Building a one-vs-all classifier for spatial prediction of detected pathogens

Peter Maskell [a], **Matt Ryan** [a], **Anjana Karawita** [a,b], **R.I. Hickson** [a,b]
and Maryam Golchin [a,b]

[a] Commonwealth Scientific and Industrial Research Organisation, Australia
[b] College of Public Health Medical and Veterinary Sciences, and Australian Institute of Tropical Health and Medicine, James Cook University, Townsville,Australia
Email: Maryam.Golchin@csiro.au

Abstract: More than 75% of human infectious diseases are caused by the transmission of pathogens from animals to humans (that is, zoonotic spillover). This demonstrates the importance of understanding the relative risks of each pathogen in each spatial region. In this study, we build one-vs-all classifiers to distinguish *Mycobacterium* and *Listeria* amongst all other recorded bacteria. We selected these two bacteria as they cause morbidity and fatality among humans and animals. We overcome the impact of class imbalance caused by spatial and taxonomical biases in detected pathogen occurrence data by under-sampling the majority negative samples and keeping all the minority positive samples. We further improved the prediction results by including animal richness data (number of genera present). Our findings highlight that there is a weak relationship between the predictive features and the relative occurrence of the target pathogen. We also identified that the inclusion of spatial-temporal information in the prediction process could increase generalisability. The biological study of the detected features suggests that more targeted infectious diseases surveillance data is required to validate the predicted results.

Keywords: *One Health, zoonotic spillover, one-vs-all classifier, Mycobacterium, Listeria*

1. INTRODUCTION

In a changing world one key challenge identified by the World Health Organization (Ghebreyesus 2020) has been the emergence of infectious diseases throughout the human population. Zoonotic spillover, in which pathogens successfully transmit from animals and replicate in the human population, is known to account for approximately 75% of these diseases. Moreover, spillover is thought to be increasing in prevalence, due to factors such as climate change and urbanisation (Daszak et al. 2020; Ellwanger and Chies 2021). Hence, there is a growing effort to understand the key factors causing spillover so that the risks of outbreaks can be mitigated.

Many general factors have already been identified and can be grouped into anthropogenic, ecological, and agricultural categories with machine learning (ML) emerging as a tool to test theories and possibly identify other factors. Previous works have generally focused on either the risk of spillover events occurring in general on a global scale (Allen et al. 2017) or the risk of spillover for specific pathogens locally, where data is available (Mayfield et al. 2020; Kaul et al. 2018).

We investigate a modelling approach that synthesises these two approaches, utilising global datasets to estimate the risk of zoonotic spillover across the world for all pathogens. There is the potential for a comprehensive one-versus-all classifier that would predict the occurrence of pathogens across the globe and increase understanding of factors associated with any specific pathogen. We focus on building classifiers for *Mycobacterium* and *Listeria*, two genera of bacteria, assessing four modelling methods that attempt to capture key aspects of the pathogens while mitigating spatial bias and class imbalance of the training data.

2. DATA

To predict pathogen occurrence (i.e., labels), we grouped features under three main categories: climate, land use and animal distribution (i.e., features) (Ellwanger and Chies 2021; Allen et al. 2017; Kaul et al. 2018; García-Peña et al. 2021). We briefly described the source of these data and the pre-processing step in 2.1--2.2.

2.1. Source data

Pathogen occurrence data was sourced from the Global Biodiversity Information Facility (GBIF 2021). We selected three kingdoms (fungi, bacteria, and viruses) to represent pathogens with potential to cause disease for humans. However, due to GBIF's crowdsourced approach, the spatial distribution of pathogen recordings is highly biased. The GBIF dataset also contains a relatively small amount of data for viruses and so our selected data is dominated by the bacteria and fungi classes.

To represent climate, monthly temperature and rainfall records between 1970 and 2000 at a 1km^2 spatial resolution were sourced from WorldClim (Fick and Hijmans 2017). Land use features were sourced from the Land-Use Harmonization 2 (LUH2) model (Hurtt et al. 2019a; 2019b; 2020). This dataset details global land use managements, states and transitions at a 0.25×0.25degree (approximately 25×25km) spatial resolution between the years 850 and 2100.

Animal distribution data (IUCN) was composed from mammal and bird distribution maps of mammal and bird species contained in the IUCN Red List spatial dataset (IUCN 2018). This dataset contained recordings ranging from 2008-2021 at the time of download. Each species is associated with a set of polygons indicating areas where that species is extinct, possibly extinct, possibly extant, probably extant, and extant. The latter three extant categories were used to indicate species presence, while the extinct categories were treated as absence. The number of species present is referred to as "richness". Although this data reflects temporal variation, we treated this data as static distribution of mammals and birds as of 2021.

2.2. Data pre-processing

Labels and features were converted to rasters at the 100km^2 resolution using the WGS1984 coordinate reference system and aggregated and resampled as detailed in Table 1. Since we wished to predict the presence of pathogens, the Max resampling technique was used for pathogen occurrence and IUCN data. For climate and land use data we assigned the Bilinear and the Nearest Neighbour resampling techniques, respectively. We made this decision to preserve the continuous nature of climate data and discontinuous boundaries of categorical land use data.

We aggregated pathogen occurrence data at genus level as it is identified to be useful for evaluating zoonotic spillover risk (Olival et al. 2017). However, in order to reduce computation time and complexity of the ML models, we decided to aggregate IUCN data to the next higher taxa level of family. Note that both climate and IUCN data are static while pathogen and land use are dynamic at a yearly timescale.

Table 1. High-level details of source data and resampling technique that is used based on data type

Data class	Data sub-class	Sourced format	Aggregation	Resampling technique	Number of available features	Number of selected features	Possible values
Pathogen occurrence	Fungi Bacteria Virus	Points (.csv)	100km^2, genus taxa and year	Max	4195 3536 17	0 2 0	{0, 1}
Climate		Raster (.tiff)	100km^2	Bilinear	19	11	(-inf, inf)
Land use	Management State transitions	Raster (.tiff)	100km^2, land type, year	Nearest Neighbour	13 18 95	11 12 66	[0,1] [0,1] [0,1] or [0, inf)
IUCN	Birds Mammals	Polygon (.shp)	100km^2, family taxa	Max	338 154	208 114	[0,1]

Raster data was then combined into a single dataset indexed on location-year and, given the relative nature of the modelling process, all locations without a pathogen occurrence were dropped. Finally, a correlation analysis was conducted, dropping features with high spearman correlation (>0.9) over location-year. The final data contained 12,341 observations on 422 features.

3. METHOD

We focus on the creation of one-vs-all classifiers to distinguish *Mycobacterium* and *Listeria* amongst all other recorded bacteria. *Mycobacterium* is considered one of the major causes of death among humans, and also affects animal health (Tortoli 2014). *Listeria spp.* are a group of Gram-positive bacteria which has been isolated from variety of environments, including water, plants, and soil. One species in particular, *Listeria monocyotogenes,* is known to cause food-borne illnesses in humans and animals (Mead et al. 1999).

We used XGBoost (Chen and Guestrin 2016) to build a one-versus-all classifier to predict the most likely pathogen to occur in a given location-year. Labelled data was generated by assigning positive labels to samples of the target class and negative labels to any other class. We further investigated the impact of class imbalance, as this can lead to the model neglecting the minority class.

We optimised hyperparameters of our models using a 4-fold cross validation technique with stratified sampling and a two-step random grid search. First, the key parameters max_depth, colsample_bytree and n_estimators were optimised by setting the learning rate to 0.3. Second, for the optimum key parameters, the learning rate and regularisation parameters (reg_alpha, reg_lambda and gamma) were optimised. The F1 score was used for model comparison. Descriptions and values of all hyperparameters are found in Table 2. Due to the spatial aggregation of point data to a 100km^2 resolution raster and the spatial bias of the GBIF dataset, many location-years are associated with multiple pathogen genus occurrences. This can include both the target genus and other genera, resulting in one sample having two labels which would create more uncertainty in modelling prediction. To overcome this, we explored four methods to enable the model to find distinguishing features of the target genus from other pathogens. These methods capture both temporal and spatial information implicitly.

Method 1: we dropped duplicated samples and under-sampled the negative class by dropping samples with both negative and positive labels.

Method 2: we dropped duplicated samples and dropped samples with both positive and negative labels. This approach increased the class imbalance.

Method 3: in order to negate the class imbalance caused by Method 2, we randomly sampled 10% of the majority negative label class (under-sampling) while keeping all samples of the minority target genus class.

Method 4: Method 3 with the inclusion of richness data. In this method, we replaced animal presence data at family level with genus richness at the family level.

4. RESULTS AND DISCUSSION

In the prediction of both *Mycobacterium* and *Listeria*, Method 4 (Method 3 + richness) optimally predicted pathogen occurrence as determined by the F1 score (Table 3). Figure 1 demonstrates how this model can be used to predict pathogen occurrence in less populated areas as the lack of observation is noticeable in occurrence data in unpopulated areas (Figure 1-left). We further observed that occurrence of *Mycobacterium* is underpredicted throughout Asia whereas *Listeria* is underpredicted in Europe (Figure 1). This highlights the importance of inclusion of spatial-temporal information in the prediction process to increase generalisability of the method of *Mycobacterium* and *Listeria*.

Table 2. Hyperparameter for XGBoost. The “scale_pos_weight” parameter is fixed at the ratio of negative cases (n –k) to positive cases (k)

Hyperparameter	*Description*	*Grid values*
scale_pos_weight	weights minority class in underlying algorithm to handle label imbalance	(n – k)/k
n_estimators	number of gradient boosted trees	{50, 100, 500, 1000}
max_depth	maximum depth of tree for base learners	{3, 6, 9, 12}
colsample_bytree	percentage of columns to sub-sample when creating each tree. Helps to prevent overfitting	{0.5, 0.7, 1}
learning_rate	step size of each iteration in algorithms optimization process	{0.06, 0.15, 0.30, 0.50, 0.90}
reg_alpha	L1 regularization. Can improve speed for large number of features	{0.00, 0.10, 0.20, 0.50, 0.60}
reg_lambda	L2 regularization. Can help prevent overfitting	{1, 2, 4, 8, 12}
gamma	minimum loss deduction to make further partition of leaf within tree	{0, 1, 2}

Table 3. Best parameter configuration and mean results from 4-fold cross validation for the XGBoost model trained for each genus and overlap method. Values in parentheses indicate standard error.

	Method	*scale_pos_weight*	*n_estimators*	*max_depth*	*colsample_bytree*	*learning_rate*	*reg_alpha*	*reg_lambda*	*gamma*	*Accuracy*	*F1 score*	*Precision*	*Recall*
Mycobacterium	1	13.4	50	12	0.7	0.3	0.4	1	0	0.91 (0.00)	0.30 (0.01)	0.32 (0.01)	0.38 (0.01)
	2	201.5	50	6	0.7	0.3	0.1	1	0	0.99 (0.00)	0.32 (0.08)	0.36 (0.08)	0.30 (0.08)
	3	20.2	50	12	0.7	0.3	0.0	4	0	0.96 (0.00)	0.56 (0.03)	0.61 (0.06)	0.52 (0.03)
	4 (3 + richness)	20.2	50	12	0.7	0.3	0.0	4	0	0.96 (0.00)	0.59 (0.04)	0.55 (0.04)	0.65 (0.06)
Listeria	1	46.5	500	12	0.5	0.3	0.0	1	0	0.97 (0.00)	0.34 (0.02)	0.38 (0.03)	0.32 (0.02)
	2	105.1	1000	9	0.5	0.3	0.0	8	0	0.99 (0.00)	0.35 (0.04)	0.44 (0.03)	0.30 (0.04)
	3	10.6	1000	6	0.5	0.3	0.0	2	0	0.92 (0.00)	0.42 (0.04)	0.54 (0.01)	0.36 (0.05)
	4 (3 + richness)	10.6	1000	6	0.5	0.3	0.0	2	0	0.92 (0.00)	0.46 (0.03)	0.59 (0.02)	0.38 (0.04)

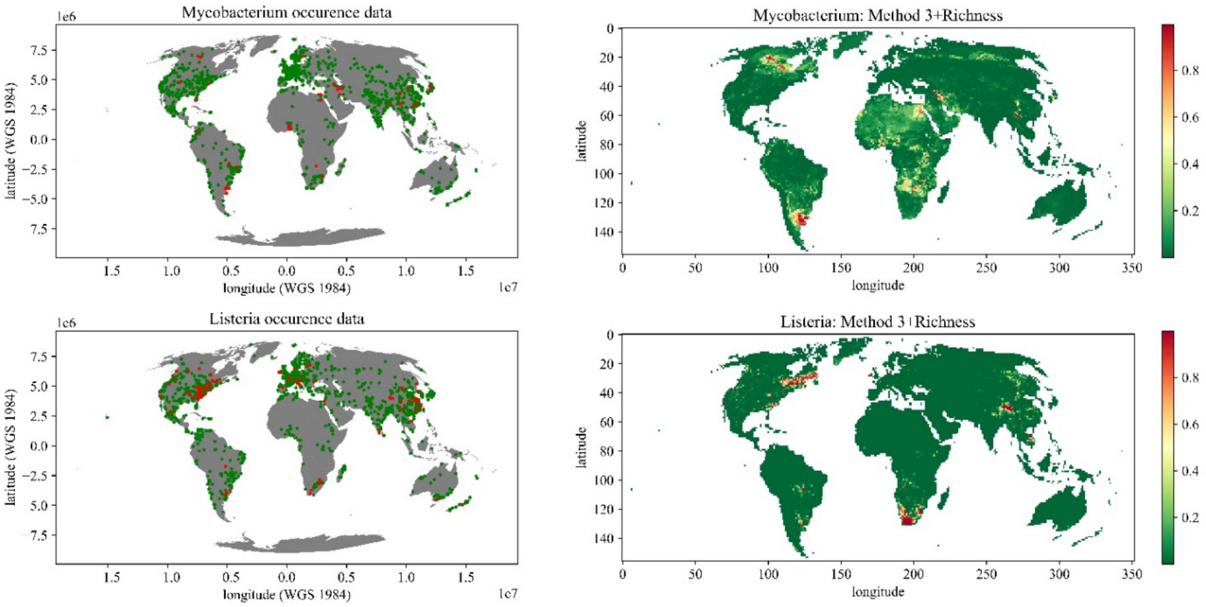

Figure 1. Training data (left) and predicted results (right) using Method 4 (Method 3 + richness) at the global scale for 2021. The top row shows the prediction for *Mycobacterium* and the bottom row shows the *Listeria* model. This illustrates the use of these models to predict bacteria occurrences in less populated areas.

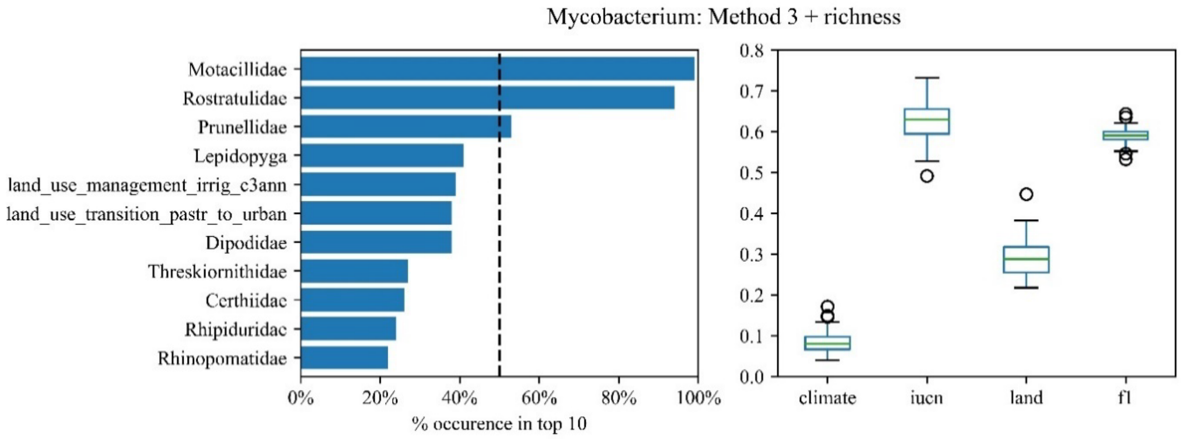

Figure 2. *Mycobacterium* Method 4 (Method 3 + richness): (Left) Percentage occurrence of features occurring in the top 10 features of the model for 100 different random seeds. The vertical line indicates occurrence in 50% of models trained. (Right) Box plot of total importance of different feature categories and F1 score.

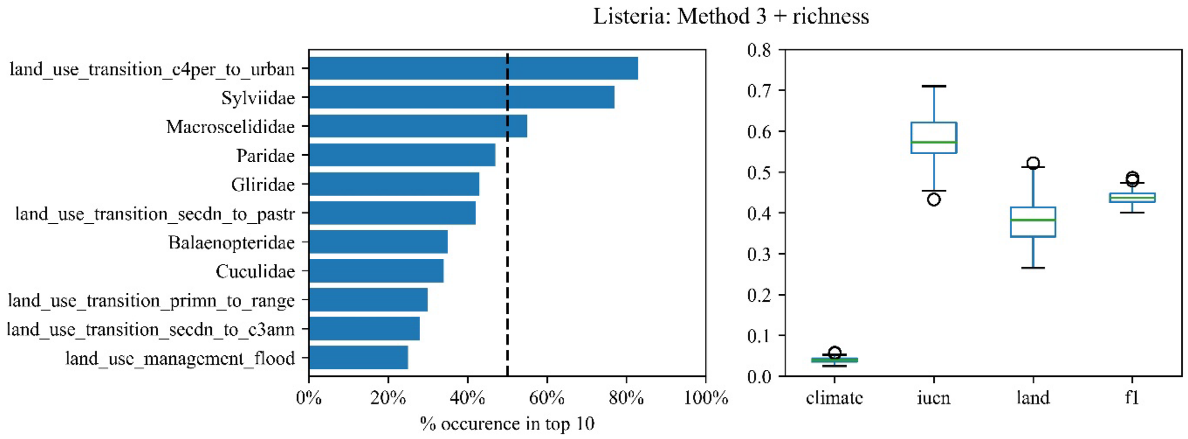

Figure 3. *Listeria* Method 4 (Method 3 + richness): (Left) Percentage occurrence of features occurring in the top 10 features of the model for 100 different random seeds. The vertical line indicates occurrence in 50% of models trained. (Right) Box plot of total importance of different feature categories and F1 score.

4.1. Sensitivity analysis

To assess the relationships between the importance of features, predicted results, and initial state of the system, we conducted a sensitivity analysis. To do this, we fitted 100 XGBoost models using Method 4 (Method 3 + richness) with a random integer between 1 and 10^6 as the initial seed. Across all 100 fits, the IUCN data was the most predictive of pathogen occurrence for both *Mycobacterium* and *Listeria* (Figure 2 and Figure 3, right). We also noticed a strong relationship between initial seed and variable importance in the final model. To quantify this relationship, we recorded how frequently each predictor was reported in the top 10 most important features for each model fit (Figure 2 and Figure 3, left). Observe that there are only three features for each pathogen that consistently appear as important features in more than 50% of the models fit. This suggests that the XGBoost decision tree modelling process is quite sensitive to its initial state (that is, the first weak learner the algorithm builds) and hence is more likely to detect spurious correlations in the data rather than true relationships. Future work will explore random forests as a remedy to address these spurious relationships.

4.2. Bias analysis

GBIF is formed through the synthesis of many reported biological surveys, thus it has been identified as carrying certain taxonomic (Troudet et al. 2017) or spatial biases (Yesson et al. 2007). To quantify the biases relating to our specific pathogen dataset, a correlation analysis was conducted comparing the spatial recording of pathogens to global population and various World Bank indicators. Not surprisingly, both Income Group and Urban population had the highest spearman correlation of 0.32 with the pathogen occurrence. When aggregating pathogen occurrences to the country level, correlation value increased with Urban Population and Population Total producing 0.59 and 0.5 spearman coefficients respectively. This means number of detected pathogens is biased towards higher populated areas.

4.3. Feature analysis

We assess the results of the Method 4 (Method 3 + richness) model for its validity in terms of current knowledge on infectious disease surveillance. Over 170 *Mycobacteria* species have been identified in the world with an extended host spectra from birds to mammals. Although certain passerines were found to be infected with *Mycobacterium* spp. (Borovská et al. 2011) there are no sufficient data to explicitly implicate *Sylvidae* as a major reservoir compared to other animal hosts. Similarly, the members of the family *Macroscelidae* were reportedly infected with *Mycobacteria* (Clancy et al. 2013), the evidence is not sufficient to conclude the validity of the results. On the other hand, *Listeria* spp. In particular, *Listeria monocytogenes* is common in wild birds (Brobey, Kucknoor, and Armacost 2017). Therefore, the prediction of the model is acceptable in terms of the class of animal involved, although inadequate data for specific families of birds makes it difficult to assess the level prediction.

5. CONCLUSION

In this study, we have discussed the concept of a comprehensive one-vs-all classification model that could provide insight into both the spatial risk and driving factors of a large number of pathogens. This concept was tested through the implementation of models trained to predict the relative risk of *Mycobacterium* and *Listeria* genera. We considered the spatial and temporal information of data in our modelling implicitly. Four different methods were assessed for dealing with the major problem faced in the spatial bias and class imbalance of the label dataset, with Method 4 (Method 3 + richness) providing the best generalisability to test data as measured by the F1 score. A sensitivity analysis was conducted to validate the suitability of features used and indicated a high variation in important features. This likely indicates a weak relationship between the predictive features and the relative occurrence of the target pathogens tested. Although the predictions of the model are theoretically possible in nature, future infectious disease surveillance targeted on these taxa will be required for informed validation of these predictions.

REFERENCES

Allen, Toph, Kris A. Murray, Carlos Zambrana-Torrelio, Stephen S. Morse, Carlo Rondinini, Moreno Di Marco, Nathan Breit, Kevin J. Olival, and Peter Daszak. 2017. ‘Global Hotspots and Correlates of Emerging Zoonotic Diseases’. Nature Communications 8 (1): 1124. https://doi.org/10.1038/s41467-017-00923-8.

Borovská, Petra, Peter Kabát, Martina Ficová, Alfréd Trnka, Darina Svetlíková, and Tatiana Betáková. 2011. ‘Prevalence of Avian Influenza Viruses, Mycobacterium Avium, and Mycobacterium Avium, Subsp. Paratuberculosis in Marsh-Dwelling Passerines in Slovakia, 2008’. Biologia 66 (2): 282–87. https://doi.org/10.2478/s11756-011-0016-3.

Brobey, Britni, Ashwini Kucknoor, and Jim Armacost. 2017. ‘Prevalence of Trichomonas, Salmonella, and Listeria in Wild Birds from Southeast Texas’. Avian Diseases 61 (3): 347–52. https://doi.org/10.1637/11607-020617-RegR.

Chen, Tianqi, and Carlos Guestrin. 2016. ‘XGBoost: A Scalable Tree Boosting System’. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–94. https://doi.org/10.1145/2939672.2939785.

Clancy, Meredith M., Margarita Woc-Colburn, Tabitha Viner, Carlos Sanchez, and Suzan Murray. 2013. ‘Retrospective Analysis of Mortalities in Elephant Shrews (Macroscelididae) and Tree Shrews (Tupaiidae) at the Smithsonian National Zoological Park, Usa’. Journal of Zoo and Wildlife Medicine 44 (2): 302–9.

Daszak, Peter, Carlos das Neves, John Amuasi, David Haymen, Thijs Kuiken, Benjamin Roche, Carlos Zambrana-Torrelio, et al. 2020. Workshop Report on Biodiversity and Pandemics of the Intergovernmental Platform on Biodiversity and Ecosystem Services. Bonn, Germany: Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). https://ipbes.net/pandemics.

Ellwanger, Joel Henrique, and José Artur Bogo Chies. 2021. ‘Zoonotic Spillover: Understanding Basic Aspects for Better Prevention’. Genetics and Molecular Biology 44 (1 Suppl 1): e20200355. https://doi.org/10.1590/1678-4685-GMB-2020-0355.

Fick, Stephen E., and Robert J. Hijmans. 2017. ‘WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas’. International Journal of Climatology 37 (12): 4302–15. https://doi.org/10.1002/joc.5086.

García-Peña, Gabriel E., André V. Rubio, Hugo Mendoza, Miguel Fernández, Matthew T. Milholland, A. Alonso Aguirre, Gerardo Suzán, and Carlos Zambrana-Torrelio. 2021. 'Land-Use Change and Rodent-Borne Diseases: Hazards on the Shared Socioeconomic Pathways'. Philosophical Transactions of the Royal Society B: Biological Sciences 376 (1837): 20200362. https://doi.org/10.1098/rstb.2020.0362.

GBIF. 2021. 16 September 2021. https://www.gbif.org/.

Ghebreyesus, Tedros. 2020. 'Urgent Health Challenges for the next Decade'. 13 January 2020. https://www.who.int/news-room/photo-story/photo-story-detail/urgent-health-challenges-for-the-next-decade.

Hurtt, George, Louise Chini, Ritvik Sahajpal, Steve Frolking, Benjamin L. Bodirsky, Katherine Calvin, Jonathan C. Doelman, et al. 2020. 'Harmonization of Global Land-Use Change and Management for the Period 850–2100 (LUH2) for CMIP6'. Preprint. Climate and Earth System Modeling. https://doi.org/10.5194/gmd-2019-360.

Hurtt, George, Louise Chini, Ritvik Sahajpal, Steve Frolking, Benjamin Leon Bodirsky, Kate Calvin, Jonathan Doelman, et al. 2019a. 'Harmonization of Global Land Use Change and Management for the Period 850-2015'. Earth System Grid Federation. https://doi.org/10.22033/ESGF/input4MIPs.10454.

———. 2019b. 'Harmonization of Global Land Use Change and Management for the Period 2015-2300'. Earth System Grid Federation. https://doi.org/10.22033/ESGF/input4MIPs.10468.

IUCN. 2018. https://www.iucnredlist.org.

Kaul, RajReni B., Michelle V. Evans, Courtney C. Murdock, and John M. Drake. 2018. 'Spatio-Temporal Spillover Risk of Yellow Fever in Brazil'. Parasites & Vectors 11 (1): 488. https://doi.org/10.1186/s13071-018-3063-6.

Mayfield, Helen J., Hugh Sturrock, Benjamin F. Arnold, Ricardo Andrade-Pacheco, Therese Kearns, Patricia Graves, Take Naseri, Robert Thomsen, Katherine Gass, and Colleen L. Lau. 2020. 'Supporting Elimination of Lymphatic Filariasis in Samoa by Predicting Locations of Residual Infection Using Machine Learning and Geostatistics'. Scientific Reports 10 (1): 20570. https://doi.org/10.1038/s41598-020-77519-8.

Mead, P. S., L. Slutsker, V. Dietz, L. F. McCaig, J. S. Bresee, C. Shapiro, P. M. Griffin, and R. V. Tauxe. 1999. 'Food-Related Illness and Death in the United States'. Emerging Infectious Diseases 5 (5): 607–25. https://doi.org/10.3201/eid0505.990502.

Olival, Kevin J., Parviez R. Hosseini, Carlos Zambrana-Torrelio, Noam Ross, Tiffany L. Bogich, and Peter Daszak. 2017. 'Host and Viral Traits Predict Zoonotic Spillover from Mammals'. Nature 546 (7660): 646–50. https://doi.org/10.1038/nature22975.

Tortoli, Enrico. 2014. 'Microbiological Features and Clinical Relevance of New Species of the Genus Mycobacterium'. Clinical Microbiology Reviews 27 (4): 727–52. https://doi.org/10.1128/CMR.00035-14.

Troudet, Julien, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, and Frédéric Legendre. 2017. 'Taxonomic Bias in Biodiversity Data and Societal Preferences'. Scientific Reports 7 (1): 9132. https://doi.org/10.1038/s41598-017-09084-6.

Yesson, Chris, Peter W. Brewer, Tim Sutton, Neil Caithness, Jaspreet S. Pahwa, Mikhaila Burgess, W. Alec Gray, et al. 2007. 'How Global Is the Global Biodiversity Information Facility?' PLOS ONE 2 (11): e1124. https://doi.org/10.1371/journal.pone.0001124.