STAT3302 Project

Group E: Jiaxin Yang, Tingwei Guan, Jike Zhong, Yahui Zhou, Jiangyue Zhu

4/18/2022

Introduction

The paper that we discuss analyzes whether activity engagements and value factors are statistically associated with some demographic characteristics based on responses from 3,204 adult residents of the U.S.A. about their relationship with the non-human world during the pandemic. To focus on measuring the impact of the COVID-19 pandemic on human-nature interactions, we analyzed reported changes in 6 outdoor activities and 5 demographic characteristics by using the paper's processing and thoughts. Within the 3,204 valid responses, we found that the Covid-19 has resulted in more people engaging in nature-related, self-oriented, and easily accessible activities, rather than social-gathering activities or those that are not so easily accessible. It may suggest that people tend to stay more at home and be alone during the pandemic. We mainly apply multinomial logistic regression to find the relationship between different variables and replicate results.

Methods

In the paper, the author utilizes several statistics derived from multinomial logistic regression to explain the results of analyzing COVID data. Here are the explanations of these statistics used below.

To process the analysis, we generated the multinational logistic regressions to identify socio-demographic patterns for the six most common activities among respondents. In selecting activities to model, we used the most common activities because we were interested in the impact of COVID-19 on access and the possible equity dimensions of this impact, and we considered the more-widely-engaged-in activities the likeliest to show evidence of socio-demographic differences. The whole processing is beginning from multinomial distribution, to hypothesis test, and derive multinomial logistic regression to get the results.

Firstly, we identified the multinomial distribution among our data. We considered the top six popular activities separately as six random variable Y among the respondents. Here are the variables name: Gardening, Hiking, Relaxing Socially Increased, Relaxing Socially Decreased, Relaxing Alone, Walking, Wildlife Watching. We take Gardening as an example to illustrate the distribution model, since we analysis the six respondents individually.

We considered a random variable Gardening Y with 5 demographic characteristics, geography, gender, race/ethnicity, income, employment. Let $\pi_1, \pi_2, \ldots, \pi_5$ denote the respective probabilities with $\pi_1 + \pi_2 + \ldots + \pi_5 = 1$. If there are 5 independent observations

of Y which result in y_1 outcomes in category 1, y_2 outcomes in category 2, and so on, then let

$$\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_5 \end{bmatrix}, with \sum_{j=1}^{5} y_j = n.$$

The multinomial distribution is

$$f(\mathrm{y}|n) = \frac{n!}{y_1!\, y_2!\, y_3! \dots y_5!} {\pi_1}^{y_1} {\pi_2}^{y_2} \dots {\pi_5}^{y_5} \quad (1)$$

it is denoted by M(n, π_1,π_2,…,π_5), n = 2521 in gardening category.

The following relationship with the Poisson distribution ensures that generalized linear modelling is appropriate. Let Y_1,…, Y_5 denote independent random variables with distributions Y_j~Poss(λ_j). Their joint probability distribution is

$$f(\mathrm{y}) = \prod_{j=1}^{5} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}. \quad (2)$$

Let n = Y_1 + Y_2 + … + Y_5, then n is a random variable with the distribution n ~ Poss(λ_1 + λ_2+…+λ_5). Therefore, the distribution of y conditional on n is

$$f(\mathrm{y}|n) = \left[\prod_{j=1}^{5} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \right] / \frac{(\lambda_1 + \lambda_2 + \dots + \lambda_5)^n e^{-(\lambda_1+\lambda_2+\dots+\lambda_5)}}{n!}. \quad (3)$$

If $\pi_j = \lambda_j / \left(\sum_{k=1}^{5} \lambda_k\right)$, for j = 1, 2, …, 5, then (1) is the same as (3) and $\sum_{j=1}^{5} \pi_j = 1$. Therefore, the multinomial distribution can be regarded as the joint distribution of Poisson random variables, conditional upon their sum n. This result provides a justification for the use of generalized linear modeling.

In addition, we chose multinominal logistic regression as our regression modal to further extend our analysis. Multinominal logistic regression models are used when there is no natural order among the response categories. One category is arbitrarily chose as the reference category. In our model, we supposed the first category, gardening, as the reference category. Then the logits for the other 5 categories are defined by

$$logit(\pi_j) = log\left(\frac{\pi_j}{\pi_1}\right) = \mathrm{x}_j^T \beta_j,\ for\ j = 2,3,\dots,5. \quad (4)$$

The 5 logit equations are used simultaneously to parameter estimates estimate the parameters β_j. Once the b_j have been obtained, the linear predictors $\mathrm{x}_j^T b_j$ can be calculated. From (4), $\widehat{\pi_j} = \widehat{\pi_1} exp\left(\mathrm{x}_j^T b_j\right)$, for j = 2, 3, …, 5. Since $\widehat{\pi_1}$ + $\widehat{\pi_2}$ + … + $\widehat{\pi_5}$ = 1, so

$$\widehat{\pi_1} = \frac{1}{1 + \sum_{j=2}^{5} exp\left(\mathrm{x}_j^T b_j\right)}$$

and

$$\widehat{\pi_j} = \frac{exp\left(\mathrm{x}_j^T b_j\right)}{1 + \sum_{j=2}^{5} exp\left(\mathrm{x}_j^T b_j\right)}, for\ j = 2,3,\ldots,5.$$

Fitted values for each covariate pattern can be calculated by multiplying the estimated probabilities $\widehat{\pi_j}$ by the total frequency of the covariate pattern.

Then, the paper uses Chi-Square Test Hypothesis to set up the potential conclusion and build multinomial logistic regression model to compute statistics to gain results, but what is the Chi-Square Test Hypothesis?

Chi-Square Test Hypothesis:

To process the significant test, we need to set up a hypothesis for Chi-Square goodness of fit test for each individual category. Under each individual category of activities, we set the null hypothesis assumes that there is no significant difference between the null model and the reduced model among five different factors. On the other hand, we set the alternative hypothesis assumes that there is a significant difference between the null model and the reduced model among five different factors.

In this case, various statistics help a lot. We mainly compute Chi-squared statistic, Pseudo R_2, and log ratio of multinomial logistic regression to test whether respondents' preferences for changes in six activity engagements are statistically associated with five demographic characteristics.

Pearson chi-squared residuals

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}$$

where o_i and e_i are the observed and expected frequencies for i = 1, . . .

Chi-squared statistic

$$X^2 = \sum_{i=1}^{N} r_i^2$$

A chi-squared test is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, which also refers that the sampling distribution (if the null hypothesis is true) of the test statistic approximates a chi-squared distribution more and more closely as sample sizes increase. When $\alpha < 0.05$, we can say that our training data locates in the right-tail of our model, which can illustrate the relationship between activities and demographic characteristics better.

Deviance, defined in terms of the maximum values of the log-likelihood function for the fitted model, l(b), and for the maximal model, $l(b_{max})$.

$$D = 2[l(b_{max}) - l(b)]$$

Likelihood ratio chi-squared statistic, defined in terms of the maximum value of the log likelihood function for the minimal model, $l(b_{min})$, and l(b).

$$C = 2[l(b) - l(b_{min})]$$

Therefore, we get the formula to compute $PseudoR_2$

$$PseudoR^2 = \frac{l(b_{min}) - l(b)}{l(b_{min})}$$

The odds ratio OR_j is another convenient statistic to interpret the effects of explanatory factors than the parameters β. Based on the multinomial logistic regression simple model above, we get

log odds:

$$log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) = \beta_{0j}$$

when x = 0, indicating the relation is absent,

$$log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) = \beta_{0j} + \beta_{1j}$$

when x = 1, indicating the relation is present.

the logarithm of the odds ratio:

$$logOR_j = log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) = \beta_{1j}$$

Hence, $OR_j = exp(\beta_{1j})$.

Based on those methods, we generate the following codes to gain final results.

Results

In Figure 1, we compare the overall socio-demographic data of our sample with the Vermont population. By calculating the various percentages of the Vermont population and comparing it with the percentage of citizens of the United States, we concluded that the percentage of the Vermont population matches with the percentage of U.S. citizens, and the sample is broadly similar to the Vermont population, so that we can predict the trends across the United States by analyzing the outdoor activities and the values associated with human-nature relationships across geographic areas and demographic characteristics.

Within the 3,204 valid responses, we discuss the characteristics in 6 aspects: age, gender, race, ethnicity, household income, and ZIP code. In Table1, we compare the aggregated demographic data of our sample to the Vermont population (published by official U.S.A. census). Our sample is approximately close to the Vermont population. The mean age of the sample is around 54.7 years old. Nearly one-third of the participants (63.2%) are female, almost all of the participants (91.6%) are white and (99.6%) are not Hispanic or Latino. The household income in 2019 of most individuals reported ranges from $25,000 to 149,999 evenly. Half of the participants come from urban areas.

Table 1. Demographic characteristics of survey respondents, tabulated for comparison to the 2014–2018 U.S.A. Census Bureau's American Community Survey

Characteristic	Sample	U.S.A. Census
Mean age	54.7 years old	50.5 years old
Gender		
Female	63.2% (2,013)	50.7%
Male	35.8% (1,139)	49.3%
Non-binary	1.0% (32)	
Race		
American Indian or Alaskan Native	0.4% (13)	0.3%
Asian	0.5% (15)	1.7%
Black or African American	0.1% (2)	1.3%
Middle Eastern or North African	0.1% (3)	
Two or more races	3.3% (105)	1.9%
White	91.6% (2,936)	94.3%
Other	4.1% (130)	
Ethnicity		
Hispanic or Latino	0.4% (13)	1.9%
Not Hispanic or Latino	99.6% (3,191)	98.1%
2019 Household Income		
< $10,000	1.2% (35)	4.9%
$10,000 - $24,999	7.2% (217)	14.7%
$25,000 - $49,999	18.4% (556)	22.1%
$50,000 - $74,999	22.3% (673)	18.8%
$75,000 - $99,999	18.9% (572)	14.0%
$100,000 - $149,999	20.3% (613)	15.3%
$150,000 - $199,999	6.7% (203)	5.1%
> $200,000	5.1% (154)	5.0%
Zip Code within Census urban-rural classification		
Urbanized Area	26.0% (823)	17.4%
Urban Cluster	25.9% (820)	28.2%
Rural	48.2% (1,528)	54.4%

Figure 1: Sample to Population

In order to visualize the impact of Covid-19 on human-nature relationships, we produced a stacked barplot showing the level of impact ("Less", "Same", "More", and "Don't Do This Activity), normalized into the percentage of response, Covid-19 had on each of the 15 specific activities (selected based on categories put forth in the Vermont Statewide Comprehensive Outdoor Recreation Plan). For simplicity and clarity, we utilized matplotlib library with Python environment to help us produce the figure, but other methods such as using Matlab or R should produce the identical result.

The figure shows that Covid-19 has had a rather significant impact on the human-nature relationships, as the union of "more" and "less" responses largely outweighs the response "same" for most of the activities we investigated. More specifically, on one hand, our figure shows that Covid-19 has resulted in more people engaging in nature-related, self-oriented, and easily-accessible activities such as "Walking", "Wildlife", and "Hiking", "Gardening", and "Relaxing Oneself". On the other hand, Covid-19 has also resulted in people engaging less in social-gathering activities or those that are not so easily accessible such as "Gathering", "Hunting", "Camping", and "Boating". We think the result is reasonably owing to the fact that during Covid-19, people tend to stay more at home (as a result of the global-wide implementation of lock down) and alone.

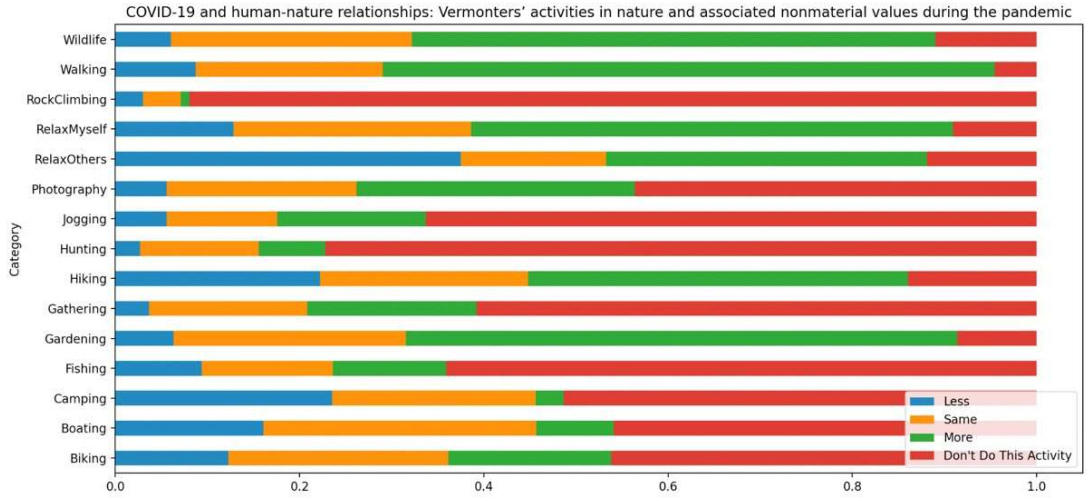

Figure 2: Human-nature relationships

In order to discover the socio-demographic factors associated with increased engagement in activities, we use the multinomial logistic regression (Figure 3) on each 6 activities. In the model, for the independent variables, we take rural, female, above median, lost job as the reference level for geography, gender, race/ethnicity, income, and employment. We divide each 6 activity (dependent variable) into 4 levels: less, same, more, don't do this activity. To find the factors (independent variables) which would significantly increase the odds ratio for reporting increased activity engagement, we choose "same" as the baseline and discuss how the factors affect the odds ratio of "more" over "same". If the odds ratio is close to 1, the factor (independent variable) does not significantly increase people's participation rate in the activity. If the odds ratio is much higher than 1, then we can

conclude that the factor (independent variable) leads to a higher participation rate in the activity.

	Geography	Gender	Race / Ethnicity	Income	Employment
Gardening	Rural: 1.27	Female: 1.64			Lost Job: 1.69
Hiking		Female: 1.70			
Relaxing Socially Increased		Female: 2.81		Above Median: 1.22	Lost Job: 1.71
Relaxing Socially Decreased	Rural: 1.45	Female: 1.52			
Relaxing Alone		Female: 1.93			
Walking	Rural: 1.32	Female: 2.90			Lost Job: 1.61
Watching Wildlife		Female: 2.09			Lost Job: 1.65

Figure 3: Demographic variables that showed significantly higher odds for increased activity engagement compared to no change and were significant in the model

Based on our results, we find that

- females who lost their job and live in the rural tend to have more gardening activity; females are more likely to hike and relax alone;
- females who lost their job but have a higher income than the average tend to relax socially, while females who live in the rural are less likely to relax socially and tend to walk and watch wildlife instead.

Take gardening activity as an example. Based on our data, people living in the rural increase the odds of increased engagement as opposed to the same engagement in gardening by 1.27 times, controlling for gender and employment level; Female increase the odds of increased engagement as opposed to the same engagement in gardening by 1.64 times, controlling for geography and employment level; people who lost their job increase the odds of increased engagement as opposed to same engagement in gardening by 1.69 times, controlling for geography and gender.

The results suggest that nature may play an important role in copying during the COVID-19 pandemic, but the interactions and related values that people perceive may be different among populations. The findings emphasize the significance of both the planning of natural resources and understanding how and why people interact with nature during crises like the COVID-19 pandemic.

Comments

Some aspects need to be improved in further similar topic studies. First, the sample in this study took random convenience of internet users on a platform called Front Porch Forum, which led to a higher proportion of females than that males, and a higher salary group, which led to the skew towards higher income ranges among respondents, which may have larger statistical bias rather than more appropriate random selection. Besides, the sample excluded the people under the age of 18, which is also an important group target.

Second, considering the possibility that some effects we observed might be influenced by special events other than the pandemic, the activity engagement questionnaire asked respondents to compare their activity levels during the COVID-19 to those one year before. Accordingly, special events that are not related to COVID-19, such as developing a new hobby, having a baby, or injury in an accident, may explain changes in people's interactions with nature. In all, however, we can assume that most of the changes taken during the pandemic were directly related to the COVID-19 itself.

Nevertheless, overall, the study indicates how the COVID-19 made an impact on Vermont residents' interactions with nature among different demographic characteristics in the early stages of the pandemic in the U.S.A. The results in this research add to our knowledge of the potential for change within the relationship between human and natural crises, as well as how these kinds of changes may vary related to different populations of diversity. Lastly, the findings are helpful for practical implications for the planning of natural resources during special natural disaster periods in advance.

Appendix

Here are the GitHub link for our coding based files. The figure 2 is generated by Python using run.py file, and the figure 3 is computed by R using demographic.Rmd file.

https://github.com/Jike338/STAT3303_Research_Project.git