JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 02/06 pp128−138
Volume 3, Number 2, June 2022
DOI: 10.23919/JSC.2021.0021

The Hidden Sexual Minorities: Machine Learning Approaches to Estimate the Sexual Minority Orientation Among Beijing College Students

Yunsong Chen*, Guangye He, and Guodong Ju*

Abstract: Based on the fourth-wave Beijing College Students Panel Survey (BCSPS), this study aims to provide accurate estimation of the percentage of the potential sexual minorities among the Beijing college students by using machine learning methods. Specifically, we employ random forest (RF), an ensemble learning approach for classification and regression, to predict the sexual orientation of those who were not willing to disclose his/her inherent sexual identity. To overcome the imbalance problem arising from far different numerical proportion of sexual minority and majority members, we adopt the repeated random sub-sampling for training set by partitioning those who expressed heterosexual orientation into different number of splits and further combining each split with those who expressed sexual minority orientation. The prediction from 24-split random forest suggests that youths in Beijing with sexual minority orientation amount to 5.71%, almost two times that of the original estimation 3.03%. The results are robust to alternative learning methods and covariate sets. Besides, it is also suggested that random forest outperforms other learning algorithms, including AdaBoost, Naïve Bayes, support vector machine (SVM), and logistic regression, in dealing with missing data, by showing higher accuracy, F1 score, and area under curve (AUC) value.

Key words: sexual minority orientation; imbalanced missing data; random forest; machine learning

1 Introduction

Across the globe, rights movements of sexual minority† population have advanced at an unprecedented rate over the past few decades in Western countries and many other regions[1]. Despite this, individuals with sexual minority identity in most cultures are relegated to lower status and even marginalized as a distinct social group. Compared to sexual majorities, members of sexual minorities are segregated geographically[2], socially, and psychologically, and are constantly exposed to the discriminatory tastes of the heterosexual majority[3−5].

While most of the existing literature on sexual minorities has been conducted in the Western world, research relevant to sexual minorities in China is comparatively scarce[6]. Owing to negative stereotypes and devaluation, sexual minority identities often result in antipathy (i.e., disgust), and even blame attributions and justification of unfair social treatment[3, 6, 7]. Such long-standing stigmas and moral condemnation make individuals with sexual minority orientation reluctant to disclose their inherent sexual orientation[8], leading to a high percentage of missing value and downward-biased estimation in obtaining information from traditional social surveys. For example, in the USA, the relevant statistics from the Centers for Disease Control and

• Yunsong Chen and Guangye He are with the Department of Sociology, Nanjing University, Nanjing 210023, China. E-mail: yunsong.chen@nju.edu.cn; hgy.gloria@nju.edu.cn.
• Guodong Ju is with the Department of Social Policy, London School of Economics and Political Science, London, WC2A 2AE, UK. E-mail: G.Ju@lse.ac.uk.
* To whom correspondence should be addressed.
Manuscript received: 2021-01-14; revised: 2021-11-24; accepted: 2021-11-25

† In this paper, we use "sexual minorities" to cover people who are not heterosexual individuals, which typically include gay men, lesbians, bisexuals, and queers. But our dataset generally covers people who identify them as gay men, lesbians, and bisexuals.

Prevention (CDC) have shown that an estimated 4% of adults identify themselves as members of a sexual minority. However, if we look at items relevant to same-sex sexual behavior and same-sex attraction, the number of individuals who reported having same-sex sexual behavior experience, and same-sex attraction has increased to about 8.2% and 11%, respectively.[8] This, to some degree, implies that the question relevant to sexual minority orientation may suffer from severe underestimation. A report from the American College Health Association showed that in 2006, 10% out of 33 000 undergraduates are identified as sexual minorities in the USA.[9] In China, the paucity of reliable data that can identify the sexual minority group and its associated needs, contribute in no small measure to the invisibility of sexual minority population in the eyes of academia as well as critical social programs[10, 11].

To overcome the barriers to estimating the size of China's sexual minority population, this research, based on the Beijing College Students Panel Survey (BCSPS), aims to estimate the genuine percentage of sexual minorities among Chinese youth in Beijing using machine learning methods. We use random forest (RF), an ensemble learning approach for classification and regression, to predict the sexual orientation of those who did not respond to the survey question. Given that the size of sexual minorities is far smaller than that of heterosexual people, the standard RF classifier tends to be biased towards the majority class. To deal with this, we adopt repeated random sub-sampling methods[12]. To evaluate model performance, the result of bagging, boosting, Gaussian naïve Bayes (Gaussian NB), support vector machine (SVM), and logistic regression classifiers will also serve to provide a comparison. Results from multiple machine learning classifiers and different sets of covariates generally show that the genuine percentage of sexual minorities among Beijing college students is almost two times larger than that obtained from traditional social survey data such as the BCSPS. Through our research, we paid particular attention to protect individuals' information security. All results that presented in the paper reflect the group trend rather than personal traits.

2 Materials and Methods

2.1 Data

The analysis is based on the Beijing College Students Panel Survey (BCSPS) conducted by the Survey and Data Center of Renmin University in China. The first-wave data were collected in 2009. Using stratified random sampling college students from 15 universities in Beijing were selected, with a response rate around 93%. Since then, the follow-up surveys are conducted each year until 2013. In the fourth-wave survey, there included a question relevant to respondents' sexual orientation. It asked the surveyed college students to report their sexual orientation and provided four items to choose from ("heterosexual", "homosexual", "bisexual", and "not sure"). As Table 1 shows, among the 4043 samples, 3684 students identified themselves as the heterosexual and 115 identified themselves as sexual minorities (homo or bisexual), or belonging to sexual minorities. According to this, the percentage of sexual minorities among college students in Beijing is around 3.03%, if we deleted 244 samples with missing values on sexual orientation (i.e., 156 declined to respond and 88 chose " not sure"). However, as mentioned above, many of the 244 students who did not explicitly respond to the question about their sexual orientation are very likely to be sexual minorities, given the lurking stigma towards the sexual minorities in China[4].

The fourth-wave of BCSPS contains 910 variables for 4043 respondents providing comprehensive information about demographic characteristics, academic performance, and socioeconomic attributes, which offer us an opportunity to predict missing values

Table 1 Frequency and percentage for sexual orientation in the BCSPS.

Answer	Frequency	Percentage (%)	LGBQ rate (%)
Heterosexual	3684	91.12	
Homosexual	49	1.21	3.03 (*N*=3799)
Bisexual	66	1.63	
Not sure	88	2.18	Missing value
Refuse to answer	156	3.85	
Sum	*N*=4043	100	

on sexual orientation. However, including all 910 variables in the predicting models would introduce a lot of redundant information. Accordingly, we used a Lasso regression algorithm[13, 14] to select 337 sexuality-related variables to perform prediction, including individual family background, educational performance, and mental and physical health, as well as indices of social behavior, which, to a large degree, ensures the comprehensiveness of variables that could capture the differences in characteristics between individuals with sexual majority and minority orientation. For a robustness check, we also ran prediction models using all 910 predictors. Because of space constraints, we do not present the statistics of all predictors of sexual orientation, which are available upon request. Instead, in Tables 2−7 we present the major demographics of the samples and other selected attributes.

2.2 Imbalanced problem

As shown, the dataset is highly imbalanced in terms of sexual orientation. This is not surprising because the percentage of sexual minorities is far smaller than that of heterosexual people worldwide. In our data, the imbalance rate is 3.12% (115/3684) and the percentage of sexual minority is 3.03% (115/3799). If one is using a standard prediction algorithm, the highly imbalanced data may lead to under-prediction of the minority group because most classifiers work on data drawn from the same distribution as the training set. In this vein, it is not easy to prepare appropriate data for training and testing, which leads to a wrong prediction[15]. Taking sexual orientation as an example, if 99% of people declare themselves to be heterosexual, a standard machine learning algorithm (be it a naïve Bayesian classifier or a decision tree) can hardly do better than the 99% accuracy achieved by the trivial classifier that labels everyone as heterosexual. That is, when applying machine learning on highly imbalanced datasets, the built-in goal to maximize the accuracy of the learning algorithm will inevitably under-predict the minority class, as this is the "intelligent" thing to do. To deal with this, a common practice is to rebalance the training sets by resampling, boosting, bagging, or conducting repeated random sub-sampling[16].

In this paper we used the repeated random sub-sampling approach to address the imbalanced data problem. That is, we partitioned the training data into sub-samples using 6-split, 12-split, 24-split, and 32-split of the original sexual majority respondents, respectively. By doing so, each sub-sample ends up containing a less imbalanced dataset than the original one (the one with an imbalance rate equal to 3.03%). Specifically, for the 6-split sub-sample, the number of sexual majority participants is 614 (the original number =3684/6). As a result, the imbalance rate is 18.73% (=115/(3684/6)). Similarly, for the 12-split sub-sample, the number of sexual majority participants is 307, and the imbalance rate is 37.46% (=115/(3684/12)). For the

Table 2 Descriptive analysis of biological sex factor.

Biological sex	Frequency	Percentage (%)
Male	2124	52.54
Female	1919	47.46

Table 3 Descriptive analysis of race factor.

Race	Frequency	Percentage (%)
Hanchu	3588	88.99
Manchu	133	3.3
Huichu	94	2.33
Mongolian	48	1.19
Others	180	4.45

Table 4 Descriptive analysis of political status factor.

Party	Frequency	Percentage (%)
Communists	1490	37.05
Youth league	2223	55.27
Other parties	4	0.1
No party	305	7.58

Table 5 Descriptive analysis of university tier factor.

University level	Frequency	Percentage (%)
Top 3	1150	28.44
985 Level	750	18.55
211 Level	686	16.97
Others	1457	36.04

Table 6 Descriptive analysis of age factor.

Current age	Frequency	Percentage (%)
⩽20 (Minimum=19)	24	0.59
21−23	2198	54.37
24−26	1793	44.35
⩾27 (Maximum=29)	28	0.69

Table 7 Descriptive analysis of hukou status factor.

Hukou	Frequency	Percentage (%)
City	2894	71.58
Village	1128	27.9
Blank	21	0.52

24-split sub-sample, the number of sexual majority participants is 153 and the imbalance rate is 74.92% (=115/(3684/24)). Finally, the 32-split sub-sample contains an equal number of instances from the sexual minority and sexual majority (115:115), with the imbalance rate being 1. Note that the 32-split approach is 1:1-resampling, following Khalilia et al.[12] by partitioning the training data into sub-samples with each sub-sample containing half the sexual minority and half the sexual majority, except for the last sub-sample (in some cases). In fact, we also tried repeated random sub-sampling. That is, we replicated the cases reporting sexual minority with 6-times, 12-times, and 24-times. By doing so, each sub-sample ends up with a less imbalanced dataset compared with the original datasets. However, due to the presence of repeated samples, the overfitting problem emerged, so we did not present the relevant results in this paper.

In RF, a forest itself consists of an ensemble of decision tress which could output a classification, where the output is predicted using mode of observations in the terminal nodes. In this paper, the splitting decision is based on the *Gini* index, a measure of node purity. It is given by the following formula:

$$Gini = 1 - \sum_0^k P_k{}^2 \quad (1)$$

where P_k is the probability of being classified as class k in a node. Thus, the *Gini* index takes values in [0, 1], and 0 means all elements are of the same class. The decision will be made when it gets the lowest *Gini* index. The *Gini* index measures the distribution of class label in nodes. A smaller value of the *Gini* index suggests a purer node. For a split to occur, the *Gini* index for a child node should be less than that for the parent node.

RF has the reputation of being insensitive to the training parameters[17]; to choose the maximum number of features for a split, we followed standard practice by taking the square root of the number of our variables ($18\approx \sqrt{337}$) for each individual tree. We followed the rule of thumb to set 50 as the number of trees for the RF algorithm.

2.3 Sample splitting

In this paper, we developed an N-fold cross validation using the whole available BCSPS data. We chose 6, 12, 24, and 32 (1:1) as split times (T_i) to create an upgrade trend for the sexual minority rate. For each splitting turn, we first separated the samples for learning (N_l=3799) and the samples for predicting (N_p=244). And then, the learning sample was further divided into N_l sexual majority set ($N_{\text{l}0}$=3684) and sexual minority set ($N_{\text{l}1}$=115). For the sexual majority set, we randomly sorted the $N_{\text{l}0}$ set and partitioned them into T_i sub-samples. Every $N_{\text{l}0}$ subsample has an equal number of unrepeated samples except the last $N_{\text{l}0}$ sub-samples of 24 and 32 splits (they contain 165 and 119 samples accordingly). For the sexual minority set, the data points are sampled without replacement.

To generate subsamples for prediction, we combined each sub-sample from $N_{\text{l}0}$ with $N_{\text{l}1}$. By doing this, each sub-sample of the sexual majority sample was selected once, while the sexual minority sample was selected T_i times. For each sub-sampling set, we randomly sorted the data and took 30% as the validation set, and the remaining 70% as the training set, where the 30/70 ratio was chosen by trial and error.

For each round, we trained the model on the training set and validated it on the validation set. Based on this, we predict the missing sexual orientation in N_p samples using a "major voting" approach. That is, an individual would be labeled as a member of the sexual minority if that individual received more than half the votes among T_i subsamples for prediction. Otherwise, the individual would be classified as heterosexual. To make the final decision, the aforesaid process would be repeated 100 times, and each time a different random seed would be selected. In the end, to choose between the numbers of splits, the average receiver operating characteristic (ROC) curve and the area under curve (AUC) for the classifier would have been calculated and compared. The analytical scheme is drawn in Fig. 1.

As mentioned before, we do two sets of predictions with RF, one using 337 predictors obtained from Lasso, and the other using all 910 predictors. To evaluate the performance of the RF model, we compared the result of 24-split sub-sampling with the other five algorithms. We used the default parameters for bagging, boosting, Naïve Bayes, and logistic regression. For SVM, the linear kernel was used. We performed machine learning classification using Scikit-Learn toolkit driven by Python; specifically, we used RandomForestClassifier (for random forest), BaggingClassifier (for Bagging), AdaBoostClassifier (for AdaBoost), GaussianNB (for Gaussian Naïve Bayes), SVC (for Support Vector Machine), and LogisticRegression (for logistic) packages supported in Scikit-Learn.

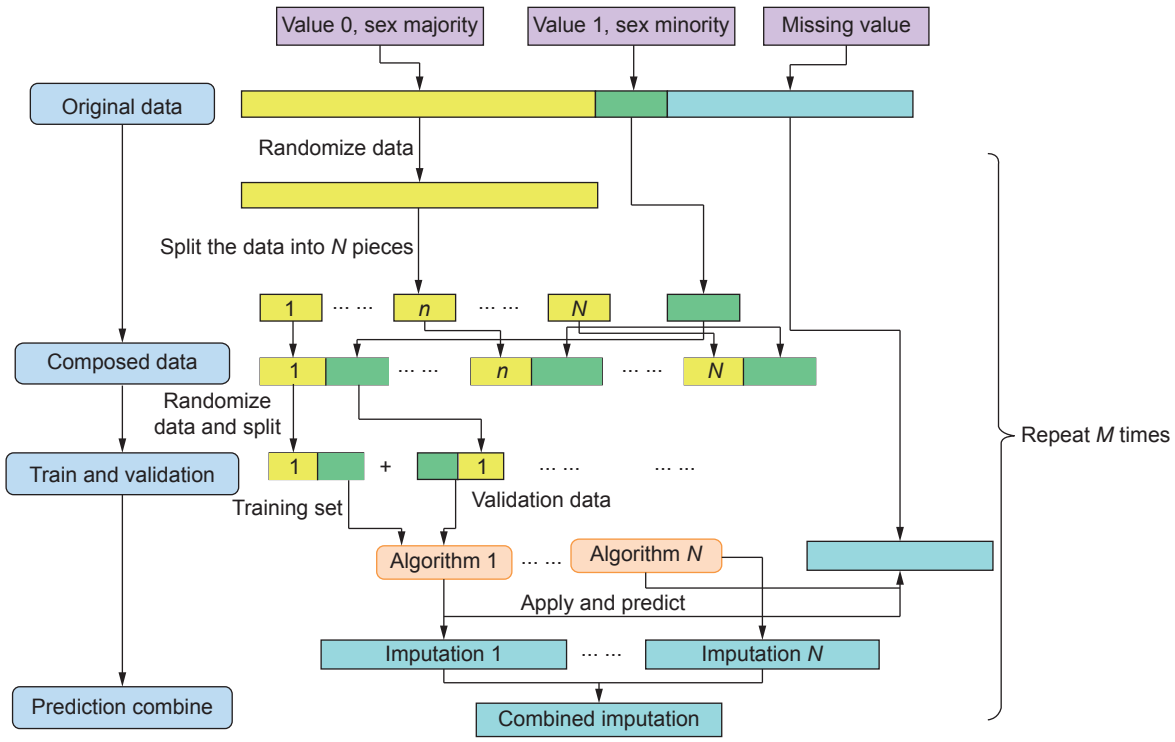

Fig. 1 Missing value imputation of sexual minority orientation.

Considering that sexual minorities are more likely to experience victimization and bullying and mental health difficulties globally due to the heavy stigma that surrounds them[18], students with sexual minority orientation may also have higher level of psychological distress if comparing them to students with heterosexual orientation. Based on the newly imputed data, we then examined the psychological wellbeing and gender attitudes between college students with sexual majority orientation and sexual minority orientation.

3 Result

3.1 Primary results

We present the results below in Tables 8 and 9 using

Table 8 Result from RF predictors using different splitting strategies (337 control variables).

Random forest	Maj:Min (Mean)	Number of samples					
		Majority=0 (*N*=3684)		Minority=1 (*N*=115)		Vacancy (*N*=244)	
		0	1	0	1	0	1
No split	3684:115	3684	0	34	81	223	21
6-split	614:115	610	4	26	89	201	43
12-split	307:115	301	6	14	101	166	78
24-split	153:115	145	8	8	107	128	116
32-split (1:1)	115:115	107	8	9	106	113	131

Table 9 Algorithm performance of RF using different splitting strategies (337 control variables).

Random forest	Train Acc	Val Acc	Train F1	Val F1	AUC
No split	0.9947	0.9763	0.9537	0.4940	0.9568
6-split	0.9902	0.8322	0.9810	0.5647	0.9735
12-split	0.9968	0.7843	0.9960	0.6327	0.9788
24-split	0.9954	0.7134	0.9953	0.6865	0.9858
32-split (1:1)	0.9985	0.6334	0.9985	0.6276	0.9822

Note: (1) Train: The abbreviation of the Train set; (2)Val: The abbreviation of the Validation set.

337 variables to predict sexual orientation using a resampling RF approach. As seen from Tables 8 and 9, when using the 6-split subsample, the results of the RF prediction show 43 out of the 244 respondents who did not report their sexual orientation are likely to be LGBQ. As for the 12-split and 24-split subsamples, the RF approach inferred that 78 and 116 out of 244 vacancy samples are likely to be sexual minorities. Note that the results from the 24 splits experiment are obtained from a relatively balanced number of sexual minority and majority cases across all training sets. When it comes to the 32 splits experiment using fully balanced data (115 vs. 115), the RF approach inferred that 131 of them are likely to be sexual minorities.

According to the metrics, the RF predictors tend to obtain higher accuracy for sexual majority samples in the presence of highly imbalanced data. For example, the validation accuracy is above 0.976 when fitting the RF model onto the original dataset, while its counterpart of the 32-split subsample is merely 0.633. Meanwhile, the failed prediction of sexual orientation for the sexual minority group decreased as the split number increased. When it comes to the none-split experiment, the ratio of mistakenly predicting sexual minority people as sexual majority people are as high as 29.6% (=34/115). In contrast, the ratio decreased to 7.9% (=9/115) when using 32-split data.

Considering the fact that many sexual minorities were most likely to decline to admit to their sexual orientation, and that the RF algorithm relies more on the information of control variables rather than adjusting strategy to make accurate predictions when the data are more balanced in terms of sexual orientation, we adopt the results obtained from 24-split. This choice is further supported by key statistics, including a higher AUC and F1 score of validation. In this vein, the corresponding percentage of sexual minorities among Beijing college students is 5.71% (=(115+116)/4043) rather than 3.03%. Therefore, the original estimation of the rate of sexual minorities among Beijing college students based on traditional surveys is nearly half of the genuine value, which is substantially downward biased.

Based on the newly created data, we further conduct the analysis to compare the subjective wellbeing between students with sexual majority orientation and students with sexual minority orientation. The results show that relative to students with sexual majority orientation, students with sexual minority orientation have higher level of depression, lower level of happiness, and they are more likely to hold resistant attitudes towards love when facing disapproval from family members and are more likely to approve homosexual behaviors (see Table A1 in the Appendix), consistent with existing literature[5, 10, 19].

3.2 Robustness check

To check the robustness of the results, we also used all 910 variables to predict sexual orientation for the sake

Table 10 Result from RF predictors using different splitting strategies (910 control variables).

Random forest	Maj:Min (Mean)	Number of samples					
		Majority=0 (*N*=3684)		Minority=1 (*N*=115)		Vacancy (*N*=244)	
		0	1	0	1	0	1
No split	3684:115	3684	0	37	78	244	0
6-split	614:115	614	0	28	87	218	26
12-split	307:115	305	2	18	97	190	54
24-split	153:115	145	8	10	105	136	108
32-split (1:1)	115:115	145	8	10	105	116	128

Table 11 Algorithm performance of RF using different splitting strategies (910 control variables).

Random forest	Train Acc	Val Acc	Train F1	Val F1	AUC
No split	0.9967	0.9645	0.9690	0.4910	0.9305
6-split	0.9920	0.8379	0.9845	0.4818	0.9723
12-split	0.9953	0.7824	0.9942	0.5536	0.9824
24-split	0.9980	0.6611	0.9980	0.6177	0.9809
32-split (1:1)	0.9997	0.6076	0.9997	0.6032	0.9787

of robustness, and report the relevant results in Tables 10 and 11. Comparing the results obtained from the 910 variables in Tables 10 and 11 with those from the 337 variables in Tables 8 and 9, we found that, overall, the predicted frequency of suspected sexual minorities is a bit smaller than that reflected in Table 8 and 9. As Tables 10 and 11 show, in the 6-split subsample there are 26 suspected sexual minorities among 244 respondents who did not report their sexual orientation. When further using the 12-split, 24-split, and 32-split subsamples, the RF approach inferred that 54, 108, and 128 of 244 vacancy samples were likely to be sexual minorities, respectively.

As to the validation accuracy of the RF model and the ratio of the false positive of heterosexual people, we observe that compared with the RF result using 337 variables, the validation accuracy of the RF predictors using the entire variable set on the original dataset is smaller, and reaches 0.965. Correspondingly, the ratio of mistakenly predicting suspected sexual minorities as sexual majorities is a bit higher and reaches 32.2% (=37/115). Furthermore, looking at the RF results on the 32-split subsamples, these two numbers are 66.11 and 8.7% (=10/115), respectively.

By comparing the AUC and Val F1 results across different splits of subsamples, the 24-split is still preferred. Although the AUC of 24-split is slightly smaller than the 12-split, the difference (0.0015) is negligible. Moreover, AUC is problematic in highly imbalanced data, under this circumstance, so F1 score should be given more weight. Therefore, based on the RF results obtained from the 24-split, the corresponding percentage of suspected sexual minorities among Beijing college students is 5.52% (=(115+108)/4043).

To cross-validate the results, we also used and compared the results from Bagging, AdaBoost, Gaussian NB, SVM, and logistic regression methods with the results from the RF approach. To save space, we only present the results obtained from the 24-split subsamples with 377 control variables (the results from the other types of subsamples are available upon request). As shown in Tables 12 and 13, by applying these competitive classifiers, the estimated percentage of suspected sexual minorities among Beijing college students ranges from 3.03% to 5.96%. The smallest estimate is generated by SVM. As mentioned earlier, the data are highly imbalanced in terms of sexual orientation; an SVM classifier trained on a dataset as such often produces models that are biased towards the majority class. Accordingly, as anticipated, almost no people are classified into a sexual minority among the 244 non-respondents using SVM. The largest estimate is generated by a boosting classifier, followed by a bagging classifier, where an estimated 126 and 116

Table 12 Comparing bagging, boosting, Gaussian NB, and SVM (24-split subsample 377 control variables).

Algorithm	Maj:Min (Mean)	Number of samples					
		Majority=0 (*N*=3684)		Minority=1 (*N*=115)		Vacancy (*N*=244)	
		0	1	0	1	0	1
Random forest	153:115	145	8	8	107	128	116
Bagging	153:115	145	8	9	106	128	116
AdaBoost	153:115	131	22	24	91	118	126
Gaussian NB	153:115	138	15	99	16	218	26
SVM	153:115	153	0	19	96	244	0
Logistic	153:115	129	24	68	47	176	68

Table 13 Algorithm performance of bagging, boosting, Gaussian NB, and SVM (24-split subsample with 337 control variables).

Algorithm	Train Acc	Val Acc	Train F1	Val F1	AUC
Random forest	0.9954	0.7134	0.9953	0.6865	0.9858
Bagging	0.9954	0.7134	0.9953	0.6865	0.9829
AdaBoost	0.8720	0.6555	0.8697	0.6391	0.9087
Gaussian NB	0.5680	0.6101	0.4254	0.3836	0.5560
SVM	1	0.6494	1	0.3937	0.9812
Logistic	0.6049	0.5786	0.5832	0.5183	0.6243

individuals are likely to be sexual minorities, respectively.

Among five alternative classifiers, bagging prevails, as it has the highest validation accuracy, the lowest false positive of heterosexual people, and is high in both ROC and Val F1. Note that random forest, which combines the concepts of bagging and random selection of features, is an extension of bagging[12]. Such results are thus in accordance with expectations. Based on the results from the bagging classifier, 116 out of 244 individuals are likely to be sexual minorities, which is consistent with the RF classifier. In this analysis, we choose RF over bagging because, compared with bagging, RF can attenuate tree correlation by injecting more randomness into the tree-growing process, which largely increases predictive power. Moreover, RF also has a better out-of-box performance. As Probst et al.[20] have shown, random forests have the least variability in their prediction accuracy when tuning among popular machine learning algorithms. To further show the performance of classification models, the ROC curves of RF with different split strategies and different algorithms are drawn in Fig. 2.

Furthermore, considering that some LGB individuals may deliberately report they are heterosexual people, we excluded samples who declared themselves as heterosexual people but were predicted as sexual minorities in the test set during the first round. Note that this does not necessarily indicate that these samples are LGB individuals who provide fake answers but only suggest they have higher probability of being LGB people compared to others. We then took the remained samples to make the prediction on missing samples. The 24-split random forest model with 337 variables shows that the percentage of LGB becomes 5.61%(we only took a one-round test as a robust check to back up our major results), which changed little compared to our general finding. The convergent trend remains consistent.

4 Discussion

In the past decade, profound social transformations have caused the emergence of a broad socio-political climate that is gradually more accepting of sexual minorities in many countries[21]. Despite the overall shift toward reduced discrimination and greater acceptance, it would be premature to claim the imminent demise of societal stigma and individual prejudice against sexual minority population. Being "invisible" in mainstream society has become a living strategy for many sexual minorities in China. The perceived stigma may drive biased responses, particularly when using traditional survey methods, as sexual minorities often feel reluctant to disclose their sexual orientation, leading to a high percentage of missing values. To overcome such barriers, we adopt random forest imputation, a machine learning approach to infer the sexual orientation of non-respondents based on survey questions from the Beijing College Student Panel Survey. After filling out those missing values,

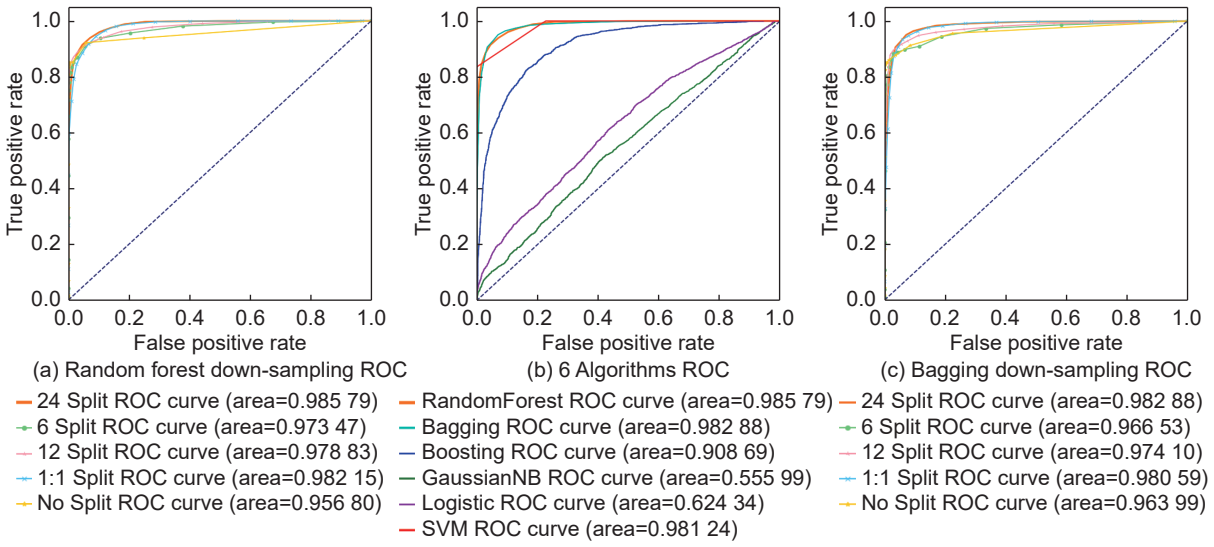

Fig. 2 ROC curves of RF with different split strategies and different algorithms.

the percentage of college students with sexual minority orientation increased from 3.03% to 5.71%, which is a bit higher than official estimation (2%−4%) based on the first nationwide survey of sexual behavior in China conducted between 1998 and 1999[22]. The results remain consistent if alternative classifiers are adopted.

Regarding our result, one doubt may raise as we did not distinguish "not sure" and "refuse to answer" but simply taking them as "missing value". While different implications may link to these answers※, we argue that this does not challenge our findings. First, as respondents in the fourth-wave BCSPS are all senior-year college students, people who do not understand the meaning of terms are rare. Second and most importantly, our models are trained to give the best predict for samples' sexual orientation without predefining their sexual preferences. If based on the training set, these models get the ability to predict an individual's sexual orientation, then ideally, they can figure out whether an individual is a sexual/gender majority or minority in any cases within the dataset. A valid prediction model independently predicts individuals' sexual orientations without researchers' influence.

Still, there are some limitations worthy of discussion. First, this analysis only provides crude estimation of percentage of college students with sexual minority orientation without further differentiating specific types, especially, it does not capture gender minority (e.g., transgender) and some subtypes of sexual minorities, namely pansexual or asexual[23]. Our coding scheme which only contains two values 1-yes and 0-no, ensures that we still could capture those people but with lower precision. Considering sensitivity of the sexual orientation and LGBQ related issue, the data that involve sexual orientation are rare in China. Besides, most survey data relevant to sexual minorities are collected using the snowball sampling strategy, the data created are the non-probability sample which have no societal representativeness. BCSPS data are actually the only public available and regional representative dataset that has information about sexual orientation. It is the one of the few data that have potential to provide the accurate inferring of LGBQ community, though the measurement of sexual orientation is far from ideal.

Second, extrapolating the sexual minority proportion in this analysis is based on the assumption that respondents reported their real sexual orientation. Note that due to the stigmatized social status of sexual minorities, respondents' answers might be biased towards prevalent social norms, which are more acceptable to mainstream society. This means there may be some sexual minorities who deliberately hide their true sexual orientation by intentionally giving inaccurate responses. It is therefore likely that our estimation of the percentage of sexual minorities is still underestimated. In other words, our results provide a conservative estimation of sexual minorities of college students in Beijing. Nevertheless, as the BCSPS is a large-scale social survey with stratified-random sampling and clear commitment to protect respondents' information security, LGB individuals generally have little necessity to deliberately pretend they are heterosexual people in an anonymous questionnaire. Furthermore, our robust check by expelling samples who declared heterosexual preference but were predicted as sexual minorities in the first round shows little change occurred compared with general findings.

Third, as the results show, the training accuracy is much higher than the validation accuracy, which implies that our RF classifier may suffer from an overfitting problem. However, Breiman[24] claimed that random forests do not overfit as they can generate an internal unbiased estimate of generalization error when more trees are added to the model. There is significant controversy surrounding this subject. Some other scholars have shown that as the forest building progresses, it is the generalization error variance rather than the bias itself that would decrease to zero in the RF[25]. In this research, by adopting a Lasso regression algorithm, we largely reduce the size of the variable sets for estimation, which, to some degree, have down-weighted bias arising due to an overfitting problem.

Finally, another potential source that may lead to biased estimation the fact that our assessment is based on a one-item question: "What is your sexual orientation?" Considering the sensitivity of this question, substantial inaccuracies may incur. A more accurate estimate might be achieved by incorporating prevalent sexual orientation tests (including, the Kinsey scale test, the Epstein sexual orientation inventory (ESOI), and Storm's (1980) sexual orientation test with quadrants, and even

※ Although "refuse to answer" clearly belongs to missing value, people who choose "not sure" may be also because they are not familiar with the meaning of terms including "heterosexual", "homosexual" , and "bisexual".

list experience) into our approach. It is also worth noting that this study is limited to assessing the sexual orientation of college students in Beijing, but does not seek to estimate the number of gender minorities.

In spite of some deficiencies, the methods we used to predict the potential sexual minorities have undoubtedly been a breakthrough. Our effort helps increase the visibility of sexual minorities, which in turn will enhance the health and wellbeing of the sexual minorities and serve to protect their rights, which are so important in achieving equality in society and before the law.

Appendix

All relevant data and coding are available from the authors. As an indirect robust check, we select four popular well-being indicators including personal subjective depression, happiness, resistance to family pressure, and whether take liberatory sexual behaviour (see Table A1). Existing studies generally report that compared to heterosexual people, LGB people are more likely to endure subjective depression, have a lower level of happiness, resist family pressure, and keep more liberal sexual behaviour. Suppose that, the regression with our imputed dataset gets counterintuitive results not only different from the dataset before the imputation but also deviate from existing findings, then the confidence in our prediction should be carefully doubted. However, we got regression results consistent with both existing findings and the pre-imputed dataset. A caveat is that this check can only suggest our prediction is "not wrong" rather than prove we are "correct".

Table A1 Sexual orientation, psychological well-being, and gender attitude of Beijing college students (odds ratio).

Sexual orientation	Depression		Happiness		Resistance		Liboratory sexual behaviors	
	Before imputation	After imputation	Before imputation	After imputation	Before imputation	After imputation	Before imputation	After imputation
Sexual minority	**1.861*** **		**0.692***		**1.433****		**5.870*** **	
	(0.382)		**(0.143)**		**(0.262)**		**(1.444)**	
Sexual minority		**2.056*** **		**0.672****		**1.625*** **		**5.697*** **
		(0.326)		**(0.109)**		**(0.231)**		**(1.008)**

Note: (1)*** p<0.01, ** p<0.05, * p<0.1. (2) Biological sex, age, ethnicity, hukou status, school rank, party member, religion, familial income, and province of origin are controlled. Cut points are omitted. (3) Numbers in parentheses are robust standard errors.

References

[1] A. Ghaziani, V. Taylor, and A. Stone, Cycles of sameness and difference in LGBT social movements, *Annual Review of Sociology*, vol. 42, no. 1, pp. 165–183, 2016.

[2] L. P. Gross, *Up from Invisibility: Lesbians, Gay Men, and the Media in America*. New York, NY, USA: Columbia University Press, 2012.

[3] USAID, Being LGBT in Asia: China country report, https://www.undp.org/sites/g/files/zskgke326/files/publications/Being%20LGBT%20in%20Asia%20-%20China%20Country%20Report%20.pdf, 2014.

[4] Y. Y. Wang, Z. S. Hu, K. Peng, Y. Xin, Y. Yang, J. Drescher, and R. S. Chen, Discrimination against LGBT populations in China, *Lancet Public Health*, vol. 4, no. 9, pp. E440–E441, 2019.

[5] J. H. Lee, K. E. Gamarel, K. J. Bryant, N. D. Zaller, and D. Operario, Discrimination, mental health, and substance use disorders among sexual minority populations, *Lgbt Health*, vol. 3, no. 4, pp. 258–265, 2016.

[6] UNDP, Being LGBTI in China: a national survey on social attitudes towards sexual orientation, gender identity and gender expression, https://www.asia-pacific.undp.org/content/rbap/en/home/library/democratic_governance/hiv_aids/being-lgbti-in-china-a-national-survey-on-social-attitudes-towa.html, 2018.

[7] W. O'Donohue and C. E. Caselles, Homophobia: Conceptual, definitional, and value issues, *Journal of Psychopathology and Behavioral Assessment*, vol. 15, no. 3, pp. 177–195, 1993.

[8] G. J. Gates, How may people are lesbian, gay, bisexual and transgender? https://williamsinstitute.law.ucla.edu/publications/how-many-people-lgbt/, 2011.

[9] P. N. P. Institute, LGBTQ students in higher education, https://pnpi.org/wp-content/uploads/2021/05/LGBTQStudentsinHigherEducation_PNPI_May2021.pdf, 2018.

[10] W. J. Xu, L. J. Zheng, Y. Xu, and Y. Zheng, Internalized homophobia, mental health, sexual behaviors, and outness of gay/bisexual men from Southwest China, *International Journal for Equity in Health*, doi: https://doi.org/10.1186/s12939-017-0530-1.

[11] Y. Hu, Sex ideologies in China: Examining interprovince differences, *The Journal of Sex Research*, vol. 53, no. 9, pp. 1118–1130, 2016.

[12] M. Khalilia, S. Chakraborty, and M. Popescu, Predicting disease risks from highly imbalanced data using random forest, *BMC Medical Informatics and Decision Making*, doi: 10.1186/1472-6947-11-51.

[13] B. Lantz, *Machine learning with R*. Birmingham, UK: Packt Publishinig, 2015.

[14] R. Tibshirani, Regression shrinkage and selection via the

Lasso: A retrospective, *Journal of the Royal Statistical Society. Series B*: *Methodological* , vol. 73, no. 3, pp. 273–282, 2011.
[15] F. Provost, Machine learning from imbalanced data sets 101, presented at the AAAI'2000 Workshop on Learning from Imbalanced Data Sets, Austin, TX, USA, 2000.
[16] N. Japkowicz and S. Stephen, The class imbalance problem: A systematic study, *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.
[17] D. S. Palmer, N. M. O'Boyle, R. C. Glen, and J. B. O. Mitchell, Random forest models to predict aqueous solubility, *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 150–158, 2007.
[18] P. R. Sterzing, W. F. Auslander, and J. T. Goldbach, An exploratory study of bullying involvement for sexual minority youth: Bully-only, victim-only, and bully-victim roles, *Society for Social Work and Research*, vol. 5, no. 3, pp. 321–337, 2014.
[19] L. Zeeman, N. Sherriff, K. Browne, N. McGlynn, M. Mirandola, L. Gios, R. Davis, J. Sanchez-Lambert, S. Aujean, N. Pinto, et al., A review of lesbian, gay, bisexual, trans and intersex (LGBTI) health and healthcare inequalities, *European Journal of Public Health*, vol. 29, no. 5, pp. 974–980, 2019.
[20] P. Probst, B. Bischl, and A. L. Boulesteix, Tunability: Importance of hyperparameters of machine learning algorithms, *Journal of Machine Learning Research*, vol. 20, no. 53, pp. 1–32, 2019.
[21] J. E. Lane, A new cultural cleavage in post-modern society, *Brazilian Journal of Political Economy*, vol. 27, no. 3, pp. 375–393, 2007.
[22] D. Wong, Sexual minorities in China, in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, J. D. Wright, ed. Amsterdam, the Netherlands: Elsevier, 2015, pp. 734–739.
[23] Y. T. Suen and R. C. H. Chan, A nationwide cross-sectional study of 15,611 lesbian, gay and bisexual people in China: Disclosure of sexual orientation and experiences of negative treatment in health care, *International Journal for Equity in Health*, vol. 19, p. 46, 2020.
[24] L. Breiman, Random forests, *Machine Learning*, vol. 45, pp. 5–32, 2001.
[25] F. Tang and H. Ishwaran, Random forest missing data algorithms, *Statistical Analysis and Data Mining*, vol. 10, no. 6, pp. 363–377, 2017.

Yunsong Chen is a professor at the Department of Sociology, Nanjing University. He earned the PhD degree from Oxford University in 2012. His research focuses on social capital, social network, causal inference, and computational social science. He has published in *British Journal of Sociology, Social Science & Medicine, Social Networks, Poetics, Chinese Sociological Review, etc*.

Guangye He is an associate professor at the Department of Sociology, Nanjing University. She earned the PhD degree from the Hong Kong University of Science and Technology in 2016. Her research focuses on family sociology, social stratification, and quantitative methodology in sociology. She has published in *Social Science Research, Poetics, Chinese Sociological Review, China Review, and Journal of Contemporary China*.

Guodong Ju is a PhD candidate in the Department of Social Policy, London School of Economics and Political Science. His research interests cover social equality, minority rights, and social process. He applies quantitative approaches including causal inference, computational social science, and social network analysis.