HOSTED BY
ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jides

Mining online political opinion surveys for suspect entries: An interdisciplinary comparison

Costantinos Djouvas[a,*], *Fernando Mendez*[b], *Nicolas Tsapatsoulis*[a]

[a] *Cyprus University of Technology, Department of Communication and Internet Studies, 94 Anexartisias St. Iakovides building, 3rd Floor, Limassol 3040, Cyprus*
[b] *Electronic Democracy Centre (e-DC), Zentrum für Demokratie Aarau ZDA, Kuttigerstr. 21 CH - 5000 Aarau, Switzerland*

ARTICLE INFO

Article history:
Published online 2 December 2016

Keywords:
Voting advice applications
Data cleaning
Machine learning
Data mining
Anomaly detection
Psychometric Likert scale

ABSTRACT

Filtering data generated by so-called Voting Advice Applications (VAAs) in order to remove entries that exhibit unrealistic behavior (i.e., cannot correspond to a real political view) is of primary importance. If such entries are significantly present in VAA generated datasets, they can render conclusions drawn from VAA data analysis invalid. In this work we investigate approaches that can be used for automating the process of identifying entries that appear to be suspicious in terms of a users' answer patterns. We utilize two unsupervised data mining techniques and compare their performance against a well established psychometric approach. Our results suggest that the performance of data mining approaches is comparable to those drawing on psychometric theory with a fraction of the complexity. More specifically, our simulations show that data mining techniques as well as psychometric approaches can be used to identify truly 'rogue' data (i.e., completely random data injected into the dataset under investigation). However, when analysing real datasets the performance of all approaches dropped considerably. This suggests that 'suspect' entries are neither random nor clustered. This finding poses some limitations on the use of unsupervised techniques, suggesting that the latter can only complement rather than substitute existing methods to identifying suspicious entries.

1. Introduction

In this paper we draw on data generated by an EU-wide Voting Advice Application (VAA), called EUvox. VAAs are freely available web tools that match the preferences of voters to that of candidates or political parties [1–5]. The mechanism of the VAA is simple. A set of experts compile a collection of important issues or policy statements, $Q = \{Q_1, Q_2, \ldots, Q_k\}$, upon which users have to express their opinion by selecting one of several categories (or answers). In the majority of cases the number of policy statements (referred to also as questions or items) is 30. Furthermore,

Peer review under responsibility of Qassim University.

* *Corresponding author.*
E-mail addresses: costas.tziouvas@cut.ac.cy (C. Djouvas), fernando.mendez@zda.uzh.ch (F. Mendez), nicolas.tsapatsoulis@cut.ac.cy (N. Tsapatsoulis).

http://dx.doi.org/10.1016/j.jides.2016.11.003

the possible answers are defined as a Likert scale with the following options: 'Strongly Disagree', 'Disagree', 'Neither Agree nor Disagree', 'Agree', 'Strongly Agree', as well as a 'No opinion' category that are encoded as follows: 1 = 'Strongly Disagree', 2 = 'Disagree', 3 = 'Neither Agree nor Disagree', 4 = 'Agree', 5 = 'Strongly Agree', and 6 = 'No opinion'. Thus, every answer can take values in the set $A = \{1, 2, 3, 4, 5, 6\}$. Moreover, we define S_i to be the sequence of answers of the ith user to the policy statement (e.g. $S_i = 2, 4, 3, \ldots, 1, 6, 5$), where $|S_i|$ = number of policy statements. Let us denote with a_i^j to be the jth element ($j \in \{1, 2, \ldots, |S_i|\}$) of the ith sequence (typically in a VAA setting this sequence corresponds to the filled in questionnaire of the ith VAA user).

In the absence of any prior information the probability to get any of the values in A is uniform, i.e.,:

$$p(a_i^j = k) = \frac{1}{|A|}, \tag{1}$$

where $k \in A$. In practice, however, because the sequences are generated by voters that express a finite (actually a small) number of political views some answers are more probable than others. For instance, the positive and negative answers in the middle categories of the scale (i.e., Disagree, and Agree) are more probable than the extreme ones (i.e., Strongly Disagree, and Strongly Agree); this is due to the tendency of humans to avoid taking extreme positions in political statements such as those included in a VAA [6]. Furthermore, the meaning of a 'No opinion' response category attracts itself a special research interest, while its probability of appearance is in general lower than other response options [6]. Fig. 1 depicts the distribution of responses for each of the 30 policy statements for one of the datasets – Ireland – that is used in this paper.[1]

Thus far, we have not taken into consideration the content of the policy statements. In a real setting, however, where users are completing a VAA questionnaire to receive a vote recommendation, there is a highly influential factor that conditions the selection of answers to policy questions: the political view or ideology of a user. It is precisely for this reason that analysts have used VAA generated data to study the dimensionality of the political space by identifying configurations of latent dimensions, such as a left–right or liberal–conservative [7–10]. The fact that most users have a political ideology ensures that there is some degree of consistency in answer responses. For instance, if a user has right-wing political preferences for policy A, B and C they are also likely to have a right-wing political preference for policy D. Evidently, this assumes that A, B, C and D hang together in the policy space. VAA designers have some a-priori knowledge about established dimensions of political competition and design their questionnaires to include groups of policy items that are related to each other [11,9,10]. For instance, the EUvox had three categories related to (1) Europe, (2) economy and (3) broader societal issues. Most users' response patterns would therefore entail some degree of ideological consistency across these dimensions. This has important implications since it makes some sequences of answering patterns more probable than others, which renders any assumption concerning the independence of VAA policy statements invalid. In a mathematically rigorous way this can be stated as follows: let $S = \{S_1, S_2, \ldots, S_n, \}$ be a set of sequences produced by n VAA users. The conditional probability $P(S_i|S)$ is very different than the probability $P(S_i)$. If we assume independence between the VAA statements then $P(S_i)$ and $P(S_i|S)$ can be computed as follows:

$$P(S_i) = \prod_{j=1}^{n} p(a_i^j) \tag{2}$$

and

$$P(S_i|S) = \prod_{j=1}^{n} p(a_i^j|S). \tag{3}$$

If a 'legitimate' VAA user is likely to answer the questionnaire in an ideologically structured manner, and this applies to most users attracted to a VAA, then randomly and inconsistently generated sequences are likely to be quite rare in a real VAA setting. If such sequences in S exist, then they would constitute suspicious entries and should be removed prior to performing any data analysis. We know from various studies that VAA data do contain rogue entries although to our knowledge no filtering is undertaken using structured pattern analyses of responses (see [12,13] for a review). Current best practices [14,3,15–17,13] remove entries by collecting and utilizing some para-data (such as total time taken to complete questionnaire and IP filters). These have sometimes been complemented with the removal of entries with dubious answer patterns, such as too many 'No opinions' or many consecutive same answer responses. Such filters are easy to implement and do not require any sophisticated analysis—though they can become quite arbitrary. If researchers are serious about their cleaning methods, a powerful case has been made for the need to collect data based on individual item response timers [12]. As we shall see below, this paper builds on this key insight.

The identification of suspect entries in questionnaires is a field of study within psychometrics (see [18] for a review). Specifically, there is a literature derived from Item Response Theory that focuses on identifying such inconsistencies in test scores such as exams [19,20]. In data mining terms these inconsistencies are usually described as anomalous and there exists within the field a number of unsupervised anomaly detection techniques that are applied in different areas and disciplines. For example unsupervised anomaly detection can be used for intrusion detection [21,22], for fraud detection [23,24], in medicine for disease detection [25] and for prescription control [26], in image processing for identifying foreign object [27,28], and in text data for novelty detection [29]. However, to the best of our knowledge, there is very limited literature on applying data mining techniques to data consisting of responses to Likert items, as is the case in VAA questionnaires. [30] applied a number of both supervised and unsupervised techniques for extracting patterns in students' evaluations of their instructors where Davier in [31] used data mining techniques to address the problem of testing quality control.

[1] See Section 3 for more information about the datasets used.

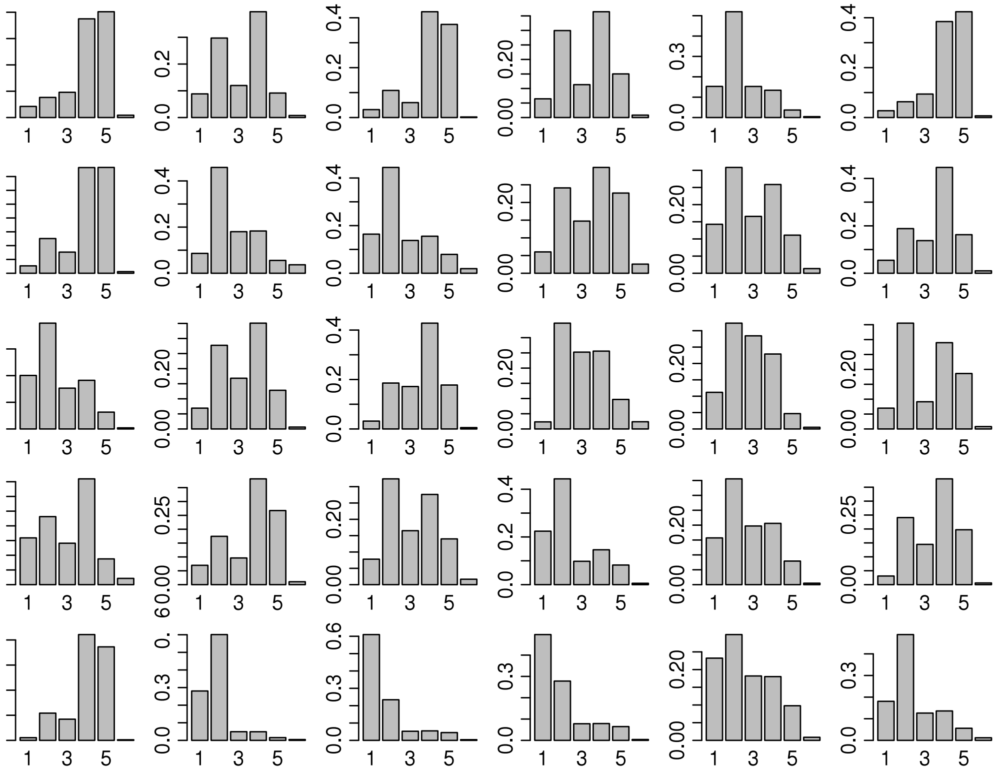

Fig. 1 – Distribution of answer for the 30 statements in the Ireland dataset—EUVox 2014 (see Section 3 for more information about the dataset).

2. Background

Below we describe two approaches that could be applied to the problem of suspect detection in online self-administered surveys. The first outlines the basic intuition behind the psychometric approaches before moving on to some specific data mining techniques.

Table 1 – Policy statements per dimension.

Country	Dimension		
	EU	Economy	Culture
Ireland	1–4, 6, 7	11–15, 20	24, 25, 27–30
Greece	1, 2, 4, 6, 7, 10	10–16, 18, 19	21, 22, 25–30
England	1–6, 9, 10	11–16, 19, 20	21, 23, 24, 25, 27–30

2.1. *Psychometric*

The psychometric approach to identifying suspect answering responses involves two steps (this procedure draws on the approach of [19]). The first step is to reduce the dimensionality of the data to yield meaningful scales—in the case of VAA generated data these are scales measuring the same underlying, latent ideological trait. We draw on Mokken Scaling Analysis (MSA) [32,33], which is a psychometric method of data reduction that has been applied to VAA data by political scientists to identify unidimensional scales that consist of hierarchically ordered items measuring the same latent concept, such as a left vs. right scale or a socially conservative vs. socially liberal scale [11,8–10].[2] MSA is a non–parametric technique derived from Item Response Theory within the field of psychometrics. It is a scaling technique in which the probability of an answer to a particular item (question) depends on the characteristics of the item, such as its 'difficulty' (popularity). The MSA yields scales (groups of items) that are hypothesized traits or attributes of a non-directly observed latent construct. A scale is assessed with two tests that offer scalability coefficients: the item-specific H_i and the overall H. According to common rule of thumb these should both >0.30. For the analysis performed in this paper, the MSA was implemented in R using the `mokken` package. The MSA yielded three substantively meaningful scales related to (1) attitudes towards Europe (pro EU versus anti EU); (2) the economy (left vs. right) and a (3) cultural scale (socially conservative versus socially liberal attitudes) that satisfied the empirical tests. Table 1 shows the policy questions that were grouped per dimension for each of the datasets.

After scaling analysis is performed to identify three unidimensional scales we can use person-fit statistics to identify aberrant answering patterns. These statistics all draw on the number of so-called Guttman errors in a hierarchically ordered scale (i.e., such as those yielded by the MSA). The basic intuition can be illustrated with a simple example. Consider a latent dimension or scale that measures social liberalism, where 0 means socially conservative and a high

[2] Specifically, we use the Monotone Homogeneity Model (MHM) originally proposed by Mokken in 1971 for dichotomous items and extended by Molenaar [33] to polytomous items.

score means socially liberal attitudes. Furthermore, assume that the scale consists of three items with dichotomous answer categories (no/yes) and that the items are ordered in terms of their expected 'difficulty' or popularity as follows: Q1 Gay couples should enjoy basic rights; Q2 Gay couples should be allowed to marry; Q3 Gay couples should be allowed to adopt children. Now consider the following answer patterns: user 1 [1,1,1]; user 2 [0,0,0]; user 3 [1,0,0] and user 4 [0,0,1]. User 1 possesses a high 'test' score (3) on the latent trait of social liberalism, whereas the score of user 2 (0) would be indicative of extremely socially conservative attitudes. While on different sides of the scale, both answer patterns are consistent. Now consider users 3 and 4, both of whom have the same score of 1 on the latent trait being measured. Although their scores are the same, user 4 clearly exhibits an aberrant answer pattern. The difference can be expressed in the number of Guttman errors, which refers to answering a more 'difficult' (in practice the less popular) item correctly and an 'easier' (more popular) item incorrectly. In our illustration users 1, 2 and 3 have 0 Guttman errors while user 4 has 2 Guttman errors.[3]

2.2. *Data mining*

As mentioned earlier, the goal of this work is to identify entries in S that exhibit irregular behavior and deviate much from the norm using unsupervised machine learning techniques.[4] For doing that, a two step process was introduced. The first was to use a clustering method that clusters S into m clusters. This can be formally defined as:

$$C = \{C_1, C_2, \dots, C_m\}, \tag{4}$$

where $C_i \subset S, C_i \cap C_j = \varnothing, |\bigcup_{i=1}^{m} C_i| = |S|$. Then, an anomaly detection approach for identifying outliers in each generated cluster will be applied; identified entries will be marked as suspicious.

The unsupervised clustering method selected was the Expectation Maximization (EM) [34,35]. This is because EM is a probability based approach assuming gaussian mixture observations, and also it does not require the specification of any hard to extract parameters (e.g., k in k-means, *epsilon* and *minPts* in DBScan).

EM is an efficient method that solves the Maximum Likelihood Estimation (MLE) problem; that is, given a set of parameters Θ, assign values to Θ so that the likelihood function is maximized. For example, in the case of a gaussian mixture model $\Theta = (a, \{\mu_k\}, \{C_k\})$, where a is a probability mass function over the different models, $\{\mu_k\}$ is the set of means one for each model, and $\{C_k\}$ a set of covariance matrices one for each model. Given the aforementioned model and Θ, the likelihood function estimates the model to which an observation belongs to.

Estimation Maximization is an iterative process that is repeated until convergence. During the fist step (known as E-step) EM computes a probability distribution Q over a set of latent variables z given a set of observation $x = \{x_1, x_2, \dots, x_m\}$, using Θ. In the second (known as M-step) a probability distribution of new parameters are calculated using the distributions computed during the first step. The above can be formally defined as follows:

$$Q(z_i) = p(z_i|x_i; \Theta), \quad \text{where } i \in \{1, 2, \dots |x|\} \ (E\text{-}step) \tag{5}$$

and

$$\Theta = \arg\max_{\Theta} \sum_{i=1}^{m} \sum_{z_i} Q(z_i) \log \frac{p(x_i, z_i; \Theta)}{Q(z_i)} \ (M\text{-}step). \tag{6}$$

In our case, the likelihood function assigns a cluster to each S_i.

Clusters created by the EM function could potentially represent user groups with different ideologies. The next step is to use an outlier detection technique in order to identify anomalous entries inside each of the generated clusters. All such entries will be marked as *suspicious*. For doing so, two techniques where adopted: (1) a simple Mahalanobis distance [36], and (2) a hi-dimensional outlier identification called PCOut [37].

The Mahalanobis distance [36] is an estimator that calculates the distance between an observation $x = \{x_1, x_2, \dots, x_m\}$ and a distribution D defined by its mean, μ, and covariance, C, matrices. Formally this is defined as:

$$d_M(x) = \sqrt{(x - \mu)^T C^{-1} (x - \mu)}. \tag{7}$$

Having calculated the distance of a point x in terms of how many standard deviations away x is from the mean of D, identifying outliers only requires the definition of a distance threshold that separates normal from anomalous observation. This threshold is application specific and is defined in the next section.

The second method is called PCOut [37] and it consists of two distinct phases; the first deals with location outlier detection (data point that lay far way from the average of a model) and the second with scatter outlier detection (data point that emanates from a different model). Both phases utilize robust distance estimators[5] (robust Mahalanobis and Euclidean distances respectively) for producing two weights w_1 and w_2 respectively. The final weight of an observation x is defined as:

$$w_x = \frac{(W_{1_x} + s)(W_{2_x} + s)}{1 + s^2} \tag{8}$$

where s is a scaling contact. An observation x is an outlier if $w_x < s$.

[3] The Guttman errors are calculated by summing the item pairs that are [0,1]. Item pairs [0,0], [1,1] or [1,0] do not represent violations. For instance an answer pattern for a user 5 [1,0,1] has 1 Guttman error.

[4] Using unsupervised rather than supervised machine learning techniques was a deliberate decision due to fact that we would like to avoid the dubious process of labeling the data.

[5] Estimators, like Mahalanobis distance, can become biased in the case of highly contaminated datasets (i.e. contain many outliers). Robust estimators try to alleviate this by reducing the influence of the outliers on the estimator through a weighting function.

3. Datasets

The datasets used in this paper derived from a special transnational VAA, the EUvox, which was used during the 2014 European Parliament elections. The uniqueness of this VAA derives from the fact that EUvox was essentially a collection of VAAs that corresponded to a different EU member state[6] with a high overlap (above 70%) of identical policy statements. The EUvox was running for a period of one month (end of April until end of May 2014) during which the datasets presented in this work were collected. Data from three instances of EUvox that generated datasets of varying sizes have been selected: (1) the datasets of Ireland (7k entries), (2) Greece (55k entries), and (3) England (115k entries). Detailed analysis of all data collected by the EUvox VAA is well beyond the scope of this paper and we therefore concentrate on these three datasets of varying sizes—small, medium and large.

We can formally define the dataset in terms of users' answers. Consider a dataset as a set $S = \{S_1, S_2, \ldots, S_n\}$ (i.e., the answers of all users). $S_i = \{a_i^1, a_i^2, \ldots, a_i^m\}$ is the answers of a particular user, $|S_i| = 30$, and $a_i^j \in A$, where $A = \{1, 2, 3, 4, 5, 6\}$. Furthermore, each answer is associated with a time interval designating the time required for a user to answer a policy statement. We formally define this as $T = \{T_1, T_2, \ldots, T_n\}$, where $T_i = \{t_i^1, t_i^2, \ldots, t_i^m\}$, $|T_i| = 30$, and $t_i^j \in \mathbb{Z}^*$ (i.e., the granularity is in seconds), and $t_i^j \widehat{=} a_i^j$. In our case, T was utilized for assigning a label[7] $Y_i = \{\text{'}\textit{Valid}\text{'}, \text{'}\textit{Suspect}\text{'}\}$ to each entry in a dataset. For formally defining the labeling process, we should first provide some domain specific definitions.

Definition 1. A timer violation is defined as an entry in T_i, where $a_i^j \le 2$.

The above definition derives from the assumption that it is humanly impossible to read a policy statement, process the information, and provide an answer in less than two seconds.

Definition 2. An entry S_i is labeled as 'Suspect' if the number of timer violations lies above the 97%-quantile of the average violations of the particular instance.

The intuition behind Definition 2, is that only entries with many violation, well over the average number of violations in the instance are marked as 'Suspect'. Thus for each instance the distribution of timer violations is calculated for extracting a number that corresponds to the 97%-quantile of the instance's timer violations.

After applying the aforementioned procedures on the three datasets, each dataset was saved as a csv file consisting of 31 columns. The first 30 columns represented S, and the last column Y, contained the label of each entry. Table 2 summarizes some of the statistical information of the three datasets.

Table 2 – Datasets information.

Country	97%-quantile	Suspects	Valid	Total
Ireland	8	220	7 014	7 216
Greece	5	1528	54 019	55 547
England	8	3393	111 128	114 521

An additional step was the creation of a simulated dataset. More specifically, for each of the datasets we created a simulated copy where all 'Suspect' entries were replaced by a random sequence. This resulted in a total of 6 datasets, three original and three simulated. The purpose of the simulated datasets was the creation of suspicious entries that exhibited completely random behavior (i.e., $P(a_i^j) = \frac{1}{|L|}$). By running the simulation we are therefore able to evaluate whether it is, in principle, possible to identify true 'rogue' data. Furthermore, by comparing the performance of the different approaches on each pair of datasets (i.e., the simulated and the original) one can extract some information regarding the pattern (i.e., randomness) of 'Suspect' entries in the original datasets.

4. Methodology

4.1. *Psychometric*

Psychometricians use person-fit statistics, which are an extension of the Guttman error logic, to identify atypical answer patterns in tests or surveys [18]. We test two person-fit statistics that have been recently developed to deal with polytomous data and which are therefore ideally suited for EUvox data in which the number of answer options is not dichotomous [19,20]. The analysis is carried out using the R package, `PerFit` [38], which implements two person-fit statistics suited to data that has been preprocessed using a non-parametric Item Response Theory model such as MSA. We use the Gnormed.poly function, which calculates the number of Guttman errors for polytomous items, and the U3poly person-fit statistic, which generalizes van der Fliers [39] dichotomous U3 person-fit statistic to polytomously scored items. We derive three person-fit statistics for every user, one for each of the three scales identified through the MSA (see Section 2.1 for details). The coefficients are calculated to range from 0–1, where a high score indicates aberrant answer patterns. Because falsely rejecting the null-hypothesis in most applications of person-fit statistics will not have large consequences Meijer et al. [20] have suggested using a 5% or 10% Type 1 error rate or cut off point to identify the most aberrant answer patterns. We use the more conservative 5% cut off point, i.e., all users above the 95%–quantile for any one of the three scales are flagged as suspect entries.

4.2. *Data mining*

For identifying suspicious behavior in a VAA generated datasets based on data mining approaches, we use two high dimensional anomaly detection methods; that is mahalanobis distance and PCOut.

Mahalanobis distance was utilized for measuring the distance in the high dimensional space of a data entry (i.e., a

[6] Some countries were split into regions.

[7] Please note, that labels are only used for performance validation and not for clustering since only unsupervised methods are considered in this work.

Table 3 – Summary of evaluation for the simulated dataset.

County		Ireland		Greece		England	
Method		Sens	Spec	Sens	Spec	Sens	Spec
Mahalanobis	Clustered	0.17	0.95	0.37	0.96	0.00	0.96
	Un-clustered	0.66	0.95	0.53	0.95	0.73	0.95
PCOut	Clustered	0.63	0.83	0.13	0.82	0.50	0.81
	Un-clustered	1.00	0.78	1.00	0.78	1.00	0.80
Psychometric	Gnormed	0.84	0.99	0.67	0.89	0.63	0.90
	U3 fit	0.81	0.89	0.63	0.89	0.67	0.90

Table 4 – Summary of evaluation for the original dataset.

County		Ireland		Greece		England	
Method		Sens	Spec	Sens	Spec	Sens	Spec
Mahalanobis	Clustered	0.17	0.96	0.09	0.96	0.04	0.96
	Un-clustered	0.16	0.95	0.16	0.95	0.07	0.96
PCOut	Clustered	0.30	0.83	0.30	0.80	0.21	0.79
	Un-clustered	0.39	0.74	0.33	0.76	0.21	0.78
Psychometric	Gnormed	0.25	0.87	0.23	0.88	0.12	0.88
	U3 fit	0.26	0.87	0.23	0.88	0.12	0.88

point in the hypersphere) from the centroid of its group. Then for each group the mean and the standard deviation of the distance were calculated; outliers are defined as all points whose distance deviate more than two standard deviations from the average. For the second method used, PCOut, there is no requirement to specify the number of parameters for the identification of outliers. What has to made clear however, is that the final output of both methods is similar: that is a dataset where every entry is classified as either 'Valid' or 'Suspect'.

Both of the aforementioned methods were implemented in R. For the former approach, an in-house built code was implemented facilitating the classification. For the latter, the Multivariate outlier detection based on robust methods (mvoutlier) [37] was utilized.

The performance of the two approaches was evaluated on twelve datasets, the three original datasets and the three datasets that contained simulated rogue entries. In addition the EM clustering technique produced a further six new datasets. This is driven by the assumption that clustering will result in the creation of more coherent subsets where anomalies would be easier to identify.

5. Results and analysis

This section presents a summary of the results generated by the different approaches. We use the 'Sensitivity' and 'Specificity' metric to evaluate performance in terms of 'Suspect' detection. The metrics are formally defined below:

- True positive (TP): The number of entries labeled as 'Suspect' and are classified as 'Suspect'.
- False positive (FP): The number of entries labeled as 'Valid' and are classified as 'Suspect'.
- False negative (FN): The number of entries labeled as 'Suspect' and are classified as 'Valid'.
- True negative (TN): The number of entries labeled as 'Valid' and are classified as 'Valid'.

Using the above definitions, Sensitivity and Specificity can be formally defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{10}$$

Sensitivity returns the fraction of correctly predicted 'Suspect' entries from the entire 'Suspect' population while Specificity relates to the fraction of correctly predicted 'Valid' entries from the entire 'Valid' population. Tables 3 and 4presents the Sensitivity (Sens) and Specificity (Spec) of all the approaches for the simulated and original datasets respectively (more detailed results can be found in the confusion matrices presented in the Appendix).

Beginning with the simulated datasets we find quite varied performance among the approaches insofar as the sensitivity rate is concerned. Of the two data mining approaches the PCOut clearly outperforms the simpler Mahalanobis approaches. Furthermore, it is clear that un-clustered approaches are superior to culstered approaches. In relation to the two psychometric based approaches relying on person-fit statistics, they both perform very similarly. The simulation shows that detection of what are truly 'rogue' entries can in principle be achieved, especially using the un-clustered PCOut data mining technique and the two psychometric approaches. When we look at the performance on the real dataset however, a very different picture emerges as can be seen in Tables 3 and 4.

5.1. *Area under the ROC curve*

In this section we take a closer look at the better performing approaches and use the area under the ROC (receiver operator

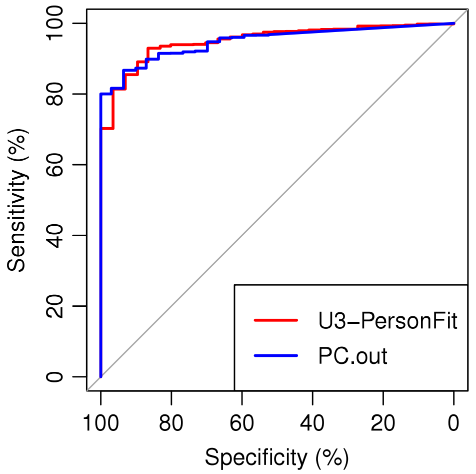

Fig. 2 – Ireland simulated.

Table 5 – Area under the ROC curve (in %) for simulated data with 95% confidence intervals.

Case	Class	AUC	Lower	Upper
Ireland	U3	95.13	94.2	96.07
	PC.out	94.91	94.13	95.69
Greece	U3	92.97	92.47	93.48
	PC.out	93.06	92.77	93.36
England	U3	97.19	97.08	97.31
	PC.out	94.83	94.67	94.99

chracteristics) curve as our evaluation metric for comparing the approaches. ROC curves and the associated area under the curve (AUC) metric is a common way of evaluating binary classification problems [40]. There are two good reasons for focussing on the area under the ROC curve values: (1) it is insensitive to unbalanced datasets and thus commonly used for fraud detection (which makes it ideally suited to the task at hand) and (2) it operates directly on the classification scores and does not require the predictions to be thresholded (e.g. assigned to one class or the other). This will be useful for visualizing the trade-off between sensitivity (true positive rate) and specificity (true negative rate).

The ROC curves in Figs. 2–7 present the simulated and original dataset pairs. We plot the best performing method from the data mining approach (PCOut) and one of the psychometric approaches, the U3 person-fit statistic (although we could have also used the Gnormed statistic since they essentially perform the same). For the PCOut we can use the probabilities directly outputted by the package to plot the curve. In the case of the psychometric approaches we run a logistic regression with the suspect versus valid class as the outcome variable and use the three person-fit statistics as predictors. We can then extract the predicted probabilities from the model and use this vector for plotting the ROC curve. In addition to the plotted ROC curve, to better gauge the performance we can compute for the area under the curve values, as well as the 95% confidence intervals for the point estimate (this is done in Tables 5 and 6)

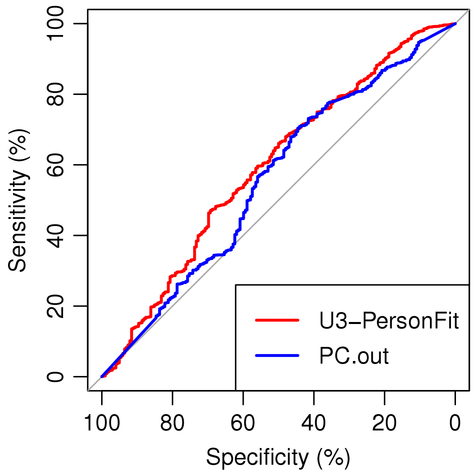

Fig. 3 – Ireland original.

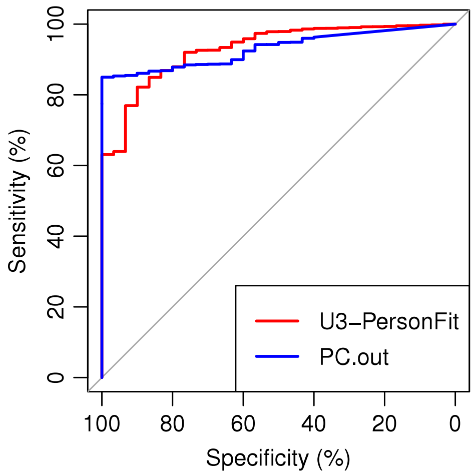

Fig. 4 – Greece simulated.

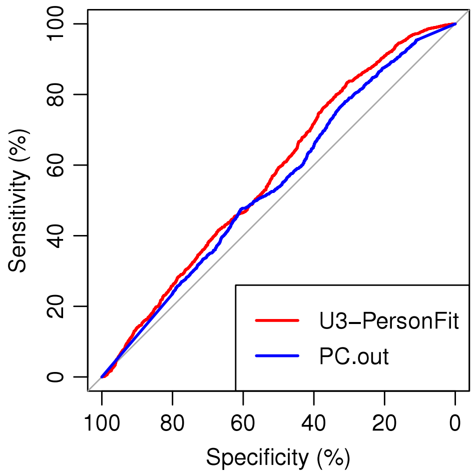

Fig. 5 – Greece original.

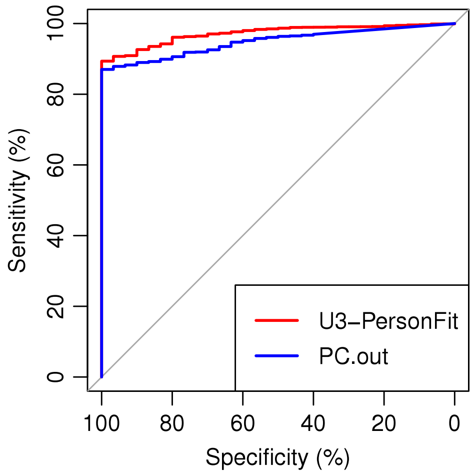

Fig. 6 – England simulated.

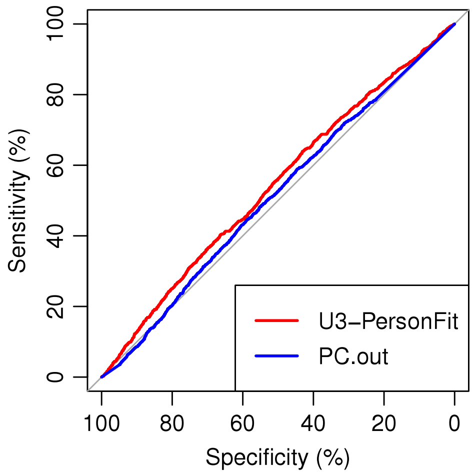

Fig. 7 – England original.

Table 6 – Area under the ROC curve (in %) for original data with 95% confidence intervals.

Case	Class	AUC	Lower	Upper
Ireland	U3	59.44	55.23	63.66
	PC.out	55.81	51.51	60.1
Greece	U3	57.63	56.07	59.2
	PC.out	54.87	53.35	56.39
England	U3	54.16	53.18	55.14
	PC.out	51.36	50.37	52.35

5.2. *Results summary*

As shown in the various tables and figures, the most obvious finding is the poor performance on the original datasets of the proposed methods compared to the performance on simulated data. Nonetheless, we can draw a number of conclusions regarding both the nature of the data used and the adequacy of the approaches used for suspect detection:

1. Timer based 'Suspect' entries, i.e., entries labeled as 'Suspect' using Definition 2 do not exhibit random behavior. This has been demonstrated by the radically different performance when using suspect detection approaches on the simulated datasets compared to the original dataset.
2. Among the Data Mining approaches, PCOut that adopts a hybrid approach (identifies anomalous data using both location outlier detection and scatter outlier detection), outperform other approaches in terms of the total number of suspect entries identified. This suggests that two types of suspect entries exist, those that lie far away from the centroid of the model and those that do not fit the model regardless of their position. Furthermore, the performance of a general data mining approach (i.e., PCOut), is comparable to more computationally intensive psychometric approaches specifically designed for this purpose.
3. All approaches work better on smaller original datasets. This is due to the fact that all datasets regardless of the number of their entries have a similar multidimensional space. Thus datasets with more entries are more dense, something that degrades the performance since the complexity of identifying distinct subsets is increased.
4. Despite the fact that our original intuition was that more coherent clustered subsets would boost the performance of anomaly detection techniques, this was not the case. The explanation for this is similar to that of the above finding; clusters are more dense subsets of the original datasets.
5. Techniques exploited in this work can supplement existing techniques but not substitute them. This is because existing techniques identify suspicious behavior, e.g., timer violations, multiple submissions, etc. Since this suspicious behavior is not always reflected in terms of inconsistent response patterns, such violations cannot be identified by the approaches described in this paper. This points towards the use of combined approaches to suspect detection.

6. Discussion and future work

VAA generated datasets contain suspect entries that correspond to click through behaviour of users that is irrespective of the content of the questionnaire. This has been demonstrated in the literature by collecting individual item (question) timers [12,13]. Furthermore, it is a well-known phenomenon in survey research, referred to as 'satisficing' by psychologists [41]. Current best practices identify suspect VAA entries through para-data such as the time taken to answer questions [16,17], the number of repeated attempts [15–17], the answers provided to some additional, opt-in questions [14,3,10] as filtering methods. Techniques relying on users' answer patterns are at a very primitive stage, consisting mainly of filtering out users with a predetermined number of consecutive same answer patterns (e.g. more than 10) or users with many 'No Opinions' [42,17,43,44]. To our knowledge there has been no attempt to apply more sophisticated methodologies to identifying suspect entries on the basis of users' answer patterns [an exception is [13]].

In this paper we investigated a number of classification techniques that automatically identify inconsistent entries

Table A.7 – Clustered dataset using Mahalanobis distance.

Country	Labeled as	Original dataset		Simulated dataset	
		Classified as			
		Valid	Suspect	Valid	Suspect
Ireland	Valid	6 711	303	6 639	375
	Suspect	168	34	168	34
Greece	Valid	51 833	2186	51 731	2288
	Suspect	1 392	136	968	560
England	Valid	106 631	4497	106 291	4837
	Suspect	3 241	152	3 393	0

Table A.8 – Un-clustered dataset using Mahalanobis distance.

Country	Labeled as	Original dataset		Simulated dataset	
		Classified as			
		Valid	Suspect	Valid	Suspect
Ireland	Valid	6 689	325	6 695	319
	Suspect	169	33	68	134
Greece	Valid	51 418	2601	51 426	2593
	Suspect	1 279	249	713	815
England	Valid	106 185	4943	106 107	5021
	Suspect	3 171	222	906	2487

Table A.9 – Clustered dataset using PCOut.

Country	Labeled as	Original dataset		Simulated dataset	
		Classified as			
		Valid	Suspect	Valid	Suspect
Ireland	Valid	5 793	1 221	5 811	1 203
	Suspect	142	60	75	127
Greece	Valid	43 185	10 834	44 128	9 891
	Suspect	1 066	462	1 324	204
England	Valid	87 924	23 204	89 492	21 636
	Suspect	2 682	711	1 697	1 696

without the use of any paradata. We first drew on the logic of psychometric approaches that are used to address similar problems of suspect entries in exam or test scenarios based on inconsistent answer patterns. We compared the former with techniques drawn from data mining. Both sets of approaches shared the characteristic that no prior knowledge of what is a suspect entry is assumed. However, given that we possessed the relevant paradata we were able to validate the results of our suspect detection methods against labeled datasets where the labels belonged to the set $Y = \{\text{'}\mathit{Valid}\text{'}, \text{'}\mathit{Suspect}\text{'}\}$ and were assigned to each entry based on known timer violation criteria. We first applied the methods to simulated datasets and found that performance was surprisingly robust in terms of using answer patterns to correctly identify truly rogue data that had been simulated. Indeed, the best performing data mining technique performed very similarly to the psychometric approaches—in both cases virtually perfect classification in terms of ROC curve analysis. Most importantly, the data mining technique was able to accomplish this without additional layers of analysis, such as sophisticated preprocessing and the use of domain specific knowledge.

When we tested the performance of both approaches on a dataset comprised of real 'suspect' entries rather than randomly simulated ones, the results were poor. Indeed, an analysis of the area under the ROC curve showed that both psychometric and data mining classification was little better than random chance. For the moment, therefore, our results suggest that the proposed techniques can complement and not replace existing methods since they identify a different set of violations: that is, users that exhibit suspect behavior in terms of their ideological profile that need not be associated with timer response violations. Insofar as the data mining techniques are concerned, and in the absence of any benchmarks, forthcoming research will apply supervised machine learning techniques on VAA generated datasets. This will further extract insights on the performance of machine learning techniques in the area of Likert-scale data cleaning.

Appendix. Detailed results

Tables A.7–A.12 present the results acquired. They are grouped by dataset type (i.e., clustered, and un-clustered) and by the approach used i.e., Data Mining (Mahalanobis and PCOut), and Psychometric. Each of these tables represent the

Table A.10 – Un-clustered dataset using PCOut.

Country	Labeled as	Original dataset		Simulated dataset	
		Classified as			
		Valid	Suspect	Valid	Suspect
Ireland	Valid	5 219	1 795	5 445	1 569
	Suspect	123	79	0	202
Greece	Valid	40 797	13 222	42 036	11 983
	Suspect	1 019	509	0	1 528
England	Valid	87 020	24 108	88 716	22 412
	Suspect	2 681	712	0	3 393

Table A.11 – Psychometric based using Gnormed.

Country	Labeled as	Original dataset		Simulated dataset	
		Classified as			
		Valid	Suspect	Valid	Suspect
Ireland	Valid	6 134	880	6 274	33
	Suspect	152	50	33	169
Greece	Valid	47 372	6 647	48 100	5 919
	Suspect	1 172	356	510	1 018
England	Valid	97 506	13 622	99 695	11 433
	Suspect	2 975	418	1 245	2 148

Table A.12 – Psychometric based using U3 fit.

Country	Labeled as	Original dataset		Simulated dataset	
		Classified as			
		Valid	Suspect	Valid	Suspect
Ireland	Valid	6 122	892	6 264	750
	Suspect	149	53	39	163
Greece	Valid	47 307	6 712	48 112	5 907
	Suspect	1 178	350	561	967
England	Valid	97 535	13 593	99 824	11 304
	Suspect	2 982	411	1 132	2 261

confusion matrix generated by calculating the frequencies of the labels assigned to each entry (as described in Section 3) against the class assigned by the aforementioned techniques.

REFERENCES

[1] D. Garzia, S. Marschall, Matching Voters with Parties and Candidates: Voting Advice Applications in Comparative Perspective, ECPR Press, 2014.

[2] M. Rosema, J. Anderson, S. Walgrave, The design, purpose, and effects of voting advice applications, Electoral Stud. 36 (2014) 240–243.

[3] R.M. Alvarez, I. Levin, A.H. Trechsel, K. Vassil, Voting advice applications: How useful and for whom? J. Inf. Technol. Polit. 11 (1) (2014) 82–101.

[4] T. Fossen, J. Anderson, What's the point of voting advice applications? competing perspectives on democracy and citizenship, Electoral Stud. 36 (2014) 244–251.

[5] T. Louwerse, M. Rosema, Design effects of voting advice applications, Acta Polit. 49 (3) (2014) 286–312.

[6] A. Baka, V. Triga, L. Figgou, Neither agree, nor disagree: a critical analysis of the middle answer category ivoting advice applications, Int. J. Electron. Governance 5 (41732) (2012) 244–263.

[7] J. Wheatley, Identifying latent policy dimensions from public opinion data: An inductive approach, J. Elections Public Opin. Parties (2014) 1–19. (ahead-of-print).

[8] J. Wheatley, C. Carman, F. Mendez, J. Mitchell, The dimensionality of the scottish political space results from an experiment on the 2011 holyrood elections, Party Polit. 20 (6) (2014) 864–878.

[9] M. Germann, F. Mendez, J. Wheatley, U. Serdült, Spatial maps in voting advice applications: The case for dynamic scale validation, Acta Politica (2014).

[10] M. Germann, F. Mendez, Dynamic scale validation reloaded, Qual. Quant. (2015) 1–27.

[11] K. Gemenis, Estimating partiespolicy positions through voting advice applications: Some methodological considerations, Acta politica 48 (3) (2013) 268–295.

[12] I. Andreadis, Matching Voters with Parties and Candidates. Voting Advice Applications in Comparative Perspective, ECPR Press, 2014, (Chapter) Data Quality and Data Cleaning.

[13] Cleaning Out Rogue Entries in Voting Advice Application Datasets, 2014.

[14] R.M. Alvarez, I. Levin, P. Mair, A. Trechsel, Party preferences in the digital age: The impact of voting advice applications, Party Polit. 20 (2) (2014) 227–236.

[15] I. Katakis, N. Tsapatsoulis, F. Mendez, V. Triga, C. Djouvas, Social voting advice applications - definitions, challenges, datasets and evaluation, IEEE Trans. Cybern. 44 (7) (2014) 1039–1052.

[16] F. Mendez, Matching voters with political parties and candidates: an empirical test of four algorithms, Int. J. Electron. Governance 5 (3–4) (2012) 264–278.

[17] J. Wheatley, Using vaas to explore the dimensionality of the policy space: experiments from Brazil, Peru, Scotland and Cyprus, Int. J. Electron. Governance 5 (3–4) (2012) 318–348.

[18] G. Karabatsos, Comparing the aberrant response detection performance of thirty-six person-fit statistics, Appl. Meas. Educ. (2003).

[19] W. Emons, Nonparametric person-fit analysis of polytomous item scores, Appl. Psychol. Meas., 2008.

[20] R.R. Meijer, A.S.M. Niessen, J.N. Tendeiro, A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics examples and a computer program. Assessment, 2015, 1073191115577800.

[21] G. Kim, S. Lee, S. Kim, A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, Expert Syst. Appl. 41 (4, Part 2) (2014) 1690–1700.

[22] K. Leung, C. Leckie, Unsupervised anomaly detection in network intrusion detection using clusters, in: Proceedings of the Twenty-eighth Australasian Conference on Computer Science - Volume 38, ACSC'05, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2005, pp. 333–342.

[23] L. Akoglu, C. Faloutsos, Anomaly, event, and fraud detection in large network datasets, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM'13, ACM, New York, NY, USA, 2013, pp. 773–774.

[24] M. Anderka, T. Klerx, S. Priesterjahn, H. Kleine Büning, Automatic ATM fraud detection as a sequence-based anomaly detection problem, in: ICPRAM 2014 - Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods, ESEO, Angers, Loire Valley, France, 6–8 March, 2014, 2014, pp. 759–764.

[25] K. Das, J. Schneider, Detecting anomalous records in categorical datasets, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'07, ACM, New York, NY, USA, 2007, pp. 220–229.

[26] X. Hu, M. Gallagher, W. Loveday, J. Connor, J. Wiles, Detecting anomalies in controlled drug prescription data using probabilistic models, in: S. Chalup, A. Blair, M. Randall (Eds.), Artificial Life and Computational Intelligence, in: Lecture Notes in Computer Science, vol. 8955, Springer International Publishing, 2015, pp. 337–349.

[27] B. Du, L. Zhang, D. Tao, D. Zhang, Unsupervised transfer learning for target detection from hyperspectral images, Neurocomputing 120 (2013) 72–82. image Feature Detection and Description.

[28] Y. Fujiki, N. Haas, Y. Li, C. Otto, B. Paluri, S. Pankanti, Anomaly detection in images and videos. US Patent 8,724,904, 2014.

[29] A. Mahapatra, N. Srivastava, J. Srivastava, Contextual anomaly detection in text data, Algorithms 5 (4) (2012) 469–489.

[30] E. Fokoue, N. Gündüz, Data mining and machine learning techniques for extracting patterns in studentsevaluations of instructors, 2013. URL http://scholarworks.rit.edu/article/1746.

[31] A.A. von Davier, The use of quality control and data mining techniques for monitoring scaled scores: An overview, ETS Res. Rep. Ser. 2012 (2) (2012) pp. 1–18.

[32] R.J. Mokken, A Theory and Procedure of Scale Analysis: With Applications in Political Research, Vol. 1, Walter de Gruyter, 1971.

[33] I.W. Molenaar, Nonparametric models for polytomous responses, in: Handbook of Modern Item Response Theory, Springer, 1997, pp. 369–380.

[34] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–38.

[35] G. McLachlan, T. Krishnan, The EM algorithm and extensions, second ed., in: Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 2008.

[36] P.C. Mahalanobis, On the generalized distance in statistics, Proc. Natl. Inst. Sci. (Calcutta) 2 (1936) 49–55.

[37] P. Filzmoser, R. Maronna, M. Werner, Outlier identification in high dimensions, Comput. Statist. Data Anal. 52 (3) (2008) 1694–1711.

[38] J.N. Tendeiro, Package perfit, 2014.

[39] H. Van Der Flier, Deviant response patterns and comparability of test scores, J. Cross-Cultural Psychol. 13 (3) (1982) 267–298.

[40] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, proc: an open-source package for r and s+ to analyze and compare roc curves, BMC Bioinformatics 12 (1) (2011) 77.

[41] J.A. Krosnick, Survey research, Annu. Rev. Psychol. 50 (1) (1999) 537–567.

[42] F. Mendez, Matching voters with political parties and candidates: An empirical test of four algorithms, Int. J. Electron. Governance 5 (3) (2012) 264–278.

[43] F. Mendez, Matching Voters with Parties and Candidates: Voting Advice Applications in Comparative Perspective, ECPR Press, 2014, (Chapter). Whatbehind a matching algorithm? A critical assessment of how voting advice applications produce voting recommendations.

[44] F. Mendez, J. Wheatley, Matching Voters with Parties and Candidates: Voting Advice Applications in Comparative Perspective, ECPR Press, 2014, (Chapter). Using VAA-Generated Data for Mapping Partisan Supporters in the Ideological Space.