MASTER OF INFORMATION AND DATA SCIENCE

W210 SYNTHETIC CAPSTONE

PREDICTING BAT CORONAVIRUS POSITIVITY

Ryan Kim

Isaac Law

Timothy Quek

Peter Wang

Fall 2020

INTRODUCTION

About 20% of the approximately 5,000 species of mammals are bats (order Chiroptera [meaning “hand wing”]). Among mammals, bats are unique in their ability to fly, and are found on all continents except Antarctica. Bats serve an important role in insect control and pollination of food crops, and their excrement (“guano”) is used as fertilizer and in the manufacture of soaps, gasohol and antibiotics. Unfortunately, bats are also increasingly being recognized as reservoir hosts for viruses which can cross species barriers to cause infections in humans and other animals. [1]

The ability of bats to fly allows for efficient virus transmission and bats have been suspected in a number of major emerging infectious disease outbreaks. These include outbreaks of the SARS (Severe Acute Respiratory Syndrome) virus, MERS (Middle East Respiratory Syndrome) virus, Ebola virus, Nipah virus and others. Lately, there has been significant interest in the scientific literature about bat associated viruses, which culminated in an online browsable database of bat associated viruses (DBatVir). [2]

The recent global COVID-19 pandemic has acutely raised the importance of studying bat related coronaviruses. The current hypothesis is that the SARS-CoV-2 virus was derived from a bat coronavirus, perhaps through an intermediate host. [3] Therefore, elucidating the predisposing factors that make it more likely for a bat to carry coronavirus would be important to help us understand how the current COVID-19

pandemic started and spread, and find ways to reduce the risk of such pandemics in the future.

The use of data science and machine learning in the study of emerging infectious diseases is not new. For example, Rulli *et al.* used land cover change data in conjunction with Ebola virus disease outbreak records to demonstrate elegantly that in Ebola virus outbreaks, the index cases in humans occurred mostly in hotspots of forest fragmentation [4]. More specifically related to bats, Plowright *et al.* used published records of Nipah virus surveillance globally, and applied a trait based machine learning approach to a subset of bat species in order to identify species that had a high likelihood to carry Nipah virus for targeted surveillance [5]. On a broader level, Allen *et al.* used a global database of emerging infectious diseases (EID) and merged it with other datasets to produce a global hotspot map of spatial variation in zoonotic EID risk [3].

In our study, we collected published information on coronavirus positivity from various bat species and merged this information with publicly available datasets on bat characteristics, geographical, and environmental features. Models were then fit with the final aim of predicting the prevalence of coronavirus positivity of a particular bat species. Factors associated with a higher coronavirus positivity among bats were identified and plausible mechanisms discussed.

METHODS

Database and Features

Our data collection is assembled from five sources: 1) PanTheria; [6] 2) Eltontraits; [7] 3) Prevalence data assembled from various academic research papers (see Appendix 1); 4) Bat ecology / viral diversity database from Canadian study; [8] 5) Zoonotic infectious diseases database. [9] Each of these sources provided information on features included in our merged dataset.

PanTheria [6] is a global-level species dataset compiled from various literature sources, covering ecological and geographical characteristics of all known living mammals, as well as those recently extinct. In addition, the PanTheria database contains databases of geographic distribution and global climatic and anthropogenic variables. The following ecological and life history variables are covered by the PanTheria database:

1. Activity Cycle; 2. Age at Eye Opening; 3. Age at First Birth; 4. Average Lifespan; 5. Body Mass; 6. Diet; 7. Dispersal Age; 8. Adult Limb Length; 9. Gestation Length; 10. Group Composition & Size; 11. Growth Data; 12. Habitat Layer; 13. Head-Body Length; 14. Interbirth Interval; 15. Litter size; 16. Litters Per Year; 17. Maximum Longevity; 18. Metabolic Rate; 19. Migratory Behaviour; 20. Mortality Data; 21. Population Density; 22. Ranging Behaviour; 23. Sexual Maturity Age; 24. Teat Number; and 25. Weaning Age.[4]

Next, the Eltontraits database [7] contains information on species according to their physiological, behavioral, and ecological attributes. The database is based on global species-level compilation of key attributes of 9,993 and 5,400 extant bird and mammal species. For the diet and foraging stratum, the database translates verbal descriptions into relative importance of different categories. Coupled with body size (a continuous variable) and activity time (a categorical variable), this creates a more precise description of the given species.

Thirdly, the bat prevalence data (Appendix 1) is collected from various academic literature from which we searched through various academic journals and literature. The following ecological and life features are included: 1. Bat species; 2. The sample number; 3. The total number positive; 4. Percentage positive; 5. Overall positive / negative; 6. Year of publication; 7. Continent; 8. Location; and 9. Reference.

Fourthly, we included information from a viral diversity and reservoir status study from the University of Toronto which contained further information on bat traits as well as phylogenetic information. [8]

Lastly, we expanded our database to include data fields from the Zoonotic Infectious Diseases database which includes geographical and ecological features, such as information on land use, poultry and livestock, and mammalian species diversity. [9]

We merged the features from the data sources above and assembled a dataset to facilitate further analysis. From this merged database, we conducted further exploratory data analysis using Python and R.

Exploratory Data Analysis

Regarding missing values, we implemented four criteria to process missing data:

1. Take away the features with more than 40% of the data missing (based on the calculation that the mean proportion of missing data was 40% across all features)

2. Take away features where more than 50% of the species have missing data

3. Take away those species with more than 50% missing data and where less than 10 species have available data; and 4. Impute the median for the genus for a particular feature and use the median value to replace the missing values.

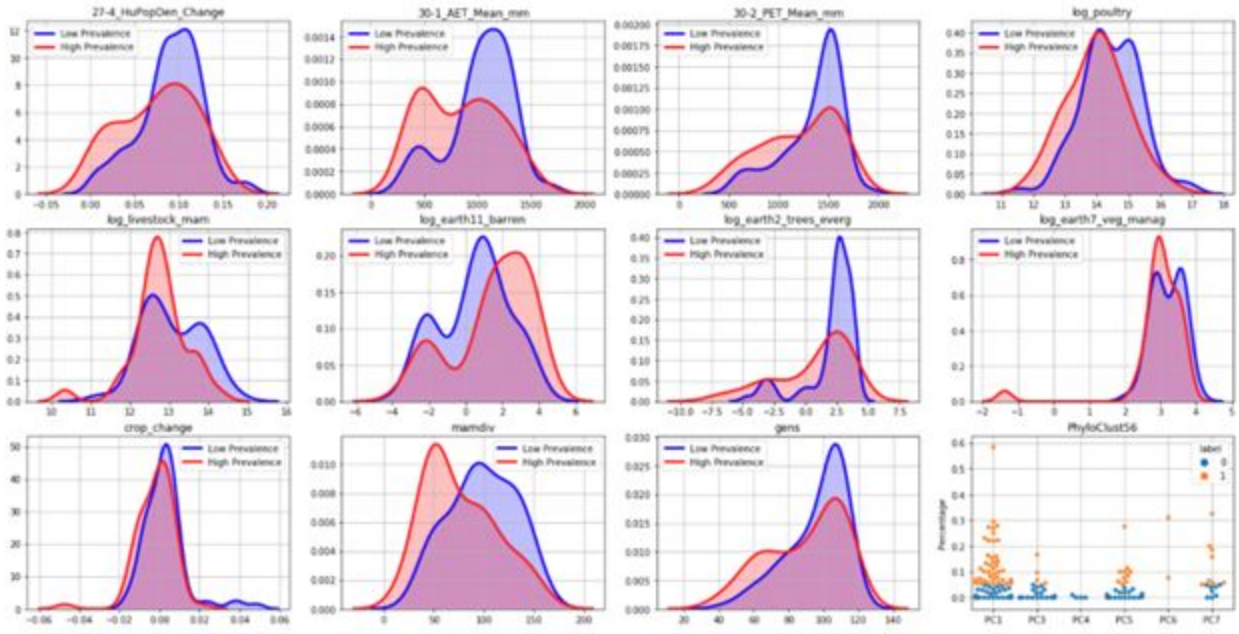
After resolving issues around missing values, we proceeded to plot histograms and scatterplots for various features in relation to positive / negative coronavirus percentages of various species. In addition, we performed feature transformation for those features that displayed skewness by applying log transformations. The histograms and scatter plots provide initial understanding of the relationship among various ecological and behavioral features and positive / negative coronavirus percentages. Subsequently, we performed dimensionality reduction for selective features, such as the AET (actual evapotranspiration) over PET (potential evapotranspiration) feature.

Next, we proceeded to perform binary classification and prevalence rate analysis on the merged dataset. The dataset is first split into training and testing datasets by splitting 80% to training and 20% to testing categories. Statistical classification is a form of machine learning and a type of supervised learning where categories are predefined

and used to categorize probabilistic observations into predetermined categories. In the case of binary statistical classification, there are two scenarios involved; namely, coronavirus positive (defined as “1”) or coronavirus negative (defined as “0”). We plotted this binary “Covid status” against different features.However, we found a strong association between binary classification and number of bat samples in each species, that is, the binary classification outcome depended heavily on the number of bats in each species sample set. Given this confounding variable in the relationship, we decided to forgo the binary classification method and enhance our analysis by focusing on the classification of high vs. low coronavirus prevalence rates.

Instead of focusing on binary classification methods, we turned our attention to studying the relationship between classifiers of high vs. low coronavirus prevalence rates. We divided the prevalence rates into low prevalence and high prevalence groups based on *5%* coronavirus positive rate threshold. By plotting the clustering of low and high prevalence groups, we were able to identify significant features that showed significant differences between low and high prevalence groups. Through feature selection, we filtered the significant features according to the following: 1. Human population density change; 2. Weather (mean precipitation, mean weather, actual / potential evapotranspiration rate); 3. Phylogenetic features (56 million years ago factorized cluster); 4. Land use and environment (land-barren (barren land) , evergreen broadleaf, managed vegetation, crop change); 5. Ecology of mammals (mammalian diversity, number of poultry, number of mammals livestock); 6. Location cluster (geographical cluster). In the following results section, we will delve into the specific features and their respective results.

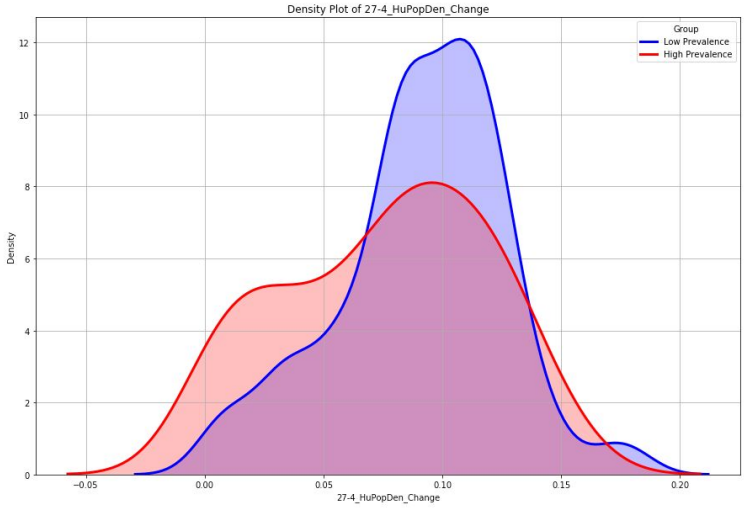
Figure 1: Significant features post feature selection process



RESULTS

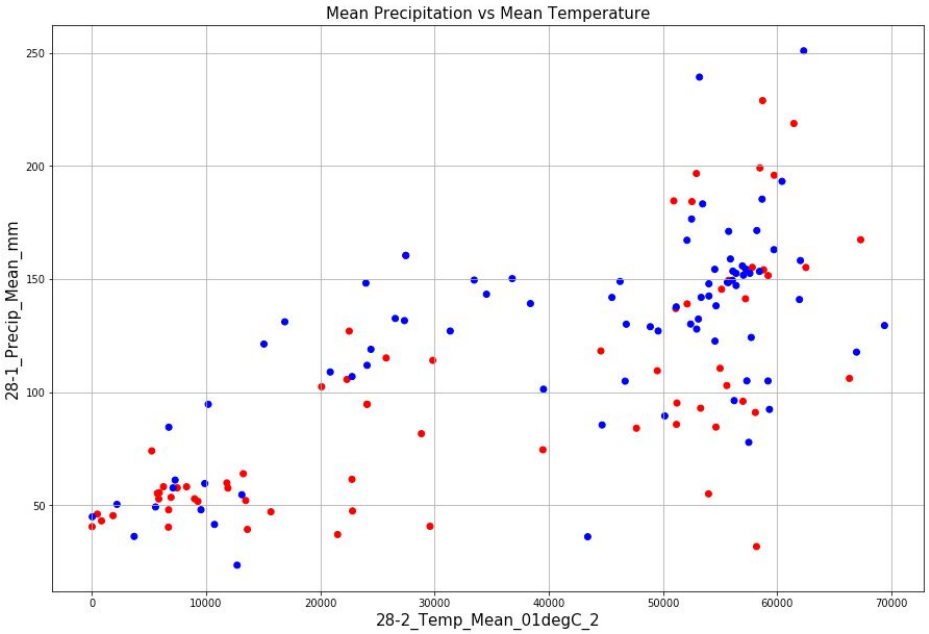
Human population density change refers to the rate of change of human population density between 1990 and 1995. Interestingly, a lower change in human density feature tends to be associated with higher bat coronavirus prevalence rate.

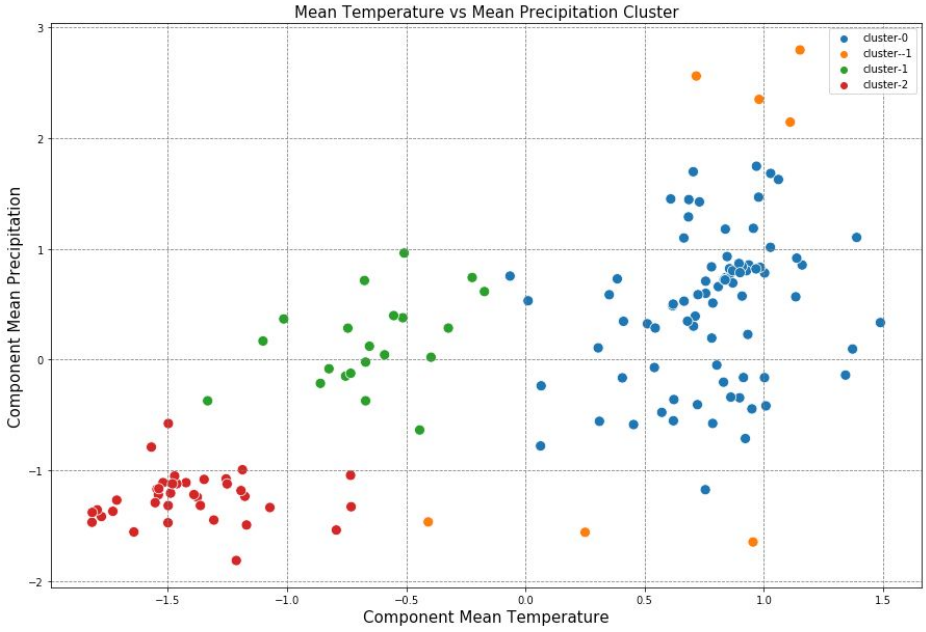
Figure 2: Human population density feature for high and low coronavirus prevalence rates



For the three weather variables: mean precipitation, mean temperature (squared) and ratio of actual to potential evapotranspiration from ground from 1920 to 1980, we observe that lower temperature and precipitation are associated with higher bat coronavirus prevalence rate.

Figure 3: Weather features for high and low coronavirus prevalence rates

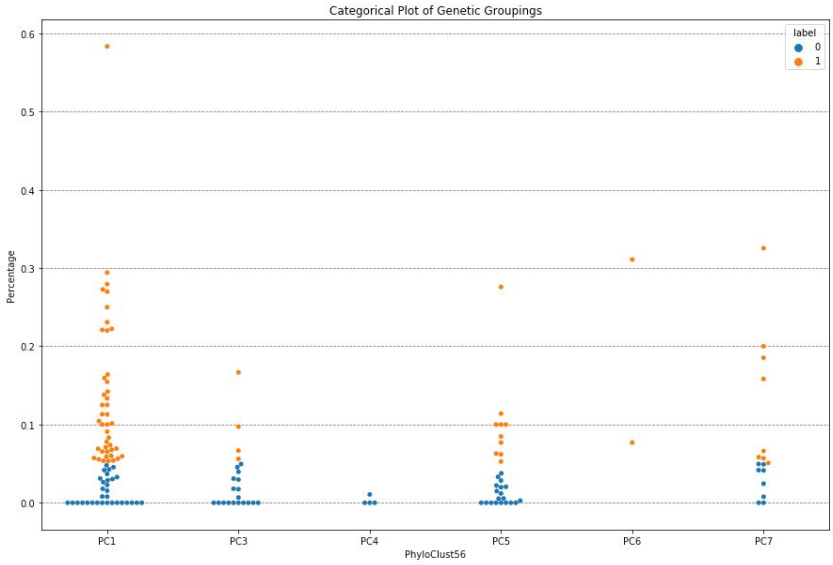




Phylogenetic cluster features refers to evolutionary relationships among bats from 56 million years ago. This categorical feature shows a meaningful distribution and statistical significance, especially in the first “PC1” group which denotes high prevalence

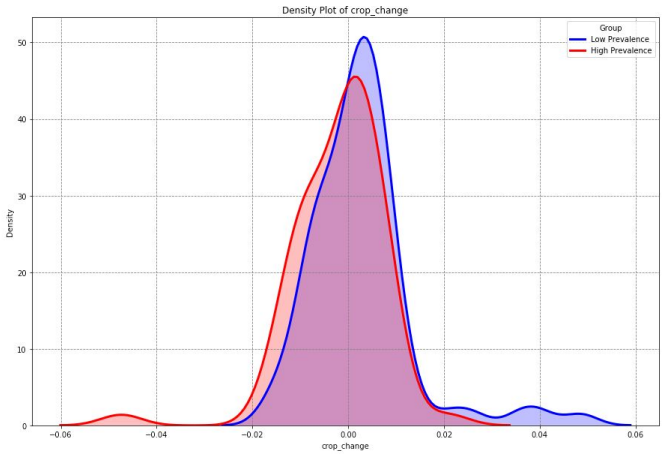
species. In contrast, the phylogenetic cluster from 41 million years ago did not show a significant relationship.

Figure 4: Phylogenetic features for high and low coronavirus prevalence rates



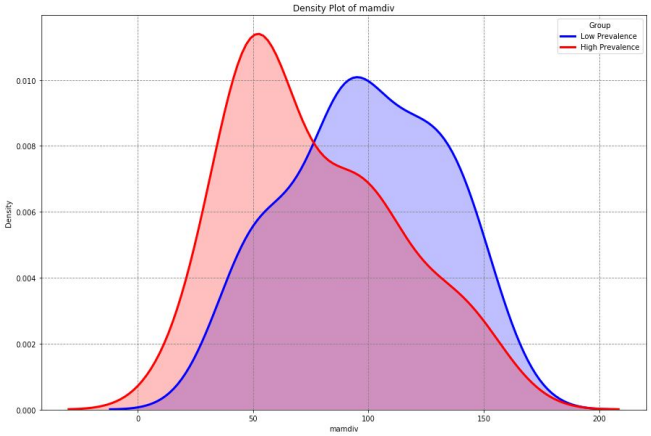
Land use and environmental factors such as land-barren, evergreen broadleaf, managed vegetation and crop change show significance in relationship.

Figure 5: Land and environmental features for high and low coronavirus prevalence rates



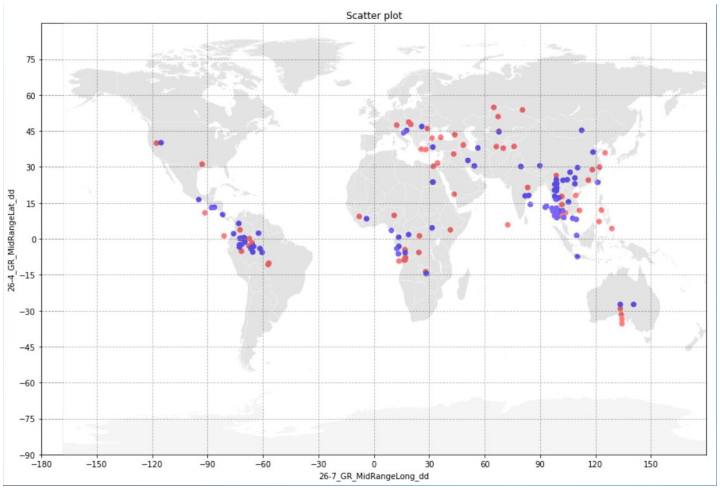
Mammalian diversity, which refers to the average number of mammal species per grid cell and poultry show statistically significant relationships with bat coronavirus prevalence after logarithmic transformation.

Figure 6: Mammalian diversity feature for high and low coronavirus prevalence rates



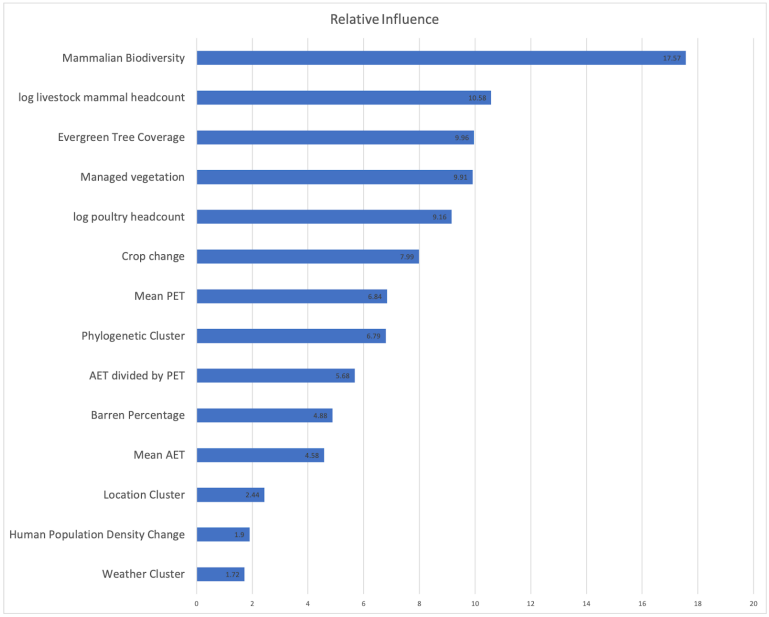
Lastly, location cluster features, specifically factorized cluster 3 is statistically significant and corresponds to the Western African region.

Figure 7: Location feature for high and low coronavirus prevalence rates



Generalized boosted model was adopted to analyze the features above. A generalized boosting classifier model was used with Bernoulli loss function and cross validation. A generalized boosting model is a form of gradient boosting, which is a machine learning technique for regression and classification problems which produce prediction models in stepwise fashion. Our generalized boosted model based on binary classification yielded a model accuracy of approximately 74%. While the generalized boosted model does not provide p-values or coefficients, it ranks variables by relative influence and ranks mammal and poultry ecological variables such as mammal biodiversity and livestock mammal headcount as the highest influencing variables. That is, higher mammal biodiversity corresponds to lower bat coronavirus prevalence rate. Other relative influence factors are displayed below.

Figure 8: Relative influence of selected features



We proceeded to assess the performance improvement of our feature-predicting model through the following methods as shown below: linear regression, logistic regression, random forest, generalized boosting and Poisson regression. In order to cross-compare the different regression methods, we calculated the root mean square error (RMSE). As shown below in the performance improvement chart, Poisson regression yielded the lowest RMSE among all the regression methods.

Next, we proceeded to analyze the relationships between features and prevalence rate using Poisson-type regressions. Poisson regressions were chosen because the Poisson regression yielded the lowest RMSE out of all the regression methods. Subsequently, Three kinds of regressions were adopted: count response Poisson regression, negative binomial Poisson regression and zero-inflated Poisson regression. We based our poisson regression model on count response modeling which models the number of positive bats out of 100 bats. The model is a stepwise forward inclusion based on AIC with RMSE of approximately 5.5. This count response model

poisson regression shows reasonable fit except a degree of under fitting on zero counts and high extreme counts. Similar to our binary classification study, we find mammal biodiversity as a statistically significant and important feature from the poisson regression model.

Figure 9: Regression outputs

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.880e+00  1.385e+00   2.803 0.005070 **
PhyloClust56PC3            2.176e-01  3.600e-01   0.604 0.545567
PhyloClust56PC4           -3.355e+00  1.012e+00  -3.316 0.000914 ***
PhyloClust56PC5           -5.079e-01  1.479e-01  -3.435 0.000593 ***
PhyloClust56PC6            2.729e+00  4.059e-01   6.723 1.78e-11 ***
PhyloClust56PC7           -2.911e-01  1.570e-01  -1.854 0.063733 .
crop_change               -3.543e+01  1.341e+01  -2.643 0.008227 **
X30.2_PET_Mean_mm          3.278e-03  4.008e-04   8.178 2.89e-16 ***
cluster_weather_DBSCAN0    6.128e-01  3.092e-01   1.982 0.047484 *
cluster_weather_DBSCAN1    1.296e+00  3.280e-01   3.953 7.71e-05 ***
cluster_weather_DBSCAN2    1.500e+00  3.607e-01   4.160 3.19e-05 ***
log_livestock_mam         -2.955e-01  9.927e-02  -2.976 0.002917 **
X27.4_HuPopDen_Change     -6.106e+00  2.765e+00  -2.208 0.027236 *
cluster_locationAmerica   -9.248e-01  3.301e-01  -2.801 0.005087 **
cluster_locationAsia      -1.132e+00  2.203e-01  -5.139 2.76e-07 ***
cluster_locationAustralia -1.866e+00  4.084e-01  -4.568 4.92e-06 ***
cluster_locationEurope    -5.355e-01  3.503e-01  -1.529 0.126278
mamdiv                    -2.031e-02  4.013e-03  -5.060 4.19e-07 ***
earth11_barren            -1.727e-02  5.322e-03  -3.245 0.001176 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Finally, based on regression and classification steps from above, we constructed a 95% confidence interval with Poisson regression results and cross checked with a generalized boosted model. The resulting model flags a species to be high coronavirus risk when both of these models converge and yield common results. Leveraging this model prediction, we tried to deploy our model on unstudied Rhinolophus bats which are commonly identified as major reservoir bats hosts for SARS related coronavirus. In the table below, from a list of Rhinolophus bats, we are able to predict the potential coronavirus reservoir hosts based on Poisson regression and generalized boosted model predictions.

Figure 10: Table of predicted Rhinolophus bats with high coronavirus risk

Species	Estimate	ci.lower	ci.upper	GBM_Prediction
Rhinolophus acuminatus	7.65	5.07	11.42	0.37
Rhinolophus alcyone	5.16	3.63	7.19	0.37
Rhinolophus arcuatus	5.42	2.76	9.88	0.50
Rhinolophus beddomei	10.75	7.26	15.95	0.48
Rhinolophus blasii	12.15	6.58	22.03	0.66
Rhinolophus borneensis	2.01	0.95	3.92	0.36
Rhinolophus capensis	8.02	4.60	13.87	0.57
Rhinolophus celebensis	13.88	8.83	21.57	0.48
Rhinolophus clivosus	6.81	4.26	10.74	0.63
Rhinolophus coelophyllus	3.74	2.17	6.34	0.39
Rhinolophus cornutus	9.35	5.36	16.33	0.58
Rhinolophus creaghi	1.90	0.76	4.31	0.36
Rhinolophus darlingi	4.32	2.68	6.82	0.49
Rhinolophus denti	2.64	1.35	4.86	0.39
Rhinolophus eloquens	4.24	2.57	6.86	0.58
Rhinolophus euryotis	18.37	8.72	36.85	0.62
Rhinolophus formosae	1.58	0.85	2.89	0.48
Rhinolophus fumigatus	6.82	4.77	9.58	0.42
Rhinolophus inops	27.56	17.73	42.47	0.57
Rhinolophus maclaudi	2.38	1.15	4.81	0.39
Rhinolophus marshalli	1.75	1.10	2.76	0.31
Rhinolophus megaphyllus	3.85	1.86	7.90	0.56
Rhinolophus mehelyi	9.12	6.59	12.46	0.51
Rhinolophus paradoxolophus	0.66	0.35	1.23	0.33
Rhinolophus philippinensis	23.00	13.39	39.39	0.61
Rhinolophus robinsoni	2.88	1.43	5.68	0.34
Rhinolophus rufus	27.86	17.98	42.80	0.57
Rhinolophus ruwenzorii	0.35	0.14	0.85	0.28
Rhinolophus sedulus	1.97	0.85	4.13	0.36
Rhinolophus siamensis	4.62	3.33	6.33	0.52
Rhinolophus simulator	5.29	3.57	7.68	0.43
Rhinolophus stheno	3.19	2.14	4.70	0.30
Rhinolophus subrufus	27.99	18.06	43.00	0.57
Rhinolophus swinnyi	4.57	2.80	7.31	0.59
Rhinolophus trifoliatus	1.93	1.00	3.46	0.25
Rhinolophus virgo	26.00	16.62	40.32	0.61

DISCUSSION AND CONCLUSIONS

Our study pooled data from published studies looking at the prevalence of coronavirus in samples obtained from various bat species. Subsequently, we merged this data with datasets containing information on biological traits, species distribution, foraging information, as well as geographical, ecological, human population and land use attributes. Using this merged dataset, we used factors found to be associated with coronavirus prevalence on univariate analysis to fit a Poisson regression model and a generalized boosting classifier model. These two models were used to predict the prevalence of coronavirus positivity among bats with unknown coronavirus prevalence

rates. Finally, applying our model to Rhinolophus bat, we found five Rhinolophus bat species were predicted to have a high prevalence of coronavirus positivity.

We have found a number of interesting associations.

1. Areas in the globe with lower temperature and precipitation / evapotranspiration rates were associated with a higher coronavirus prevalence amongst bats. This is consistent with the findings of previous research that shows the coronavirus infectivity is higher at lower temperatures and humidity [10].
2. Areas with high mammalian biodiversity appear to have a lower prevalence of bat coronavirus positivity. This finding is consistent with the findings of previous research, which show that biodiversity loss increases disease transmission. The mechanism of this is unclear, as it would be intuitive that greater species biodiversity would increase the risk of interspecies transmission of diseases. However, one speculation is that some species that are better at buffering disease transmission tend to be affected most with a reduction in biodiversity. Conversely, other species, who may have higher rates of reproduction (and spend less resources on host immunity), survive longer during such reductions in biodiversity.[11]
3. Africa appears to have a higher prevalence of bat coronavirus positivity compared to the other continents, independent of the effect on weather.
4. Certain phylogenetic clusters are associated with lower bat coronavirus positivity, and it is tempting to speculate if these could be related to inherited genetic resistance to viral infection.

5. Change in human population density, change in cropland, number of poultry / mammalian livestock, and proportion of barren land were associated with a lower bat coronavirus positivity. One may suspect that with greater human presence and development in a particular area, one would have a lower bat population density, making it less likely that bats would be able to spread coronavirus to one another.

Our study took pioneering steps to look at predictive factors for bat coronavirus positivity. Given that the current COVID-19 pandemic is postulated to be related to zoonotic transmission of a bat coronavirus, it is important to understand the factors that make it more likely for bat coronaviruses to spread. This may allow us to predict where the next outbreak may strike.

At the same time, our work has a number of limitations. For one, there are significant methodological inconsistencies when comparing the various studies that were used in the dataset for bat coronavirus prevalence. Some studies used bat stool samples, while others used respiratory samples. In addition, the studies were carried out over a long period (published between 2005 and 2020), and factors which affected bat coronavirus positivity at one time point may be different from those at another time point. There was also a significant amount of missing data in the other datasets which could have introduced bias into our conclusions. For example, bats with a significant amount of missing data may have been different from bats without missing data in important ways that could be related to coronavirus positivity. Despite all the limitations, we note that our findings were quite consistent with those found in other studies.

Looking to the future, we hope that we can combine further refinements to our methodology with more information on emerging infectious diseases in humans. In doing so, we may not only predict which bat species are more likely to carry coronavirus, but also where and when the transmission between bats and humans may be most likely to occur in the future. This would enable us to prevent the next coronavirus pandemic.

References

[1] Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V., and Schountz, T. (2006). Bats: important reservoir hosts of emerging viruses. Clin. Microbiol. Rev. 19, 531–545. doi: 10.1128/CMR.00017-06

[2] Chen, L., Liu, B., Yang, J., and Jin, Q. (2014). DBatVir: the database of bat-associated viruses. Database 2014:bau021. doi: 10.1093/database/bau021.

[3] Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, et al. Global hotspots and correlates of emerging zoonotic diseases. Nat Commun. 2017 Oct 24;8(1):1124. doi: 10.1038/s41467-017-00923-8.

[4] Maria Cristina Rulli, Monia Santini, David T S Hayman, Paolo D'Odorico. The nexus between forest fragmentation in Africa and Ebola virus disease outbreaks. Sci Rep. 2017 Feb 14;7:41613. doi: 10.1038/srep41613.

[5] Plowright RK, Becker DJ, Crowley DE, Washburne AD, Huang T, Nameer PO, et al. Prioritizing surveillance of Nipah virus in India. PLoS Negl Trop Dis. 2019 Jun 27;13(6):e0007393. doi: 10.1371/journal.pntd.0007393. eCollection 2019 Jun.

[6] Pantheria (2010) https://ecologicaldata.org/wiki/pantheria

[7] Wilman, H. et al. (2014), EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals, Ecological Archives E095-178. https://doi.org/10.1890/13-1917.1

[8] Guy C, Ratcliffe JM, Mideo N. The influence of bat ecology on viral diversity and reservoir status. Ecol Evol. 2020 May 8;10(12):5748-5758. doi: 10.1002/ece3.6315. eCollection 2020 Jun.

[9] Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, et al. Global hotspots and correlates of emerging zoonotic diseases. Nat Commun. 2017 Oct 24;8(1):1124. doi: 10.1038/s41467-017-00923-8.

[10] A Chan KH, Malik Peiris JS, Lam SY, Poon LLM, Yuen KY, Seto WH. The Effects of Temperature and Relative Humidity on the Viability of the SARS Coronavirus. Adv Virol. 2011;2011:734690. doi: 10.1155/2011/734690. Epub 2011 Oct 1.

[11] Keesing F, Belden LK, Daszak P, Dobson A, Harvell CD, Holt RD, et al. Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature. 2010 Dec 2;468(7324):647-52. doi: 10.1038/nature09575.