MSSP Portfolio

Jingjian Gao

May 2023

Department of Statistics
Boston University

Table of Contents

Butterfly Batesian Mimicry

Introduction

In the summer of 2022, a project was designed to investigate the idea that birds can remember the appearance of chemically defended prey species and their mimics over extended periods by our client. The experiment aimed to train birds to recognize the toxic butterfly Battus and assess how long they could memorize it. The objective was to examine how time affects the way birds learn aposematic signals and generalize them to palatable mimics. The study took place at Quabbin Reservoir in Massachusetts and consisted of four experiments in distinct geographical areas.

Data and Methods

The study involved the creation of 7500 fake butterfly models, with 3000 used for learning and 4500 for testing. The models belonged to three different groups of butterflies, Battus, Limenitis, and Junonia (control group). The experimenters placed the fake butterfly models on large foliage to simulate resting butterflies in the natural environment.

In the first experiment, the "model" and the "control" were set out for four days, and evidence of bird attacks was recorded on a daily basis before being taken down. After a four-week learning gap, the "mimic" and the "control" were set out, and evidence of bird attacks was recorded on a daily basis for four days. Attack marks were easily distinguishable since the facsimiles were made from clay. All data was collected between 9 am and noon since most attacks happen in the morning from 6-9 am.

The experiment was repeated at other locations, but with different time gaps of two weeks, one week, and simultaneously. For the simultaneous experiment, all three types of butterfly facsimiles were presented at the same time. A total of 30,000 observations were collected.

Results and Conclusion

Unfortunately, the client decided to work on this project independently before we delivered our exploratory data analysis (EDA) report, and no actual results were produced. However, the data collected could provide valuable insights into how birds learn and retain information about chemically defended prey species and their mimics over long periods of time.

BMC CT Scanning Follow up

Introduction

Our client's project focuses on analyzing the data from the lung cancer screening program at Boston Medical Center. We have 267 patient records and will be using the Lung-RADS score, a radiologic grading system, to determine how urgently someone should get repeat imaging. Rads-3 has a 1-2% chance of lung cancer, Rads-4a has a 5-15% chance, and Rads-4b has a greater than 15% chance.

Our goal is to investigate factors associated with obtaining follow-up CT scans of Lung Rads 3-4 findings and whether reminder letters sent by lung cancer screening programs increase follow-up or timely follow-up scans. However, our client has not yet provided additional information to help us achieve these goals.

Data and Methods

The data includes demographic information such as race, ethnicity, gender, language, insurance status, and smoking history, as well as risk categories. The data was recorded between May 2019 and May 2020 and may have been influenced by COVID-19. As we have a limited amount of data, we have several concerns for our client.

We asked our client to determine how early they send out the reminder letters after the first screening, whether COVID-19 will have an impact on the data collected, whether BMC or patients' insurance will cover the CT scan, and if there is any confounding in the data set, such as demographic information.

Results

We produced an EDA report for our client, which includes bar plots and tables to visualize our findings. Based on our graphs, we found associations between some of the factors in the data and the number of follow-up checks. Smoking history was a significant factor to consider. We also produced a Ridge graph to estimate that the reminder letter may be sent around day 550.

Conclusion

While we await additional information from our client, our preliminary analysis shows that smoking history is an important factor to consider when determining follow-up CT scans. We recommend further investigation into the impact of COVID-19 on the data collected and the coverage of the CT scan by BMC or patients' insurance. We also suggest examining potential confounding variables, such as demographic information, that may affect the results.

Children’s modality of communication

Introduction

Our client is working on several models to investigate the interaction between children and parents.
Model 1 explores the relationship between children's mode of communication (gesture, speech, gesture-speech combination, AAC) and parent responsivity (contingent, non-contingent, no response).
Model 2 examines how the precision of communication (precise, imprecise) affects parent responses.
Model 3 focuses on multiple forms of communication (reach point gesture, intelligible speech, etc.) and their impact on parent responses.
Finally, Model 4 looks at the relationship between children's mode of communication and parent responsivity based on the specific mode used (speech, gesture, gesture-speech, AAC, AAC+speech, no response).

Data and Methods

The study participants included 47 minimally verbal autistic children (10 females; 48-95 months) who produced 20 or fewer different spontaneous words in a 15-minute sample, according to the definition by Butler et al. (2021). The study was conducted remotely between December 2020 and November 2021, during the coronavirus pandemic (Ritchie et al., 2020). Participants were recruited through social media advertising and the Simons.

All the data collected is count data, such as how often specific modes of communication were used and how often parents responded to them within a 15-minute session. Our client approached us for advice on selecting the correct analysis approach for her data. During the intake meeting, we provided some initial guidance to her.

Results

Unfortunately, due to our schedule, we were not available to provide further advice to our client on her analysis after April. We regret that we could not offer more assistance, but we wish her the best of luck with her research.

Genetic Counseling Safety-Net vs Non-Safety-Net

Introduction

Our client's project aimed to investigate carrier screening practices in various settings by collecting data from genetic counselors through survey responses. Carrier screening guidelines and clinical practices were not standardized, with factors such as the person ordering the testing, panel size, methodology, and other variables varying among different medical centers.

Our client's project explored this variability by collecting information from genetic counselors through surveys to gain insights into carrier screening practices in different settings. The primary objective of this project was to compare responses from genetic counselors working in Safety Net and non-Safety Net hospitals across multiple variables related to genetic counselors and carrier screening.

Data and Methods

Data was collected from the survey and entered into an Excel sheet (which we did not have access to). The data cleaning process involved handling missing data and renaming columns. The survey included nine questions from our client.

We used plots to contrast different proportions of Safety Net and non-Safety Net hospitals across the variables to answer the research questions and made additional graphs to investigate multiple variables. To determine statistical significance, we used the chi-squared test and logistic regression to compare the proportion of Safety Net and non-Safety Net hospitals across different factors.

Results

We analyzed the proportions of Safety Net and non-Safety Net hospitals across different variables. After conducting statistical tests and analyzing the plots, we found no significant differences between Safety Net and non-Safety Net hospitals in terms of demographics and variables related to carrier screening practices.

Conclusion

Overall, our analysis suggested that there were no substantial differences between Safety Net and non-Safety Net hospitals in terms of carrier screening practices. Our findings may contribute to the development of standardized guidelines for carrier screening practices in medical centers.

Chloris Geospatial

Introduction

Chloris Geospatial provides customers with insights and analytics to support the transition towards a net-zero and nature-positive economy. As part of this effort, our team developed a simple, intuitive R Shiny application that enables visualization of changes in forests around the world, contributing to the reduction of greenhouse gas emissions as outlined in the Paris Agreement.

Data and Methods

We utilized open-source data from the U.S. Geological Survey (USGS) and specifically focused on the Landsat 8 satellite, which contains the OLI (operational land imager) and TIRS (thermal infrared sensor) sensors. To store the data, we set up a temporary cloud server.

We gathered carbon emission data from USGS and Google sources and used R functions, such as number2binary, getrcl, and LS_band_graph, to create custom queries that effectively classify each pixel in an image based on its corresponding integer value, representing natural conditions. These queries allowed us to develop a highly versatile method that can be applied to various locations globally, providing clients with valuable insights into carbon emission patterns.

Results

Our R Shiny application allows users to select any combination of start date, band number, and mask conditions, resulting in customizable and dynamic data visualizations. Clients can now effectively monitor and measure their investments in nature-based solutions and accelerate their efforts to achieve their sustainability goals at scale.

Conclusion

By leveraging the power of open-source data and R programming, our team successfully developed an intuitive, customizable, and versatile application that empowers clients worldwide to take effective action towards reducing greenhouse gas emissions. Our method can be applied to multiple locations around the world, allowing for accurate monitoring and measurement of carbon emissions, supporting the development of a net-zero and nature-positive economy.

BPS CHESS Project

Introduction

The CHESS project aimed to improve indoor air quality and school sustainability by analyzing environmental monitoring data from Boston Public Schools (BPS). While many schools have increased their monitoring efforts, there is a challenge in transparently communicating the data while balancing indoor air quality, energy management goals, and the transmission of SARS-CoV-2.

Our project focused on analyzing indoor air quality data and identifying patterns and trends in CO2, CO, PM2.5, temperature, humidity, and other variables.

Data and Methods

The project utilized approximately 4500 CSV files obtained from BPS. We created a database and cleaned the data, identifying patterns of problematic sensors through exploratory data analysis in R. We then visualized the data and presented our findings to the client.

Results

After conducting our analysis, we identified certain patterns in CO2, CO, and other variables. Unfortunately, due to the non-disclosure agreement we signed, we are unable to share our results in detail.

Conclusion

Our findings will assist BPS in identifying potential issues that cause poor indoor air quality in their schools. By improving indoor air quality and sustainability, we hope to create a healthier and more sustainable environment for students and faculty. The CHESS project is an important step towards achieving this goal.