Using Machine Learning Techniques on Satellite Data to Predict the Effect of Urbanization on Avian Biodiversity

by

Savannah Tynan

S.B., Computer Science and Engineering, Massachusetts Institute of Technology, 2022

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Department of Electrical Engineering and Computer Science
Sept 2, 2022

Certified by . . . .
John E. Fernandez
Professor of building technology in the Department of Architecture
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Using Machine Learning Techniques on Satellite Data to Predict the Effect of Urbanization on Avian Biodiversity

by

Savannah Tynan

Submitted to the Department of Electrical Engineering and Computer Science
on Sept 2, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Assessing the relationship between environmental and socio-economic conditions of an urban area and local urban biodiversity loss is integral to informing policy decisions, urban design, and community action plans.

Though some previous research explores the relationship between urban areas and biodiversity loss, the limited available studies are often specific to only one city or region. Those that do model this phenomena in multiple cities are limited to environmental variables, and rarely examine the socio-economic conditions of a city, such as average GDP or population density. To our knowledge, no studies analyze the complex underlying relationship between socio-economic as well as environmental conditions within urban areas and biodiversity, which is necessary for strategically protecting the most at risk regions.

This work aims to leverage satellite datasets to predict cities' risk exposure for bird biodiversity loss. This research aspires to develop an analytical approach that can be used for various types of biodiversity, though we restrict our initial analysis to birds, as they offer a broad range of data and can be used as an indicator for other species.[1] We aim for our approach to be applicable to all urban areas, so this research leverages a globally representative sample of cities with robust survey data. The over-arching goal of this project is to design a methodology to better advise resource allocation for the protection of global biodiversity.

We process 9 publicly available satellite datasets to create environmental and socio-economic features for every functional urban area (FUA), as classified by the OECD, totalling over 9,000 FUAs. We train and test 3 models: linear regression, random forest regression, and a hybrid supervised and unsupervised model. We analyze the predictive power of these approaches as well as relative importance assigned to each input feature. We find that all 3 of the approaches have the ability to accurately predict biodiversity loss and all of them find that the maximum land modification

[1] Sara Fraixedas et al., "A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions," *Ecological Indicators* 118 (2020): 106728.

value of each FUA is the most useful feature in determining biodiversity loss. Finally, we discuss the implications of these findings and our models ability to inform resource allocation.

Thesis Supervisor: John E. Fernandez
Title: Professor of building technology in the Department of Architecture

Acknowledgements

I am extremely grateful to everyone who made my time working with MIT's Environmental Solutions Initiative so rewarding. I would like to thank my advisor, John Fernandez, whose guidance made this project possible.

I would also especially like to thank Norhan Bayomi and Matias Williams for their constant support and extremely generous mentorship. Their help was absolutely indispensable. I acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this thesis.

Finally, I would like to thank my dad who encouraged and supported me to keep pursuing my goals.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and problem statement

Human manipulation of the global environment has caused extensive changes to the distribution of organisms on earth, namely an immense loss of biodiversity. Understanding and mapping biodiversity loss is a fundamental conservation priority as threats from the growing human population continue to surge, and stakeholders, such as international agreements to reduce biodiversity loss (e.g. the Aichi Biodiversity Targets, Convention on Biological Diversity 2010), call for a data driven basis to prioritize their response.[1][2]

More than 4.3 billion people live in urban areas, comprising more than 56% of the global population, whereas in 1900 only 10% of the population resided in cities.[3] Clearly, the human population is continuing its rapid trend of urbanization. The world is now hosting cities of unprecedented size.

As shown in Figure 1, a large proportion of this growth will occur in the global south, a term often used to refer to developing countries–not the southern region of the globe, potentially encroaching on many areas of rich biodiversity.[4] The rapidly

[1] Kevin J Gaston, “Global patterns in biodiversity,” *Nature* 405, no. 6783 (2000): 220–227.

[2] Biosafety Unit, *Aichi Biodiversity targets*, September 2020, https://www.cbd.int/sp/targets/.

[3] Hannah Ritchie and Max Roser, “Urbanization,” https://ourworldindata.org/urbanization, *Our World in Data*, 2018,

[4] Andrea Hollington et al., “Introduction: Concepts of the global south,” *Voices from around the World* 1 (2015).

changing urban environment is known to affect local biodiversity, yet comprehensive global studies of the relationship between environmental variables, socio-economic variables and avian biodiversity are lacking. In order to carry out a quantitative analysis capable of representing this complex relationship, it is necessary to leverage a multidisciplinary perspective. The combination of computer science and urban studies is crucial to leverage the novel environmental data available.

Biodiversity is declining globally, and the aforementioned rapid urbanization poses many risks to biodiversity, as well as many potential solutions.[5] The growth in population will undoubtedly lead to continued changes in land use, resource consumption, and economic circumstances.

In past research, many studies have lacked socio-economic variables such as the per capita GDP of cities and population density only examining environmental factors such as change in land use. Leveraging socio-economic as well as environmental data is key in understanding the complex underlying relationship of urban areas and biodiversity. Furthermore, we do not limit our analysis to politically defined city limits, as these boundaries are often political and do not represent the practical limits of a city. As a result, using the official city limits can distort measures of population as they do not consider individuals who may live right outside of the official boundaries, but commute to the city everyday. Rather, we examine the OECD's functional urban area (FUA), or the area including the urban core, as well as the commuting zone.[6]

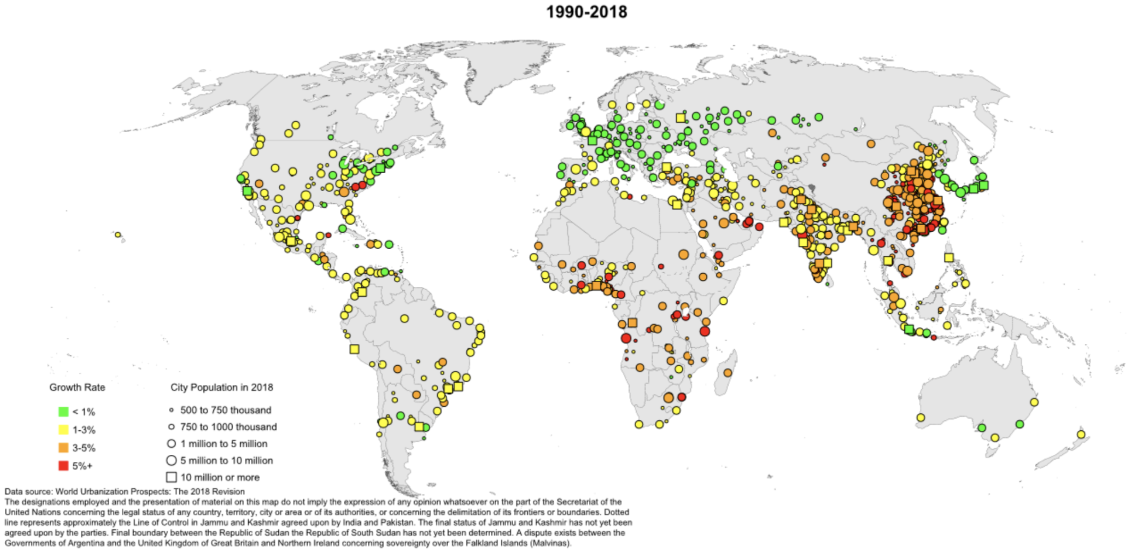
Assessing what factors are most important in determining biodiversity loss is integral to informing policy decisions, urban design, and community action plans. Global biodiversity targets set for 2020 by the Convention on Biological Diversity were not achieved, so it is essential that we are able to determine where to focus resources in order to minimize global biodiversity loss.[7]

[5] Anne Larigauderie and Harold A Mooney, "The Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services: moving a step closer to an IPCC-like mechanism for biodiversity," *Current opinion in environmental sustainability* 2, nos. 1-2 (2010): 9–14.

[6] Lewis Dijkstra, Hugo Poelman, and Paolo Veneri, "The EU-OECD definition of a functional urban area," 2019,

[7] Unit, *Aichi Biodiversity targets.*

Figure 1-1: Growth Rates of Cities



Data source: World Urbanization Prospects: The 2018 Revision
The designations employed and the presentation of material on this map do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted line represents approximately the Line of Control in Jammu and Kashmir agreed upon by India and Pakistan. The final status of Jammu and Kashmir has not yet been agreed upon by the parties. Final boundary between the Republic of Sudan the Republic of South Sudan has not yet been determined. A dispute exists between the Governments of Argentina and the United Kingdom of Great Britain and Northern Ireland concerning sovereignty over the Falkland Islands (Malvinas).

1.2 Research goal and methodology

This work seeks to leverage numerous satellite datasets to create environmental and socio-economic parameters of global cities in machine learning models to predict which cities are at highest risk for rapid avian biodiversity loss, and determine which parameters most influence local avian biodiversity. In addition to the main contribution, this research presents a novel, more comprehensive set of features for urban areas that can be used in future biodiversity research. We believe this work will help the environmental planning community gain a deeper understanding of the drivers of avian biodiversity loss that is not currently available in the literature. We restrict our initial analysis to birds, as avian research and available data is most plentiful, and bird biodiversity loss can be used as an indicator for other species.[8] However, we aim to create a feature set and analysis that can be generalized to other types of biodiversity, such as plant diversity.

In this work, we aim to train machine learning models that can generalize to all urban areas, though we initially focus on a subset of 54 cities which have robust avian

[8] Fraixedas et al., "A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions."

biodiversity survey data and have been used in previous research.[9] These cities are geographically diverse, representing 36 countries on 6 continents and 6 bio-geographic realms, including representation of the global south. See figure 4-4 for a visualization of the cities. We believe this selection of cities to be a representative set of the worlds cities, and as far as we are aware it is the largest global compilation of urban biodiversity data available.[10]

One of the major contributions of this work is the use of machine learning algorithms to understand the drivers of biodiversity loss and predict where future loss will occur. A novel combination of socio-economic and environmental variables derived from satellite datasets are used to train machine learning algorithms that have not, to our knowledge, been used to model these complex relationships. The over-arching goal of this project is to design a methodology to better advise resource allocation for the protection of global biodiversity.

1.3 Thesis outline

This research is divided into 6 chapters. In chapter 2 we discuss works related to urbanization and biodiversity, the use of satellite and spatial data to predict biodiversity loss, and finally we discuss the key works and concepts related to machine learning techniques used in this research. In chapter 3, we detail the modelling and analysis methods used to obtain our results.

In chapter 4 we describe the sources we used to collect necessary data as well as the methodology we used to process that data and create features.

Chapter 5 presents the results of our analysis. Chapter 6 discusses implications of the work presented and outlines areas for future research. See table 1.1 for a summary of chapter content and key contributions.

[9] Myla FJ Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers," *Proceedings of the royal society B: biological sciences* 281, no. 1780 (2014): 20133330.

[10] Aronson et al.

Chapter and section	**Summary**	**Contribution to the field**
3.2	This section describes how a random forest classifier is used to predict avian biodiversity loss over a global sample of cities	The use of ML for predicting biodiversity loss is limited
3.3	This section describes how a hybrid model, combining both supervised and unsupervised ML, is used to predict avian biodiversity loss over a global sample of cities	The use of ML for predicting biodiversity loss is limited
4	This chapter describes the feature creation and processing of 10 novel datasets	Novel socio-economic and environmental feature set for all OECD FUAs
5.1	A summary and analysis of our models ability to predict avian biodiversity loss	Novel results presenting the ability of ML models to use satellite data measuring both socio-economic and environmental variables to predict avian biodiversity loss across a global sample of urban areas
5.2	An analysis of feature importance in predicting urban avian biodiversity loss	A novel understanding of which environmental and urban conditions are most related to urban avian biodiversity loss

Table 1.1: Contributions by chapter and section

Chapter 2

Back ground and related works

2.1 Urbanization and biodiversity

According to Diaz et al. (2019), "The number of species currently threatened with extinction is unprecedented in human history: an estimated 1 million species of animals and plants."[1] Since the World's first national park, Yellowstone, was created in 1872, a conservation based approach to biodiversity has dominated.[2] Conservation, however, has not been shown to work in all countries or regions, and has even led to damaging social challenges.[3] In order to quell the immense human impact on our planet, we must make informed, strategic decisions to best promote a bio-diverse population.

More than half of humanity currently resides in cities or urban areas, which is a drastic increase from 1900 when city dwellers comprised only 10% of the population.[4] Overall, the world's urban population is growing by about 1% per year, with much of that growth concentrated in the global south.[5] Furthermore, it has been shown

[1] Sandra Diaz et al., "Pervasive human-driven decline of life on Earth points to the need for transformative change," *Science* 366, no. 6471 (2019): eaax3100.

[2] John E Fernández and Marcela Angel, "Ecological city-states in an era of environmental disaster: Security, climate change and biodiversity," *Sustainability* 12, no. 14 (2020): 5532.

[3] Paige West, James Igoe, and Dan Brockington, "Parks and peoples: the social impact of protected areas," *Annu. Rev. Anthropol.* 35 (2006): 251–277.

[4] Diaz et al., "Pervasive human-driven decline of life on Earth points to the need for transformative change."

[5] Stuart Basten, "Re-Examining the fertility assumptions for Pacific Asia in the UN's 2010 World Population Prospects," *University of Oxford Department of Social Policy and Intervention, Barnett*

cities are likely to overlap with important biodiversity hot spots and naturally species rich zones.[6] In order to make key decisions for biodiversity conservation, we must determine which factors impact urban biodiversity.

Over the past 20 years, there have been numerous studies evaluating the effect of urban environmental conditions on biodiversity.[7] Multiple studies examine the impact of cities on global biodiversity, but limit their analysis of biodiversity loss to habitat loss quantification, calling for more research to examine loss caused by other aspects of urbanization. For example, the study done by Powers and Jetz examines land use projections to 2070 to evaluate potential losses in suitable habitat.[8] Similarly, Seto et al. 2012 forecast global urban land cover change and examine its possible effects on biodiversity hotspots and carbon biomass, and found that the projected land use changes in urban area alone would cause concerning loss of biodiversity.[9] However, the main limitation with most of these studies is they do not examine other effects of increased urban population such as light pollution, local temperature change, etc.

Overall, the study of ecology has focused on features of climate and geography as determinants of biodiversity, neglecting to consider the local socio-economic conditions. Consequently, there are limited studies examining the relationship between the socio-economic features of a city and biodiversity loss. Those that do often limit their scope to a small region. For example, Hope et al. 2003 examines the socioeconomic impact on plant biodiversity, but their analysis is limited to the Central Arizona–Phoenix region, and uses a local biodiversity sampling approach.[10] In the Arizona-Phoenix region, their analysis found that more affluent neighborhoods host

Papers in Social Research 1 (2013).

[6] Richard P Cincotta, Jennifer Wisnewski, and Robert Engelman, "Human population in the biodiversity hotspots," *Nature* 404, no. 6781 (2000): 990–992.

[7] Sonja Knapp et al., "A research agenda for urban biodiversity in the global extinction crisis," *BioScience* 71, no. 3 (2021): 268–279.

[8] Ryan P Powers and Walter Jetz, "Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios," *Nature Climate Change* 9, no. 4 (2019): 323–329.

[9] Karen C Seto, Burak Güneralp, and Lucy R Hutyra, "Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools," *Proceedings of the National Academy of Sciences* 109, no. 40 (2012): 16083–16088.

[10] Diane Hope et al., "Socioeconomics drive urban plant diversity," *Proceedings of the National Academy of Sciences* 100, no. 15 (2008): 8788–8792, ISSN: 0027-8424, https://doi.org/10.1073/pnas.1537557100, eprint: https://www.pnas.org/content/100/15/8788.full.pdf, https://www.pnas.org/content/100/15/8788.

higher levels of plant biodiversity, and they name this phenomenon the "luxury effect." The luxury effect on plant biodiversity has been recognized in various urban centers: Beijing, Bujumbura, Las Angeles, etc.[11][12][13] Furthermore, some studies examine the luxury effect on local bird diversity, as well as other abiotic factors such as median housing age. For example, Loss et al. 2009 found that newer neighborhoods were more likely to house more species in Chicago, Illinois, and they assert that their " results suggest that investigating a combination of abiotic and environmental features of the built landscape, rather than focusing solely on environmental features, may provide a more complete understanding of the factors influencing avian diversity in human-dominated landscapes."[14] Despite the various studies examining the luxury effect, there is a limited understanding of the generalizability of this relationship to all cities, since so much of the research had limited geographic scope. Studies including multiple cities are necessary to examine what socio-economic factors drive biodiversity globally rather than at the local level.[15]

Aaronson et al. recognize that "comparative studies of urban biodiversity leading to robust generalities of the status and drivers of biodiversity in cities at the global scale are lacking," and compiled a dataset of avian diversity in 54 cities around the world. They also analyzed the loss of density in cities by modelling a number of anthropogenic and non-anthropogenic features of the city, including land cover, city age, climate and topography, but they do not consider socio-economic factors, such as gross domestic product (GDP).[16]

[11] Joseph Bigirimana et al., "Domestic garden plant diversity in Bujumbura, Burundi: Role of the socio-economical status of the neighborhood and alien species invasion risk," *Landscape and Urban Planning* 107, no. 2 (2012): 118–126.

[12] Lorraine Weller Clarke, G Darrel Jenerette, and Antonio Davila, "The luxury of vegetation and the legacy of tree biodiversity in Los Angeles, CA," *Landscape and Urban Planning* 116 (2013): 48–59.

[13] Hua-Feng Wang et al., "A basic assessment of residential plant diversity and its ecosystem services and disservices in Beijing, China," *Applied Geography* 64 (2015): 121–131.

[14] Scott R Loss, Marilyn O Ruiz, and Jeffrey D Brawn, "Relationships between avian diversity, neighborhood age, income, and environmental characteristics of an urban landscape," *Biological Conservation* 142, no. 11 (2009): 2578–2585.

[15] Misha Leong, Robert R Dunn, and Michelle D Trautwein, "Biodiversity and socioeconomics in the city: a review of the luxury effect," *Biology Letters* 14, no. 5 (2018): 20180082.

[16] Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers."

In short, many of these studies limit their scope to small regions or certain aspects of biodiversity, rather than analyzing the complex underlying relationship between urban environments and biodiversity, which is necessary for allocating limited resources to strategically protect the most at risk biodiversity hot-spots. Therefore, this research addresses this gap by examining a representative sample of global cities and creating a broad set of parameters for each including both socio-economic as well as environmental variables.

2.2 Spatial data

The novel abundance of spatial data, data that is largely collected by satellites, presents an opportunity to develop a scalable and sophisticated model of biodiversity. Tuia et al. state that "these new technologies and the data they generate hold great potential for large-scale environmental monitoring and understanding."[17] Furthermore, many previously inaccessible areas of interest can now be analyzed through high-resolution remote sensing technology.

Remote sensing data has been shown to be effective in measuring a number of environmental conditions, such as vegetation cover, photosynthetic capacity, leaf density, humidity, rainfall, etc.[18] An example of satellite data can be seen in figure 2, which depicts Normalized Difference Vegetation index (NDVI), a measure of greenery on the surface of the earth. See section 4.1 for more details on how NDVI has been used in previous research.

The work to use this data to map biodiversity is up and coming, but studies are generally limited to specific regions, such as West Africa. For example, Vaglio et al. (2014) explore whether airborbe hyperspectral imagery can be used to predict the diversity of upper canopy trees in a West African forest. Their research finds that airborn hyperspectral data is able to predict upper canopy diversity in this

[17] Devis Tuia et al., “Perspectives in machine learning for wildlife conservation,” *Nature Communications* 13, no. 1 (February 2022), https://doi.org/10.1038/s41467-022-27980-y, https://doi.org/10.1038%2Fs41467-022-27980-y.

[18] Shunlin Liang, *Quantitative remote sensing of land surfaces* (John Wiley & Sons, 2005).

area, however, they call for additional research to examine whether this technique is effective in different ecosystems.[19] Guo et a. (2017) present another example of a study which leverages airborne lidar data to model vegetation structure, but limits the geographical range of study to only Alberta, Canada. The limited geographic range of a large body of research in this field does not allow the findings to generalize to many cities, so it can only advise on limited areas.

Studies are also frequently limited to only one source of satellite data that provide a very limited measure of the factors that can affect biodiversity. For example, Powers and Jetz (2019) use land use change scenarios to predict biodiversity loss[20] While this work provides a global predictions of biodiversity loss, they fail to account for features outside of land use change that affect biodiversity, such as population density, night light, etc. As far as we know, no research has combined both socio-economic and environmental data to produce global predictions of biodiversity loss.

2.3 Machine learning

Machine learning (ML) is a rapidly growing field of artificial intelligence, changing every industry and incorporating into every facet of society. ML is able to learn patterns from large quantities of data and has been shown to be effective in predicting a number of risk profiles.[21] In fact, Christin et al. found that that deep learning, ML techniques that use neural networks, "can be beneficial to most ecological disciplines, including applied contexts, such as management and conservation."[22]

By generating data based predictions, ML techniques would allow scarce resources to be allocated in a strategic manner to maximize environmental benefits. For example, Hino et al. 2018 found that leveraging increasingly available electronic data

[19] Gaia Vaglio Laurin et al., “Biodiversity mapping in a tropical West African forest with airborne hyperspectral data,” *PloS one* 9, no. 6 (2014): e97910.

[20] Powers and Jetz, “Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios.”

[21] Alexander Y Sun and Bridget R Scanlon, “How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions,” *Environmental Research Letters* 14, no. 7 (2019): 073001.

[22] Sylvain Christin, Éric Hervet, and Nicolas Lecomte, “Applications for deep learning in ecology,” *Methods in Ecology and Evolution* 10, no. 10 (2019): 1632–1644.

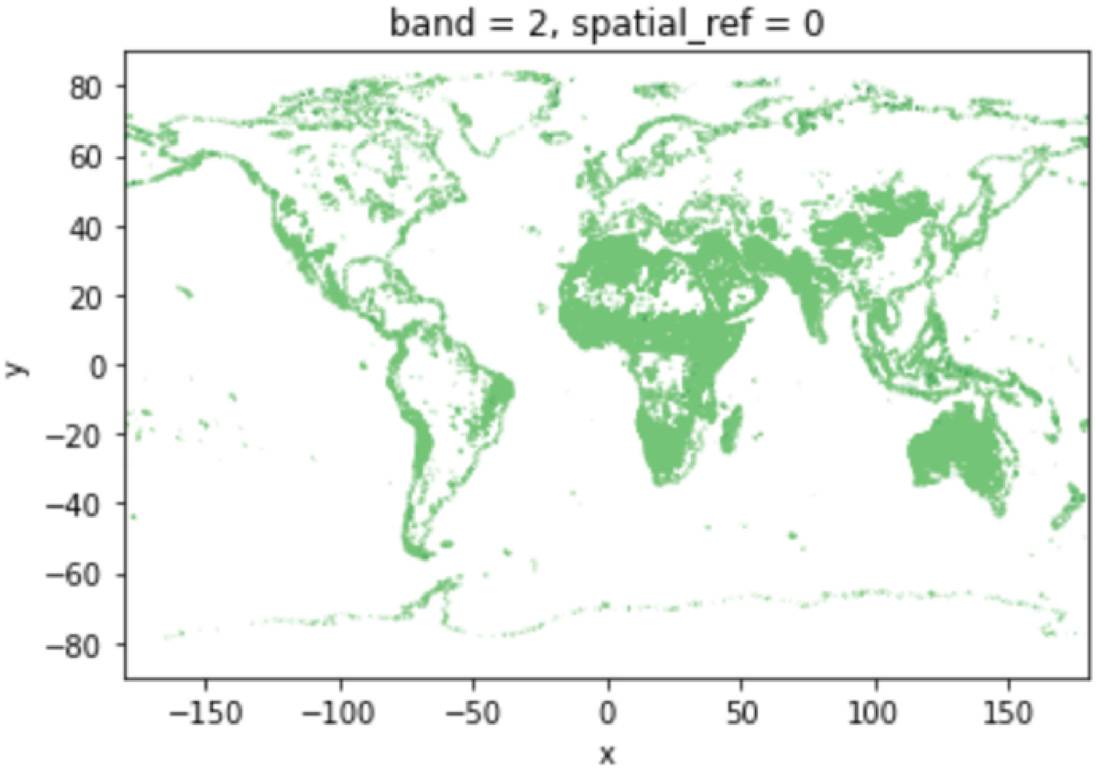


Figure 2-1: Example of Spatial Data: Normalized Difference Vegetation Index, global average in 1990 Source: Eric Vermote and NOAA CDR Program, *NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5*, 2019, https://doi.org/10.7289/V5ZG6QH9

allowed them to more appropriately distribute resources and prevent environmental harms, doubling their rate of effectiveness.[23] The allocation of scarce resources is particularly relevant to the topic of biodiversity. For example, if the US tries to create cities that are able to preserve biodiversity, it is crucial to know what cities are most important to focus on and what aspects of urban environments most affect the local biodiversity.

Many studies that analyze the impact of urbanization on biodiversity do not leverage ML techniques and instead rely on more traditional statistical techniques. In fact, Aronson MFJ et al. offer a novel global-scale comparative study of urban biodiversity, and use regression models to find how anthropogenic and non-anthropogenic factors

[23] Miyuki Hino, Elinor Benami, and Nina Brooks, "Machine learning for environmental monitoring," *Nature Sustainability* 1, no. 10 (2018): 583–588.

correlate to bird and plant species density in cities worldwide, however their study does not examine socio-economic factors, and they do not incorporate modern ML techniques into their analysis.[24]

ML problems can be "supervised" or "unsupervised". In supervised learning problems, training inputs are labelled with their ground truth outputs, and the model is trained on these outputs so it can learn the complex relationship between the inputs to the model and the outputs, leading to the correct outcome prediction when confronted with new, unseen inputs.[25] A key assumption of supervised learning is that the training data provided to the algorithm is representative of the examples that the model will be asked to predict. Unsupervised ML approaches are commonly used to separate data into different groups or "clusters." An unsupervised algorithm does not train on ground truth labels, but learns underlying patterns of the data to predict which outputs are most similar.

In this paper, we use 3 machine learning algorithms to learn the relationship between the socio-economic and environmental features of a city and its avian biodiversity loss: linear regression, random forest regression, and a hybrid model. In the remainder of this section we provide necessary background for the selected algorithms.

2.3.1 Linear regression

Linear regression is a simple statistical algorithm to analyze the linear relationship between input variables with an output and which variables in particular are most correlated with the output. Linear models are simple to evaluate and easy to visualize, which make them particularly helpful in understanding the relationship between environmental and socio-economic variables and avian biodiversity loss. For this reason, linear regression, or similar variations of it, has been one of the most common techniques within the environmental research community to model biodiversity loss. For example, Aronson et al (2014) use a version of linear regression to model avian

[24] Aronson et al., “A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers.”

[25] Sun and Scanlon, “How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions.”

biodiversity loss, which serves as a great baseline for which this research can compare.[26] Linear regression, however, is limited in the types of relationships it can learn between input variables and the predicted output, so it can potentially miss many of the nuances inherit in environmental relationships.

Linear regression finds the linear equation which takes in one or more input variables and outputs the best approximation of the dependent variable.[27] The relationship is modelled as follows

$$f(x_i) = \theta_0 + \theta(x_i)^T$$

Where θ_0 if the value of f when $x_i = 0$, and θ_1 is a vector representing the relationship between each variable and f

The most common method of fitting a regression line is ordinary least squares (OLS), which minimizes the sum of squares of the vertical deviations from the actual output data to the predicted line. In other words, linear regression fits a linear model with coefficients $theta = (theta_1, \dots, theta_n)$ to minimize the residual sum of squares between the observed values in the dataset, and the values predicted by the linear model.

2.3.2 Decision trees

A decision tree is a divide-and-conquer approach to classification and regression.[28] A decision tree is constructed by recursively partitioning the input features of a training set to find a set of decision rules that partition the input space to create an informative hierarchical prediction model.[29] Decision trees can be used for both classification problems, those with discrete output spaces, or regression problems, those with continuous output spaces.

[26] Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers."

[27] Evangelos C Alexopoulos, "Introduction to multivariate regression analysis," *Hippokratia* 14, no. Suppl 1 (2010): 23.

[28] Anthony J Myles et al., "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, no. 6 (2004): 275–285.

[29] Myles et al.

A classification problem takes in a training sample L = (X1, Y1), (X2 , Y2), ... , (XN, YN) of N observations, where each Xi = (Xa, . . Xk) is a k-dimensional vector of input variables and Y is an independent variable that takes one of j discrete values. The goal of classification algorithms is to learn the relationship between the values of the X vector and the given value of Y such that given a new value of X, the algorithm is able to correctly predict the corresponding value of Y.

Many classification algorithms exist, but decision trees have become increasingly popular because of their interpretability and high demonstrated prediction accuracy. Decision trees that are applied to classification problems are often called classification trees.

Decision trees form the basis of the random forest algorithm, described in the next section, and further used in this work to predict avian biodiversity loss from city features.

2.3.3 Random forest algorithm

The random forest algorithm (RF) was proposed by L. Breiman in 2001, and has since been successful as both a classification and regression method.[30][31] The algorithm samples fractions of the data, creates a randomized decision tree predictor on that subset of data, then finds the average of the decision trees' predictions. Over the past several decades, random forests have become widely used in machine learning research due to their applicability to a wide range of problems. The algorithm is also recognized for its accuracy and its ability to deal with small sample sizes and high-dimensional feature spaces.

Despite its ubiquity within the ML community, RF has only seen limited application within in biodiversity research, but there are recent examples of its use. For instance, Bayat et al. (2021) use RF to model tree diversity in Hyrcanian forests in northern Iran.[32] However, to our knowledge we are the first to apply RF to predict

[30] Leo Breiman, "Random forests," *Machine learning* 45, no. 1 (2001): 5–32.

[31] Gérard Biau and Erwan Scornet, "A random forest guided tour," *Test* 25, no. 2 (2016): 197–227.

[32] Mahmoud Bayat et al., "Assessing biotic and abiotic effects on biodiversity index using machine

biodiversity loss at a global scale.

An important aspect of the random forest algorithm is its ability to return measures of variable importance. Unlike most ML techniques where it is difficult to tease out how inputs relate to the model's predictions, random forests offer the transparency and explainability by directly measuring the importance of each variable in generating its predictions. For applied environmental research, model explainability is key to using data driven research to guide real world resource allocation. This property is particularly salient in this research, so that we can also identify which factors contribute the most to loss of biodiversity, and offer targeted interventions. For example, random forest may show that night light is extremely detrimental to avian biodiversity, so that information can be used to launch targeted interventions to help curb biodiversity loss in urban areas.

2.3.4 Hybrid supervised and unsupervised learning

Machine learning techniques are typically supervised *or* unsupervised. Supervised algorithms tend to focus on finding the relationship between each input variable and the output variable, while unsupervised techniques often reveal the structural characteristics of the input data.[33]

In the last ten years, scholars have proposed a number of hybrid models that combine unsupervised and supervised learning, to improve the performance of ML models. The most commonly used method for constructing a hybrid model is to combine a clustering algorithm with a decision tree algorithm. For example, Bose et al. (2009) proposed a hybrid model consisting of an unsupervised clustering technique and a boosted C5.0 decision tree to predict customer churn.[34] First, a clustering algorithm is used to cluster the samples. The clustering labels are then added to the original dataset as a new feature. The new dataset, constaining the assigned cluster

learning," *Forests* 12, no. 4 (2021): 461.

[33] Jin Xiao et al., "A hybrid classification framework based on clustering," *IEEE Transactions on Industrial Informatics* 16, no. 4 (2019): 2177–2188.

[34] Indranil Bose and Xi Chen, "Hybrid models using unsupervised clustering for prediction of customer churn," *Journal of Organizational Computing and Electronic Commerce* 19, no. 2 (2009): 133–151.

of each datapoint, is used to train a boosted C5.0 decision tree model. Their results show that the clustering technique improves the performance of their model compared with the benchmark.[35] Furthermore, Rajamohamed et al. combined K-means clustering algorithms with five supervised models, to construct different versions of hybrid models, indicating that various supervised and unsupervised techniques can successfully be combined.[36]

[35] Bose and Chen.

[36] R Rajamohamed and J Manokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," *Cluster Computing* 21, no. 1 (2018): 65–77.

Chapter 3

Methods

In this chapter, we describe how we implement machine learning models to take in city features, such as average GDP and night light, and output the predicted amount of biodiversity loss in a FUA. The main goals of our methods are twofold: to be able to accurately predict avian biodiversity loss in cities where tedious survey data is not available, and to identify the main drivers of urban biodiversity loss so we are able to identify relevant interventions.

This chapter has 3 main sections. We discuss the machine learning methods, including linear regression, random forest, and a hybrid unsupervised and supervised model, used to create biodiversity loss predictions. These models are also used to analyze how various socio-economic and environmental features relate to biodiversity in order to increase understanding of the relationship between cities and avian biodiversity.

The first models trained on the dataset include linear regression, described in section 2.3.1, and serve as a useful baseline since regression models were used by Aronson et al. (2014).[1] Aronson et al. (2014) results offer a useful benchmark for our model, since we use the dataset of avian biodiversity they compiled to create ground truth labels, as described in 4.9

The random forest algorithm, as described in 2.3.3 was also selected due to its ease

[1] Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers."

of interpretability. Finally, we implement a hybrid model that uses both supervised and unsupervised learning methods to leverage the structural insights gained from clustering as the supervised model learns to predict on unseen data.

3.1 Linear regression

The linear regression analysis in this work was implemented in Python using scikit-learns's LinearRegression (version 1.0.2).[2] For the regression problem the intercept was calculated, because the data is not assumed to be centered. All variables were normalized so that the model would not be biased toward features with a larger scale. The coefficients were not constrained to be positive.

In all models (linear, random forest, and hybrid), dataset partitioning was done with an 80:20 scheme, so out of the $n = 54$ datapoints available, 43 were chosen uniformly and identically at random for model training, and the remaining 11 formed the test set on which the loss function was computed and reported. Training was done in the MIT supercloud with 6 cpus.

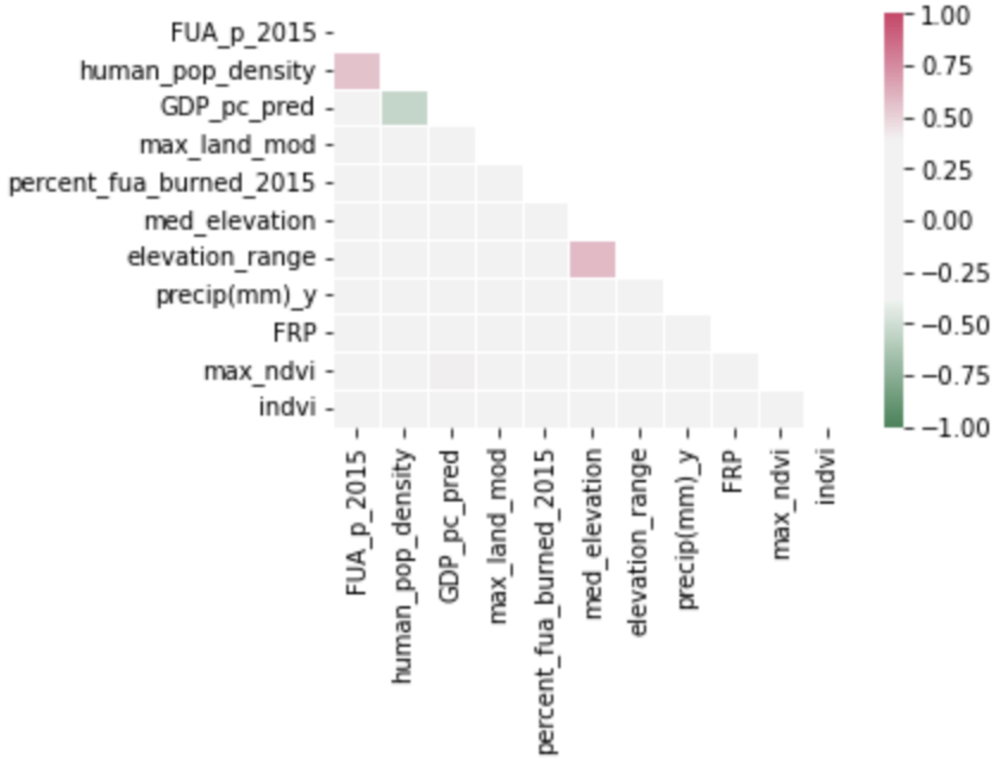
Before running the regression algorithm, we visualized the Spearman rank correlation coefficients for all pairwise combinations of our input features.[3] Large correlations between input features would imply multi-collinearity, which would not affect the accuracy of the linear regression but could confound the ranking of importance of features. By visualizing the Spearman coefficients in 3-1, we see that almost all of the coefficients are near 0, and there are no significant correlations to confound our model.[4]

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–2830.

[3] Leann Myers and Maria J Sirois, "Spearman correlation coefficients, differences between," *Encyclopedia of statistical sciences* 12 (2004).

[4] Kelly H Zou, Kemal Tuncali, and Stuart G Silverman, "Correlation and simple linear regression," *Radiology* 227, no. 3 (2003): 617–628.

Figure 3-1: Spearman rank correlation coefficients of features



3.2 Random forest

The Random forest algorithm used in this analysis was implemented in Python using scikit-learn's RandomForestRegressor (version 1.0.2).[5] For the regression problem a total of 100 individual decision trees were bagged. The trees had no max depth and split nodes until complete purity was reached. Mean squared error was chosen as the loss function to minimize.

3.3 Hybrid

Our hybrid model combines both unsupervised and supervised machine learning. For the unsupervised component, the cities are clustered using an improved version of K-means presented by Paul S Bradley and Usama M Fayyad..[6] We chose K-means

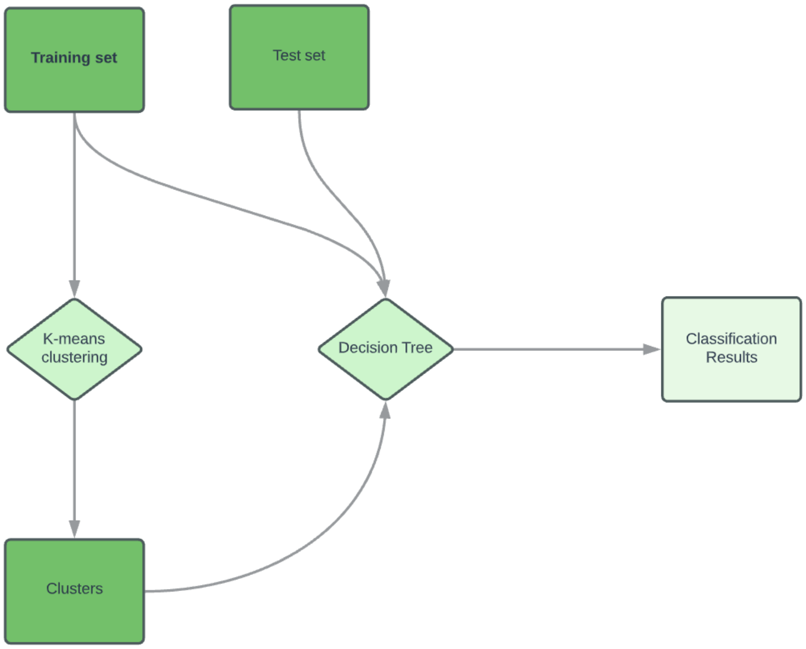
[5] Pedregosa et al., "Scikit-learn: Machine Learning in Python."

[6] Paul S Bradley and Usama M Fayyad, "Refining initial points for k-means clustering.," in *ICML*, vol. 98 (Citeseer, 1998), 91–99.

because it is a widely used algorithms, and we leverage a modification to make it less sensitive to the initialization parameters. Analyzing the output of this step is helpful in analyzing what trends the unsupervised algorithm is identifying in the data, and it allows us to create visualizations of the clusters, as shown in section 5

In the supervised classification portion of our model, we train a C4.5 decision tree on the entire training set, using the hybrid information gain ratio presented by Xiao et al. to consider both the class labels and the clusters to which they belong.[7] We chose to use C4.5 because of its good performance, but also for its extremely strong interpretability. We hope for this research to be able to inform policy decisions, so it is extremely high priority that this model is able to be understood.

Figure 3-2: Hybrid Model



[7] Xiao et al., "A hybrid classification framework based on clustering."

Chapter 4

Data collection and processing

This research aims to analyze both the environmental and socio-economic conditions present in cities to predict avian biodiversity loss. Recent research has shown that species diversity is affected by social, economic and cultural influences which are not recognized by traditional ecological theory.[1] For example, Hope et al. (2008) found that family income was one of the most useful variables in predicting plant diversity across the Arizona-Phoenix region.

In order to interpret and use the satellite data, we must find representative features for every city. For example, If a city has many elevations recorded at different points within the city, we must decide how to aggregate that into a feature for the city, e.g. maximum elevation within a city or mean elevation. The datasets used in this study are publicly available as referenced in the remainder of this section. All datasets were processed in Python using statistical packages Pandas, and Numpy as well as Geopandas and Rioxarray, libraries specific to processing spatial and geographical data.[2][3][4]

We selected variables that have been shown to affect species diversity in previous

[1] Hope et al., "Socioeconomics drive urban plant diversity."

[2] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, ed. Stéfan van der Walt and Jarrod Millman (2010), 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.

[3] The pandas development team, *pandas-dev/pandas: Pandas*, v. latest, February 2020, https://doi.org/10.5281/zenodo.3509134, https://doi.org/10.5281/zenodo.3509134.

[4] Kelsey Jordahl et al., *geopandas/geopandas: v0.8.1*, v. v0.8.1, July 2020, https://doi.org/10.5281/zenodo.3946761, https://doi.org/10.5281/zenodo.3946761.

research, including Normalized Difference Vegetation Index, Wildfires, Land Surface Temperature, Land Modification, Precipitation, Night Light, Elevation, Population Density, and GDP. The remainder of this chapter describes the selection and creation of our features, citing relevant literature to justify why each feature is useful in predicting biodiversity loss. Furthermore, the description and source of the data for features used in our analysis in summarized in table 4.1

4.1 NDVI

Normalized Difference Vegetation Index (NDVI) describes the amount of vegetation cover on an area of land.[5] NDVI ranges from -1 to +1 and quantifies vegetation by measuring the difference between the near-infrared spectrum, which vegetation strongly reflects, and red light, which vegetation absorbs. Using The National Oceanic and Atmospheric Administration's (NOAA) Advanced Very High Resolution (AVHRR) imaging, scientists have been measuring the wavelengths and intensity of visible (VIS) and near-infrared (NIR) light reflected by the surface of the earth back into space.[6]

The NOAA AVHRR instrument has five sensors two of which measure light with wavelengths ranging from 0.55 to 0.70 and 0.73 to 1.0 micrometers. Using these AVHRR sensors, we can calculate the amount of light reflected by earth in visible and near infrared spectra. To quantify the density of vegetation, each pixel corresponds to 1 square km of earth's surface, and each NDVI value is given by near-infrared radiation minus visible radiation divided by near-infrared radiation plus visible radiation, see figure 4 for illustration.[7]

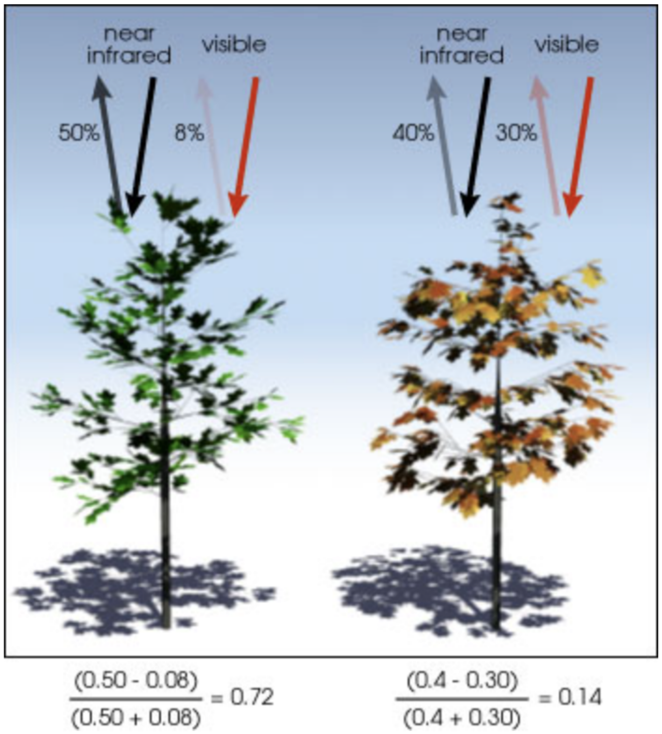
$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{VIS}}{\mathrm{NIR} + \mathrm{VIS}}$$

[5] Eric Vermote and NOAA CDR Program, *NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5*, 2019, https://doi.org/10.7289/V5ZG6QH9.

[6] Vermote and Program.

[7] John Weier and David Herring, *Measuring Vegetation (NDVI and EVI)*, https://earthobservatory.nasa.gov/features/MeasuringVegetation.

Figure 4-1: Measuring NDVI



NDVI is calculated from the visible and near-infrared light reflected by vegetation. Healthy vegetation (left) absorbs most of the visible light that hits it, and reflects a large portion of the near-infrared light. Unhealthy or sparse vegetation (right) reflects more visible light and less near-infrared light. The numbers on the figure above are representative of actual values, but real vegetation is much more varied. (Illustration by Robert Simmon).

Using this dataset, we found two features based on NDVI for each city: maximum NDVI and Integrated NDVI (INDVI). INDVI is defined as the number of positive NDVi values over a year. It can also be thought of as the number of "green days," because it counts the number of days when the greenery is above 0. These indices were chosen because Pettorelli et al. (2005) identify them as the most useful indicators for overall productivity and bio-mass, both of which are extremely influential on local biodiversity.[8]

[8] Nathalie Pettorelli et al., “Using the satellite-derived NDVI to assess ecological responses to environmental change,” *Trends in ecology and evolution* 20, no. 9 (2005): 503–510.

4.2 Wildfires

Wildfires cause a massive amount of destruction to ecosystems and are known to cause catastrophic biodiversity loss.[9] Not only do the fires scorch necessary vegetation, they also release immense amounts smoke and change local environments for decades. Furthermore, Silveira et al. (2015) found that recurrent fires exhibit strong effects on species richness and composition across all sample groups, and recurrent fires had more severe environmental effects when compared to instances of non-recurrent fires in the Amazon Forest.[10]

Wildfires play a crucial role in understanding how local bird populations may be at risk for species loss. We have obtained data on wild fires over the past two decades from NASA Earth Observations Active Fires dataset.[11] This data captures whether a wildfire was present on an area of land at a given time, so it is able to tell us the frequency of fires as well as the intensity of the burn. Using the global raw satellite data from 2015, we computed the percent of the functional urban area that was burned throughout the course of the year by dividing the number of pixels that were burned, by the total number of pixels contianed within the FUA. This metric represents the spread of fire and how much of the FUA 's environment was affected. We also calculated the average fire radiative power (FRP) within a FUA over the course of 2015 by taking the average of daily observations, and taking the average of all of the pixels within the city. FRP is a measure of fire intensity, similar to how hot a fire burns, and it is indicative of the amount of emissions released by a fire as well as the degree of damage done.[12] Both of these metrics were shown to impact avian biodiversity in Bolivia in pervious research.[13]

[9] Letıcia Couto Garcia et al., "Record-breaking wildfires in the world's largest continuous tropical wetland: integrative fire management is urgently needed for both biodiversity and humans," *Journal of environmental management* 293 (2021): 112870.

[10] Juliana M Silveira et al., "A multi-taxa assessment of biodiversity change after single and recurrent wildfires in a Brazilian Amazon forest," *Biotropica* 48, no. 2 (2016): 170–180.

[11] Louis Giglio et al., "Collection 6 modis burned area product user's guide version 1.3," *NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA*, 2020,

[12] O Maillard et al., *Impact of Fires on Key Biodiversity Areas (KBAs) and Priority Bird Species for Conservation in Bolivia. Fire 2022, 5, 4*, 2022.

[13] Maillard et al.

4.3 Land surface temperature

Land surface temperature measures the temperature of an area of the earth's surface, which is related to many factors such as green house gases and land use.[14] For example, cities with black asphalt may have an increased surface temperature and may be less hospitable to bird species. This effect is commonly referred to as the urban heat island (UHI) effect, and is it a result of the increased heat storage capacity of urban surfaces.[15] In urban areas where the local temperature is abnormally hot, food crops may die, ruining entire ecosystems and also affecting bird populations. In fact, UHI has been identified as one of the major problems in the 21st century, as it presents extremely detrimental effects to urban areas.[16] Land surface temperature is an important indicator of the UHI and, therefore, how the climate and local weather is changing within a city.

We source data from NASA Earth Observation's Land Surface Temperature data set.[17] We find all data points from 2015 that lie within the bounds of each FUA, and create an average value per FUA by averaging each time observation, and averaging all data points within the FUA. This metric is key to or research, because the model may show that a higher average land surface temperature within a city is correlated with a higher rish of avian biodiversity loss, suggesting that an effective potential biodiversity intervention is reducing the UHI effect.

[14] Duy X Tran et al., "Characterizing the relationship between land use land cover change and land surface temperature," *ISPRS Journal of Photogrammetry and Remote Sensing* 124 (2017): 119–132.

[15] Tran et al.

[16] Ahmed Memon Rizwan, Leung YC Dennis, and LIU Chunho, "A review on the generation, determination and mitigation of Urban Heat Island," *Journal of environmental sciences* 20, no. 1 (2008): 120–128.

[17] *MOD11C2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 0.05Deg CMG V006.*, 2015, https://doi.org/10.5067/MODIS/MOD11C2.006.

4.4 Land modification/land use changes

The human modification of terrestrial systems is known to be one of the most significant drivers of habitat and, therefore, biodiversity loss.[18] In fact, Sala et al. state that for terrestrial ecosystems, land-use change will liely have the largest effect on biodiversity loss.[19] It is important to understand how the land in a city has been changed from its natural state to be able to assess how that may affect local species populations.

We leverage the The Global Human Modification (gHM) data set which estimates the degree of human modification on all lands excluding Antarctica at a 1km resolution.[20] This dataset provides a continuous 0-1 metric at a resolution of 1 km that reflects the proportion of a landscape modified. gHM is comprised of 5 types of input variables: human settlement, agriculture, transportation, mining and energy production, and electrical infrastructure. See figure 4-2 for a visualization of this data. The Nature Conservancy Global Development Risk Assessment (GDRA) has stated that this data set can be used to "improve conservation strategies in landscape level mitigation, laws and regulations, lending requirements, and protection," such as biodiversity preservation.[21] To create a gHM value per city, we take the gHM for each $1km^2$ of land in the FUA, and calculate the average throughout the city as well as the maximum value within the city. These metrics are useful inputs in our biodiversity analysis, as they gauge how much land modification really affects the biodiversity loss within a FUA.

4.5 Precipitation

Precipitation, or the amount of rainfall that a city receives, is extremely important in contextualizing the local environmental conditions. Many avian species rely on

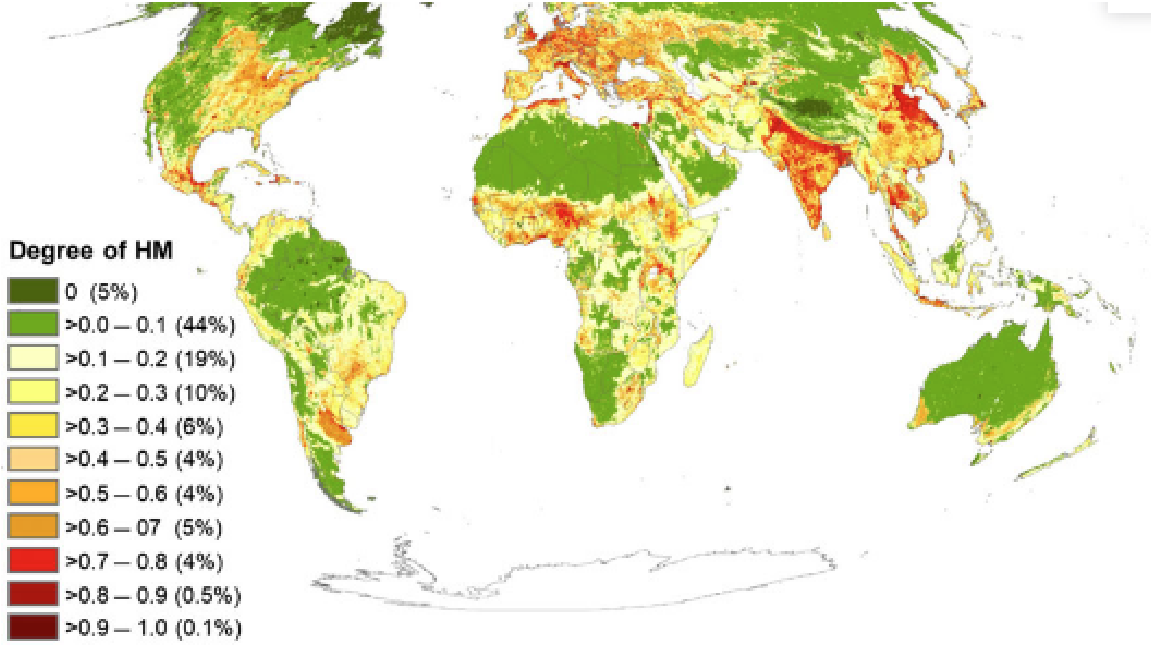
[18] Osvaldo E Sala et al., “Global biodiversity scenarios for the year 2100,” *science* 287, no. 5459 (2000): 1770–1774.

[19] Sala et al.

[20] Christina M Kennedy et al., “Managing the middle: A shift in conservation priorities based on the global human modification gradient,” *Global Change Biology* 25, no. 3 (2019): 811–826.

[21] *Current land modification*, http://gdra-tnc.org/current/.

Figure 4-2: Cumulative human modification across global terrestrial lands. Source: Christina M Kennedy et al., “Managing the middle: A shift in conservation priorities based on the global human modification gradient,” *Global Change Biology* 25, no. 3 (2019): 811–826



predictable precipitation, and overly dry or wet conditions can threaten their population. In fact, McCain and Colwell (2011) modelled local population extirpation risk for a range of temperature and precipitation scenarios and found that under the driest conditions (minimum predicted precipitation), local extirpation risks increased drastically (50–60%).[22]

The Microwave Climate Data Center offers high quality geophysical data from satellite microwave sensors, providing accurate measures of liquid water precipitation, rain rate, at a resolution of .25 x .25 degrees. These Special Sensor Microwave/Imager (SSM/I) and Special Sensor Microwave Imager Sounder (SSMIS) data products are produced as part of the NASA’s MEaSUREs Program.[23]

We isolate all precipitation within each functional urban area and take the spatial

[22] Christy M McCain and Robert K Colwell, “Assessing the threat to montane biodiversity from discordant shifts in temperature and precipitation in a changing climate,” *Ecology letters* 14, no. 12 (2011): 1236–1245.

[23] F. J. Wentz, K. A. Hilburn, and D K. Smith, *Remote Sensing Systems DMSP SSMI / SSMIS Environmental Suite on 0.25 deg grid, Version 7*, 2012, https://www.remss.com/missions/ssmi/.

average followed the sum over year 2015, to find the average yearly precipitation. Including this feature in our model allows us to gauge how the average precipitation in a city affects the local avian population, which may indicate that droughts are an issue and suggest possible interventions.

4.6 Night light development Index

Light pollution is increasingly recognized as a driver of biodiversity loss, as it may cause a disruption in sleeping and mating patterns in many species.[24] Brei et al. 2016 studied the effect of light pollution on Sea Turtles in the Caribbean and found that nighttime light significantly reduces the number of sea turtle nests.[25] Furthermore, night light can serve as an empirical measurement of human development, where more developed areas tend to emit more artificial night light.[26] As a result, night light has been shows to be negatively correlated to rates of poverty.

the Joint Polar-orbiting Satellite System (JPSS), the Visible and Infrared Imaging Suite (VIIRS) Day Night Band (DNB) on board of JPSS satellites has been leveraged by Colorado School of Mines' Earth Observation Group to produce global Nighttime Light maps. This data set is a cloud-free composite, as using a cloud free dataset is integral to ensuring that areas with cloud cover are not biased toward lower emitted light due to cloud blocking. Then, following the method outlined by Elvidge et al. (2012), to create the night light development index (NLDI) we combine the night light radiance values with FUA population count to form a Lorenz and obtain the Gini coefficient for spatially at a resolution of a a quarter degree.[27]

This index acts as a simple, objective measure of human development. Not only does it directly represent the amount of night light in a FUA, it also acts as a measure

[24] Jari Lyytimäki, "Nature's nocturnal services: Light pollution as a non-recognised challenge for ecosystem services research and management," *Ecosystem Services* 3 (2013): e44–e48.

[25] Michael Brei, Agustin Perez-Barahona, and Eric Strobl, "Environmental pollution and biodiversity: Light pollution and sea turtles in the Caribbean," *Journal of Environmental Economics and Management* 77 (2016): 95–116.

[26] C. D. Elvidge et al., "The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data," *Social Geography* 7, no. 1 (2012): 23–35, https://doi.org/10.5194/sg-7-23-2012, http://www.soc-geogr.net/7/23/2012/.

[27] Elvidge et al.

of wealth and income among a FUA, as the NLDI has been shown to be correlated with both the overall human wellfare and the environmental quality of an urban area.[28] As a result, this metric is useful in our research for two reasons: NLDI helps us gauge the impact of nightlight on biodiversity, and NLDI offerns insights into how wealth and income within a FUA affect avian biodiversity.

4.7 Elevation

Elevation has been shown to be an important correlate of global patterns of species.[29] While we do not expect elevation to significantly change due to urbanization (though landscaping may cause some elevation difference), it is possible that elevation changes the way avian species interact with an environment, and so we believe it is important to include in a model predicting avian species richness.

To measure the elevation within our FUAs, we leverage data from the U.S. Geological Survey (USGS) and the National Geospatial-Intelligence Agency (NGA), who have developed a global elevation model called Global Multi-resolution Terrain Elevation Data (GMTED2010)[30] For each FUA, we calculate the minimum, maximum, and median elevation, which serve to represent the elevation most commonly found within each FUA, as well as the extremes that the avian population may experience.

4.8 FUA population statistics

In 2010, 757 million people resided in the world's 101 largest cities, and urban populations are growing faster than ever.[31] As urban populations expand, cities develop new structures and strains on the environment. We will consider cities population size by examining the European Commission's Global Human Settlement (GHS) dataset,

[28] Elvidge et al.

[29] Gaston, "Global patterns in biodiversity."

[30] Jeffrey J Danielson and Dean B Gesch, *Global multi-resolution terrain elevation data 2010 (GMTED2010)* (US Department of the Interior, US Geological Survey Washington, DC, USA, 2011).

[31] Daniel Hoornweg and Kevin Pope, "Population predictions for the world's largest cities in the 21st century," *Environment and Urbanization* 29, no. 1 (2017): 195–216.

which allows us to obtain human populations of each city.[32]

In addition to total population, it is also important to consider how dense that population is within a city, as a high population density can put unique strains on the local environment. We will also leverage the European Commission's GHS dataset to obtain this information.

4.9 Ground truth labels

Following the method outlined by Aronson et al. 2014, we construct ground truth labels by calculating the pre-urbanization species density and finding the current species density within the urban area, then finding the density loss.[33] The density of species of extant birds is calculated as the number of species per km2 for each urban using estimates of the city area given by the OECD.[34]

We first estimate the pre-urbanization bird species density by leveraging BirdLife International range maps.[35] As Aronsons et al. state, "Because range maps provide representations of species' extent of occurrence at coarse resolutions with little or no consideration of changes in occupancy owing to land-use change, they are ideally suited to estimate non-urban density of bird species within larger areas such as cities."[36]

To measure the post-urbanization population density, we leverage the dataset compiled by Aronson et al. (2014). This dataset was created by researchers who comprehensively reviewed literature, databases, and expert surveys to compile an estimate of bird species present in 54 cities globally. All species recorded in surveys conducted since 1990 are counted for each city. According to Aronson et al., some of the datasets were based on intensive surveys conducted for 1 or 2 years, whereas others

[32] M Schiavina et al., "GHS-FUA R2019AGHS functional urban areas, derived from GHS-UCDB R2019A,(2015)," *R2019A. edited by Joint Research Centre (JRC) European Commission*, 2019,

[33] Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers."

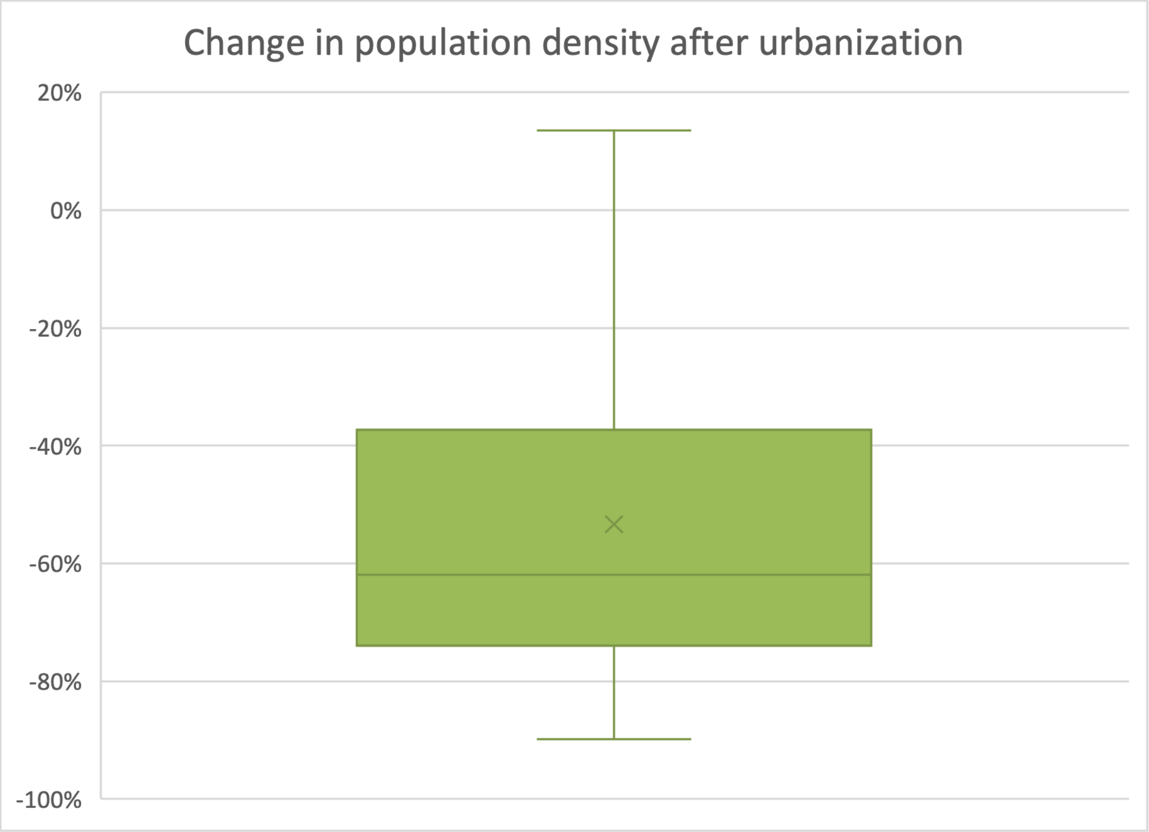
[34] Dijkstra, Poelman, and Veneri, "The EU-OECD definition of a functional urban area."

[35] BirdLife International and Handbook of the Birds of the World, *Bird species distribution maps of the world.*, 2021, http://datazone.birdlife.org/species/requestdis..

[36] Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers."

represented data collected over multiple time periods. Furthermore, the datasets were complete lists from within the administrative boundary of a city, including inner urban as well as the peri-urban areas which makes this data consistent with using the FUA, rather than the official political boundaries of a city, which often do not include peri-urban areas.

Figure 4-3: Change in population density in 54 test cities. Data source: Myla FJ Aronson et al., "A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers," *Proceedings of the royal society B: biological sciences* 281, no. 1780 (2014): 20133330



Finally, to find the density loss we subtract the pre-urbanization population density from the post-urbanization and then divide by the pre-urbanization to find that percent change in bird diversity. The distribution of our labels is depicted in figure 4-3 , which shows that bird diversity change in our labelled cities ranges from -90% to over 10%. So, this distribution represents a wide range of avian biodiversity change, which is helpful in training a model. The change in avian population density values are also geospatially visualized in figure 4-4, which presents the geographical range of our labels as well.

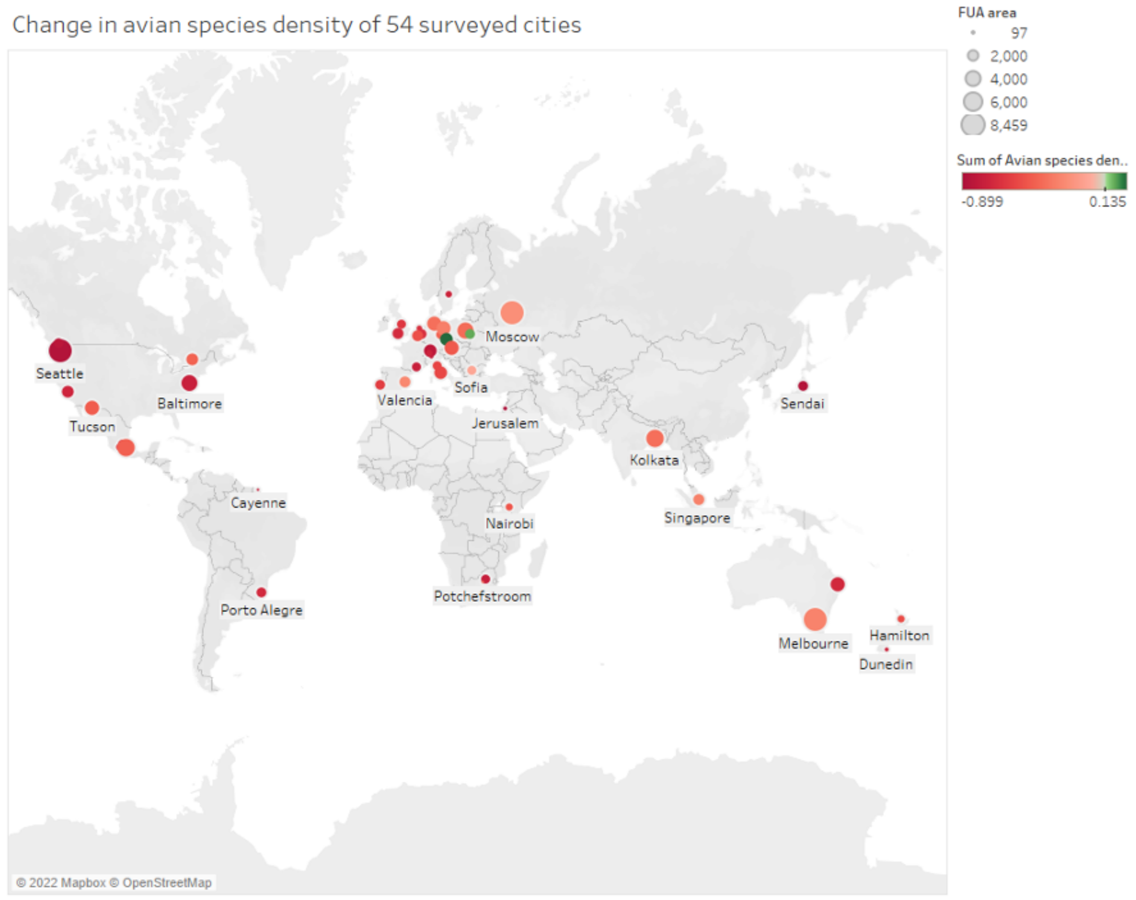


Figure 4-4: World map of change in population density in 54 test cities. Sources: Aronson et al. (2014) and BirdLife International and Handbook of the Birds of the World

Table 4.1: Description and sources for variables used in our analysis

Variable	**Description**	**Source**
NDVI	Amount of vegetation cover of an area of land	Vermote and Program, *NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5*
Wildfires	Percent of FUA burned by wildfires and intensity of fires as FRP	Giglio et al., "Collection 6 modis burned area product user's guide version 1.3"
Land Surface Temperature	Annual mean temperature	*MOD11C2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 0.05Deg CMG V006.*
Land Modification	Average Global Human modification and Maximum Global Human Modification	Kennedy et al., "Managing the middle: A shift in conservation priorities based on the global human modification gradient"
Precipitation	Average yearly rainfall	Wentz, Hilburn, and Smith, *Remote Sensing Systems DMSP SSMI / SSMIS Environmental Suite on 0.25 deg grid, Version 7*

Continued on next page

Table 4.1 – *Continued from previous page*

Variable	**Description**	**Source**
Night Light	Average night light	Elvidge et al., "The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data"
Elevation	Minimum, maximum and median elevation	Danielson and Gesch, *Global multi-resolution terrain elevation data 2010 (GMTED2010)*
FUA population statistics	Total population, population density, and change in population density	Dijkstra, Poelman, and Veneri, "The EU-OECD definition of a functional urban area"
GDP	GDP per capita	Mo Elhabashy

Chapter 5

Results

In this section we discuss the results of the 3 models, linear regression, random forest, and hybrid, on our avian biodiversity dataset. This chapter is split into 2 parts: 5.1 discusses how well our models were able to fit to the data and predict on unseen datapoints, and section 5.2 presents the importance that each model placed on the input features.

5.1 Accuracy and predictive performance

To assess the models, we score their classification accuracy on a test dataset that the model was not trained on, so that we can gauge the models ability to generalize to unseen data. We calculated severeal key metrics to gauge the performance of our models: mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). These metrics are common in data science literature in evaluating regression models, and the results are summarized in figure 5.1

The average absolute difference between a set of predicted (p_i) and observed (o_i) values, or the mean absolute error (MAE), was calculated as follows

$$MAE = N^{-1} \sum_{i=1}^{N} |p_i - o_i|$$

The MSE represents the square of the difference between a set of predicted (p_i)

and observed (o_i) values, and it is useful in penalizing predictions that are very far from observed values. RMSE is simply the square root of SME

$$MSE = N^{-1} \sum_{i=1}^{N} (p_i - o_i)^2$$

Model	**MAE**	**MSE**	**RMSE**
Linear regression	**13.49**	360.78	18.99
Random forest	13.97	261.23	16.16
Hybrid	13.95	**261.2**	**16.16**

Table 5.1: Model key performance indicators

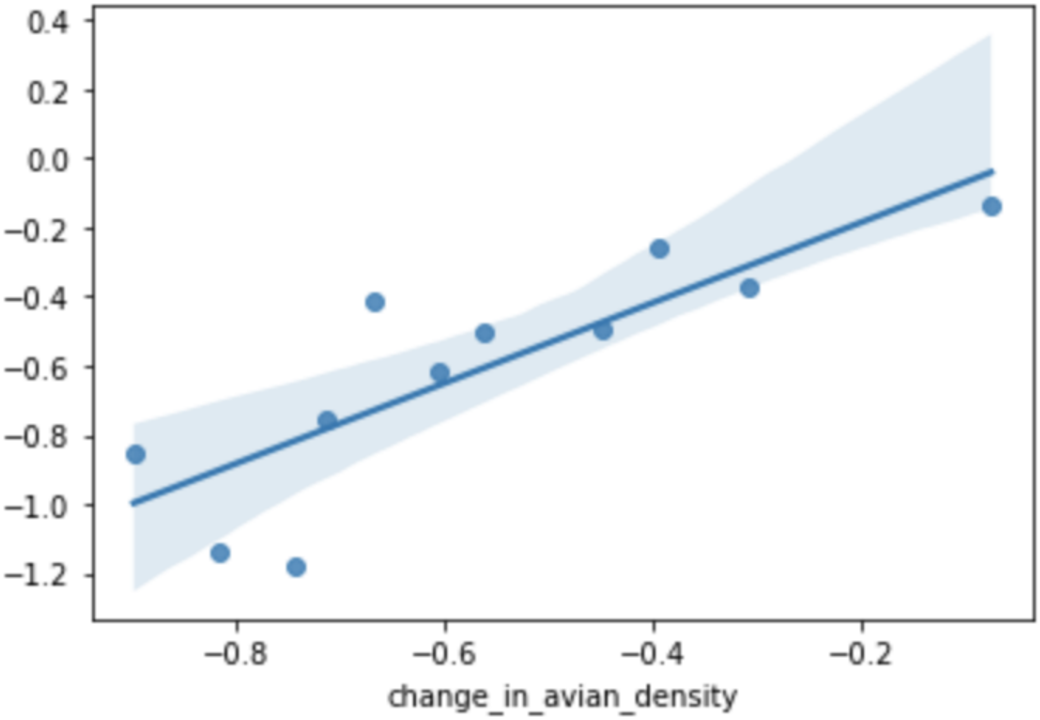
The lowest MAE achieved of 13.49 indicates that, on average, the linear regression was 13.49% off in predicting the percent change in avian biodiversity. Notably, the MAE was within one standard deviation of the change in avian species density. Therefore, any differences or prediction errors between the predicted and observed values are within the range of natural variability. Despite the fact that our linear regression performed the most accurately based on MAE, the random forest and hybrid model both achieved similarly low MAE's. The Hybrid model obtained the best accuracy when examining the MAE and MSE, though the RF achieved extremely similar results.

Furthermore, the performance of each model on the test set can be visualized in figures 5-1, 5-2, and 5-3. A visual analysis of these figures indicates that all three models were able to successfully extract a predictive signal from the data.

5.2 Feature importance analysis

A key contribution of this research is not only the ability to predict which urban areas are most at risk for avian biodiversity loss, but also to understand the key drivers of that loss. The models we implemented offer insights into how they used each input variable to predict the avian biodiversity loss. The remainder of this section will analyze the feature importance of each model.

Figure 5-1: Linear regression predicted vs. actual change in avian species density

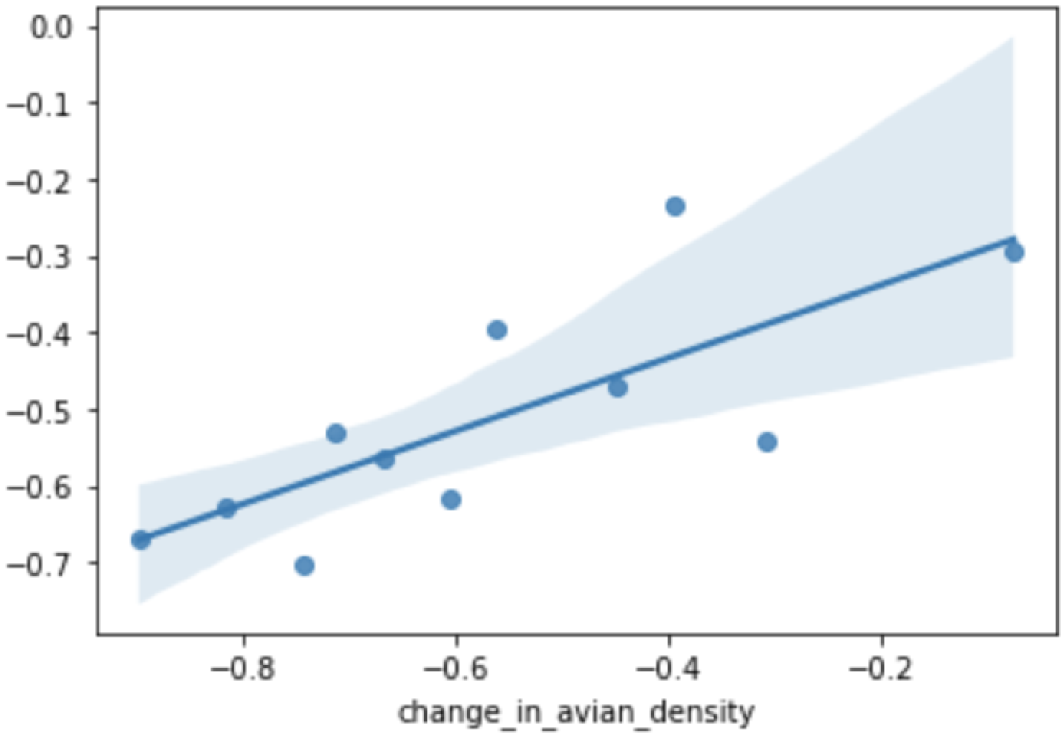


5.2.1 Linear regression

Regression analysis describes how the changes in the variables affect the value of the dependent variable. It does, however, assume the input variables are perfectly independent, which, from our Pearson analysis, we know is not perfectly true. Though the lack of independence in the input variables does not affect the performance of the model, it does confound the interpretability of the models coefficients. In other words, the coefficients should denote how much the dependent variable is expected to increase when that input variable increases by one, while holding all the other variables constant. However, since the input variables are not perfectly independent, this statement is not entirely accurate. As a result, the implications of the assigned coefficients of our input variables should be analyzed with this limitation in mind.

Nevertheless, it is notable that all three of our models consider the maximum land modification of a FUA to be the most significant feature. Linear regression assigns maximum land modification the highest coefficient, then maximum NDVI, followed

Figure 5-2: Random forest predicted vs. actual change in avian species density



by perfect of the FUA burned.

5.2.2 Random forest

RF offers a more interpretable view of the importance it places on each variable. It assigns each variable a measure of importance based on how perturbing each value changes the predicted outcome.[1] This measure does not require all features to be independent in order to get an accurate ranking. The results can be visualized in figure 5-5. Max land modification is again the most important variable in predicting the change in avian biodiversity. GDP per capita is the next most important feature, followed by median elevation, precipitation, population, elevation range, and FRP respectively.

[1] Breiman, "Random forests."

Figure 5-3: Hybrid predicted vs. actual change in avian species density

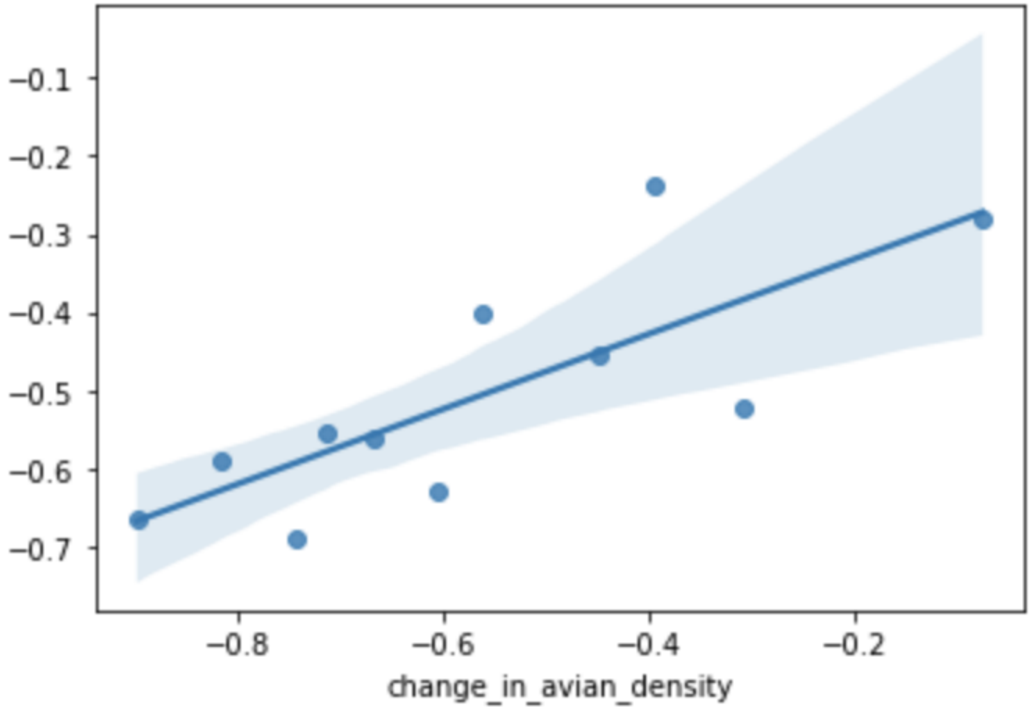
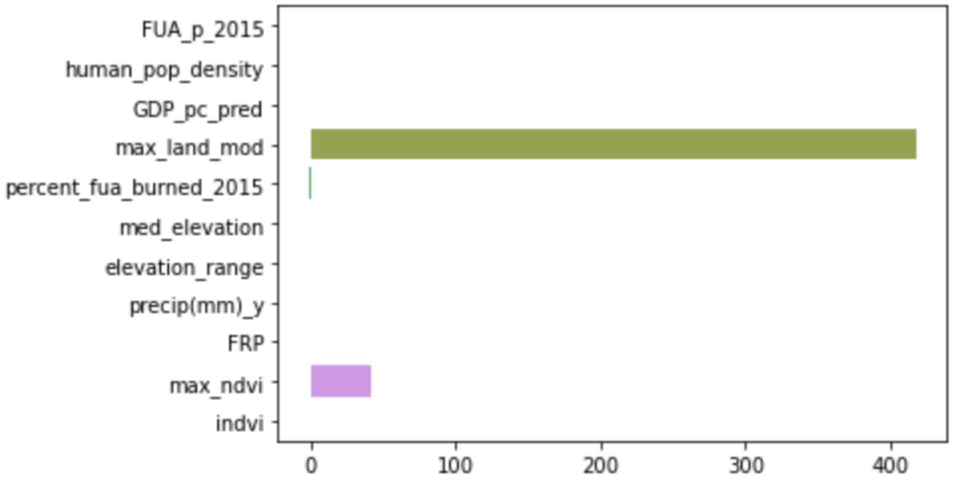


Figure 5-4: Linear regression coefficients



5.2.3 Hybrid

Our hybrid model was designed with interpretability in mind. It inherits the same feature importance algorithm as the random forest model that it uses for supervised

Figure 5-5: Random forest feature importance

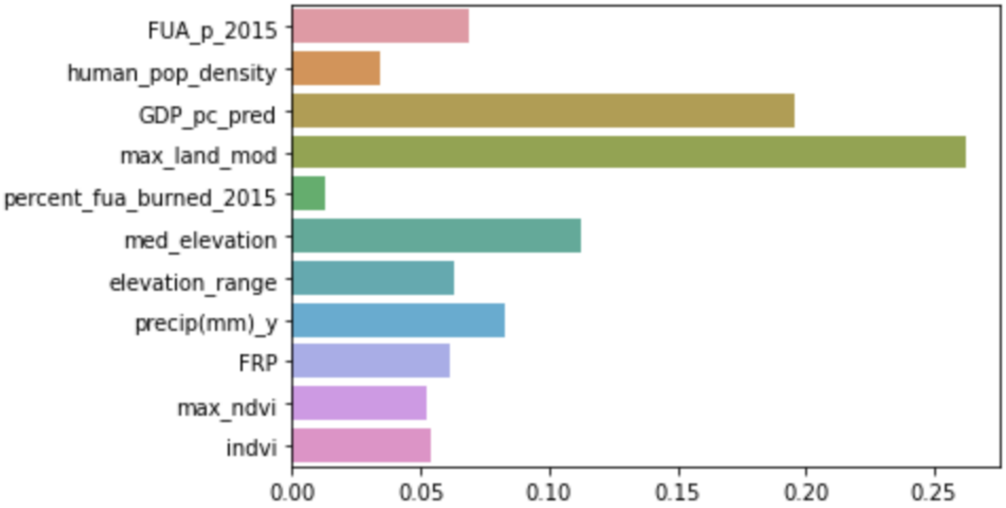
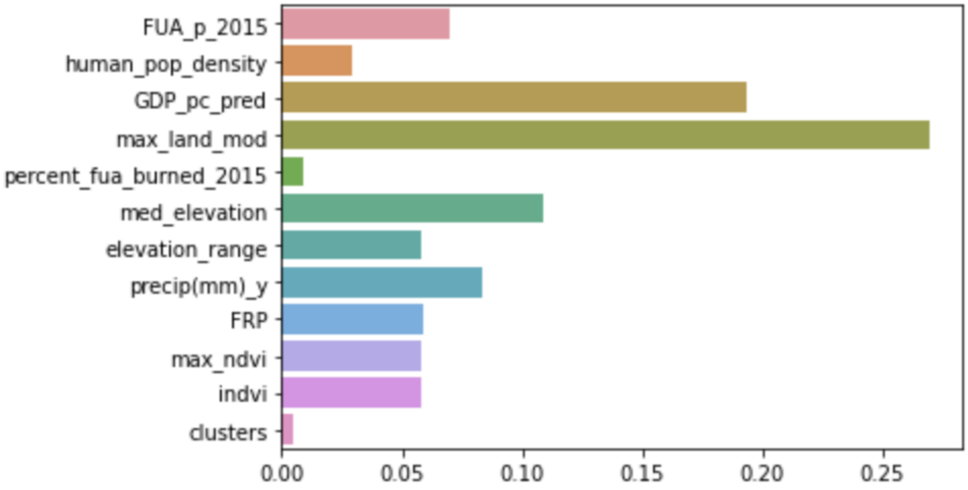


Figure 5-6: Hybrid feature importance



learning. Similar to the previous models, our hybrid model ranked the maximum land modification as the most important feature in predicting change in avian biodiversity. GDP per capita is the next most important feature, followed by median elevation, precipitation, and population respectively. After the top 5 features, the hybrid model diverges from the random forest a bit, and it ranks evevation range, FRP, maximum ndvi, and indvi almost identically.

Notably, the clusters input from the unsupervised learning portion of the model is ranked last, implying that the clusters were not helpful in predicting the change in avian diversity. This explains why the RF and hybrid model achieved almost identical performance metrics: the unsupervised portion of the model was simply not helpful for this task.

Chapter 6

Conclusions

6.1 Discussion

The results in the previous section suggest that machine learning methods are useful in predicting avian diversity loss. In MAE, the ML methods performed similar to linear regression, and they outperformed in MSE and RMSE. In addition to the improved performance, the machine learning algorithms offered a level of explainability that a linear regression model could not. While the linear regression's coefficients were potentially confounded by the lack of independence in our input variables, the random forest and hybrid models offered a more helpful view of the relationship between the socio-economic and environmental features of a FUA and its change in avian biodiversity.

The most important feature identified by all three models was maximum land modification, an estimate of human modification of land comprised of 5 input variables: human settlement, agriculture, transportation, mining and energy production, and electrical infrastructure. This finding indicate that land use change is the most significant driver of biodiversity loss, an idea that has long been posited by the ecological research community.[1] The confirmation of this assumption is helpful in identifying potential interventions. For example, because land modification is extremely helpful in predicting avian biodiversity loss, while human population density is less so, it may

[1] Sala et al., "Global biodiversity scenarios for the year 2100."

be helpful to encourage cities to encourage building up instead of out. Taller buildings encourage greater population density, while minimizing the land that needs to be developed in order to house the population. Furthermore, this finding helps explain previous research that identified a positive correlation between human population density and bird species richness, as in Luck (2007).[2]

Notably, the random forest and hybrid model both found that GDP per capita was the second most significant driver of avian biodiversity change. This finding is particularly significant because, to our knowledge, this research is the first to find a relationship between GDP per capita and biodiversity change on a global scale. Though it has been shown to hold true in small, local studies, the consequences of a global correlation are much more informative. Hope et al.(2008) found that income was extremely correlated with biodiversity in the Phoenix area, coining the term the "luxury effect", but they called for more research into the connection.[3] By examining the relationship between GDP per capita and urban biodiversity in a group of 54 cities, we have confirmed that individual income does influence avian biodiversity. This finding suggests that urban environmental interventions cannot be designed in a vacuum: they have to take into account the socio-economic conditions of an urban area as well.

Finally, our findings show that it is possible to accurately predict avian biodiversity in areas where expensive, tedious survey data may not be available. By training machine learning models and evaluating them on unseen data, we demonstrated their effectiveness in generalizing to new datapoints across diverse geographies. The ability to predict avian biodiversity loss is extremely important for stakeholders to be able to plan interventions and resource allocations. For example, if a country wanted to address biodiversity loss in its urban areas, it could use the feature set and model presented in this research to predict in which cities avian biodiversity loss would be most severe in order to respond most efficiently. We hope for this research to be helpful for urban planners and policy makers in mitigating biodiversity loss in urban

[2] Gary W Luck, "A review of the relationships between human population density and biodiversity," *Biological Reviews* 82, no. 4 (2007): 607–645.

[3] Hope et al., "Socioeconomics drive urban plant diversity."

areas, where many biodiversity hot spots are severely at risk.

6.2 Future research

We hope our work can inform future research regarding biodiversity in urban areas. A key area of future research is to expand the ML approach presented in this work to other species. Though we began with avian biodiversity, we believe the feature set we produced from satellite data should be applicable to any species. Future work should examine the relationship between these features and plant, terrestrial, aquatic and other forms of biodiversity. Though birds have been shown to be an indicator species for biodiversity loss, there is still a lot of work to be done to understand the complex underlying relationships between cities and urban biodiversity.

Furthermore, new features could easily be added and combined with our feature set to better advise urban biodiversity interventions. For example, invasive species are considered to be a leading threat to biodiversity.[4] By including invasive species in our feature set, future researchers could expand on our approach and continue to develop a deeper understanding of the drivers of biodiversity loss.

There is also ample opportunity to incorporate our findings and models into existing biodiversity research. For example, Powers and Jetz (2019) used global decadal land-use projections to year 2070 to examine suitable habitat loss, but their research could be expanded upon by incorporating the land use change projections into our trained models to create a data driven simulation of where avian biodiversity loss may be most severe in year 2070.[5] There are many ways to expand upon our research to increase understanding of biodiversity loss and how to combat it.

[4] Jennifer L Molnar et al., "Assessing the global threat of invasive species to marine biodiversity," *Frontiers in Ecology and the Environment* 6, no. 9 (2008): 485–492.

[5] Powers and Jetz, "Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios."

Bibliography

Alexopoulos, Evangelos C. “Introduction to multivariate regression analysis.” *Hippokratia* 14, no. Suppl 1 (2010): 23.

Aronson, Myla FJ, Frank A La Sorte, Charles H Nilon, Madhusudan Katti, Mark A Goddard, Christopher A Lepczyk, Paige S Warren, Nicholas SG Williams, Sarel Cilliers, Bruce Clarkson, et al. “A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers.” *Proceedings of the royal society B: biological sciences* 281, no. 1780 (2014): 20133330.

Basten, Stuart. “Re-Examining the fertility assumptions for Pacific Asia in the UN’s 2010 World Population Prospects.” *University of Oxford Department of Social Policy and Intervention, Barnett Papers in Social Research* 1 (2013).

Bayat, Mahmoud, Harold Burkhart, Manouchehr Namiranian, Seyedeh Kosar Hamidi, Sahar Heidari, and Majid Hassani. “Assessing biotic and abiotic effects on biodiversity index using machine learning.” *Forests* 12, no. 4 (2021): 461.

Biau, Gérard, and Erwan Scornet. “A random forest guided tour.” *Test* 25, no. 2 (2016): 197–227.

Bigirimana, Joseph, Jan Bogaert, Charles De Cannière, Marie-José Bigendako, and Ingrid Parmentier. “Domestic garden plant diversity in Bujumbura, Burundi: Role of the socio-economical status of the neighborhood and alien species invasion risk.” *Landscape and Urban Planning* 107, no. 2 (2012): 118–126.

Bose, Indranil, and Xi Chen. “Hybrid models using unsupervised clustering for prediction of customer churn.” *Journal of Organizational Computing and Electronic Commerce* 19, no. 2 (2009): 133–151.

Bradley, Paul S, and Usama M Fayyad. “Refining initial points for k-means clustering.” In *ICML*, 98:91–99. Citeseer, 1998.

Brei, Michael, Agustin Perez-Barahona, and Eric Strobl. “Environmental pollution and biodiversity: Light pollution and sea turtles in the Caribbean.” *Journal of Environmental Economics and Management* 77 (2016): 95–116.

Breiman, Leo. “Random forests.” *Machine learning* 45, no. 1 (2001): 5–32.

Christin, Sylvain, Éric Hervet, and Nicolas Lecomte. “Applications for deep learning in ecology.” *Methods in Ecology and Evolution* 10, no. 10 (2019): 1632–1644.

Cincotta, Richard P, Jennifer Wisnewski, and Robert Engelman. “Human population in the biodiversity hotspots.” *Nature* 404, no. 6781 (2000): 990–992.

Clarke, Lorraine Weller, G Darrel Jenerette, and Antonio Davila. “The luxury of vegetation and the legacy of tree biodiversity in Los Angeles, CA.” *Landscape and Urban Planning* 116 (2013): 48–59.

Danielson, Jeffrey J, and Dean B Gesch. *Global multi-resolution terrain elevation data 2010 (GMTED2010).* US Department of the Interior, US Geological Survey Washington, DC, USA, 2011.

Diaz, Sandra, Josef Settele, Eduardo S Brondizio, Hien T Ngo, John Agard, Almut Arneth, Patricia Balvanera, Kate A Brauman, Stuart HM Butchart, Kai MA Chan, et al. “Pervasive human-driven decline of life on Earth points to the need for transformative change.” *Science* 366, no. 6471 (2019): eaax3100.

Dijkstra, Lewis, Hugo Poelman, and Paolo Veneri. “The EU-OECD definition of a functional urban area,” 2019.

Elvidge, C. D., K. E. Baugh, S. J. Anderson, P. C. Sutton, and T. Ghosh. "The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data." *Social Geography* 7, no. 1 (2012): 23–35. https://doi.org/10.5194/sg-7-23-2012. http://www.soc-geogr.net/7/23/2012/.

Fernández, John E, and Marcela Angel. "Ecological city-states in an era of environmental disaster: Security, climate change and biodiversity." *Sustainability* 12, no. 14 (2020): 5532.

Fraixedas, Sara, Andreas Lindén, Markus Piha, Mar Cabeza, Richard Gregory, and Aleksi Lehikoinen. "A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions." *Ecological Indicators* 118 (2020): 106728.

Garcia, Letıcia Couto, Judit K Szabo, Fabio de Oliveira Roque, Alexandre de Matos Martins Pereira, Catia Nunes da Cunha, Geraldo Alves Damasceno-Júnior, Ronaldo Gonçalves Morato, Walfrido Moraes Tomas, Renata Libonati, and Danilo Bandini Ribeiro. "Record-breaking wildfires in the world's largest continuous tropical wetland: integrative fire management is urgently needed for both biodiversity and humans." *Journal of environmental management* 293 (2021): 112870.

Gaston, Kevin J. "Global patterns in biodiversity." *Nature* 405, no. 6783 (2000): 220–227.

Giglio, Louis, Luigi Boschetti, David Roy, Anja A Hoffmann, and Michael Humber. "Collection 6 modis burned area product user's guide version 1.3." *NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA*, 2020.

Current land modification. http://gdra-tnc.org/current/.

Hino, Miyuki, Elinor Benami, and Nina Brooks. "Machine learning for environmental monitoring." *Nature Sustainability* 1, no. 10 (2018): 583–588.

Hollington, Andrea, Oliver Tappe, Tijo Salverda, and Tobias Schwarz. “Introduction: Concepts of the global south.” *Voices from around the World* 1 (2015).

Hoornweg, Daniel, and Kevin Pope. “Population predictions for the world’s largest cities in the 21st century.” *Environment and Urbanization* 29, no. 1 (2017): 195–216.

Hope, Diane, Corinna Gries, Weixing Zhu, William F. Fagan, Charles L. Redman, Nancy B. Grimm, Amy L. Nelson, Chris Martin, and Ann Kinzig. “Socioeconomics drive urban plant diversity.” *Proceedings of the National Academy of Sciences* 100, no. 15 (2008): 8788–8792. ISSN: 0027-8424. https://doi.org/10.1073/pnas.1537557100. eprint: https://www.pnas.org/content/100/15/8788.full.pdf. https://www.pnas.org/content/100/15/8788.

International, BirdLife, and Handbook of the Birds of the World. *Bird species distribution maps of the world.*, 2021. http://datazone.birdlife.org/species/requestdis..

Jordahl, Kelsey, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, et al. *geopandas/geopandas: v0.8.1.* V. v0.8.1, July 2020. https://doi.org/10.5281/zenodo.3946761. https://doi.org/10.5281/zenodo.3946761.

Kennedy, Christina M, James R Oakleaf, David M Theobald, Sharon Baruch-Mordo, and Joseph Kiesecker. “Managing the middle: A shift in conservation priorities based on the global human modification gradient.” *Global Change Biology* 25, no. 3 (2019): 811–826.

Knapp, Sonja, Myla FJ Aronson, Ela Carpenter, Adriana Herrera-Montes, Kirsten Jung, D Johan Kotze, Frank A La Sorte, Christopher A Lepczyk, Ian MacGregor-Fors, J Scott MacIvor, et al. “A research agenda for urban biodiversity in the global extinction crisis.” *BioScience* 71, no. 3 (2021): 268–279.

Larigauderie, Anne, and Harold A Mooney. "The Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services: moving a step closer to an IPCC-like mechanism for biodiversity." *Current opinion in environmental sustainability* 2, nos. 1-2 (2010): 9–14.

Leong, Misha, Robert R Dunn, and Michelle D Trautwein. "Biodiversity and socioeconomics in the city: a review of the luxury effect." *Biology Letters* 14, no. 5 (2018): 20180082.

Liang, Shunlin. *Quantitative remote sensing of land surfaces.* John Wiley & Sons, 2005.

Loss, Scott R, Marilyn O Ruiz, and Jeffrey D Brawn. "Relationships between avian diversity, neighborhood age, income, and environmental characteristics of an urban landscape." *Biological Conservation* 142, no. 11 (2009): 2578–2585.

Luck, Gary W. "A review of the relationships between human population density and biodiversity." *Biological Reviews* 82, no. 4 (2007): 607–645.

Lyytimäki, Jari. "Nature's nocturnal services: Light pollution as a non-recognised challenge for ecosystem services research and management." *Ecosystem Services* 3 (2013): e44–e48.

Maillard, O, SK Herzog, RW Soria-Auza, and R Vides-Almonacid. *Impact of Fires on Key Biodiversity Areas (KBAs) and Priority Bird Species for Conservation in Bolivia. Fire 2022, 5, 4*, 2022.

McCain, Christy M, and Robert K Colwell. "Assessing the threat to montane biodiversity from discordant shifts in temperature and precipitation in a changing climate." *Ecology letters* 14, no. 12 (2011): 1236–1245.

McKinney, Wes. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. 2010. https://doi.org/10.25080/Majora-92bf1922-00a.

MOD11C2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 0.05Deg CMG V006., 2015. https://doi.org/10.5067/MODIS/MOD11C2.006.

Molnar, Jennifer L, Rebecca L Gamboa, Carmen Revenga, and Mark D Spalding. "Assessing the global threat of invasive species to marine biodiversity." *Frontiers in Ecology and the Environment* 6, no. 9 (2008): 485–492.

Myers, Leann, and Maria J Sirois. "Spearman correlation coefficients, differences between." *Encyclopedia of statistical sciences* 12 (2004).

Myles, Anthony J, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. "An introduction to decision tree modeling." *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, no. 6 (2004): 275–285.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825–2830.

Pettorelli, Nathalie, Jon Olav Vik, Atle Mysterud, Jean-Michel Gaillard, Compton J Tucker, and Nils Chr Stenseth. "Using the satellite-derived NDVI to assess ecological responses to environmental change." *Trends in ecology and evolution* 20, no. 9 (2005): 503–510.

Powers, Ryan P, and Walter Jetz. "Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios." *Nature Climate Change* 9, no. 4 (2019): 323–329.

Rajamohamed, R, and J Manokaran. "Improved credit card churn prediction based on rough clustering and supervised learning techniques." *Cluster Computing* 21, no. 1 (2018): 65–77.

Ritchie, Hannah, and Max Roser. "Urbanization." Https://ourworldindata.org/urbanization, *Our World in Data*, 2018.

Rizwan, Ahmed Memon, Leung YC Dennis, and LIU Chunho. “A review on the generation, determination and mitigation of Urban Heat Island.” *Journal of environmental sciences* 20, no. 1 (2008): 120–128.

Sala, Osvaldo E, FIII Stuart Chapin, Juan J Armesto, Eric Berlow, Janine Bloomfield, Rodolfo Dirzo, Elisabeth Huber-Sanwald, Laura F Huenneke, Robert B Jackson, Ann Kinzig, et al. “Global biodiversity scenarios for the year 2100.” *science* 287, no. 5459 (2000): 1770–1774.

Schiavina, M, A Moreno-Monroy, L Maffenini, P Veneri, and Paolo. “GHS-FUA R2019AGHS functional urban areas, derived from GHS-UCDB R2019A,(2015).” *R2019A. edited by Joint Research Centre (JRC) European Commission*, 2019.

Seto, Karen C, Burak Güneralp, and Lucy R Hutyra. “Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools.” *Proceedings of the National Academy of Sciences* 109, no. 40 (2012): 16083–16088.

Silveira, Juliana M, Julio Louzada, Jos Barlow, Rafael Andrade, Luiz Mestre, Ricardo Solar, Sébastien Lacau, and Mark A Cochrane. “A multi-taxa assessment of biodiversity change after single and recurrent wildfires in a Brazilian Amazon forest.” *Biotropica* 48, no. 2 (2016): 170–180.

Sun, Alexander Y, and Bridget R Scanlon. “How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions.” *Environmental Research Letters* 14, no. 7 (2019): 073001.

team, The pandas development. *pandas-dev/pandas: Pandas.* V. latest, February 2020. https://doi.org/10.5281/zenodo.3509134. https://doi.org/10.5281/zenodo.3509134.

Tran, Duy X, Filiberto Pla, Pedro Latorre-Carmona, Soe W Myint, Mario Caetano, and Hoan V Kieu. “Characterizing the relationship between land use land cover change and land surface temperature.” *ISPRS Journal of Photogrammetry and Remote Sensing* 124 (2017): 119–132.

Tuia, Devis, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, et al. “Perspectives in machine learning for wildlife conservation.” *Nature Communications* 13, no. 1 (February 2022). https://doi.org/10.1038/s41467-022-27980-y. https://doi.org/10.1038%2Fs41467-022-27980-y.

Unit, Biosafety. *Aichi Biodiversity targets*, September 2020. https://www.cbd.int/sp/targets/.

Vaglio Laurin, Gaia, Jonathan Cheung-Wai Chan, Qi Chen, Jeremy A Lindsell, David A Coomes, Leila Guerriero, Fabio Del Frate, Franco Miglietta, and Riccardo Valentini. “Biodiversity mapping in a tropical West African forest with airborne hyperspectral data.” *PloS one* 9, no. 6 (2014): e97910.

Vermote, Eric, and NOAA CDR Program. *NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5*, 2019. https://doi.org/10.7289/V5ZG6QH9.

Wang, Hua-Feng, Salman Qureshi, Sonja Knapp, Cynthia Ross Friedman, and Klaus Hubacek. “A basic assessment of residential plant diversity and its ecosystem services and disservices in Beijing, China.” *Applied Geography* 64 (2015): 121–131.

Weier, John, and David Herring. *Measuring Vegetation (NDVI and EVI).* https://earthobservatory.nasa.gov/features/MeasuringVegetation.

Wentz, F. J., K. A. Hilburn, and D K. Smith. *Remote Sensing Systems DMSP SSMI / SSMIS Environmental Suite on 0.25 deg grid, Version 7*, 2012. https://www.remss.com/missions/ssmi/.

West, Paige, James Igoe, and Dan Brockington. "Parks and peoples: the social impact of protected areas." *Annu. Rev. Anthropol.* 35 (2006): 251–277.

Xiao, Jin, Yuhang Tian, Ling Xie, Xiaoyi Jiang, and Jing Huang. "A hybrid classification framework based on clustering." *IEEE Transactions on Industrial Informatics* 16, no. 4 (2019): 2177–2188.

Zou, Kelly H, Kemal Tuncali, and Stuart G Silverman. "Correlation and simple linear regression." *Radiology* 227, no. 3 (2003): 617–628.