SSA 200
Strategic Use of Data

Lecture slides, activities, and additional supplementary materials are available online at:
ssa200.auburn.edu

What is a model?

The purpose of modeling

- Statistical analysis of data
- Use statistical analysis to predict the future
- Explaining variation
- Using data analysis to understand ecological processes
- Predict patterns in the future
- Evaluate competing hypothesis about how the system works

Conor P. McGowan
Anna M. Tucker
Nicole F. Angeli
Kylee Dunham

U.S. FISH & WILDLIFE SERVICE
DEPARTMENT OF THE INTERIOR

alabama cooperative fish and wildlife research unit

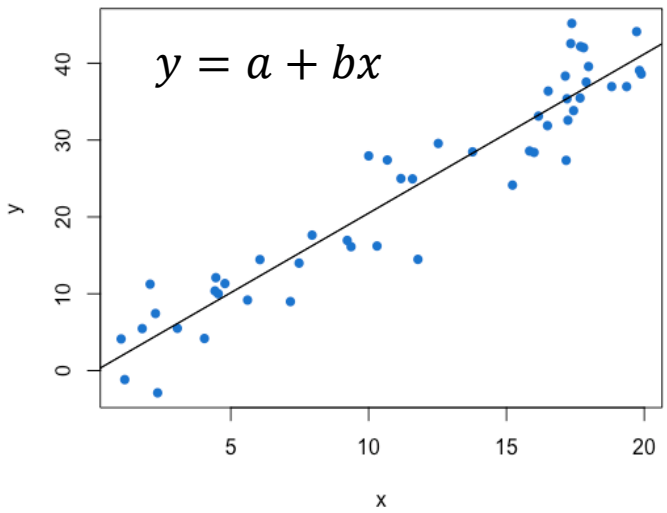
Statistical distributions

Name	Continuous or Discrete	Bounds	Common applications	Shape	Notes
Normal	Continuous	-∞, ∞	Linear regression		
Binomial	Discrete	0 or 1	Occupancy Survival		
Multinomial	Discrete	0, ∞	State transitions		
Poisson	Discrete	0, ∞	Count data		
Negative Binomial	Discrete	0, ∞	Counts with many zeros		
Log-normal	Continuous	0, ∞	Population-level productivity (projections)		
Beta	Continuous	0, 1	Population rates (projections)		
Uniform	Continuous	User-defined	Variety of applications (projections)		

Linear regression and AIC

General linear model – response variable (y) has a Normal distribution

Generalized linear model – response variable (y) has some other distribution
- **Logistic** regression – Binomial distribution
- **Poisson** regression – Poisson distribution



Parameter	Estimate	SE	*t*	p-value
Intercept	-0.22	1.29	-0.169	0.866
b	2.07	0.0978	21.16	< 0.0001

Model	AIC	ΔAIC	Np	w_i
Int + Covariate 1 + Covariate 3	345.8	0	3	0.82
Int + Covariate 1 + Covariate 2 + Covariate 3	349.1	3.1	4	0.18
Int	359.8	14.0	2	0
Int + Covariate 2	361.1	15.3	1	0

Types of uncertainty

Partial controllability – We are unable to control the exact management actions taken in a system.

Examples:
- Setting management goals – we may intend to fully restore a habitat, but may not be able to implement the exact management goals due to other logistical constraints

Observational uncertainty – We are unable to perfectly observe the state of natural systems.

Examples:
- Count data – in almost all cases, we cannot count every individual present at a specific location, but instead assume there is some probability of detecting individuals

Environmental variation – Stochastic environmental fluctuations mean that conditions typically vary randomly from year to year.

Examples:
- Predicting effects of temperature – we may estimate a relationship between temperature and survival probability that we can use to predict survival under future temperature conditions, but temperature will likely vary in a stochastic way from year to year.

Ecological uncertainty – We have an imperfect understanding of how ecological systems work.

Examples:
- Metapopulation dynamics – we think a set of populations function as a metapopulation, but have not conducted studies to explicitly estimate immigration among sites, and therefore we are unsure to what extent immigration plays a role in measured population growth rate at each site.

Some key terms

Response/dependent variable – in a statistical model, the variable that you are interested in better understanding or predicting (the "y" variable)

Predictor/independent variable – in a statistical model, the variable(s) that explain some of the observed variation in the response variable (the "x" variables)

Covariate – an environmental or ecological quantity that usually represents a stressor or species need and is included in a model as a predictor variable

Parameter – statistical quantities that are estimated to explain the relationship between predictor and response variables. Can also be used to refer to demographic vital rates of interest

Collinearity – occurs when two predictor variables in the same model are correlated with each other

Overfitting – occurs when too many predictor variables are included in the model, resulting in a model that is not very useful for prediction

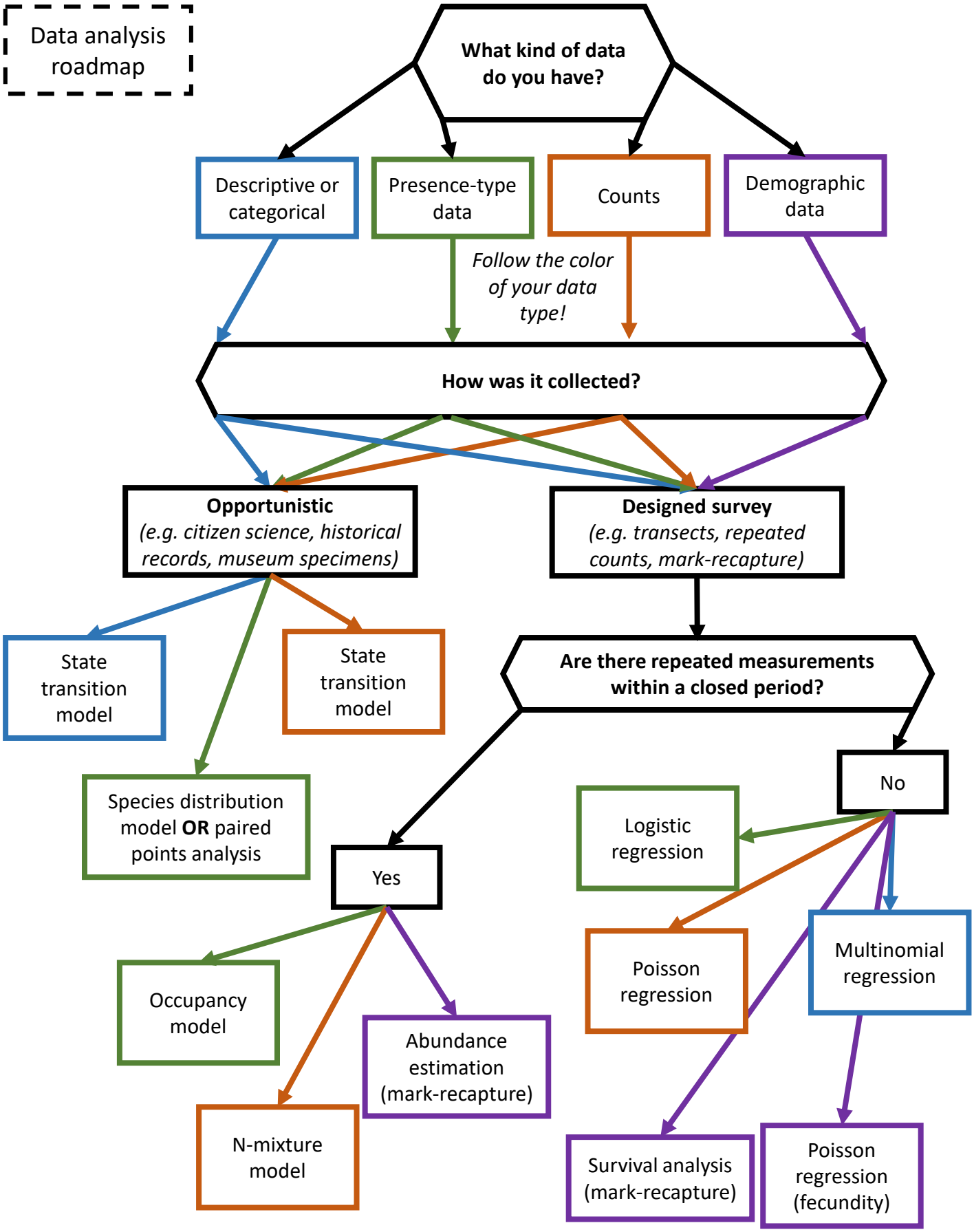
AIC – stands for Aikaike's Information Criterion – a metric used to rank models based on how well they fit the data with a penalty for the number of covariates in the model (to avoid overfitting)

Intercept – the theoretical value of the response variable if all predictors were equal to zero

Null model – the "intercept-only" model that does not include any covariates

Global model – the most complex model in the model set that includes all covariates

Population closure – an important concept for occupancy and abundance estimation, a population is considered "closed" when there are no births, deaths, immigration, or emigration



This roadmap is to serve as a general guide and is not an exhaustive list of all analysis options. Also, ***always check the specific assumptions of your planned modeling approach!***

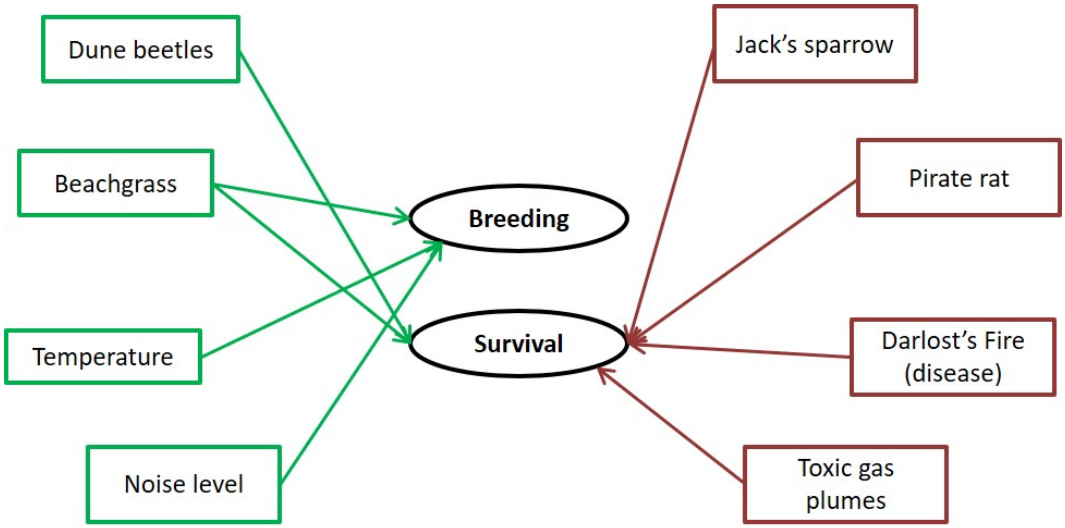
SSA 200: Strategic Use of Data
Activity 1 – Linking Conceptual Diagrams to Statistical Models

Objectives:

- Understand the link between conceptual diagrams and statistical models
- Develop ecological hypotheses about species needs and translate hypotheses into biologically-meaningful models
- Choose appropriate ecological variables from available data sources to include in models

Background Information:

The Island Mouse (*Zapus islandsonious*) is a species that lives on Darlost's Island. There are 10 current populations, although historically they occurred at 11 sites across the island. As you covered in SSA 100, the mice are adapted to the unique ecology of Darlost's Island. Their primary food source is dune beetles, which are found across the small island. They also rely on beach grass that grows along the coasts to build their nests and create shelter from the sun and cold winters. They require warm spring temperatures for successful breeding, and in years with exceptionally cold trade winds there is increased winter mortality. They are also very sensitive to noise, and loud noises can lead to depressed breeding activity and, in extreme cases, death from cardiac arrest. There are also several stressors on the island that influence their population dynamics. The two main predators of Island mice are Jack's sparrow, which are found across the island, and the pirate rat, which are only found along the coasts. Darlost's fire is a disease that causes mange and can be fatal, and toxic gas plumes from the island's volcanoes can also be deadly. This conceptual diagram summarizes the key species needs (green boxes) and threats (red boxes) and how they influence breeding and survival.



Using this conceptual model and what you know about the ecology of the mice, you are tasked with developing models that will inform how these factors influence Island mouse occurrence, abundance, and breeding success across their range.

The goal of model selection is to find the model that best explains the data using the fewest number of parameters, which will hopefully identify the most important ecological drivers of species success. It is standard practice to always include a **null model** in your model set (this may sometimes also be referred

to as the “dot model”), which is a model that does not include any covariates. Sometimes none of the data we have are good predictors of the response variable of interest. Does that mean the models are useless? This result still tells us about the system, in that in can tell us that the things we thought were important drivers of species success maybe aren’t as important, and there could be other factors that we should consider.

When lots of different data sources are available, it can be tempting to fit “all subsets”, or all possible combinations of available covariates. This kitchen sink approach is not good practice for a few reasons. First, it can be time consuming and computationally intensive to fit all possible combinations. Second, and more importantly, it can result in the selection of models that might not make a lot of sense ecologically or be hard to use in future predictions. Finally, it’s just not good science because it usually results in trying to come up with theory to explain observations *post hoc*, instead of conducting meaningful tests of biological hypotheses.

It’s also important to carefully consider the combinations of covariates to include. We want to avoid **collinearity** in model sets, which happens when one or more covariates are correlated with each other. This can be an issue for estimation of the effects of each variable alone because they are not independent of each other. For example, if we are interested in predicting grizzly bear occurrence, we may want to test a model that includes the presence of a water source or the presence of fish at a given site. Since the presence of fish is highly correlated with the presence of water, those covariates could be collinear, and we may not want to include both in the same model. The presence of fish implies the presence of water, and may be closer to the real ecological reason why bears occur near water. However, if data about fish presence isn’t available, including water in the model can be a good proxy.

For each of the following scenarios, don’t worry about the mechanics or details of fitting the models, but focus instead on the combinations of covariates that can best explain the response variable of interest. Write your models both as a formula that lists all covariates (feel free to use your own abbreviations) and as a sentence that describes the ecological hypothesis that the model represents. We’ve included space for six models, but feel free to include fewer or more candidate models in your list. For example, consider the following set of models to predict grizzly bear occurrence:

	Model (formula)	Hypothesis (sentence)
1	water + trees + rocks	Bear occurrence is dependent on the presence of water source, forest cover, and rocky substrate.
2	water + trees	Bear occurrence is dependent on the presence of a water source and forest cover.
3	water	Bear occurrence depends primarily on the presence of a water source.
4	null model	Bear occurrence is not strongly associated with water source, forest cover, or rocky substrate.

Part 1: Occupancy

The Island Mouse Recovery Team is interested in mapping the drivers of occurrence across Darlost's island in order to assess the suitability of other nearby islands for possible translocations. They have collected occurrence data from a variety of sources and want to use species distribution modeling to find the environmental covariates that best predict occurrence. Using the following data sources, develop a set of candidate models to predict occurrence of Island mice.

Response variable = Probability of island mouse presence

Source	Data Type	Variables Measured
National Land Cover Database (NLCD)	Land cover/land use in 30km grid	*Each cell is classified as one of:* Open water Perennial Ice/Snow Developed, Open Developed, Low Developed, Medium Developed, High Rock/sand/clay Deciduous forest Evergreen forest Mixed forest Dwarf scrub Shrub/scrub Grassland Sedge Lichens Moss Pasture/Hay Cultivated crops Woody wetlands Emergent wetlands
WorldClim Bioclimatic variables	Climate data in 30km grid	Annual Mean Temperature Max Temperature of Warmest Month Min Temperature of Coldest Month Temperature Annual Range (Max – Min) Mean Temperature of Wettest Quarter Mean Temperature of Driest Quarter Mean Temperature of Warmest Quarter Mean Temperature of Coldest Quarter Annual Precipitation Precipitation of Wettest Month Precipitation of Driest Month Precipitation of Wettest Quarter Precipitation of Driest Quarter Precipitation of Warmest Quarter Precipitation of Coldest Quarter
NOAA Weather Station	Daily local weather data	Maximum air temperature

	from 3 weather stations across island and 1 weather buoy offshore	Minimum air temperature Average air temperature Maximum wind speed Minimum wind speed Average wind speed Total precipitation Maximum water temperature Minimum water temperature Average water temperature

Occupancy candidate model set

	Model (formula)	Hypothesis (sentence)
1		
2		
3		
4		
5		
6		

Part 2: Abundance

A graduate student has developed a project to estimate Island mouse abundance and map the drivers of abundance across its range. His field crew conducted transect surveys at randomly-selected points across Darlost's island to estimate the abundance of Island mice while accounting for detection probability. He and his crew conducted vegetation surveys at each transect point and also recorded the presence or absence of Jack's sparrows and pirate rats each survey. Use the following data to construct a set of models to predict mouse abundance.

Response variable = Site abundance

Source	Data Type	Variables Measured
Vegetation surveys	Vegetation data collected at each transect	Percent canopy cover Percent shrub cover Percent herbaceous cover Substrate: rock, soil, sand, other Soil type: clay, silt, loam, NA Beachgrass density Average DBH Ambient noise level (decibels) Toxic gas level
Predator surveys	Presence/absence of predators during each survey	Jack's sparrow presence/absence Pirate rat presence/absence
Other transect info	Measured via GIS	Distance to nearest coast Distance to nearest volcano Ecotype: coastal, palms, mountains
NOAA Weather Station	Daily local weather data from 3 weather stations across island and 1 weather buoy offshore	Maximum air temperature Minimum air temperature Average air temperature Maximum wind speed Minimum wind speed Average wind speed Total precipitation Maximum water temperature Minimum water temperature Average water temperature
NLCD	Land cover/land use in 30km grid	*Each cell is classified as one of:* Open water Perennial Ice/Snow Developed, Open Developed, Low Developed, Medium Developed, High Rock/sand/clay Deciduous forest Evergreen forest Mixed forest

		Dwarf scrub Shrub/scrub Grassland Sedge Lichens Moss Pasture/Hay Cultivated crops Woody wetlands Emergent wetlands

Abundance candidate model set

	Model (formula)	Hypothesis (sentence)
1		
2		
3		
4		
5		
6		

Part 3: Breeding success

A study is conducted to measure breeding success in the Beach Bums, Dead Man's Dunes, and Misty Mountain populations. Over the course of five years, mouse nests are surveyed throughout breeding to estimate the number of offspring produced per female. Pitfall traps are also placed at each study population to estimate invertebrate abundance. Use the following data sources to develop models to determine factors associated with the number of offspring per female.

Response variable = Number off offspring per female

Source	Data Type	Variables Measured
Nest surveys	Data collected at each nest	Ecotype: coastal, palms, mountains Toxic gas level at nest Nest substrate: beachgrass, sand, shrub, other Female age Female disease status (test positive/negative)
Pitfall traps	Index of invertebrate diversity and abundance collected for each site in each year	Dune beetle abundance Total insect abundance Number of species detected
WorldClim Bioclimatic variables	Climate data in 30km grid	Annual Mean Temperature Max Temperature of Warmest Month Min Temperature of Coldest Month Temperature Annual Range (BIO5-BIO6) Mean Temperature of Wettest Quarter Mean Temperature of Driest Quarter Mean Temperature of Warmest Quarter Mean Temperature of Coldest Quarter Annual Precipitation Precipitation of Wettest Month Precipitation of Driest Month Precipitation of Wettest Quarter Precipitation of Driest Quarter Precipitation of Warmest Quarter Precipitation of Coldest Quarter
NOAA Weather Station	Daily local weather data from 3 weather stations across island and 1 weather buoy offshore	Maximum air temperature Minimum air temperature Average air temperature Maximum wind speed Minimum wind speed Average wind speed Total precipitation Maximum water temperature Minimum water temperature Average water temperature
NLCD	Land cover/land use in 30km grid	*Each cell is classified as one of:* Open water

		Perennial Ice/Snow Developed, Open Developed, Low Developed, Medium Developed, High Rock/sand/clay Deciduous forest Evergreen forest Mixed forest Dwarf scrub Shrub/scrub Grassland Sedge Lichens Moss Pasture/Hay Cultivated crops Woody wetlands Emergent wetlands

Breeding success candidate model set

	Model (formula)	Hypothesis (sentence)
1		
2		
3		
4		
5		
6		

SSA 200: Strategic Use of Data
Activity 2 – Interpreting Analysis Outputs

Objectives:

- Understand the use of AIC model selection to choose the best model out of a set of candidate models
- Interpret regression coefficients in terms of the ecological relationships they represent
- Use model outputs to make predictions

Background Information:

In the previous activity, you developed candidate model sets that represented competing ecological hypotheses for the drivers of Island Mouse occurrence, abundance, and breeding success. Here we will examine outputs from those models and use them to draw inference about the key stressors and needs for this species.

We will compare the relative support for competing models using a metric called Aikaike's Information Criterion, or **AIC**. We won't go into too much detail about AIC here, but know that it is a measure of how well the model fits the data while accounting for model complexity—the lower the value of AIC, the better the model.

For most analyses that use this method to rank models, you will see a table like this somewhere in the Results (where X, Y, and Z represent some ecological covariates of interest):

Model	AIC	ΔAIC	Np	w_i
Intercept + X	684	0	2	0.98
Intercept	693	9	1	0.01
Intercept + X + Y	698	14	3	0.01
Intercept + X + Y + Z	710	26	4	0

The actual AIC score for each model is not as important as the **ΔAIC ("delta AIC")**, which is the difference in AIC values between each model and the best model. This number tells us how much worse this model is than the best one. The general rule of thumb is that if ΔAIC is greater than 2, then it's not a very good model. If ΔAIC is less than 2, then we are uncertain about which model is better. If ΔAIC of the second-ranked model is ≥ 2, we would say there was a single best model and interpret the coefficient from that model. If ΔAIC of the next-best models are < 2, then we would use model averaging to averaging the coefficients from all models while accounting for model weight. Model averaging is not something that we will cover here. The third column in this table, Np, is the number of parameters in the model. The last column here, w_i, is the model weight. This corresponds to ΔAIC and is another way of seeing how "good" each model is relative to the others. In this example, the top model received 98% of the model weight, so we would be fairly confident that it is a better model than the others.

It is very important to note that **AIC is a relative measure of support only**. It only tells us how good each model is relative to the other ones in the model set. No matter how crappy the models you run, one of them will always have the lowest AIC score. This does not necessarily mean that the model with the

lowest AIC score is a good fit for our data. There are a suite of **goodness-of-fit tests** that help us determine whether a given model fits the data, and can help us identify cases where our data do not meet the assumptions of the model we think we want to use. Frequently when AIC is used, we find the goodness-of-fit of the **global model**, which is the most complex model. We will not explore the world of goodness-of-fit testing in this course, but any time you see an AIC table, be sure to check that the authors used some test to ensure the models fit the data.

Part 1: AIC model selection

A study is conducted to measure breeding success in the Beach Bums, Dead Man's Dunes, and Misty Mountain populations. Over the course of five years, mouse nests are surveyed throughout breeding to estimate the number of offspring produced per female. Pitfall traps are also placed at each study population to estimate invertebrate abundance.

We used a Poisson generalized linear model (Poisson GLM) to estimate the number of offspring produced per female as a function of several potential covariates. We fit several different models and used AIC model selection to rank models. Goodness-of-fit testing of the global model indicated adequate model fit. The results of that process were:

Model	AIC	ΔAIC	Np	w_i
Int + Ecotype + Dune beetle + Avg temp.	830	0	4	0.575
Int + Ecotype + Dune beetle + Avg temp. + Min temp. coldest month	831.2	1.2	5	0.316
Int + Avg. temp + Min temp coldest month + Max wind speed	833.5	3.5	4	0.100
Int + Dune beetle + Percent rock/sand/clay	839.4	9.4	3	0.005
Int + Ecotype	840	10.0	2	0.004
Intercept only	844.2	14.2	1	0.000

1. **Why did we use a Poisson GLM? (Why not a normal linear regression?)**

2. **Did one model receive more support than all others? If so, which one? If not, explain your answer.**

3. **In a sentence or two, interpret the model weights for the top-ranked model(s).**

4. **What are the key ecological predictors of breeding success for Island Mice?**

Part 2: Regression coefficients

A graduate student developed a project to estimate Island mouse abundance and map the drivers of abundance across its range. His field crew conducted transect surveys at randomly-selected points across Darlost's island to estimate the abundance of Island mice while accounting for detection probability. He and his crew conducted vegetation surveys at each transect point and also recorded the presence or absence of Jack's sparrows and pirate rats each survey.

We used N-mixture models to estimate Island Mouse abundance at each site as a function of several ecological covariates. We used AIC model selection to rank models and the output is below.

Model	AIC	ΔAIC	Np	w_i
Int + Beachgrass + Noise + Avg. Temp	650	0	4	0.822
Int + Noise + Avg. Temp.	654.2	4.2	3	0.101
Int + Beachgrass + Noise + Distance to volcano	655.1	5.1	4	0.064
Int + Beachgrass + Noise + Avg. Temp + Distance to volcano + Jack's sparrow presence	658.6	8.6	6	0.011
Intercept only	662.9	12.9	1	0.001
Int + Distance to volcano + Jack's sparrow presence	665.8	15.8	3	0.000

1. **Why did we use N-mixture models?**

    a. **What type of sampling design do we have?**

    b. **What are the general assumptions of N-mixture models?**

2. **Which model received the most support? Support your answer using ΔAIC and model weight. What does that model say about the ecological drivers of abundance?**

This tells us that beach grass density, ambient noise level, and average air temperature are associated with abundance, but doesn't tell us the magnitude (weak/strong) or direction (positive/negative) of that relationship. For that, we need to look at the **coefficients** (also referred to as β, "beta") for those covariates in the model.

Covariate	β coefficient	SE	p-value
Intercept	-4.3	1.2	0.004
Beachgrass density	1.4	0.21	0.0021
Ambient noise level	-2.8	0.87	0.0001
Average air temperature	0.89	0.54	0.071

The table above describes the numerical relationships between each covariate and abundance of Island Mice. The β coefficient describes the magnitude (how large is the number?) and direction (is it positive or negative?) of the relationship. The standard error (SE) tells us how precise our estimate of this relationship is, which is often a function of how much data we have. This is a measure of our uncertainty in this estimate, and is used to calculate 95% confidence intervals. Finally, the p-value tells us whether this effect is statistically "significant", in other words whether or not the 95% confidence interval for the effect contains 0. If the p-value is < 0.05, then we typically say that it is a significant effect. The **intercept** (often denoted β_0) tells us about the expected condition if all covariates equaled zero. We usually don't draw inference from the value of the intercept alone.

3. **Describe the relationship between each covariate and abundance. Does it have a positive or negative effect? Is that effect statistically significant?**

Part 3: Using models to make predictions

The Island Mouse Recovery Team is interested in mapping the drivers of occurrence across Darlost's island in order to assess the suitability of other nearby islands for possible translocations. They have collected occurrence data from a variety of sources and want to use species distribution modeling to find the environmental covariates that best predict occurrence.

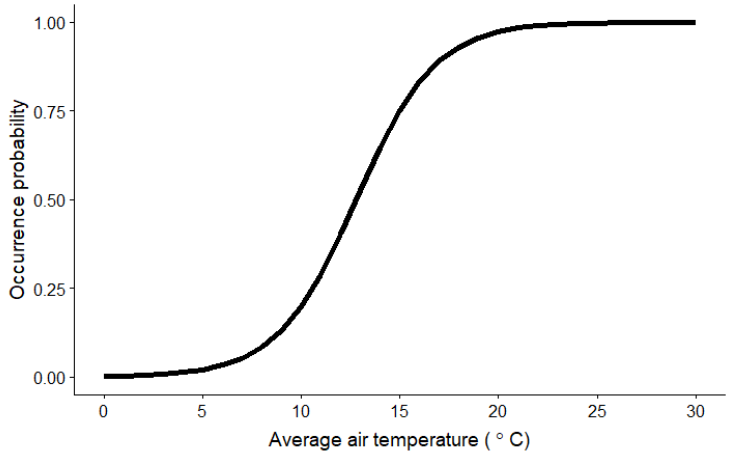
We fit a series of species distribution models and determined that the best model included average air temperature, percent cover of rock/sand/clay, and the temperature range as predictors of Island Mouse occurrence. Below are the model coefficents from the top model:

Covariate	β coefficient	SE	p-value
Intercept	-25.1	2.3	0.0001
Average air temperature	2.3	0.73	0.000026
Percent cover of rock/sand/clay	1.2	0.92	0.0032
Temperature range	-4.1	1.4	0.00085

1. **Describe the relationship between each covariate and occurrence. Does it have a positive or negative effect? Is that effect statistically significant?**

2. **Using the conceptual diagram from Activity 1, how would you interpret these results? Why are these three covariates good predictors of Island Mouse occurrence?**

From this we can see that average air temperature has a strong effect on mouse occurrence. We can visualize that relationship by plotting it (assuming all other covariates are held constant):



Use this relationship to estimate the probability of presence across Darlost's Island based on average air temperature alone. Below is a blank grid of 30x30 km squares across the island. The value in each grid is the average air temperature in °C. Use colored pencils to develop your own color scale from 0 to 1, and color in each square with corresponding probability of presence.

Probability of occurrence:

0	0.5	1

22	26	24	26	20
27	28	25	21	19
23	27	22	20	29
22	18	17	19	18
17	16	15	16	14

The Island Mouse Recovery Team is considering a few nearby islands for possible translocations of Island Mice if conditions on Darlost's Island continue to deteriorate. We can use our estimates of probability of occurrence as a suitability metric for these sites. We expect sites with high predicted probability of occurrence to be sites where Island Mice are likely to persist and thrive.

3. **Based on our analysis of Island Mouse occurrence, what should the team measure at each potential translocation site to determine if Island Mice will have a high probability of persistence there?**

Similarly to how you predicted the probability of presence across Darlost's Island above, we can use information about the new sites and the estimated relationships with occurrence to estimate the predicted probability of occurrence at each new site using the following equation:

$$p.occ = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}$$

Don't worry—you don't need to do these calculations by hand. Use this handy calculator (https://ssa200.shinyapps.io/logit-calculator) – just plug in the values for each covariate to get the probability of occurrence at that site.

For each of the sites below, calculate the predicted probability of occurrence from our model. Which one is the best option for possible translocations? Why?

Site A: Wallace Rock
Average air temperature = 23.5 °C
Percent rock/sand/clay cover = 30%
Temperature range = 15 °C

Probability of occurrence = _____

Site B: Humboldt's Atoll
Average air temperature = 20.7 °C
Percent rock/sand/clay cover = 50.5%
Temperature range = 19.7 °C

Probability of occurrence = _____

Site C: Attenborough Key
Average air temperature = 22 °C
Percent rock/sand/clay cover = 25%
Temperature range = 12 °C

Probability of occurrence = _____

Site D: Isle Lyell
Average air temperature = 25 °C
Percent rock/sand/clay cover = 5%
Temperature range = 9 °C

Probability of occurrence = _____

SSA 200: Strategic Use of Data
Activity 3 – Model-based Predictions

Objectives:

- Understand the link between data analysis and stochastic projections
- Use projection outputs to quantify the 3 R's and make predictions about future conditions for a species
- Implement different future scenarios to describe the range of likely outcomes for a species

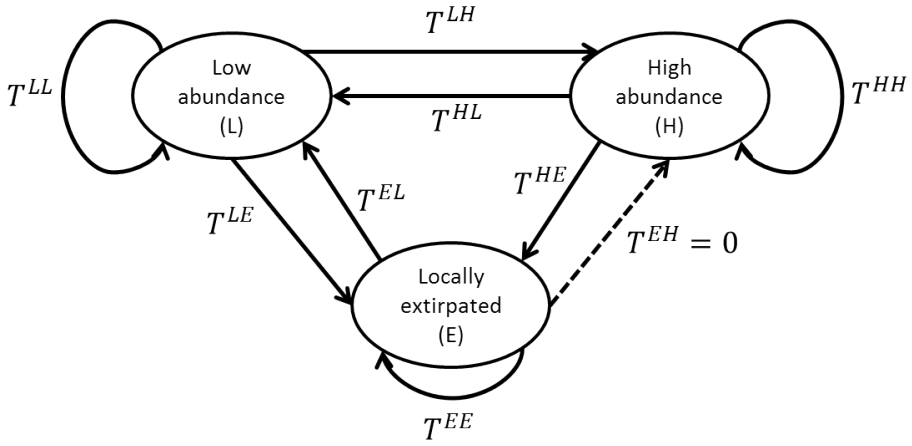
Background Information:

The first two activities focused on using available data to better understand the historic and current conditions of a species and to identify the key species needs and stressors that influence their ability to persist. Now we will use the results of those analyses to develop a stochastic population projection that we can use to estimate the probability that the species will persist until some arbitrary time horizon in the future.

We have some information about Island Mouse abundance, but transect surveys were only conducted across part of the range. We also have some anecdotal reports and information from recent satellite images that can tell us something about abundance across the island. We can pull all sources of information together using a **multistate occupancy** model. Under this framework, we will assign each population to one of three categories:

- Low abundance: mice are present but at low numbers, declining or unstable population
- High abundance: mice are consistently present, stable or increasing population
- Locally extirpated: no mice present

Populations can transition between these states with some **transition probability**, which we'll denote using ***T***. The probability T^{LH} is the probability of transitioning from the low abundance state (L) to the high abundance state (H). The probability T^{LE} is the probability of transitioning from the low abundance state (L) to the locally extirpated state (E). The time step of each transition can be any length of time that is biologically reasonable. Here we will use a one-year time step, so we are talking about the probability of transitioning between states in one year. We can visualize these transitions using a conceptual diagram:



Each arrow represents a transition probability. Note that the arrow from locally extirpated to high abundance is dashed, because here we will assume that this transition probability is equal to 0, in other words it's not possible for a population to go from locally extirpated to high abundance in one year. The arrows that loop back on each state are the probabilities of remaining in the current state.

We can also visualize the same relationships using a matrix.

$$\begin{bmatrix} T^{EE} & T^{LE} & T^{HE} \\ T^{EL} & T^{LL} & T^{HL} \\ T^{EH} & T^{LH} & T^{HH} \end{bmatrix}$$

In a transition matrix, each cell represents the probability of going *from* the column number/state *to* the row number/state. In this matrix, the columns and rows (in order) represent the extirpated, low abundance, and high abundance states.

Based on all available information, the Island Mouse Recovery Team determined the state of each population over the past 10 years. We used a multistate model to estimate each transition probability, in an analysis similar to a site-occupancy analysis.

These are the estimated transition probabilities with their corresponding standard errors:

$$\begin{bmatrix} 0.9 \pm 0.05 & 0.3 \pm 0.1 & 0.05 \pm 0.02 \\ 0.1 \pm 0.08 & 0.5 \pm 0.05 & 0.2 \pm 0.04 \\ 0 & 0.2 \pm 0.09 & 0.75 \pm 0.05 \end{bmatrix}$$

1. **What is the probability of a population that is currently low abundance remaining low abundance in the next year?**

2. **What is the probability of a high abundance population becoming extirpated in one year?**

3. **What is the probability of transitioning from locally extirpated to low abundance?**

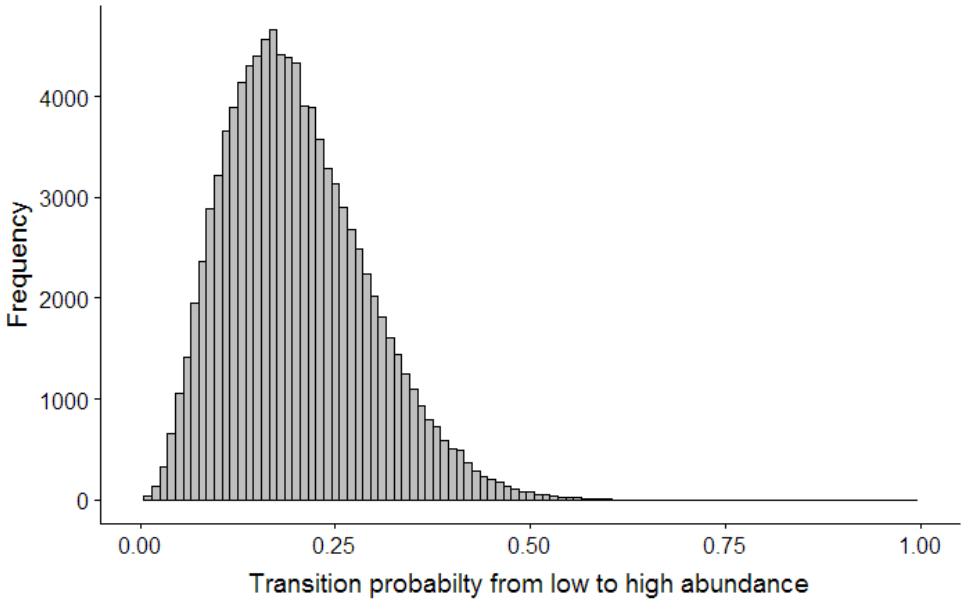
4. **Here these transition probabilities were estimated directly using a statistical model. If the data to fit such a model were not available, what are some other potential ways to determine what those values should be?**

Part 1: Projecting Future Conditions

We would like to determine the probability of Island Mice persisting 15 years in the future. We have defined three analysis units based on ecotype (coastal, paradise palms, mountain). The ecotype and current state of each population is listed below:

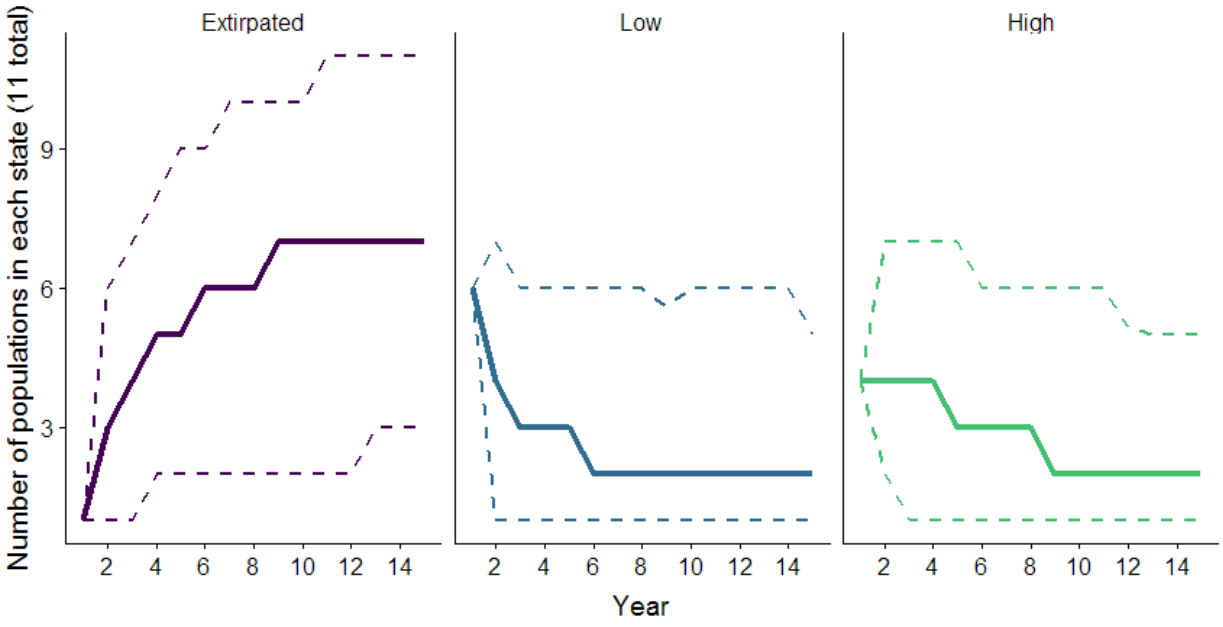
Population	Ecotype	Current state
Beach Bums	Coastal	H
Cannibal Cove	Coastal	H
Castaways	Coastal	L
Message in a Bottle	Coastal	H
Darlost's Dome	Mountain	L
Misty Mountain	Mountain	L
Skull Mountain	Mountain	E
Dead Man Dunes	Paradise palms	L
Realm of Spirits	Paradise palms	H
Snowmelt Thicket	Paradise palms	L
Treasure Grove	Paradise palms	L

Using the estimated transition probabilities above, we will project the population 15 years into the future, replicating that projection 1000 times. Our estimates of the standard error for each parameter represent our uncertainty about the true value of that parameter. To account for this **parametric uncertainty** in the projection, we can use the estimated standard error to define a Beta distribution for each transition probability. For example, the distribution of possible values for the probability of transitioning from low abundance to high abundance ($T^{LH} = 0.2 \pm 0.09$):

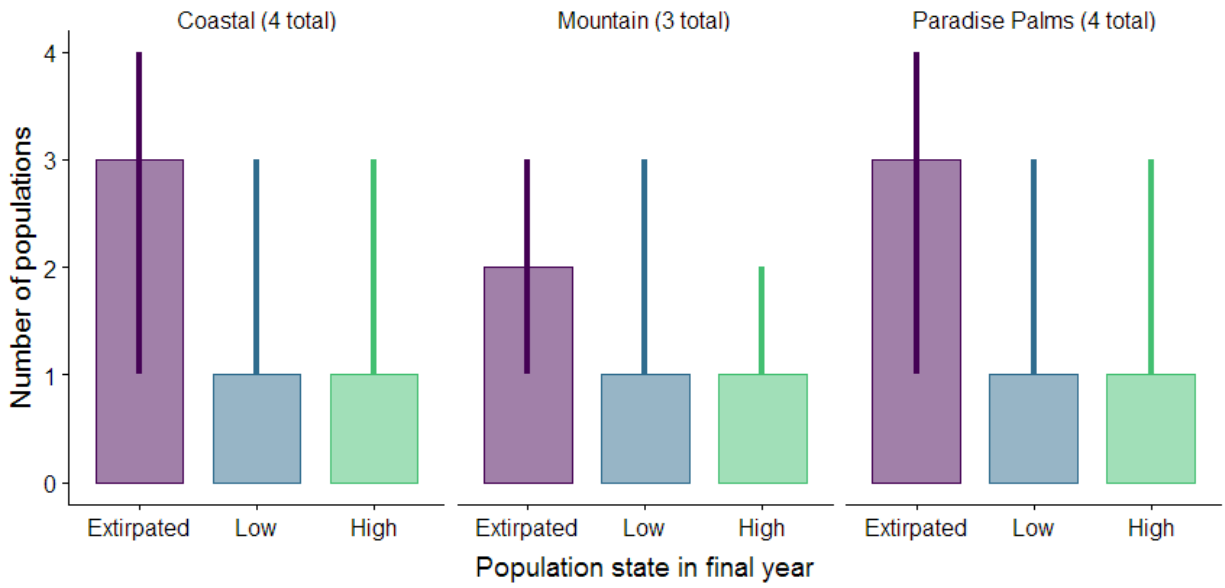


For each replicated projection, we will randomly draw a value from this distribution for the transition probability from low to high. Each transition probability will have its own distribution.

If we run that projection, we can look at the number of populations in each state over time:



Here the solid lines are the medians from 1000 replications and the dashed lines are the 95% confidence intervals. If we want to break it up by ecotype, we can look at the number of populations in each state in the final year:



Here the height of the bars show the median number of populations that ended in each state and the error bars show the 95% confidence intervals.

That same information can also be represented in table form. Here the medians are given in each cell with the 95% confidence intervals in parentheses.

Ecotype	Number of populations in each state		
	Extirpated	Low	High
Coastal	3 (1, 4)	1 (1, 3)	1 (1, 3)
Mountain	2 (1, 3)	1 (1, 3)	1 (1, 2)
Paradise Palms	3 (1, 4)	1 (1, 3)	1 (1, 3)

We can also use simulation outputs to estimate the probability that certain events will occur. We do that by calculating the proportion of replications in which that event occurred. For example, if we wanted to know the probability that at least one population from each ecotype ended in the "high abundance" state, we would count the number of replications in which all three ecotypes had at least one population in "high" and divide that by the total number of replications. Under these baseline conditions, that probability is 0.122, or 12.2%. We could also estimate the probability of extinction for the species by calculating the proportion of replications in which all populations ended in the "extirpated" state, and the probability of extinction for each ecotype.

Group	Extinction probability
Coastal	0.21
Mountain	0.37
Paradise Palms	0.24
Overall	0.05

1. **What is the most likely outcome for Island Mice in 15 years? What is the most likely outcome for each ecoregion?**

2. **Why is there uncertainty about the number of populations that will end in each state?**

3. **How would you characterize the expected resiliency, redundancy, and representation of this species? Use the above figures and tables to support your statements.**

Part 2: Alternate Scenarios

In the above projection, we assumed that baseline conditions would stay the same in the future. However, it is often useful to explore how species are expected to respond to potential changes that could occur in the future.

The first step to developing those projections is quantifying the relationships between ecological needs and/or stressors and parameters in the projection model. In this case, the parameters in the model are the transition probabilities among different states. Using the multistate model framework described above, we can include ecological covariates and determine whether they have a strong effect on transition probabilities.

We fit several models and ranked them using AIC to determine the ecological covariates that were most strongly associated with each transition probability. Three transition probabilities were found to be strongly associated with ambient noise level. T^{HL}, T^{HE}, and T^{LE} all increased as ambient noise level increased ($\beta = 1.5 \pm 0.03$, $p = 0.001$).

1. **Write your interpretation of this result in a sentence.**

Two transition probabilities were found to be strongly associated with annual temperature range. T^{HH} and T^{LH} both decreased as temperature range increased ($\beta = -1.8 \pm 0.41$, $p = 0.0023$).

2. **Write your interpretation of this result in a sentence.**

We can build these relationships into the projection model, and then see how the extinction probability and likely outcomes change if one or both of these covariates change.

3. **Under what future conditions would you expect ambient noise level to change on Darlost's Island? Would it increase or decrease?**

4. **Under what future conditions would you expect the annual temperature range to change on Darlost's Island? Would it increase or decrease?**

Use this information to develop three potential future scenarios below. Try to consider a "worst case", "best case" and "most likely" set of future conditions. Write a few sentences about why you chose that scenario, and assign expected changes in ambient noise level and temperature range (this can be positive, negative, or zero).

Scenario	Description	Change in ambient noise level (db)	Change in annual temperature range (°C)

Next you will implement your scenarios and describe the 3 R's under each. Follow this link (http://ssa200.shinyapps.io/mouse-multistate) to a web application where you can input the parameters for each scenario and see the results. Leave the number of years and number of replications on the default values for this activity.

Fill in the following table to summarize the results of your scenarios.

Scenario	Overall extinction probability	Number of populations in each state in the final year		
		Extirpated	Low	High

Part 3: Interpretation and Communication

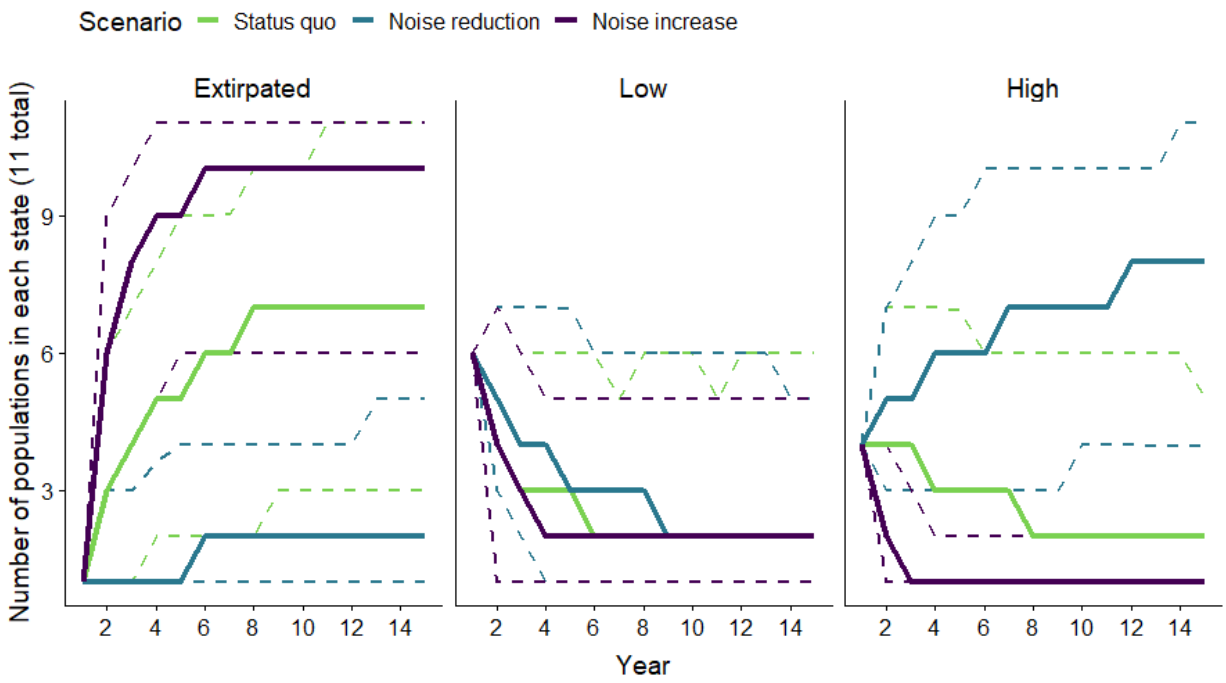
Depending on the scenarios you chose to run, you may have seen drastically varying predicted population outcomes. In addition to the baseline "status quo" scenario, we ran two additional scenarios that represented potential future climate change and management actions.

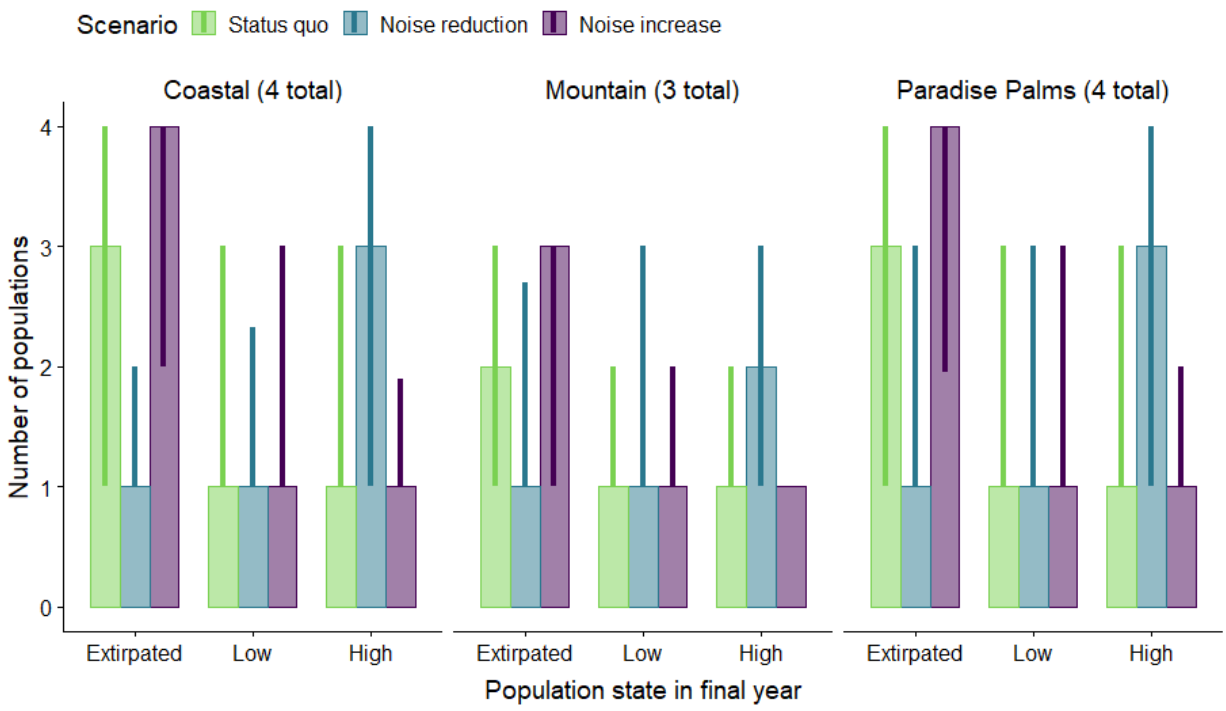
Scenario 1 – Status quo: No change in ambient noise level or annual temperature range

Scenario 2 – Noise reduction: Climate change leads to an increase in average annual temperature range of 0.5 °C, but management efforts to restrict development leads to a decrease in ambient noise level of 2 decibels.

Scenario 3 – Noise increase: Climate change leads to an increase in average annual temperature range of 0.5 °C, and further development leads to an increase in ambient noise level of 2 decibels.

The results of all three scenarios are below:





Group	Extinction probability		
	Status quo	Noise reduction	Noise increase
Coastal	0.21	0.00	0.57
Mountain	0.37	0.014	0.67
Paradise Palms	0.24	0.001	0.58
Overall	0.05	0.00	0.29

1. **Considering all three of these scenarios together, how would you characterize the expected resiliency, redundancy, and representation of Island Mice? Use the above figures and tables to support your statements.**

2. **Do you think all three of these scenarios are equally likely? Is one more likely to occur than others? How would you incorporate your uncertainty about which scenario is most likely to occur in your presentation of these results?**

3. **Which metrics and/or figures do you think are most helpful in communicating the results of this projection to decision makers? Why?**

SSA 200: Strategic Use of Data
Activity 4 – Developing an Analysis Plan

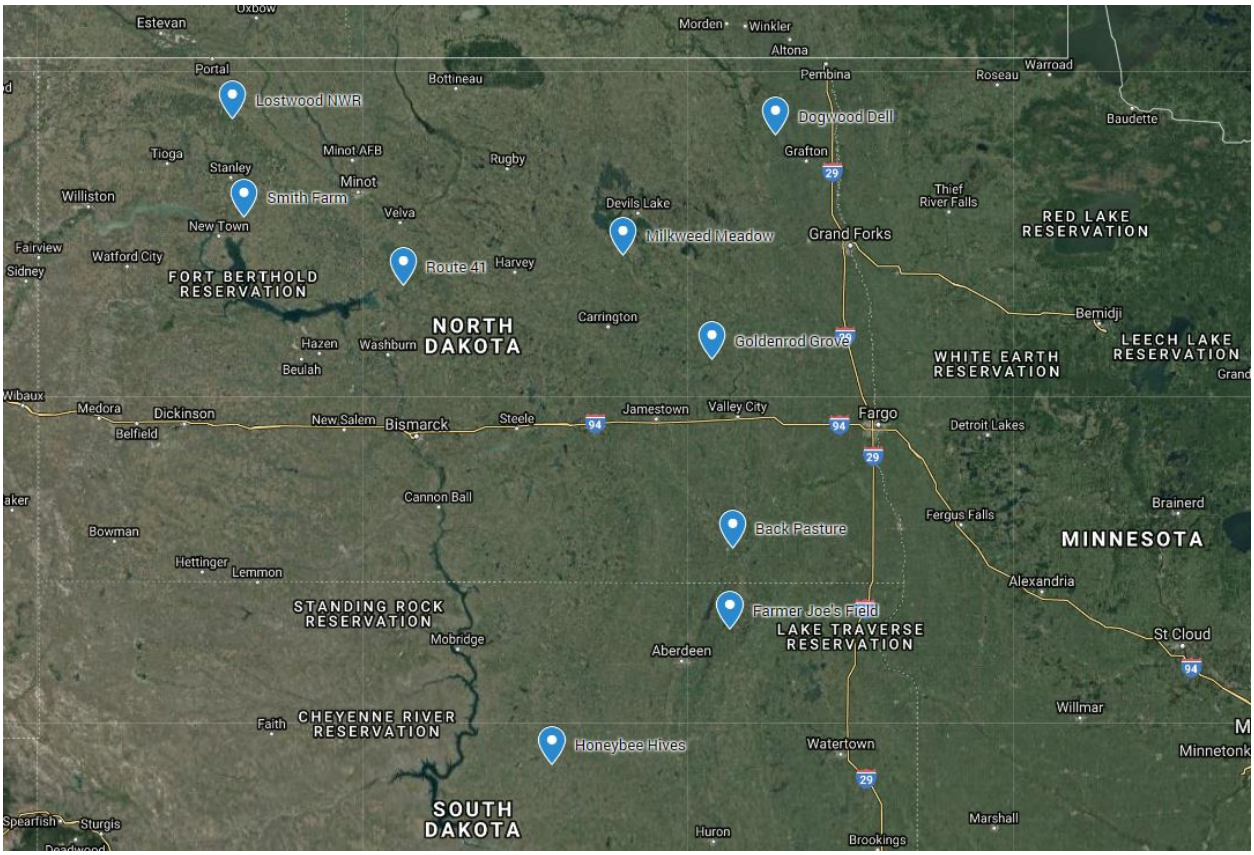
Objectives:

- Critically assess data sets to determine data types and identify appropriate analysis options
- Use all available information to develop a comprehensive analysis plan to quantitatively assess current and future conditions in the context of the 3 R's

Scenario:

The Service has been petitioned to list a small, highly range-limited butterfly species that occurs in various habitats in the Prairie Pothole region. The petition is focus on small population sizes and growing threats from habitat disturbance due to oil and gas development. The service has issued data call to management partners in the region requesting any data on the species, i.e., its habitats, its life history, population monitoring, or any other available information and data.

The service received three data sets from partners. All of the locations with survey information are mapped below:



One dataset from the Lostwood National Wildlife Refuge in North Dakota recorded presence/absence data at a number sites over 6 years and also recorded some information on habitat features and weather at the time of surveys. A second data set from the Nature Conservancy, repeatedly counted (three times per season) individuals observed on several transects at three different properties from 1999 until 2009. They did not provide information about the habitat or other covariates but that information may be available from National Land Cover Data Base or NOAA weather data archives. A third data set from the BLM spanned 2 decades in the 1980s and 90s from several sites throughout the range, but the "data" were descriptive or qualitative in nature. The observer recorded the place, time, date and whether the butterflies were "not observed", "scarce", "not many" and "many". No other covariate information was provided, but again useful information may be available from the NLCD or NOAA.

Following the data call, the service conducts their own literature search and review to find any academic or other publications available on the species and, as luck would have it, the was a now retired professor at North Dakota State that had 3 graduate students finish research thesis on this species. The students conducted field surveys and experiments and published some of their work in peer review scientific journals. The labs work concluded that the butterflies, typically have a single generation per year and the populations survive the winter as dormant eggs and that over winter survival of eggs is quite high (>90%) in the absence of early spring fires. They learned that females can lay up to 200 eggs in a single season, but that many caterpillars (>80%) die before pupating into adults.

Instructions:

The three data sets described above are available in the Excel workbook called "04_data-sets.xlsx", available from the course website. Given the available data and published literature, your task is to devise an analysis and modeling plan to conduct an SSA for this butterfly that will support the listing decision. Write up your plan as a few paragraphs or bulleted list. We will regroup to present and discuss our plans as a group.

You are encouraged to use the Data Analysis Roadmap (in your course handout and available on the course website) to guide your process of assessing the data and deciding on an analysis plan.

Some things to consider:

- Think about not only the data type (counts, presence/absence, etc) but also the methodology behind how it was collected. Be sure the data at hand meet the assumptions of your proposed analysis
- What is the response variable for each analysis? What are you trying to estimate (e.g. abundance, occupancy probability, survival, fecundity, etc.)?
- Consider how all pieces of the analysis will fit together, especially how you will use the data analysis to build a projection model
- You do not have to use all data sets or all information available in your planned analysis
- How will you define and quantify the 3 R's?

Analysis Plan: