Bias in Citizen Science: an application to the BioDiversity4All project

João Alves[1]

Instituto Superior Técnico, Av. Rovisco Pais, 1 1049-001 Lisboa, Portugal

Abstract. Citizen Science projects are growing in number and users, and their data can drive research in numerous fields, such as studies related to biodiversity monitoring and analysis of species distributions. However, most of these projects share a common challenge: the quality of their data depends heavily on their user base's preferences, skills, and observation efforts. To engage a large number of participants, Citizen Science projects raise their accessibility by lowering restrictions and simplifying methods of collecting data, which can lead to biased data. In this work, we provide an analysis of BioDiversity4All's observations. We use a Hurdle model to describe the distribution of observation counts to find possible explanatory variables that influence this distribution. Our results suggest that geographical accessibility is one of the most critical factors for the observers. We also used a SARIMA model to evaluate the impact that the recent Covid-19 restrictions had on BioDiversity4All's observation counts. Our results suggest that the BioDiversity4All project did not suffer from this pandemic; in fact, we observe a substantial increase in observations submitted to the platform during this period, including during lockdown periods.

Keywords: Citizen Science · Spatial and Temporal bias · Hurdle Model · Time Series forecasting · Covid-19.

1 Introduction

Citizen Science project consists in a collaboration between researchers, scientists and the general community, with the common goal of gathering relevant scientific information that can later be used in research or education [1, 2]. Throughout the years, many projects have been created, which spread across many fields of study, ranging from Astronomy to Ornithology. Citizen scientists is the term that describes volunteers who participate in these projects. These volunteers are not necessarily experts on the subject at hands, most of them are curious or concerned people who want to contribute to scientific investigations related to a field of their interest. Nowadays, anyone with access to the internet can become a citizen scientist. One only needs to provide observational data directly into a project's database and/or discuss its findings with fellow volunteers [3]. This is a crucial aspect of Citizen Science because it provides researchers with the possibility of gathering widely spread data about a certain field without being limited by time or budget constraints [4].

A project's participation rate and with it, its success, depends heavily on public engagement. In order to keep current volunteers interested and to capture the attention of new ones, project managers need to lower the project's restrictions and simplify protocols for gathering data. These methods prolong a project's longevity, but with lower restrictions and untrained volunteers, the bias present in gathered data may increase [5].

BioDiversity4All[1] is a Portuguese Citizen Science platform that started in 2010 and contains species' observations recorded in Portugal. It provides an insight into the Portuguese biodiversity, while stimulating the cooperation of its users. In this platform, volunteers can also add historical data, which means that they can submit observations made in the past, so we have access to data from as far back as 1970. Usually these observations contain photos as well as GPS coordinates, and the volunteer that submits them can provide an (often tentative) identification of the observed specimen. Later, other users, usually more experienced, can validate this identification. If a consensus is reached, the observation is tagged as *Research-Grade* and becomes viable to be used for research purposes. An early look into the platform shows us that it has low restrictions – volunteers only need a registered account to participate – and provides an easy submission method, via the smartphone app or website, which can lead to possible bias of submitted observations.

BioDiversity4All and other similar Citizen Science platforms target the gathering of observational data. The use of this data for research purposes comes with additional challenges, since there is no experimental design that guarantees that collected data accurately represents a species' population under study without possible distortions. These distortions can arrive via a series of factors that may not have been taken into consideration during the collection phase. These factors can induce considerable bias in the data, and apart from the original goals of possible studies, researchers have to find ways to deal with these biases, that can negatively impact the results of their study, if overlooked.

In this work, we aim to study the observations present in BioDiversity4All, and understand the impact that different geographical variables have on their distribution. At the time of producing this work, the world fell victim to the Covid-19 pandemic outbreak. In Portugal, confinement laws were established and people were restricted from going outside. Therefore, we also studied the impact that these pandemic restrictions had on BioDiveristy4All and its observations.

2 Data and Methods

2.1 Data collected

We collected a total of 495 401 observations from BioDiversity4All's platform. These observations contained GPS coordinates, species' information and ranged from 1970 to 30 Apr 2021. We used the observations until 31 Dec 2019 – 275 069 in number – for the explanatory variables analysis. The observations ranging

[1] BioDiversity4All. Available from http://www.biodiversity4all.org/

from 01 Jan 2020 to 30 Apr 2021 – 220 332 in number – were used for the analysis regarding the impact of the Covid-19 pandemic restrictions.

In order to explain the distribution of observation counts, we gathered potential explanatory variables, shown in Table 1.

Table 1. Potential explanatory variables used in the count distribution models.

Variable Name	Description	Unit	Sourced By
areaAGR	Agriculture and Agroforestry area %	km^2/km^2	COS2018 [6]
areaART	Artificial area %		
areaFLO	Forestal and scrubland area %		
areaPAST	Pasture area %		
areaWATER	Wetlands and water surface area %		
areaRES	Natural reserves area %		ICNF [7]
logPOP	Human population density	individuals/km^2; log-transformed	PORDATA [8]
densROADS	Road Density	km/km^2	Geofabrik [9]
densPATHS	Path Density		
meanALTITUDE	Mean Altitude	m	Worldclim [10]
meanTEMP	Mean Annual Temperature	ºC	
meanPRECI	Mean Annual Precipitation	mm	

QGIS [11] was used to create a map of mainland Portugal with a $5km^2$ grid, where each grid cell was associated to the count of observations made inside them, in total and for each yearly season (Spring, Summer, Autumn, and Winter). We also used QGIS to calculate the percentages/averages of each explanatory variable for their respective grid cell.

2.2 Statistical Models

To understand the importance of each explanatory variable in the distribution of observations and since there were a lot of grid cells with 0 counts, we decided to use *Hurdle* models [12], which are commonly used when dealing with excess zeros and over dispersed counts, to model observations submitted until 31 Dec 2019. To study the impact of the Covid-19 pandemic restrictions on BioDiversity4All's observation counts, we decided to use *SARIMA* models [13, 14] to perform a time series analysis that included the observations made during the pandemic.

All of the models were built using R [15]. *Hurdle* models were fitted using the package *pscl* [16] and *SARIMA* models using the packages *astsa* [17] and *forecast* [18]. For *Hurdle* models, we started by fitting them with all with the explanatory variables listed in Table 1. Then, we began removing variables that were not significant (p-value > 0.05) from both the count and zero components, thus ending only with explanatory variables from which we could extract information

regarding the factors that influence the distribution of observation counts of BioDiversity4All. Modelling considering spatial or temporal autocorrelation were left for future work.

We considered five scenarios for the *Hurdle* models: modelling all counts and modelling seasonal counts (one scenario per season). With the aim of obtaining the best model for each scenario, we ended up fitting eight different models. These models only differ in the values of the variables areaAGR, areaART, areaFLO, areaPAST, and areaWATER. For the sake of simplicity, we decided to only show the resulting coefficients for the models fitted with the original values for the areas. The results regarding the rest of the models can be seen in the thesis.

For *SARIMA* models, we started by converting weekly counts into a count time series ranging from 01 Jan 2015 to 31 Dec 2019. Then we used the *auto.arima* function from the *forecast* R package, which fits multiple variations of *SARIMA* models returning the one that has the best *AICc* value. The *SARIMA* models were fitted to the previously created count time series. We used the *forecast* function, in conjunction with the best fitted model, to predict counts from 01 Jan 2020 to 31 Apr 2021, so we could compare the predicted counts of a non-pandemic scenario with the observed counts, submitted under the Covid-19 restrictions.

3 Results and Discussion

3.1 BioDiversity4All's observations

To understand if BioDiversity4All's observations are biased, we used QGIS to create a map of mainland Portugal, where we applied a $5km^2$ grid and gathered all of the observations made inside each grid cell, associating the count of observations with the respective grid cell. You can find the resulting map, as well as the plot of observation counts grouped by taxonomic groups in Fig. 1.

By observing the map we can see that there is a higher concentration of observations in some areas with more than 1 000 observations, while the rest of the map only has at most 5 observations and many grid cells have 0 observations. By comparing the map in Fig. 1 with the distribution of explanatory variables such as logPOP, areaRES or areaART, we can see that most observations are made in regions with higher percent coverage of artificial areas, natural reserves and higher human population density.

Fig. 1 (right plot) shows us that there are 3 taxonomic groups that are most preffered by BioDiversity4All's users. The taxons *Aves*, *Insecta*, and *Plantae* have 91 656, 51 709, and 87 471 observations, respectively and are followed by the taxon *Fungi* which has 8 963 observations. This may be explained by the fact that these three taxons are more abundant or easily observable.

Given the higher observation counts on cells with higher population density, and/or better human access, this suggests that the number of observations is biased relative to the true species' abundance across the country. Further studies are needed to quantify the bias, so that the use of this data for research purposes takes this issue into account.

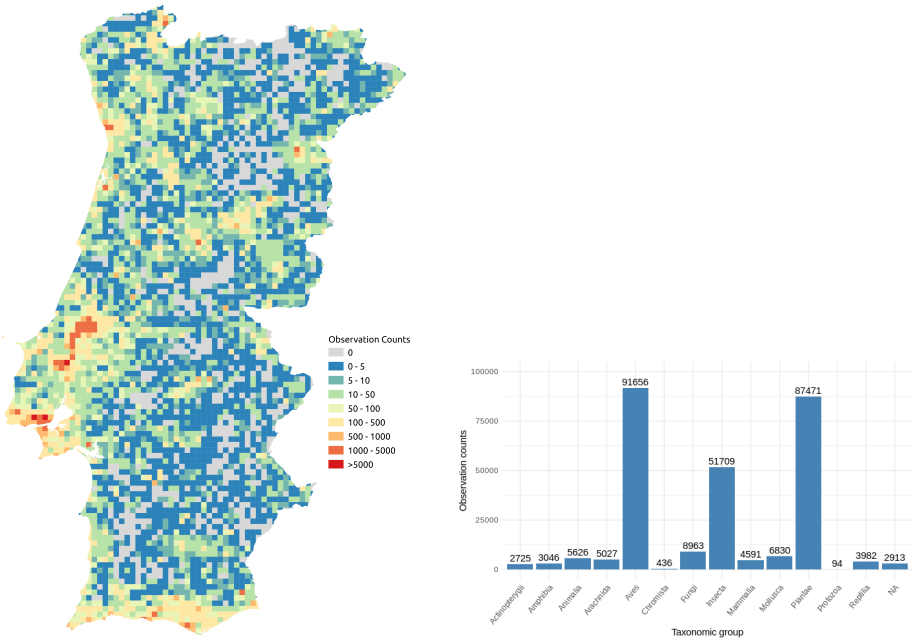

Fig. 1. On the left, a map of mainland Portugal with a 5km^2 grid. The color of a cell represents the number of observations per cell. They represent BioDiversity4All's observations ranging from 1970 to 31 Dec 2019. On the right, a bar plot that shows counts of observations grouped by their respective iconic taxon. The map on the left was created with QGIS.

3.2 Impact of geographical factors on observations' spatial distribution

With the objective of understanding the importance of each explanatory variable in the distribution of observations, five Hurdle models were fitted. One using the count of all observations up to 31 Dec 2019, and four models with the same counts, but grouped by yearly season. The resulting model coefficients are presented in Table 2.

By analysing the estimated coefficients (Table 2) regarding the zero component, which tells us which explanatory variables have the most influence on the odds in favor of an observation occurring, we can see that, across all models, logPOP and densROADS remain positive and statistically significant. This suggests that observations are more likely to be recorded in grid cells with higher population and road density. Another variable that remains negative and significant across all models is meanPRECI. Although it has a small value, it also tells us that in grid cells with higher mean annual precipitation, there is a small decrease in the odds of an observation occurring.

Regarding the count component, which represents the explanatory variables that influence the counts of observations, we can see that densPATHS and/or

densROADS are significant with positive coefficients for all models, which suggests that users tend to make more observations in easily accessible locations. Human population density and mean annual temperature are positive and significant for all models. The coefficients in this component seem to agree with the ones in the zero component, logPOP and densROADS/densPATHS have a positive influence on the odds of having an observation, and on the number of observations performed.

Table 2. Model coefficients obtained from the fitted Hurdle models, under all scenarios. For each model, the model coefficients are presented on the left column, and their corresponding Std. Errors are on the right column, inside parenthesis. The asterisks (*) represent the p-value associated with the coefficient: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Variables	**Model**	**Model Coefficients**									
	Component	All Counts		Spring		Summer		Autumn		Winter	
Intercept	Zero	1.056 ***	(0.200)	-98.759 *	(38.533)	-1.964 ***	(0.332)	-3.346 ***	(0.754)	-6.164 ***	(1.058)
	Count	-5.476 ***	(0.936)	-16.023	(24.522)	-8.893	(9.543)	-14.029	(56.869)	-6.963	(6.830)
logPOP	Zero	0.466 ***	(0.130)	0.408 ***	(0.102)	0.604 ***	(0.098)	0.931 ***	(0.098)	0.557 ***	(0.095)
	Count	0.675 ***	(0.094)	0.517 ***	(0.117)	——		0.823 ***	(0.115)	0.851 ***	(0.137)
areaAGR	Zero	——		0.947 *	(0.385)	0.008 *	(0.004)	——		——	
	Count	0.014 **	(0.005)	-0.010 ***	(0.002)	——		-0.012 ***	(0.002)	——	
areaART	Zero	——		1.000 **	(0.385)	——		——		0.024 *	(0.012)
	Count	——		-0.038 ***	(0.011)	-0.041 ***	(0.010)	-0.034 **	(0.011)	——	
areaFLO	Zero	——		0.947 *	(0.385)	0.013 ***	(0.004)	——		——	
	Count	0.025 ***	(0.004)	——		0.012 ***	(0.002)	——		0.021 ***	(0.002)
areaPAST	Zero	0.010 *	(0.005)	0.949 *	(0.385)	——		0.011 **	(0.004)	——	
	Count	——		-0.021 ***	(0.006)	——		——		——	
areaWATER	Zero	——		0.969 *	(0.386)	0.026 ***	(0.006)	0.013 **	(0.005)	0.015 **	(0.005)
	Count	0.036 ***	(0.008)	——		0.034 ***	(0.009)	——		——	
areaRES	Zero	0.019 ***	(0.002)	0.019 ***	(0.002)	0.010 ***	(0.001)	0.013 ***	(0.001)	0.012 ***	(0.001)
	Count	0.020 ***	(0.002)	0.022 ***	(0.002)	0.016 ***	(0.002)	0.015 ***	(0.002)	0.021 ***	(0.002)
densPATHS	Zero	——		——		0.634 ***	(0.147)	0.518 ***	(0.120)	0.710 ***	(0.114)
	Count	0.332 **	(0.103)	0.450 ***	(0.135)	0.695 ***	(0.133)	0.245 *	(0.108)	——	
densROADS	Zero	0.724 ***	(0.065)	0.301 ***	(0.063)	0.527 ***	(0.042)	0.378 ***	(0.035)	0.213 ***	(0.054)
	Count	0.294 ***	(0.036)	0.308 ***	(0.057)	0.417 ***	(0.052)	0.230 ***	(0.059)	0.195 ***	(0.032)
meanALTITUDE	Zero	——		0.001 ***	(0.001)	0.001 *	(0.001)	0.001 **	(0.001)	0.001 *	(0.001)
	Count	0.001 ***	(0.001)	0.002 ***	(0.001)	——		——		-0.001 ***	(0.001)
meanPRECI	Zero	-0.001 ***	(0.001)	-0.001 ***	(0.001)	-0.001 ***	(0.001)	-0.001 ***	(0.001)	-0.001 **	(0.001)
	Count	-0.001 **	(0.001)	——		——		-0.001 **	(0.001)	-0.002 ***	(0.001)
meanTEMP	Zero	——		0.233 ***	(0.039)	——		0.117 **	(0.040)	0.310 ***	(0.056)
	Count	0.300 ***	(0.049)	0.445 ***	(0.050)	0.108 ***	(0.029)	0.236 ***	(0.040)	——	

3.3 Impact of pandemic restrictions (COVID-19)

To evaluate the impact that the confinement restrictions imposed due to Covid-19 had on BioDiversity4All, we decided to analyse the weekly count of observations from 1 Jan 2015 to 30 Apr 2021. We aim to find out if there were any changes regarding the number of observations submitted. The plot of weekly observation counts can be seen in Fig. 2.

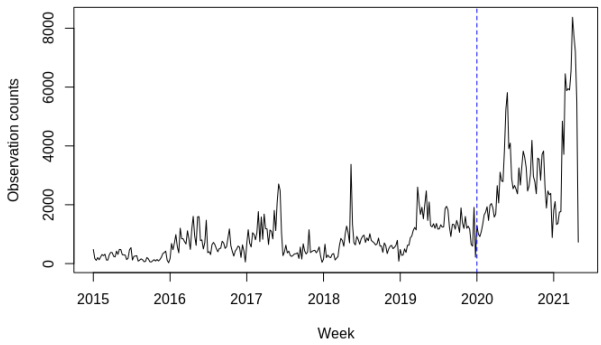

Fig. 2. Weekly counts in BioDiversity4All, ranging from 1 Jan 2015 to 30 Apr 2021. The counts to the left of the blue dashed line were used to fit the *SARIMA* model and the counts to the right were used for comparison with the predicted counts, obtained from the model.

Fig. 2 suggests that there was a considerable increase in the number of observations made past 2020. Considering that, in Portugal, confinement laws were established midst March 2020, this is a surprising result.

Although there was a significant growth in number of observations, we can see in the plot that the number has been rising through the years, which is to be expected due to the increase in popularity of Citizen Science and the inherent growth of BioDiversity4All as a project.

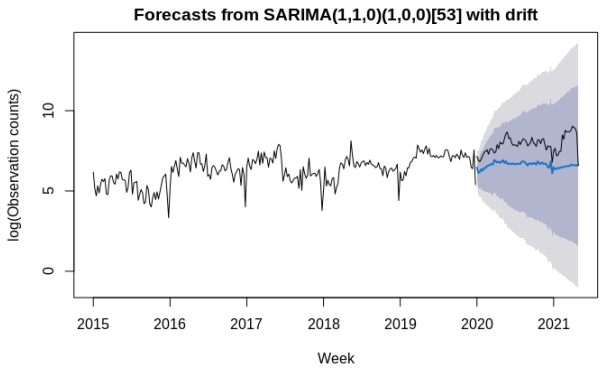

Fig. 3. Forecasted log-counts of BioDiversity4All. The black line represents observed weekly counts, log-transfomed, collected from BioDiversity4All. The blue line represents the mean values of log-counts predicted using the fitted *SARIMA* model. 95% and 80% predicted intervals are represented by the dark and light blue bands, respectively.

By fitting a *SARIMA* model to the of log-counts up to 2020, we can predict the expected log-counts from 01 Jan 2020 to 30 Apr 2021 and compare them with the observed log-counts, so we can understand if what was observed follows the patterns predicted by the *SARIMA* model or if an increase was observed. The resulting plot with the forecasted log-counts is shown in Fig. 3.

By looking at Fig. 3 we can see that the line that represents the observed counts remains consistently above the predicted counts. This tells us that there was an actual growth in observation counts, which is interesting considering the restrictions imposed during this period. However this growth is consistent with the predicted model, as suggested by the 95% predicted intervals.

4 Conclusions

The analysis regarding the distribution of observation counts in BioDiversity4All, suggests that there may be some bias present in BioDiversity4All's observations. Some taxonomic groups are overrepresented, and the observations do not seem to be evenly distributed across the country. This may be explained due to the fact that Citizen Science data heavily depends on its participants, their preferences and observational effort.

The results of the analysis regarding the relationship between certain geographical variables and the distribution of the observation counts suggest that, for all scenarios, accessibility and higher populated areas seem to be essential factors for users recording their observations.

Regarding the impact that pandemic restrictions had on BioDiversity4All, our results suggest that, despite the restrictions imposed during Covid-19, it seems that there was no impact in the submission of observations. In fact, they suggest the opposite: We have shown that there was considerable growth in the number of observations during this period, including during national lockdown.

References

1. Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., Shirk, J.: Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. BioScience 59(11), 977–984 (12 2009), https://doi.org/10.1525/bio.2009.59.11.9
2. Silvertown, J.: A new dawn for citizen science. Trends in Ecology & Evolution 24(9), 467 – 471 (2009), http://www.sciencedirect.com/science/article/pii/S016953470900175X
3. Bonney, R., Shirk, J., Phillips, T., Wiggins, A., Ballard, H., Miller-Rushing, A., Parrish, J.: Next steps for citizen science. Science (New York, N.Y.) 343, 1436–7 (03 2014)
4. Dickinson, J., Zuckerberg, B., Bonter, D.: Citizen science as an ecological research tool: Challenges and benefits. Annual Review of Ecology and Systematics 41, 149–172 (12 2010)
5. Cohn, J.P.: Citizen Science: Can Volunteers Do Real Research? BioScience 58(3), 192–197 (03 2008), https://doi.org/10.1641/B580303

6. Direção-Geral do Território: Carta de Uso e Ocupação do Solo (2018), http://mapas.dgterritorio.pt/atom-dgt/pdf-cous/COS2018/ET-COS-2018_v1.pdf
7. Instituto da Conservação da Natureza e das Florestas: Rede Nacional de Áreas Protegidas (2020), https://sig.icnf.pt/portal/home/item.html?id=02b7a03f8fbd4dada77f5f3e5f91f186
8. PORDATA: Densidade Populacional por Município (2019), https://www.pordata.pt/Municipios/Densidade+populacional-452
9. Geofabrik: OpenStreetMap – Shapefiles of Portugal (2019), https://download.geofabrik.de/europe/portugal.html
10. Fick, S.E., Hijmans, R.J.: Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37(12), 4302–4315 (2017), https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086
11. QGIS Development Team: QGIS Geographic Information System. QGIS Association (2021), https://www.qgis.org
12. Zuur, A., Ieno, E., Walker, N., Saveliev, A., Smith, G.: Mixed Effects Models and Extensions in Ecology With R, vol. 1-574 (01 2009)
13. Permanasari, A.E., Hidayah, I., Bustoni, I.A.: Sarima (seasonal arima) implementation on time series to forecast the number of malaria incidence. In: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE). pp. 203–207 (2013)
14. Shumway, R., Stoffer, D.: Time Series and Its Applications (01 2011)
15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020), https://www.R-project.org/
16. Zeileis, A., Kleiber, C., Jackman, S.: Regression models for count data in R. Journal of Statistical Software 27(8) (2008), http://www.jstatsoft.org/v27/i08/
17. Shumway, R., Stoffer, D.: Time Series Analysis and Its Applications With R Examples, vol. 9 (01 2011)
18. Hyndman, R.J., Khandakar, Y.: Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 26(3), 1–22 (2008), https://www.jstatsoft.org/article/view/v027i03