weight_w\right) \right) \#(6)$$

Thus, the average value of the retrieved ratings is computed after selecting the top-K workers with the best $cheat_w$ score and eliminating the estimated bias and maliciousness for each of those individual workers.

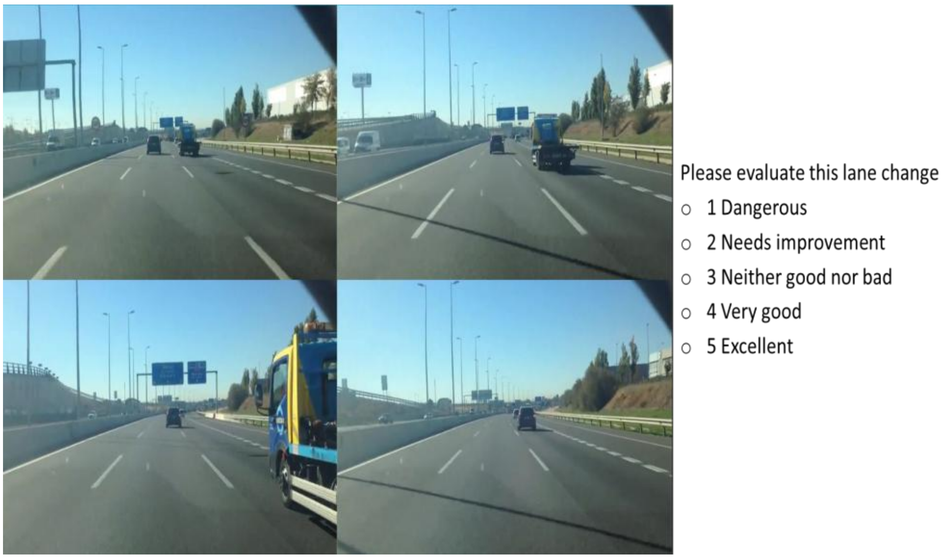
4. Experimental results

With regards to subjective crowdsourcing environments, the proposed algorithm is a promising solution. It combines linear regression techniques to negate the effect of human bias and a fuzzy logic controller with a trained fuzzy inference system to detect the low-skilled workers. In the paragraphs below, the 3-month crowdsourcing campaign conducted is described and the experimental results are discussed.

The proposed algorithm was evaluated by analyzing the results from a 3-month crowdsourcing campaign of 70 workers[1]. Specifically, the data were collected from 7th December 2020 to 19th March 2021. The campaign was performed locally, in the department's research lab in University of West Attica. Participants were all adults with

[1] More information on datasets and applied software can be found here: GitHub

a valid driving license registration. Of them, 29 (41%) were female and 41 (58%) were male, with an age ranging from 19 to 54 years. Most of the participants (72.8%) were University graduates. For the purposes of the experiments, no other personal information is published since in this study, the problem of mitigating crowdsourced data bias and malicious activity when no auxiliary information is available at the individual level is addressed.



Please evaluate this lane change

○ 1 Dangerous

○ 2 Needs improvement

○ 3 Neither good nor bad

○ 4 Very good

○ 5 Excellent

Fig. 5. Assigned lane change event

The first 132 lane change videos were extracted from the UAH-DriveSet [34] which is a public collection of data captured by their driving monitoring application, DriveSafe by various testers in different environments. Thus, workers were given lane changing acts to rate in the form of videos through a 5-point Likert scale. These lane change videos cover national highways, state highways and district roads but no rural or village roads. Additionally, videos cover all traffic volume scenarios, from light to high traffic roads.

Figure 5 shows frames of a lane-change event's video clip with the accompanying task request. A valid worker's participation required all 132 events to be completed, so the total experiment's evaluation dataset contains 9240 lane change evaluations.

In Figure 6, in a clustered bar chart, the total number of answers that were retrieved for each of the ratings for all events is presented. There are 2 noticeable observations made on the chart. Workers' most preferred response was the neutral number 3 (Neither good nor bad) which is a small indication that they were unable to decisively decide whether a lane change is more or less than safe. In addition, the total number of '1' as a feedback, which is the worst evaluation possible, exceeds the total number of '5' which indicates the best response ($1776 > 1714$); these are the video footages, city authorities need to reevaluate.

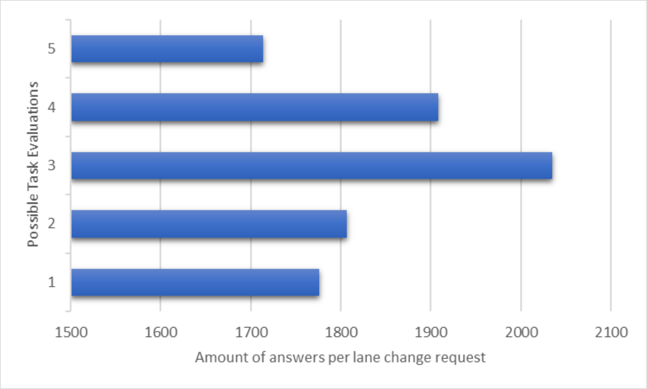


Fig. 6. Number of evaluations for each lane change event rating

In Figure 7, the average evaluations provided by the workers for each individual lane change event were presented. The first 40 lane change events (30.3%) out of 132 in total are classified as '1' (Dangerous) and '2' (Need improvement). Further actions should be planned for these drivers who performed so poorly.

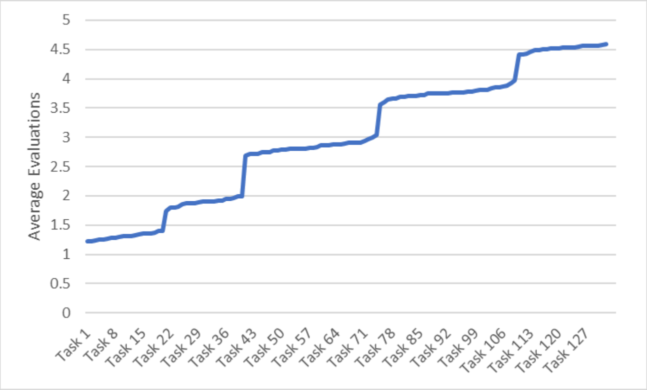


Fig. 7. Average evaluations made per lane change event

In Figure 8, the average evaluation is presented that each individual worker provides for the lane changes in total, whereas in Figure 9 the assigned weight for each worker is illustrated. It appears that in both cases, workers can be classified into two large categories. The first 32 workers (45.7%) appear to provide low responses on average while the rest (54.3%) appear to provide higher responses. In the meantime, in figure 9, the graph results show that the worker weights are clustered into the following mentioned weight classes:

- 34 workers (48.5%) have a weight in the range [0.2205 – 0.2755],
- 5 workers (7.1%) have a weight between [0.4704 – 0.4719],
- 145 workers (20%) between [0.6996-0.7603]

- 17 and the rest of the workers (24.2%) a range [0.8121 - 0.95].

Clustered results arise, due to the fuzzy logic input variables, over-under and distance scores as these are depended on the average debiased score of each task. Therefore, the variables act like a ranking scale and the better the worker's data quality is, the larger the assigned weight will be.

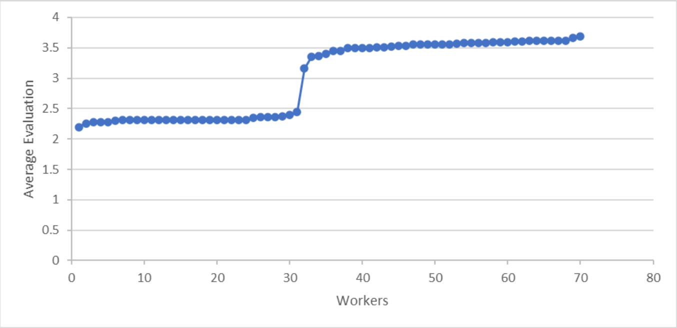


Fig. 8. Average evaluation made per worker (original dataset)

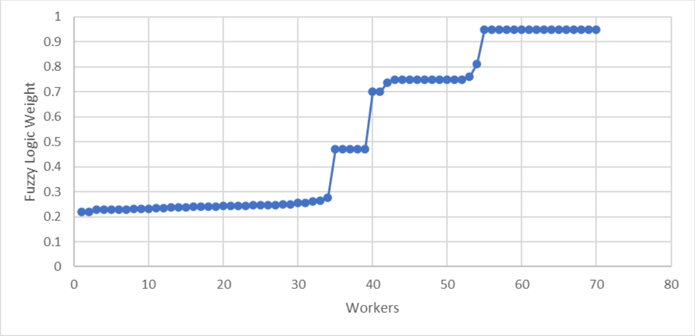


Fig. 9. Average evaluation made per worker

The root mean square error (RMSE) was used to evaluate the accuracy of the recommended algorithm in terms of reliability and effectiveness. Specifically, RMSE is a goodness-of-fit measure of how close the suggested values from different models are, to the initial values. Higher RMSE values indicate poor results, while a smaller RMSE indicates better performance. The relevant formula of RMSE is denoted in equation (7) where n is the sample size and e the difference between each evaluation to the average.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2 \,\#(7)}$$

RMSE is used under three different baseline approaches to better test the proposed algorithm. Specifically:

Random sampling, where the average evaluations from a random set of workers are computed and is denoted by the equation (8) where $a_e^R = \frac{\sum_{w \in R} a_{w,e}}{|R|}$.

$$RMSE\ (Random\ Sampling) = \sqrt{\frac{1}{E}\sum\nolimits_e \left(a_e^R - \frac{\sum_{w \in W} a_{w,e}}{|W|}\right)^2}\ \#(8)$$

Average sampling from the workers who had acquired the best individual $cheat_w$ score, where their average initial evaluation is computed. Specifically, it can be obtained as shown in equation (9), where $a_e^K = \frac{\sum_{w \in K} a_{w,e}}{|K|}$.

$$RMSE\ (Average\ from\ workers\ with\ best\ score) = \sqrt{\frac{1}{E}\sum\nolimits_e \left(a_e^K - \frac{\sum_{w \in W} a_{w,e}}{|W|}\right)^2}\ \#(9)$$

Finally, RMSE calculation based on the recommended algorithm is specified by the following formula:

$$RMSE(Recommended) = \sqrt{\frac{1}{E}\sum\nolimits_e (val_e^K - \frac{\sum_{w \in W} a_{w,e}}{|W|})^2}\ \#(10)$$

For the above equations (8, 9, 10), W represents the complete set of workers in the datasets that have evaluated lane change events e, E represents the total number of events and K the varying sample size.

Primarily, for the first set of experiments, the proposed algorithm was examined in terms of effectiveness and accuracy when the number of malicious workers in a dataset varies. For that purpose, the initial 70-worker dataset is injected with a varying number of malicious workers, denoted as $Malicious(m)$, where m is the total number of malicious workers. Truly random evaluations were submitted for each of their lane change task assignments, to simulate real life malicious workers. Figure 10 presents the RMSE(Recommended) score for the lane change events under various numbers of sample size (5-35) and m malicious workers. In all cases where the sample size varies and the injected malicious workers were in the range $m \in [0 - 50]$, the algorithm manages to perform well and keep RMSE minimal (0.112–0.2108). As variable m increases, so does RMSE(Recommended) with performance ranging in (0.5669-0.6066). Therefore, the proposed model has a limitation when the number of malicious users' ratio is known in advance and exceeds 41.6% of the total worker population.

For the rest of the experiments, two versions of datasets were compared in terms of RMSE. The first dataset is the initial, which contains the evaluations of 70 workers for 132 lane change events and the respective algorithm's data processing results. The second dataset is injected with 15 malicious workers ($\approx 17.6\%$ of the total worker population), with the same approach as in the first set of experiments. The injected malicious workers do not exceed the proposed model's ratio limitation as described before.

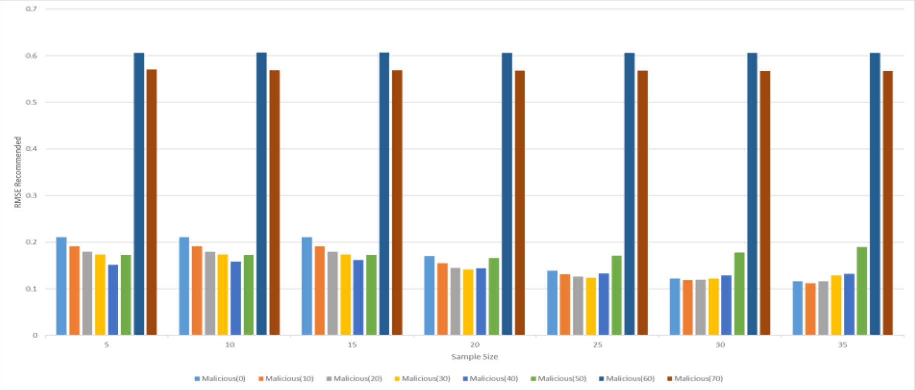


Fig. 10. Varying malicious users sample size evaluation

For both the initial dataset and the dataset with 15 added malicious workers, Figure 11 and Figure 12 respectively, present the RMSE score for the lane change events under various numbers of sample size (5-30). It makes sense that, in all cases, RMSE decreases as the sample size increases in all approaches. In addition, it is also reasonable that RMSE(Average) has high initial scores (0.71 on the initial dataset and 0.74 on the malicious) because biased workers retain better fuzzy logic scores. Furthermore, results from the charts show that the recommended algorithm outperforms all other cases, especially when the sample size is very small. So, by taking 5 workers as a sample, the original dataset results were RMSE(Random) = 0.278, RMSE(Average) = 0.717, RMSE(Recommended) = 0.21, whereas with the malicious dataset, results were even better: RMSE(Random) = 0.394, RMSE(Average) = 0.742, RMSE(Recommended) = 0.179.

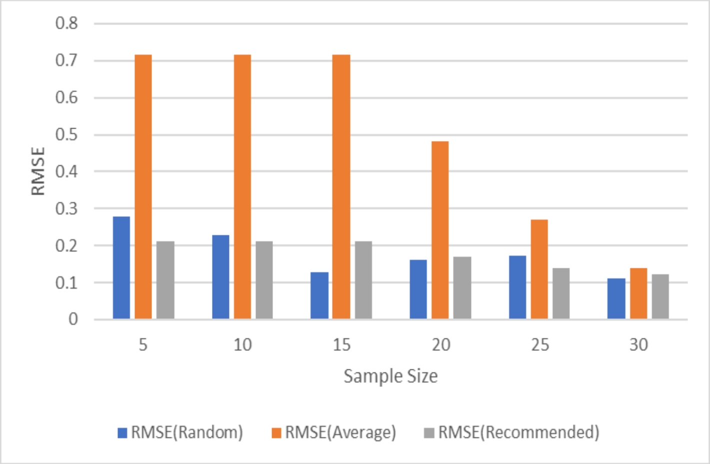


Fig. 11. Varying sample size evaluations (Original dataset)

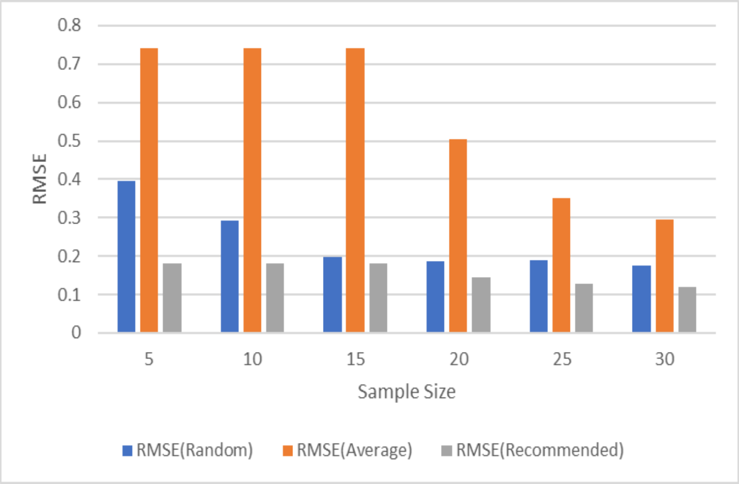


Fig. 12. Varying sample size evaluations (Malicious dataset)

Figures 13 and 14 illustrate how RMSE behaves when the sample size is kept to 15 workers, but the number of total workers varies from 30 to 50 and lastly at 70 workers. For the malicious dataset, for each worker group population, 5 valid workers are replaced with 5 malicious. Thus, the first 30 worker population had 5 malicious workers, the second 10 and the third 15. Both charts show that the recommended algorithm outperforms the two baseline approaches. In addition, although RMSE(Random) and RMSE(Average) were increased in the malicious dataset in comparison to the initial, the proposed algorithm performed similar results. Finally, the proposed algorithm is resilient to malicious workers, since RMSE values range between 0.15 and 0.192.

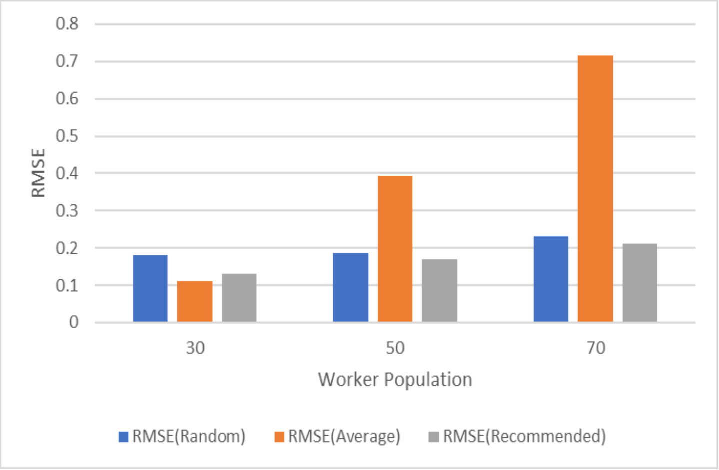


Fig. 13. Varying total workers evaluations (Original dataset)

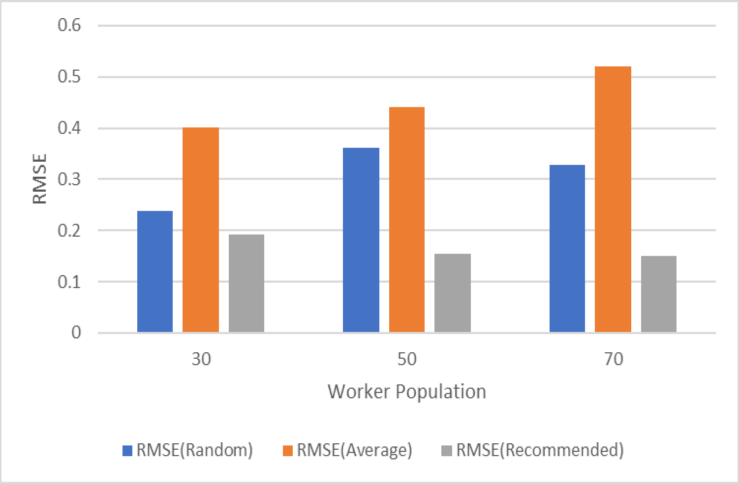


Fig. 14. Varying total workers evaluations (Malicious dataset)

Results prove that the proposed method has many benefits. The inherit human bias is estimated by exploiting linear regression for each task and negates it before the fuzzy logic controller accepts it as an input. Moreover, as a worker may not always be trustworthy, the construction of the fuzzy logic controller proved to be a suitable solution for estimating her quality. The worker's quality estimation is then used to select the best $(top - K)$ for an improved data output quality.

5. Conclusion

Crowdsourcing applications have been proposed in intelligent transportation systems for multiple case studies and undeniably it has been an effective tool in bringing people together to solve a problem that affects their community. However, crowdsourcing, like all systems, has its own set of limitations that must be resolved through proper planning and understanding of the system. Specifically, there are concerns about data quality and data management.

To address these concerns, in this paper, a novel algorithm is proposed to address the problem of mitigating crowdsourced data bias and malicious activity to evaluable subjective events when no auxiliary information is available at the individual level as a prerequisite for achieving better quality data output. Experiments involving a crowdsourcing campaign of 70 workers for three months are conducted to evaluate lane changes. Results reveal that the proposed algorithm outperforms in terms of RMSE all other baseline approaches. It should be noted that the settings for the proposed algorithm are tailored for the lane change event. Different traffic events may require a modified version of the proposed algorithm. In pursuit of a more universal approach, additional experiments with other traffic events should be performed and investigate what modifications may be required.

There are many potential enhancements regarding the current framework. Initially, since data anonymization is of crucial importance nowadays, the information recorded for each crowd worker can be protected using relevant techniques [35]. Additionally, a mobile application for real-time vehicle evaluations could automate the data gathering process. Lastly, the fuzzy logic controller has room for improvements to further improve its accuracy and assign even better worker weights.

Acknowledgment. This paper presents a part of the first author's PhD research. This research is partially supported by the Special Research Account of the University of West Attica.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases. Morgan Kaufmann, Santiago, Chile, 487-499. (1994)
2. 1. World Health OrganizationGlobal Status Report on Road Safety 2018, https://www.who.int/publications/i/item/9789241565684, (2018)
2. Barr L, Najm W (2001): Crash problem characteristics for the intelligent vehicle initiative. Presented at the
3. Hattem J, Mazureck U (2005): Good Driving! The Power of Rewarding. Presented at the November 10
4. Cao W, Lin X, Zhang K, Dong Y, Huang S, Zhang L (2017): Analysis and evaluation of driving behavior recognition based on a 3-axis accelerometer using a random forest approach. In: Proceedings - 2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN 2017, Association for Computing Machinery, Inc, pp. 303–304
5. You CW, Lane ND, Chen F, Wang R, Chen Z, Bao TJ, Montes-de-Oca M, Cheng Y, Lint M, Torresani L, Campbell AT (2013): CarSafe App: Alerting drowsy and distracted drivers using dual cameras on smartphones. In: MobiSys 2013 - Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services, ACM Press, New York, New York, USA, pp. 13–26
6. Ma X, Chau LP, Yap KH (2018): Depth video-based two-stream convolutional neural networks for driver fatigue detection. In: Proceedings of the 2017 International Conference on Orange Technologies, ICOT 2017, Institute of Electrical and Electronics Engineers Inc., pp. 155–158
7. Vander Schee BA: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business20093Jeff Howe. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business . New York, NY: Crown Business 2008. 320 pp. $26.95 . J Consum Mark 26, 305–306 (2009)
8. Hasna EE, Abdelaziz E, Zohra EF, Mohamed S: A mobile crowd sensing framework for suspect investigation: An objectivity analysis and de-identification approach. Comput Sci Inf Syst 17, 253–269 (2020)
9. Wang X, Zheng X, Zhang Q, Wang T, Shen D: Crowdsourcing in ITS: The State of the Work and the Networking. IEEE Trans Intell Transp Syst 17, 1596–1605 (2016)
10. Hube C, Fetahu B, Gadiraju U (2019): Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, New York, NY, USA, pp. 1–12

11. Gadiraju U, Kawase R, Dietze S, Demartini G (2015): Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In: Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, New York, NY, USA, pp. 1631–1640
12. Elliott Michael R., Richard Valliant: Inference for Nonprobability Samples. Stat Sci JSTOR 32, 249–264 (2017)
13. Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile KJ, Tourangeau R: Summary report of the aapor task force on non-probability sampling. J Surv Stat Methodol 1, 90–105 (2013)
14. Elliott MR: Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. Surv Pract 2, 1–7 (2009)
15. Wang W, Rothschild D, Goel S, Gelman A: Forecasting elections with non-representative polls. Int J Forecast 31, 980–991 (2015)
16. Das A, Gollapudi S, Panigrahy R, Salek M (2013): Debiasing social wisdom. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, pp. 500–508
17. Nahorniak M, Larsen DP, Volk C, Jordan CE: Using Inverse Probability Bootstrap Sampling to Eliminate Sample Induced Bias in Model Based Analysis of Unequal Probability Samples. PLoS One 10, e0131765 (2015)
18. Chen JKTUsing LASSO to Calibrate Non-probability Samples using Probability Samples, (2016)
19. Lee S: Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J Off Stat (2006)
20. Zhuang H, Parameswaran A, Roth D, Han J (2015): Debiasing crowdsourced batches. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, pp. 1593–1602
21. Ouyang RW, Kaplan L, Martin P, Toniolo A, Srivastava M, Norman TJ (2015): Debiasing crowdsourced quantitative characteristics in local businesses and services. In: IPSN 2015 - Proceedings of the 14th International Symposium on Information Processing in Sensor Networks (Part of CPS Week), Association for Computing Machinery, Inc, pp. 190–201
22. Kamar E, Kapoor A, Horvitz E: Identifying and Accounting for Task-Dependent Bias in Crowdsourcing. HCOMP (2015)
23. Fan J, Li G, Ooi BC, Tan KL, Feng J (2015): ICrowd: An adaptive crowdsourcing framework. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, pp. 1015–1030
24. Hu H, Zheng Y, Bao Z, Li G, Feng J, Cheng R (2016): Crowdsourced POI labelling: Location-aware result inference and Task Assignment. In: 2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016, Institute of Electrical and Electronics Engineers Inc., pp. 61–72
25. Khan AR, Garcia-Molina H (2017): CrowdDQS: Dynamic question selection in crowdsourcing systems. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, pp. 1447–1462
26. Liu X, Lu M, Ooi BC, Shen Y, Wu S, Zhang M: CDAS: A crowdsourcing data analytics system. Proc VLDB Endow 5, 1040–1051 (2012)
27. Boutsis I, Kalogeraki V, Guno D (2016): Reliable crowdsourced event detection in smartcities. In: 2016 1st International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), SCOPE - GCTC 2016, Institute of Electrical and Electronics Engineers Inc.
28. Eickhoff C, deVries A: How Crowdsourcable is Your Task. undefined (2011)
29. Difallah DE, Demartini G, Cudré-Mauroux P: Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. CEUR Workshop Proc 842, 20–25 (2012)

30. CHIVERS TC, ROGERS WJ, WILLIAMS ME (1974): a Technique for the Measurement of Gas-Leakage.
31.: Uncertain Rule-Based Fuzzy Systems - Introduction and New Directions, 2nd Edition | Jerry M. Mendel | Springer, https://www.springer.com/gp/book/9783319513690
32. Maksimović M, Vujović V, Perišić B, Milošević V: Developing a fuzzy logic based system for monitoring and early detection of residential fire based on thermistor sensors. Comput Sci Inf Syst 12, 63–89 (2015)
33. Khattak HA, Ameer Z, Din IU, Khan MK: Cross-layer design and optimization techniques in wireless multimedia sensor networks for smart cities. Comput Sci Inf Syst 16, 1–17 (2019)
34. Romera E, Bergasa LM, Arroyo R (2016): Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Institute of Electrical and Electronics Engineers Inc., pp. 387–392
35. Psaraftis K, Anagnostopoulos T, Ntalianis K, Mastorakis N: Customized Recommendation System for Optimum Privacy Model Adoption. Int J Econ Manag Syst 03

Konstantinos Psaraftis received the MSc degree from the Electrical and Computer Engineering Department, National Technical University of Athens (NTUA), in 2016. He is currently a PhD candidate in the Department of Business Administration at the University of West Attica. His research interests include location-based services and crowdsourcing, software engineering, and ridesharing applications. He is a member of the IEEE Technical Committee on Hyper-Intelligence since 2021.

Theodoros Anagnostopoulos received the BEng degree in informatics from the Department of Informatics and Engineering, Technical Educational Institution (TEI) of Athens, Greece, in 1997. He also received the BSc degree from the Athens University of Economics and Business (AUEB), Greece, in 2001, and the MSc degree in information systems from the Athens University of Economics and Business (AUEB), Greece, in 2002. He was a visiting PhD student at the University of Geneva (Uni Dufour), Switzerland, in 2007. He received the PhD degree in advanced location prediction techniques in mobile computing from the National and Kapodistrian University of Athens (NKUA), Greece, in 2012. In 2013, he was a postdoctoral researcher in awareness systems at the TEI of Athens, Greece. In 2014, he was a senior postdoctoral researcher in internet of things at the Information Technologies Mechanics and Optics (ITMO) University, Russia. In 2015, he was also a senior postdoctoral researcher in machine learning at the Oulu University, Finland. Currently, he is a principal research scientist in smart cities at the Research & Education at the Ordnance Survey: Britain’s Mapping Agency, United Kingdom. He is also a lecturer of internet of things at ITMO University. His current research interests are in the areas of smart cities, internet of things, connected and autonomous vehicles, cyber-security and privacy, artificial intelligence, and awareness systems for geographic information science and systems. He is a member of the ACM, the IEEE Computer and Communications Societies, and the IEEE.

Klimis S. Ntalianis received the Diploma and Ph.D. degrees from the Electrical and Computer Engineering Department, National Technical University of Athens (NTUA), in 1998 and 2003, respectively. Since 1998, he has participated in more than 25 research and development projects in different frameworks. From 2004 to 2009, he was a Senior Researcher and Projects Coordinator at the Image, Video and Multimedia Laboratory, NTUA. In 2020, he became a Professor at the University of West Attica. He has published more than 160 scientific articles (IEEE, ACM, Springer, and Elsevier) and has received more than 900 citations. He also worked as a Research Evaluator for several international journals and conferences, such as the European Union, the Romanian Executive Agency for Higher Education Research Development and Innovation Funding, the Greek Secretariat of Research and Technology, the Cyprus Promotion Foundation, the Polish National Science Center, the Natural Sciences and Engineering Research Council of Canada, the University of Jeddah (Saudi Arabia), the University of Magdeburg (Germany), the Sant Longowal Institute of Engineering & Technology (India), the Cyprus University of Technology, and other organizations. His main research interests include multimedia analysis, social computing, and new technologies for disruptive business and innovation. He has served as the General Executive Chair for the 3rd IEEE Cyber Science and Technology Congress, the 16th IEEE International Conference on Dependable, Autonomic Secure Computing, the 16th IEEE International Conference on Pervasive Intelligence and Computing, and the 4th IEEE International Conference on Big Data Intelligence and Computing.

Received: August 30, 2021; Accepted: April 03, 2022.