Identifying Irregular Respondents through Response Time

Yunlu Chen, Yuxin Wu, Changrong Xiao, Zhiyong Zhang

August 29, 2020

Abstract

Identifying and controlling for poor respondents in online surveys is a difficult and non-trivial task in study design and data collection. The goal of this project is to analyze response time data in a network study and explore ways to identify problematic responses to inform future data collection. From an online survey, we obtain response time to 1105 questions grouped into six blocks. Through initial exploratory analysis, we find that bad respondents have more missing timestamps and more change points. To compare response time between different respondents, we use several different methods, including some classification methods, semi-supervised and unsupervised clustering. First, we use overall response time and internal data consistency as our baseline models for classifying bad and good respondents. Second, cross correlation analysis is applied to measure similarities of response and response time. Then we use partial correlation to dig out the potential bad respondents. Finally, dynamic time warp is applied to perform nonlinear "warping" on the series where differences in time are not penalized. We then present the results of these methods.

1 Introduction

With the proliferation of online surveys, specialists show great concern about how to distinguish good respondents from bad ones. Those bad respondents answer the questions without revealing the true cases, and they either use impostors or answer the questions without careful thinking. The existence of such bad respondents will interfere with further analysis, and even cause bias in modeling. Intuitively, the common points of these bad respondents lie in the irregular pattern of response time. Our target is to find the features concerning response time, which could be used to

rule out those bad respondents, and further help develop methods for data screening in survey.

The bad respondents identification problem has attracted interest from researchers recently. One class of popular methods consider only responses. A frequently used design in survey is repeated question (Wilson et al., 1990) so that researchers can measure the inconsistency to decide the careless respondents. This is based on the assumption: if a respondent sticks to the real case, he or she will show similar behavior on the same question. Another class of methods take response time into consideration. For example, Huang et al. (2012) shows that due to the absence of cognitive processing, the response time for bad respondents are generally shorter than that for normal ones, but this pattern can be manipulated and therefore difficult to recognize. Our baseline models are motivated by the above two methods. We use consistency, which will be redefined later on our data, and the overall response time to classify good and bad respondents.

In terms of our data, we obtain the responses of 1105 questions from 165 students (ranging from id 1 to 180 with some students dropping halfway) and the corresponding response time. The responses are first collected and then updated two times. After the data collection of the first time, 11 students admitted that they used impostors, and these ids were marked as bad respondents. They claimed to answer the survey themselves in the following update. Through exploratory analysis, we find that the rest students who didn't admit they use impostors are not necessarily good respondents. More specifically, some of them may answer the questions without careful thinking. It motivates us to use semi-supervised and unsupervised methods besides classification methods. In this way, we can identify the potential bad respondents from those unmarked ids. Though few methods on semi-supervised and unsupervised methods for response time identification are proposed, there are existing researches on general clustering problem, of which the most popular one is dyanamic time warping clustering (Berndt and Clifford, 1994). We also explore the application of partial correlation.

The rest of the paper is organized as follows. In section 2, we present the exploratory analysis, preprocessing process and change point analysis of our data. In section 3, we provide the description of our methods, including mainly internal data consistency, cross correlation, partial correlation and dynamic time warping clustering. Results of these methods are displayed in section 4. In section 5, we conclude and make some discussions.

2 Data Exploratory Analysis

2.1 Data description and motivation

Our survey involves 1105 questions which belong to 6 blocks: infomation (6 questions), bigfive (20 questions), depression (9 questions), lonely (10 questions), happy (4 questions), friendship (176 pages with 6 questions each page). The responses we gather are from three dates in 2019: August 30th, September 2rd and October 16th. The bad id group include 11 marked respondents:

$$71, 77, 78, 82, 86, 104, 107, 125, 130, 132, 138.$$

Given the information that those bad respondents who were marked on August 30th would update their responses on either September 2rd or October 16th, while the rest respondents wouldn't change their response, we first check whether the data belong to good id group remains the same on three dates, and on which dates the bad id has different values respectively. The ids that change their responses from August to September are:

$$77, 82, 104, 125, 130, 132, 136, 138, 146,$$

and the ids that change their responses from September to October are:

$$71, 86, 107, 156.$$

Of these 13 ids, 136, 146 and 156 don't belong to the bad ids. What's more, id 78, which belongs to bad id group, is supposed to have different values on different dates, but it turns out to be the same. Before further analysis, we need to rule out these abnormal data.

Based on the analysis above, there exists many duplicated data on different dates. Therefore, we need to reorganize the raw data and divide into two groups: bad group and candidate good group. Notice that the candidate good group includes the data of good ids as well as the updated data of bad ids.

What's more, we also need to clear those ids with incomplete value vectors. Here incomplete vectors mean the recorded value vector (NA is also a recorded value) is shorter than the number of questions, 1105. So far, we have obtained 139 responses in the candidate good group and 11

responses in the bad group. In Fig. 1, we show two plots of response timestamps with one belongs to the candidate good group and the other belong to bad group.

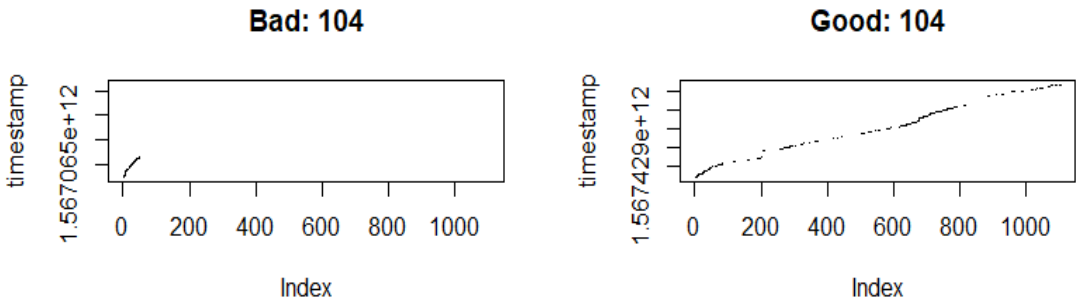

Figure 1: Response timestamps versus question index: bad id 104 (left panel) and candidate good id 104 (right panel).

From the comparison of plots of bad and good id, we can intuitively find the pattern that these individuals tend to choose the first option of the first question on every page of 'Friendship' question type. In this way, they can avoid answering the following five questions of those pages. Thus, many NAs are created in the response time of 'Friendship' question type. We try to use the NA number to distinguish the bad ids from those good ones, and we can get the following result (Fig. 2).

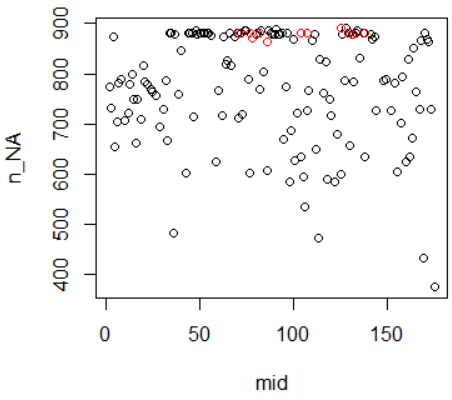

Figure 2: Number of NA versus id: red points indicate bad respondents, black points indicate candidate good respondents.

The method exposes a problem the data: we may have not marked those respondents, who treat the survey irresponsibly (using short time to answer or choose the first option in the first question of 'Friendship' question type) but haven't admitted that they use impostor (belong to the candidate good group). The question is whether these respondents truly answer the survey irresponsibly so we need to supplement those ids into the bad group, or the response time of these respondents do have some internal pattern we need to dig further. For example, for those candidate good ids having more than 880 NA, it means that they choose 175 of 176 people mentioned in 'Friendship' question as never heard of him or her. This question motivates us to do further analysis and then apply those semi-supervised and unsupervised clustering method to identify more bad ids from candidate good group.

2.2 Data preprocessing

Data completion

Before conducting the analysis of respondent quality, we need to extract some useful information from the raw data. For each respondents, the response times for all the questions are required for identification methods which will be introduced in the following sections, like cross correlation method and partial correlation method. However, we only have single timestamps in the raw data. So we extracted those timestamps from the raw data, took differences, and in this way transformed them into the response time series. The theoretical foundation to obtain the response time is that

$$\text{ResponseTime}_i = \text{Timestamp}_{i+1} - \text{Timestamp}_i.$$

In other words, response time means the time span the respondent answering a question, and it equals to the timestamp of the next question minus the timestamp of this question recorded in the raw data. For example, the timestamp of question 1 for the student whose id is 2 is 1566960992477, and that of question 2 is 1566960994356, thus the response time of question 1 is $1566960994356 - 1566960992477 = 1879$.

Now we have response time data and answer data which are essential for our further analysis. There are a certain proportion of incomplete or abnormal values in both sets of data, and we mainly deal with three types of them: missing values, negative response times, and the outliers.

For the response time data, since most time series analysis packages or functions require no missing values in the data, we first complete the missing response times with the average response time of that question. For example, if the response time of question 2 for student 1 was not recorded in the data, we fill in the empty cell with question 2's average response time for all the other students. This "fill NA values with mean values" approach is a common measure to avoid missing values in data preprocessing, and it will not distort the feature of data too much in this response time scenario.

Also, the negative response time is not possible in reality. This abnormal situation occurs because some students went back to revise the answer of former question, which leads to a higher timestamp values for the previous question, so the response times turned to negative when taking differences. Still, replacing them with average question response time is a simple but valid approach.

For the outliers, the situation is more complicated. The outlier here implies the too high or too low values among the response times for a single student, not the abnormal values for a certain question answered by different students. Some outliers are caused by the discontinuous answer of questionnaires, for example, a student answers 30 questions and pauses, after one day he/she finishes the rest of the questionnaire. However, the other explanation is that the average response time of this question is actually greater than other questions. We need to eliminate the former type of outliers but keep the latter one, since the former is a kind of error and the latter can be regarded as a feature of the data. The measure to deal with the outliers is also replacing them with the average response times. Because this method can successfully detect the former type of outliers and remain the feature of the latter ones, which satisfies our requirement.

For the answer data, we only fill the missing values with average values, since there are no unrealistic negative values and outliers in this situation.

Based on these two types of preprocessed data, we are able to further explore the data in the following paragraphs.

Good id selection

To roughly select candidate good ids for future classification methods, we applied the following criteria in the selection process. Firstly, we set the threshold for the number of NAs in the first five blocks (except "Friendship"). Since 'Info', which has 6 questions to answer, has the least questions,

we determined that no more than 6 NAs should exist in the response time of those five blocks, so that we would not have all missing values in any of the blocks. Secondly, we need to examine every page of 'Friendship' block that the response time of the first question could not be missing. With no exact information available for the friendships of each two students, we then intuitively assumed that each student should know at least 10 other students. In this case, we would further select ids through the number of NAs in response time of the rest five questions on each page.

In the above process, the candidate good ids we got are (43 in total): 2, 5, 6, 7, 8, 10, 12, 13, 14, 15, 19, 24, 25, 29, 31, 33, 39, 43, 62, 65, 67, 73, 84, 99, 101, 102, 108, 112, 114, 117, 122, 127, 144, 148, 149, 154, 156, 160, 162, 163, 167, 169, 175

2.3 Change point analysis

We applied change point analysis to determine the number of change points in the response time series and estimate the time of each change. We initially hypothesized that bad respondents may have fewer change points detected due to their invalidity.

However, contrary to our hypothesis, we found that those engaged respondents had fewer change points. Also, after re-submitting the survey, fewer change points were detected in the response time series of 11 confirmed irregular respondents.(Fig. 3)

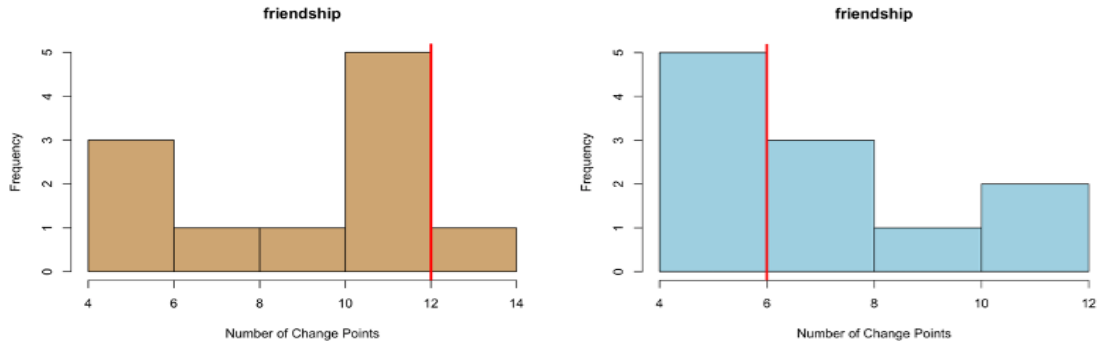

Figure 3: Number of change points in the response time series of 11 confirmed irregular respondents: August 30th (left panel) and October 16th (right panel).

3 Methods

3.1 Baseline models

In this part, we introduce two baseline models: overall response time model simply use the summation of response time to classify bad and good group, and internal data consistency model uses reciprocal properties of network data, which can form a baseline for those models concerning response time.

Overall response time

We begin by introducing the overall response time model. Intuitively, the respondents who belong to the bad group answer those questions without careful thinking, and therefore generally use shorter time. By counting the overall response time, we can get a sketchy classification model.

Internal data consistency

In this part, we focus on the reciprocal properties of network data. The relationships among students are recorded in the friendship block of the online survey, in which we try to investigate three questions: whether a student is a friend of the other student, whether a student is a contact of the other student in a social media app WeChat, and whether a student is a roommate of the other student.

In the normal cases, if student X knows student Y, then student Y should also know student X. We construct a binary adjacency A from the answer of the friendship block. Each element in the matrix takes value 0 or 1, $A_{ij} = 0$ means student i doesn't know student j, and $A_{ij} = 1$ is just the opposite. So in the form of the network's adjacency matrix, it can be explained that $A_{ij} = Aji$.

Similar to the matrix consistency in Analytic Hierarchy Process, we call this reciprocal situation in our study "consistent". We count the number of inconsistent cases for each student, in other words, compare the differences between the rows and columns of the binary adjacency matrix. And then determine the respondent quality based on the amount of inconsistent cases.

3.2 Cross correlation analysis

In this part, we introduce the cross correlation analysis. Cross correlation function is defined as follows,

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}},$$

where $\gamma_{xy}(h) = \mathrm{cov}\,(x_{t+h}, y_t) = \mathrm{E}\,[(x_{t+h} - \mu_x)\,(y_t - \mu_y)]$ and h indicates lag. It measures the similarity of two time series of different lags.

In our problem, we concern the cross correlation of response time series and response series to see if there exists a different pattern in the bad and good group.

3.3 Partial correlation analysis

We introduce the partial correlation analysis to measure the correlation between two response time series, ruling out the effect of their answer series. Partial correlation function is defined as

$$\rho_{xy.\{z_1 z_2\}} = \frac{cov\,(x, y|z_1, z_2)}{\sqrt{var\,(x|z_1, z_2) \cdot var\,(y|z_1, z_2)}}$$

This can be calculated as the correlation between the residuals of the regression of x on z_1,z_2 with the residuals of y on z_1,z_2. Here, x, y represent two response time series, while z_1,z_2 stand for their corresponding answer series. We want to use the partial correlation of response time series to explore more good/bad respondents.

3.4 Dynamic Time Warping clustering

In this part, the unsupervised method, clustering, was applied to identify respondents of bad quality. We assume that problematic respondents have similar response time pattern, so they can form a different cluster apart from the good ones.

The key of clustering method is to determine a proper similarity measure function to divide the data into clusters. Different from the usual scalar data, it's not a good way to measure the similarity between time series with Euclidean distance which ignores the sequence time structure of time series data.

We used a common criterion in time series clustering field, Dynamic Time Warping (Berndt and Clifford, 1994), as the similarity function. In many cases, the two sequences of time series data have

very similar shapes overall, but these shapes are not aligned on the x-axis. So before we compare their similarities, we need to warping one or two of the sequences on the time axis to achieve better alignment. The traditional similarity functions are not capable for this job, but Dynamic Time Warping is an effective way to achieve this warping distortion. Dynamic Time Warping calculates the similarity between two time series by extending and shortening the time series.

The following formula describe how the Dynamic Time Warping method measures the time series similarities, in which S and T are the two time series to be compared, and W is the warping path.

$$DTW(S,T) = \min_{w}[\sum_{k=1}^{p} \delta(w_k)]$$

$$\delta(i,j) = (s_i - t_j)^2$$

In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restriction and rules (Wikipedia, 2020):

- Every index from the first sequence must be matched with one or more indices from the other sequence, and vice versa.
- The first/last index from the first sequence must be matched with the first/last index from the other sequence.
- The mapping of the indices from the first sequence to indices from the other sequence must be monotonically increasing, and vice versa.

Usually dynamic programming is applied to implement this algorithm. But we are not focusing on the algorithm here, Dynamic Time Warping only serves as a similarity measurement for clustering method.

4 Results

4.1 Baseline models

Overall response time

We begin by showing the scatterplot of overall response time for different id, as is shown in Fig. 4. We can intuitively see that there are many ids which belong to good group have similar response time as bad ids. By setting the threshold as the largest response time in bad group , i.e. controlling true positive rate to be 100%, we get a false negative rate: 34.5%.

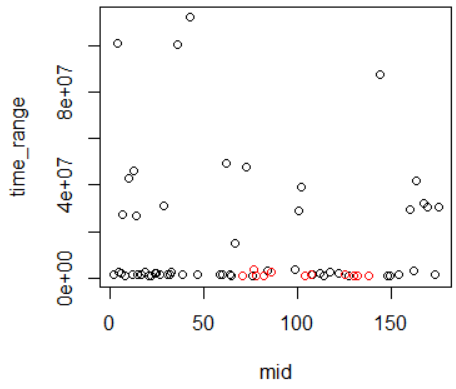

Figure 4: Overall response time versus id: red points indicate bad respondents, black points indicate candidate good respondents.

Internal data consistency

We obtain the following distribution of the number of inconsistent cases for the three types of questions.

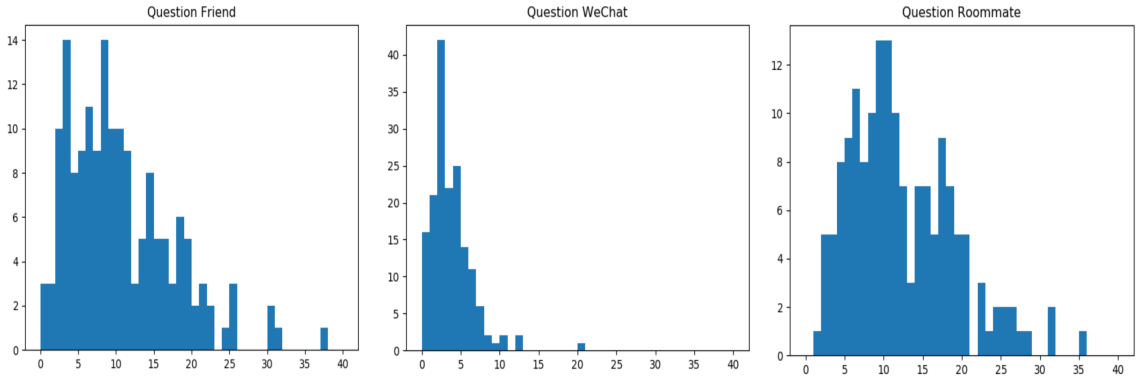

Figure 5: The distribution of the number of inconsistent cases (from left to right: question friend, question WeChat and question roommate)

Notice that a big proportion of students have more than one inconsistent cases. This doesn't mean that they are all problematic, because a problematic respondent is very likely to influence the good ones. Take a simple example, student X didn't pay attention when filling in the questionnaire and wrongly select "not a friend" for student Y, but they are friends actually, this mistake leads to one inconsistent cases for both student X and Y, even though Y did everything right. Besides, there are also some situations when student X regards student Y as a friend but Y thinks their relationships are not close enough. For these reasons, we should loosen our criterion to select careless respondents.

Then we determine the threshold of question friend as 20, question WeChat as 10 and question roommate as 20. If the number is too large, we regard this student as problematic. For question friend, it results in fifteen problematic indexes: 0, 12, 15, 23, 35, 70, 79, 90, 95, 100, 101, 106, 116, 129, 155; the problematic indexes obtained from question WeChat are: 0, 1, 38, 57, 61; and the result of question roommate is: 0, 1, 24, 30, 35, 38, 57, 61, 70, 73, 79, 81, 99, 100, 102, 106, 107, 116, 129, 143, 150, 155. We regard all these indexes as respondents of bad quality.

Different from the first two types of questions, we apply an additional approach on question roommate. In most cases, the size of a student dormitory room is no more than eight people. We assume that the university where our survey was conducted has eight-people-room dormitories, then we know logically that: when answering question roommate, one student should answer at most seven "yes", otherwise the respondent is very likely to be problematic.

Count the number of "yes" answer for each student, we obtain the distribution in Fig. 6. And then select the students whose number is greater than 7, and label them as problematic. We identified 81 problematic respondents in this way.

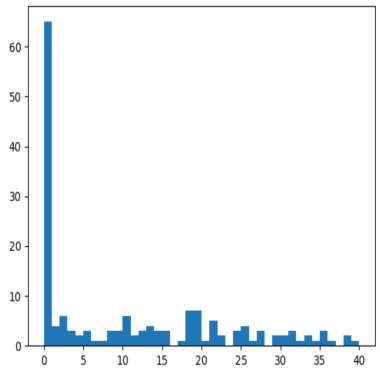

Figure 6: The distribution of the number of "yes" answer of question roommate

4.2 Cross correlation method

In the first place, we do CCF plot of each id in the bad and good group to get an intuitive understanding of group patterns, as is shown in Fig. 7. From the CCF plot, we have two observations: CCF value of lag 0 tends to be negative for bad ids Extension; CCF value tends to be negative for bad ids. To further verify our observation, we calculate the mean and median of CCF of two groups (Fig. 8).

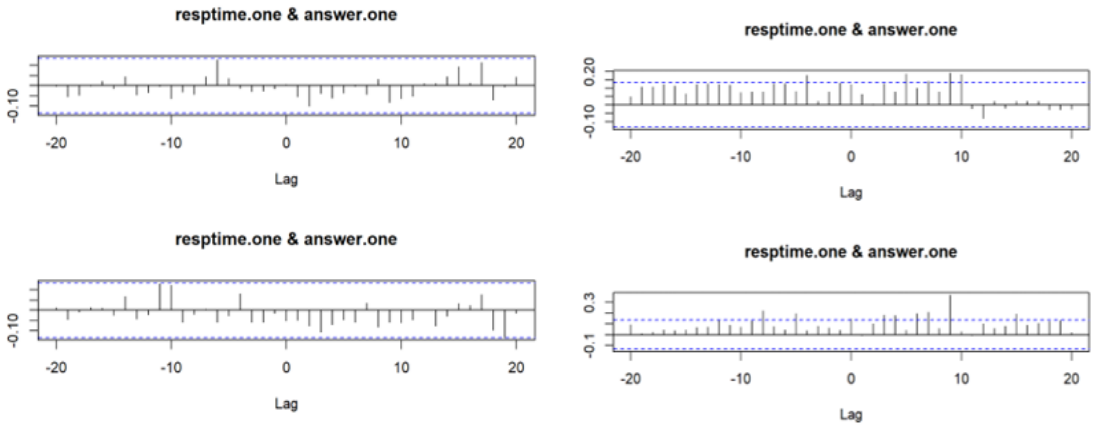

Figure 7: CCF (left panel) and number of negative CCF of lag 0, 1 (right panel) versus id: red points indicate bad respondents, black points indicate candidate good respondents.

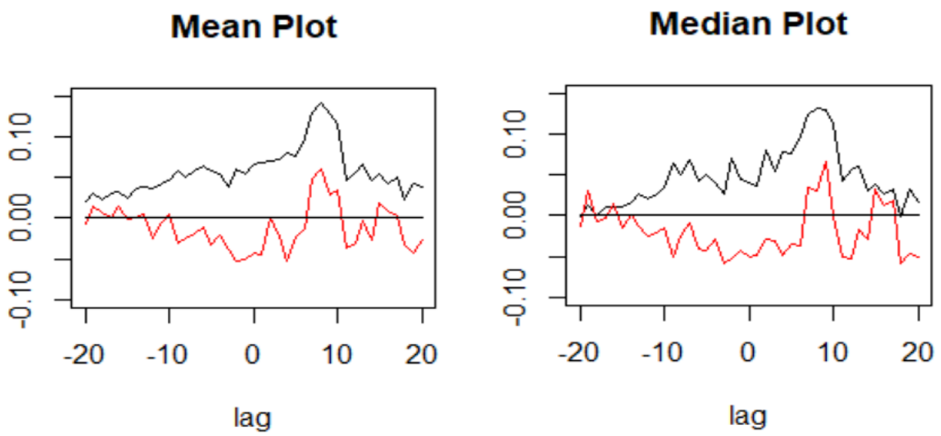

Figure 8: Mean (left panel) and median (right panel) of CCF of 20 lags: red points indicate bad respondents, black points indicate candidate good respondents.

Through the observations, we confirm that bad and good group have different patterns in terms of CCF. Therefore, we try to use this finding to do classification, and the scatterplot of our classification method is shown in Fig. 9. By using the CCF of lag 0, the accuracy of detecting the good id is 64.3% if we set the largest CCF in bad group as the threshold, i.e. controlling true positive rate to be 100%. Extending this method, we use the negative number of CCF of different lags to see if we can find a better method. We find that when lag number is 1 and we use the criterion if all 3 CCF of some id is negative, we get the accuracy of detecting the bad and the good id are 90.9% and 76.8% respectively.

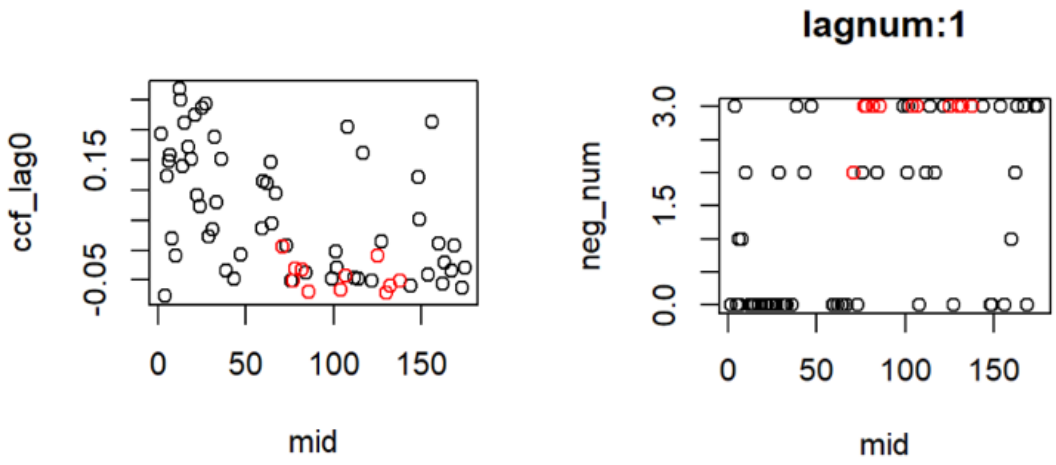

Figure 9: CCF (left panel) and number of negative CCF of lag 0, 1 (right panel) versus id: red points indicate bad respondents, black points indicate candidate good respondents.

4.3 Partial correlation analysis

Firstly, we calculated the partial correlation between each two response time series, the correlation matrix is shown in Fig. 10. A key aspect of using the partial correlation matrix is determining a threshold separating disengaged and engaged responding, which suggests, for example, if the partial correlation between an unknown respondent and a confirmed disengaged one is more than the threshold, this unknown respondent can be classified as irregular.

With no existing appropriate cutoff for the partial correlation, we tried to set the threshold at 0.9 for the correlations. We used those 43 candidate good ids, which were mentioned before in good id selection to explore new good ids with similar response time series. Identically, we made use of those 11 confirmed bad respondents to get more suspected bad ids.

Consequently, we got 25 new suspected bad ids as well as 40 more candidate good respondents. In our total candidate good group, there are two ids that belong to the confirmed 11 bad respondents.

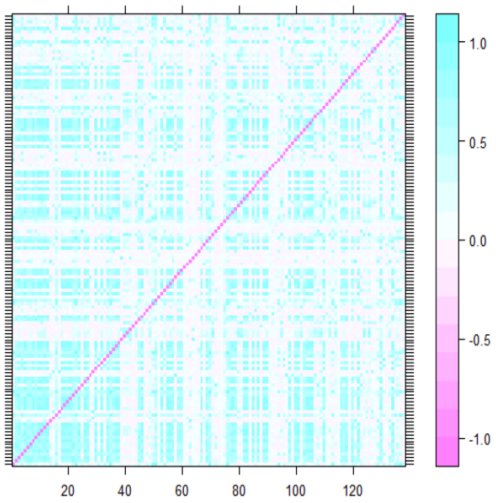

Figure 10: Correlation matrix for the response time series

4.4 Dynamic time warping clustering

We tried different sets of parameters of the Dynamic Time Warping clustering method, and obtained several clustering results. Unluckily, the problematic students we have already labeled are nearly evenly distributed in each cluster, which is quite ambiguous. We cannot say that which respondent cluster is of good quality and which is bad one. What's worse, there are no clear performance measure like classification accuracy for Dynamic Time Warping clustering method.

In short, the Dynamic Time Warping clustering provides insights on the response time series patterns, but this does not directly point to the quality of the respondents. Our assumption in the Method section may not hold, because the results of clustering show that the problematic respondents in this study may be caused by reasons apart from the time series patterns.

5 Discussion

In this paper, we explore different methods for identifying irregular respondents through response time and present their results, as is shown in Table 1. The methods we include include: overall response time method, internal data consistency method, cross correlation analysis, which belong to classification methods, partial correlation analysis, which can be seen as a semi-supervised method, and dynamic time warping clustering, which belongs to unsupervised method. To be more specific,

some discussions are listed below.

	Bad id (id in bold indicates true bad id) / Bad index	True positive	False negative
Overall response time	**71**,**77**,**78**,**82**,**86**,**104**, **107**,**125**,**130**,**132**,**138**	100%	34.5%
Data consistency	0,1,12,15,23,24,30,35,38,57, 61,70,73,79,81,90,95,99,100,101, 102,106,107,116,129,143,150,155	-	-
Roommate number	0,2,3,5,6,7,9,11,12,13,14,15,16,17,18, 19,20,21,23,24,25,26,28,30,31,32,35,38,56,57, 59,61,64,68,70,73,74,79,81,83,90,91,93,94,96, 97,99,100,101,102,103,106,107,108,110, 111,112,113,114,116,118,123,125,128,129, 131,135,139,143,145,146,147,148,149,150, 151,152,155,157,161,163	-	-
Cross correlation	4,39,47,**77**,**78**,**82**,**86**,99,102, **104**,**107**,114,122,**125**,**130**,**132**, **138**,144,154,163,167,173,175	90.9%	76.8%
Partial correlation	3,12,14,20,27,31,32,37,40,45, 53,64,65,**71**,76,**77**,**78**,**82**,**86**,89, 91,92,94,100,**104**,**107**,119,122,**125**, 126,128,**130**,129,**132**,133,**138**	100%	2.4%

Table 1: Results of different methods

- For classification methods, we can calculate the accuracy. The accuracy of data consistency is not calculated, because we are not provided the map of index in the network data to the id. We can find that cross correlation method has a better performance than the baseline

mode: overall response time model.

- Though internal data consistency method doesn't involve response time, it has intuitive interpretation. For future work, we need to consider how to better interpret other methods.

- For semi-supervised and unsupervised methods, they aim at identifying potential bad respondents, but true positive and false negative rates are not appropriate criteria to measure the performance. Applying a better performance criterion will be an interesting direction for future work.

References

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, vol. 10. Seattle, WA, USA:, 359–370.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M. and DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, **27** 99–114.

Wikipedia (2020). Dynamic time warping. Website. URL https://en.wikipedia.org/wiki/Dynamic_time_warping.

Wilson, M. A., Harvey, R. J. and Macy, B. A. (1990). Repeating items to estimate the test-retest reliability of task inventory ratings. *Journal of Applied Psychology*, **75** 158.