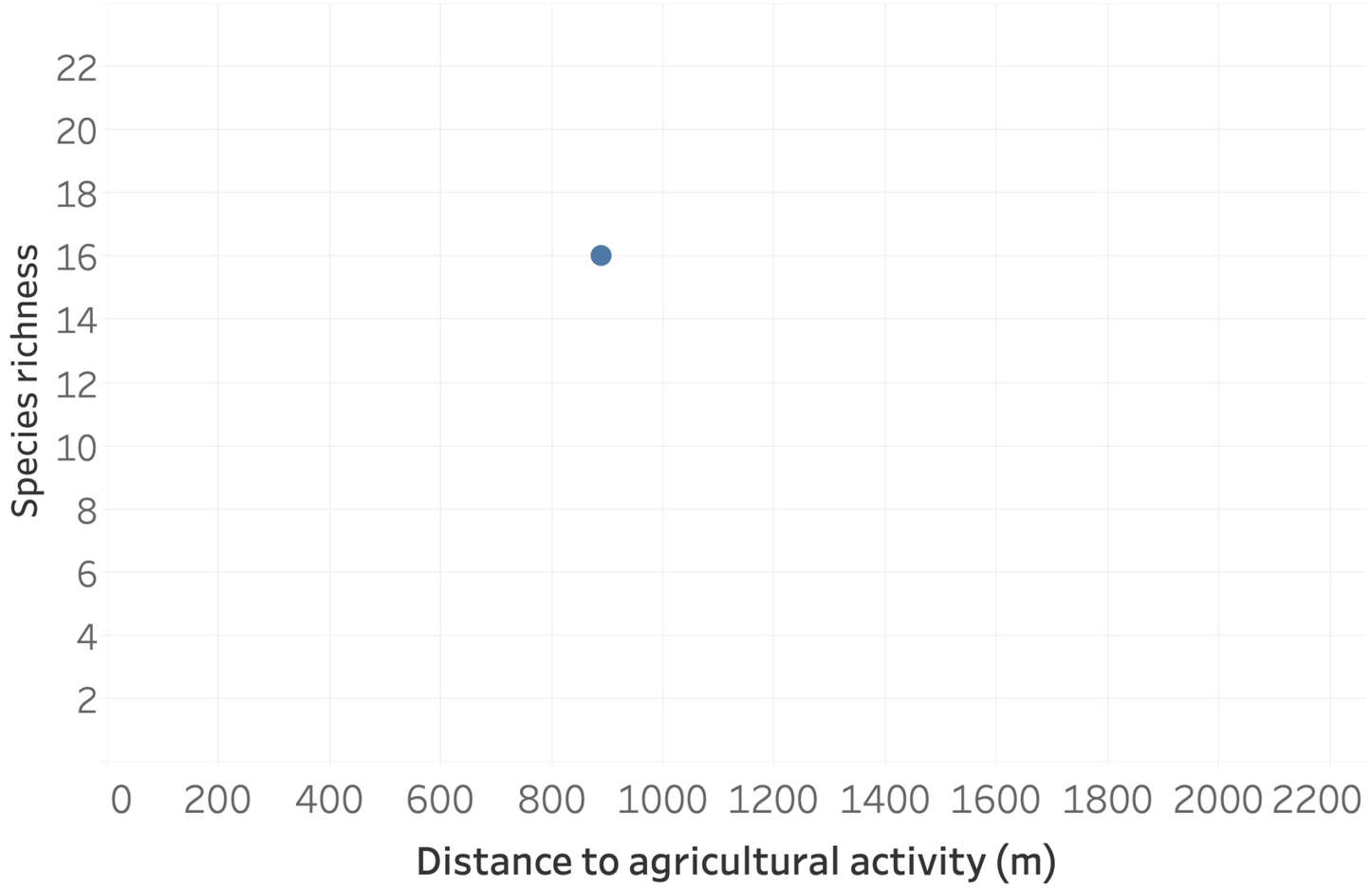
The relationship of two variables

STATISTICAL TECHNIQUES IN TABLEAU

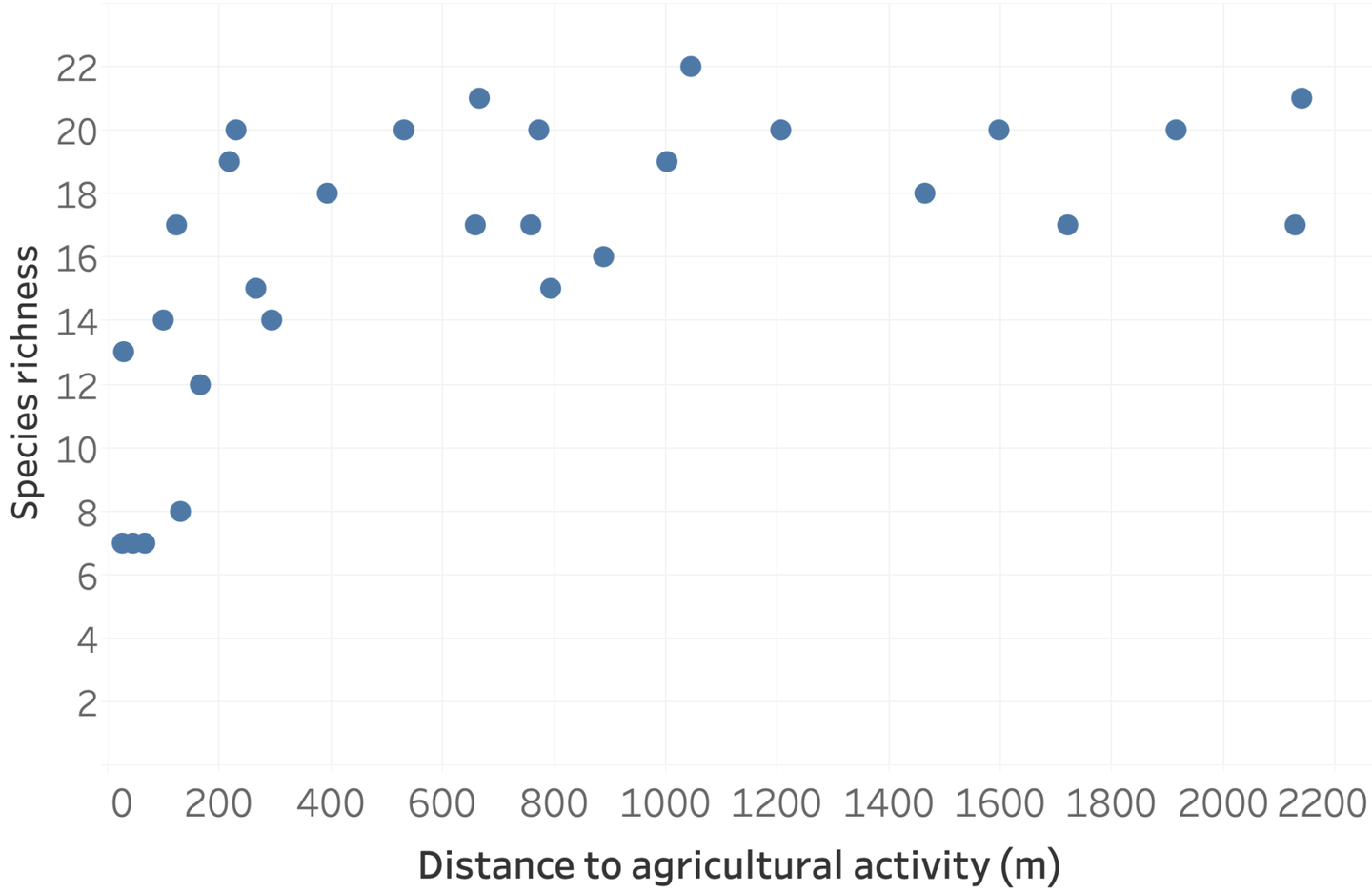
Maarten Van den Broeck
Content Developer at DataCamp

datacamp

Revisit the lake again



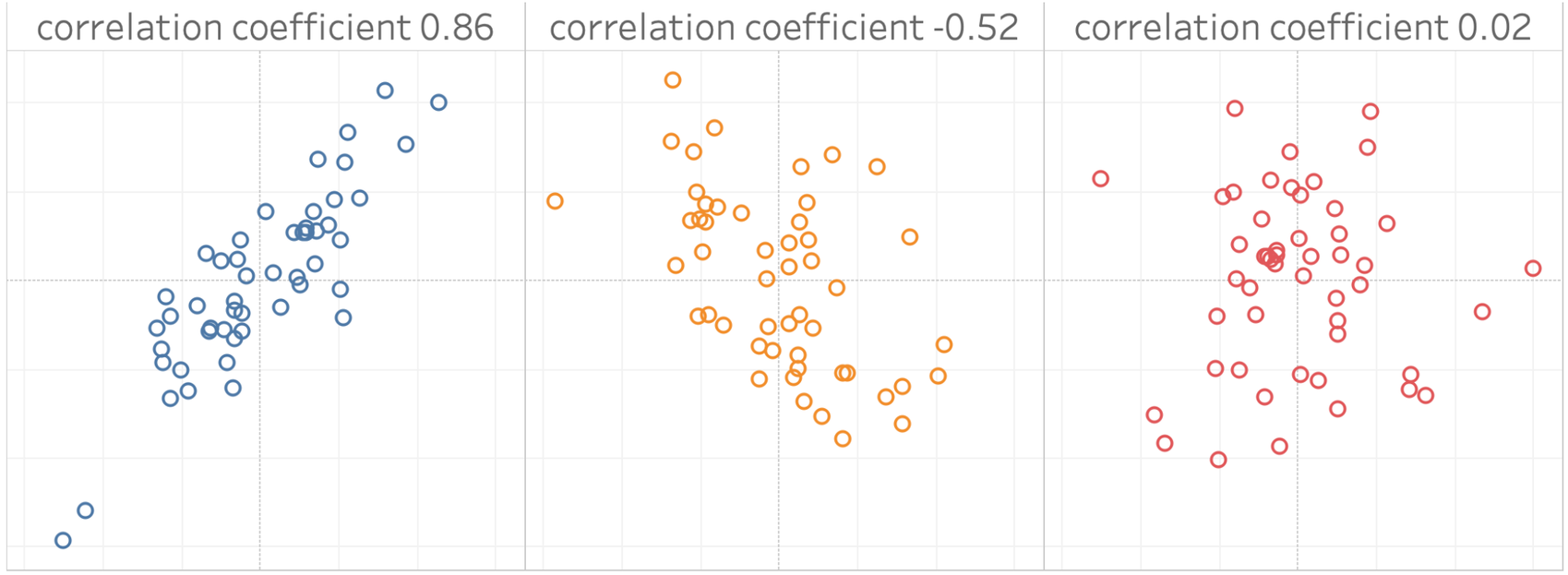
Revisit the lake again: scatter plot



- Variable of interest is placed on y-axis

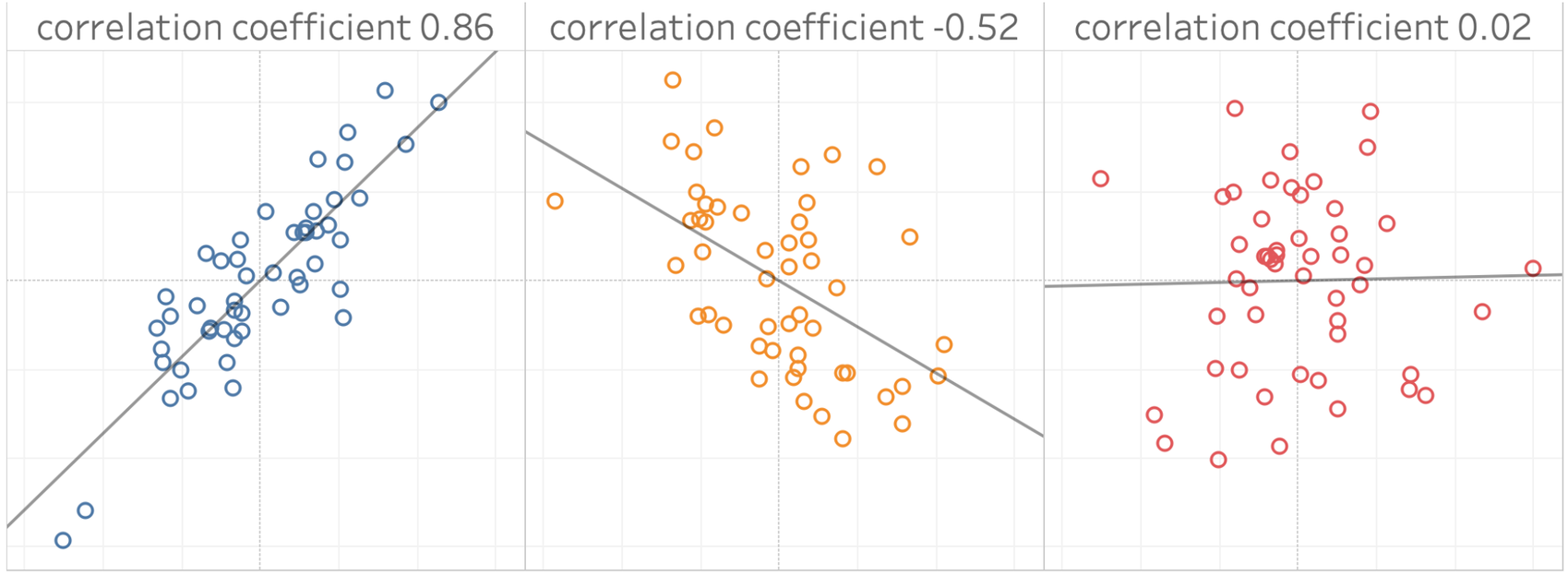
Correlation coefficient

- Quantifies the relationship between two variables
- Between `-1` and `1`
- Sign (+ or -) corresponds to direction of relationship
- Magnitude (absolute value) corresponds to strength of the relationship



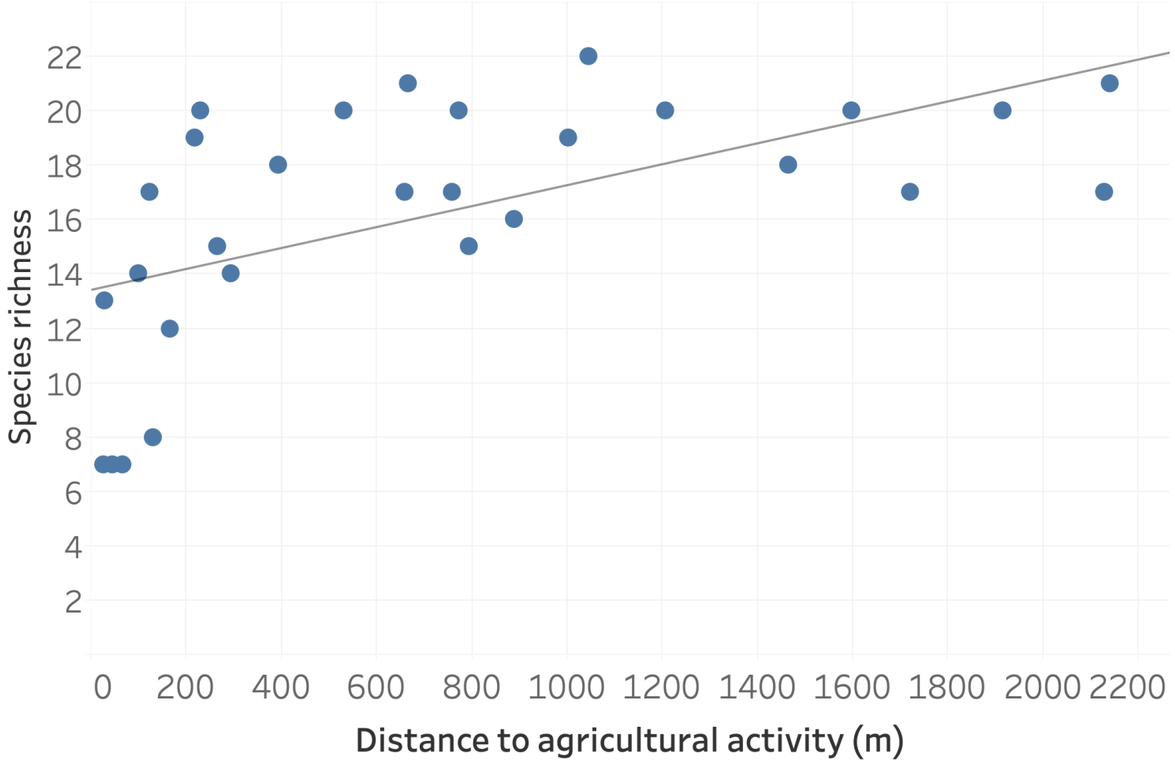
Trend lines

- Visualizes the relationship between two variables
- As close as possible to each data point
- A high (absolute) correlation value has all point close or on the trend line
- Not necessarily a straight line

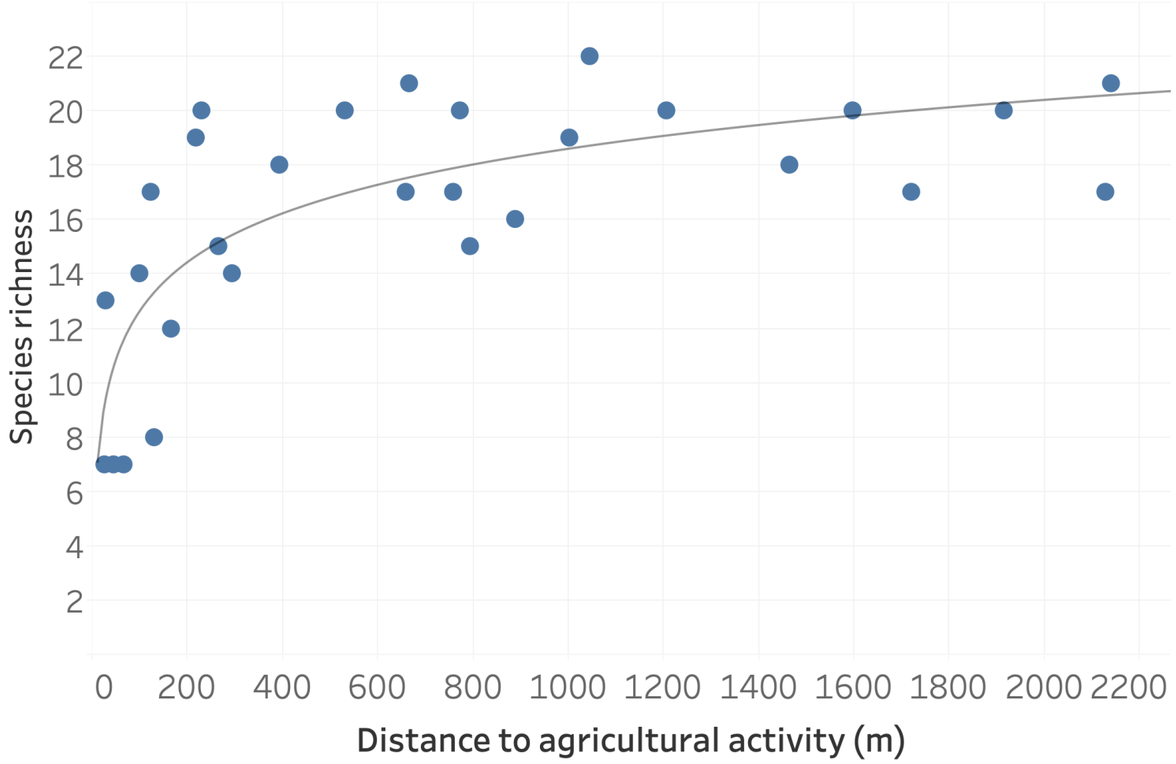


Trend lines: linear vs. logarithmic

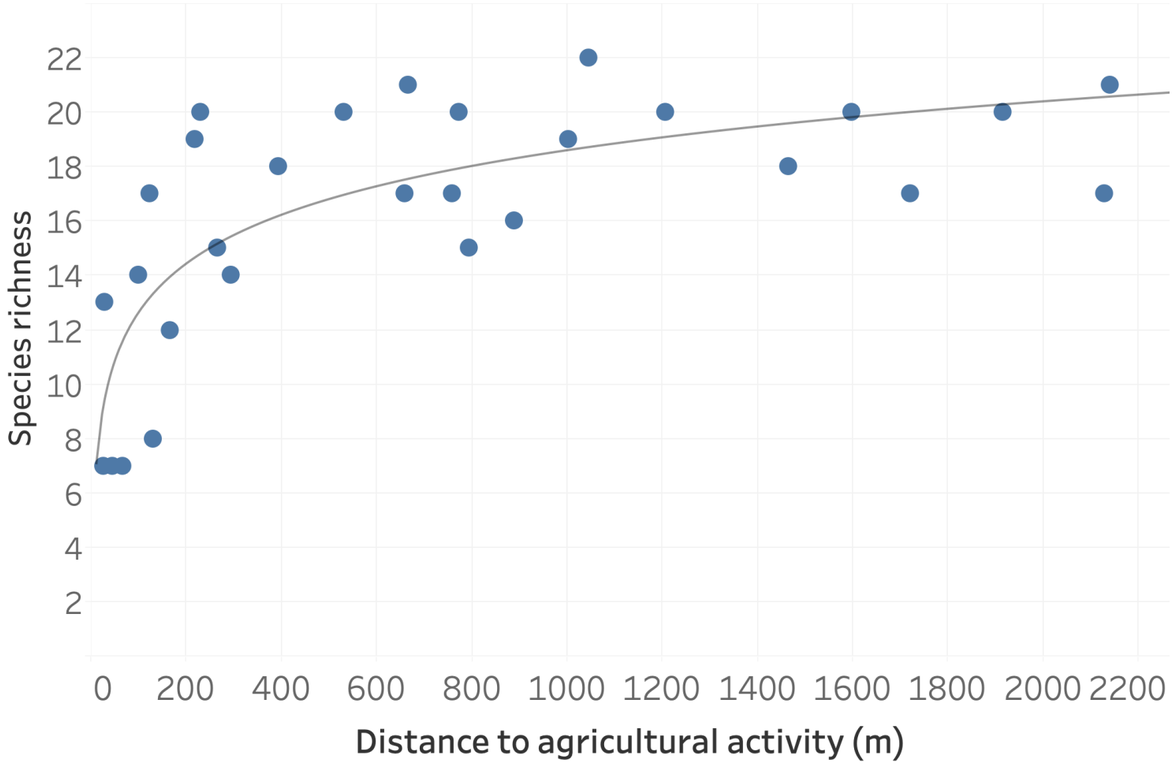
Linear trend line

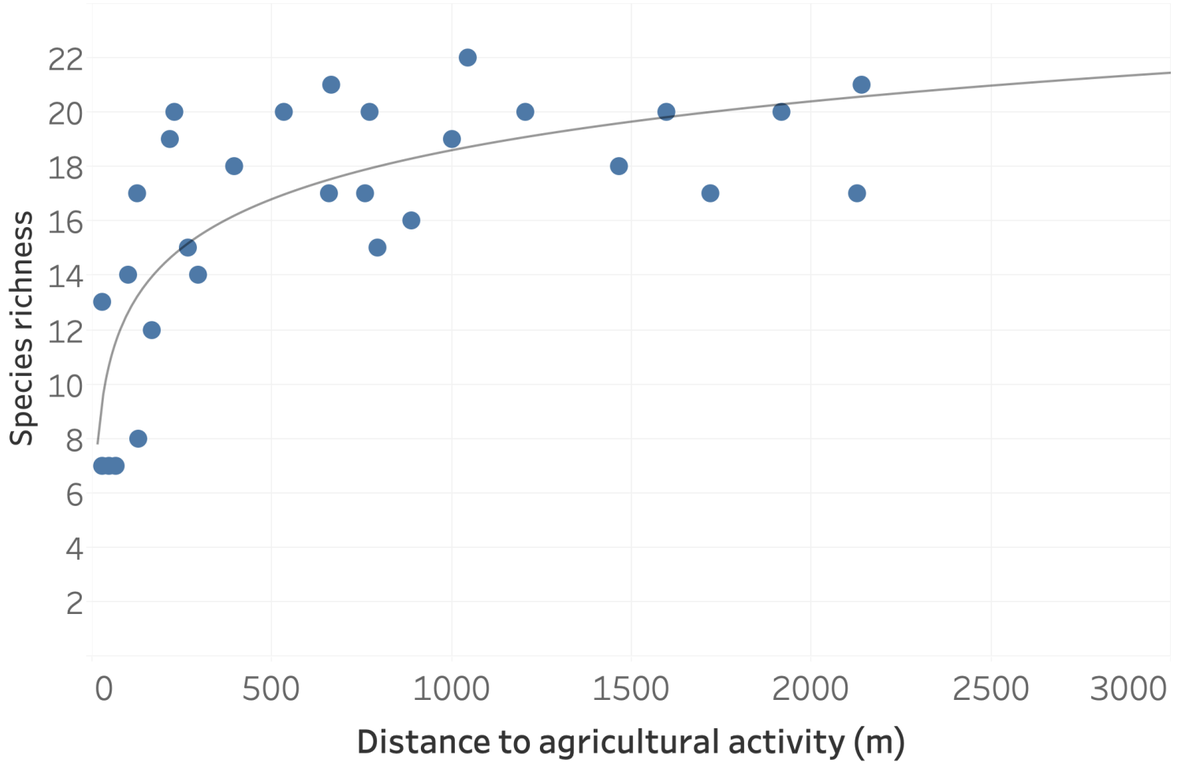


Logarithmic trend line



Trend lines: predicting and extrapolating





- Distance of 1400 m corresponds to approx. 19 species

- Distance of 3000 m corresponds to approx. 21 species

Other types of trend lines

- Exponential: inverse of logarithmic
- Power or log-log: both variables follow logarithmic trend
- Polynomial: n-degree, from second degree till eight degree

Let's practice!

STATISTICAL TECHNIQUES IN TABLEAU

Tableau: trend lines

STATISTICAL TECHNIQUES IN TABLEAU

Maarten Van den Broeck
Content Developer at DataCamp

datacamp

Let's practice!

STATISTICAL TECHNIQUES IN TABLEAU

Assessing a trend line

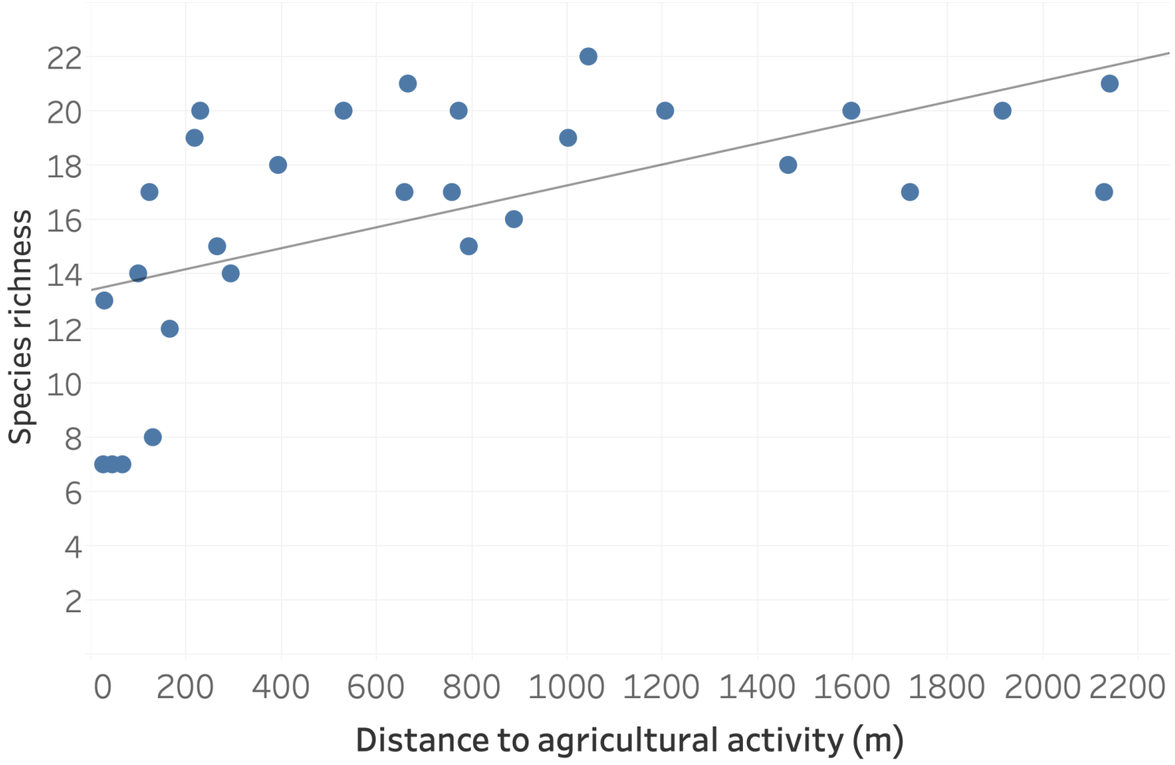
STATISTICAL TECHNIQUES IN TABLEAU

Maarten Van den Broeck
Content Developer at DataCamp

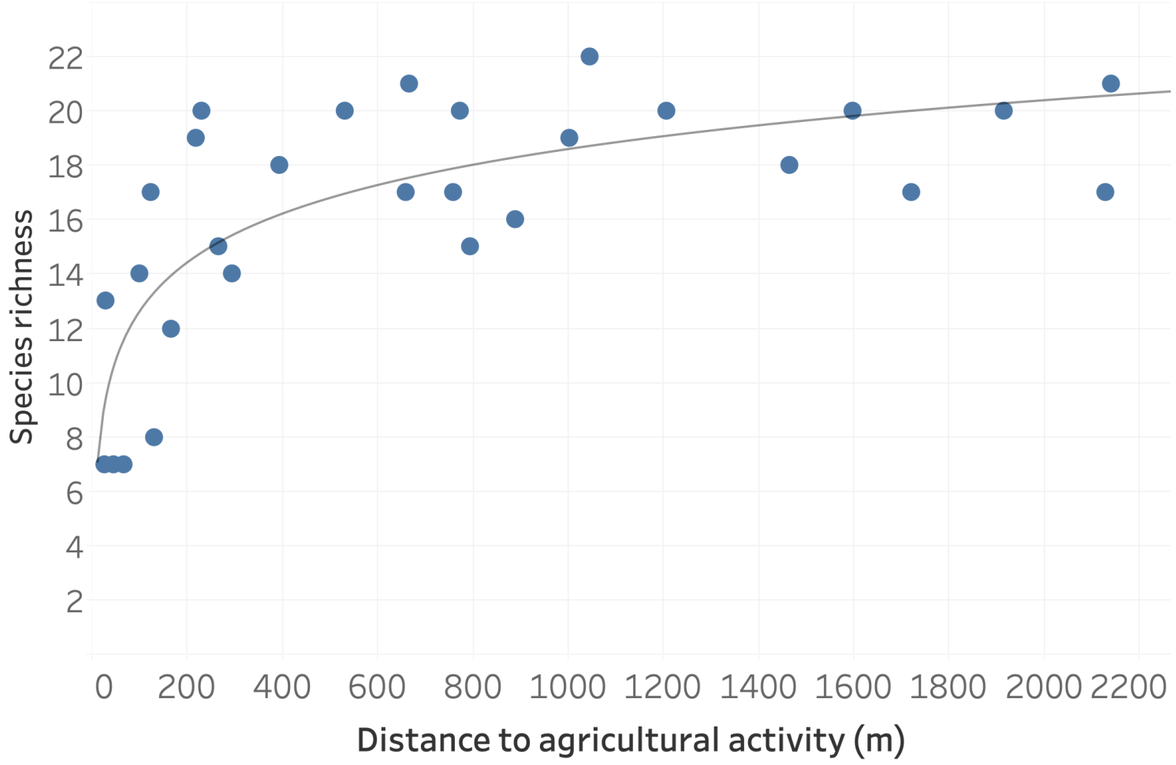
datacamp

Linear and logarithmic models

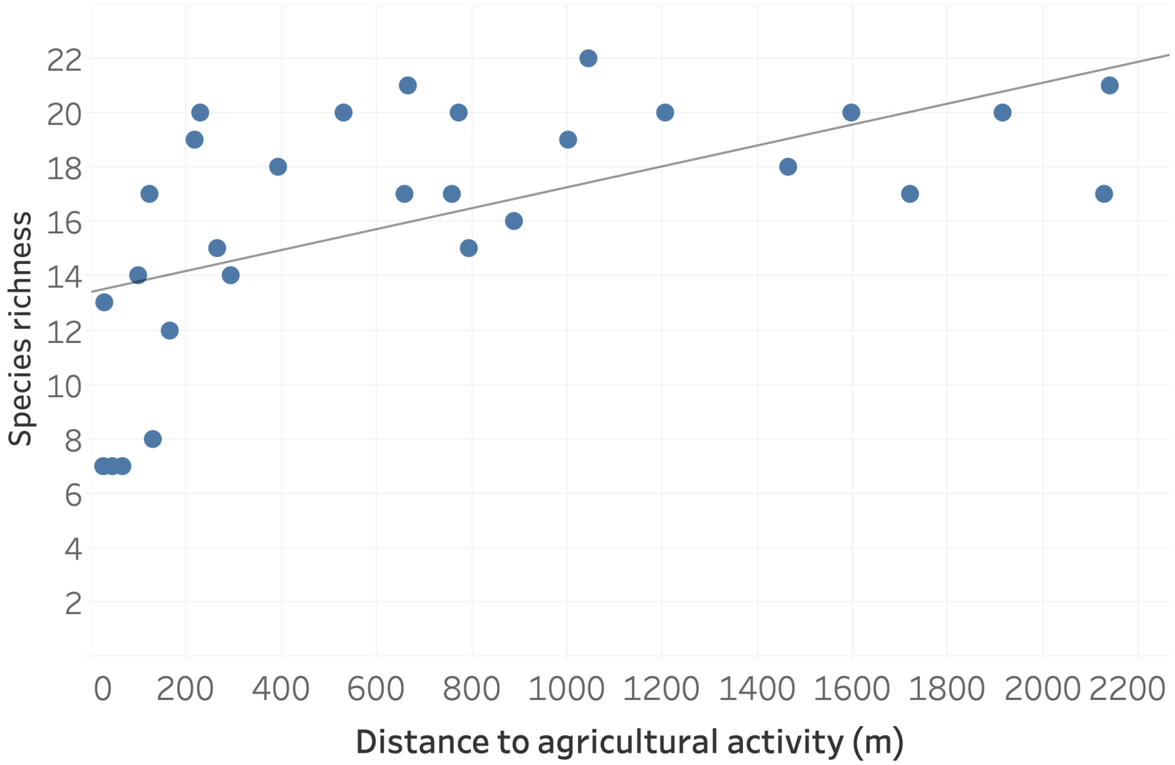
Linear



Logarithmic

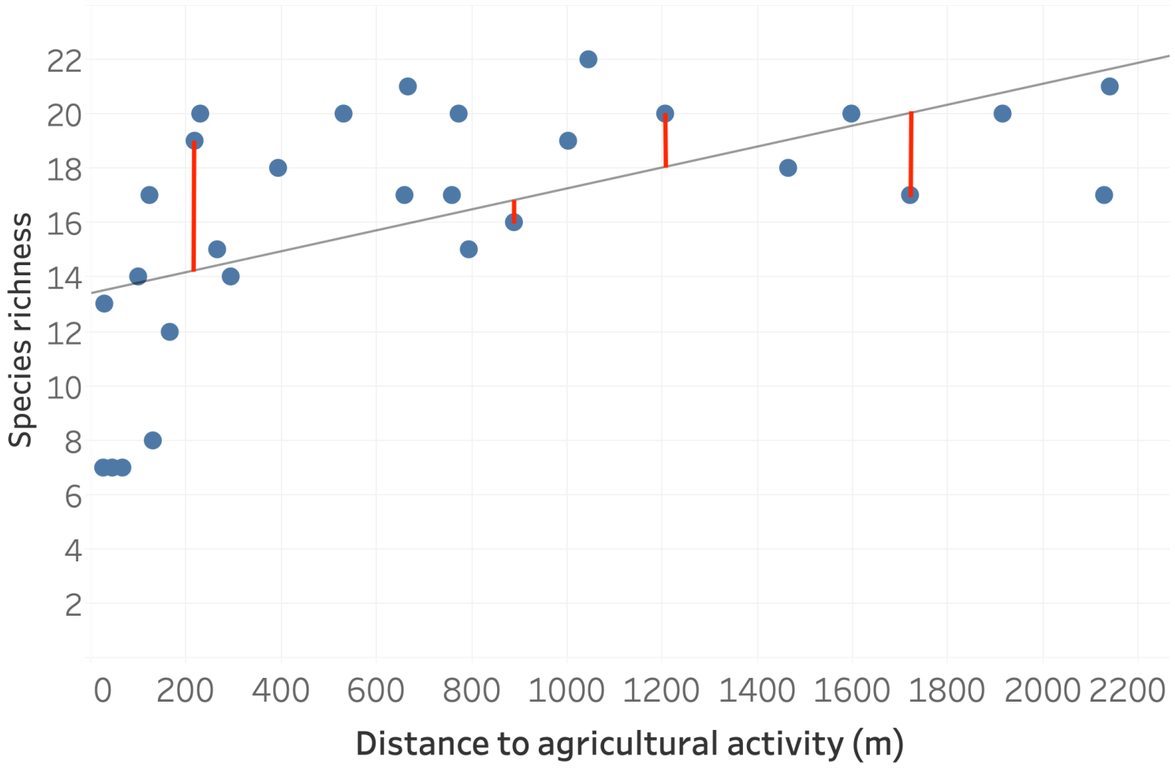


Linear model



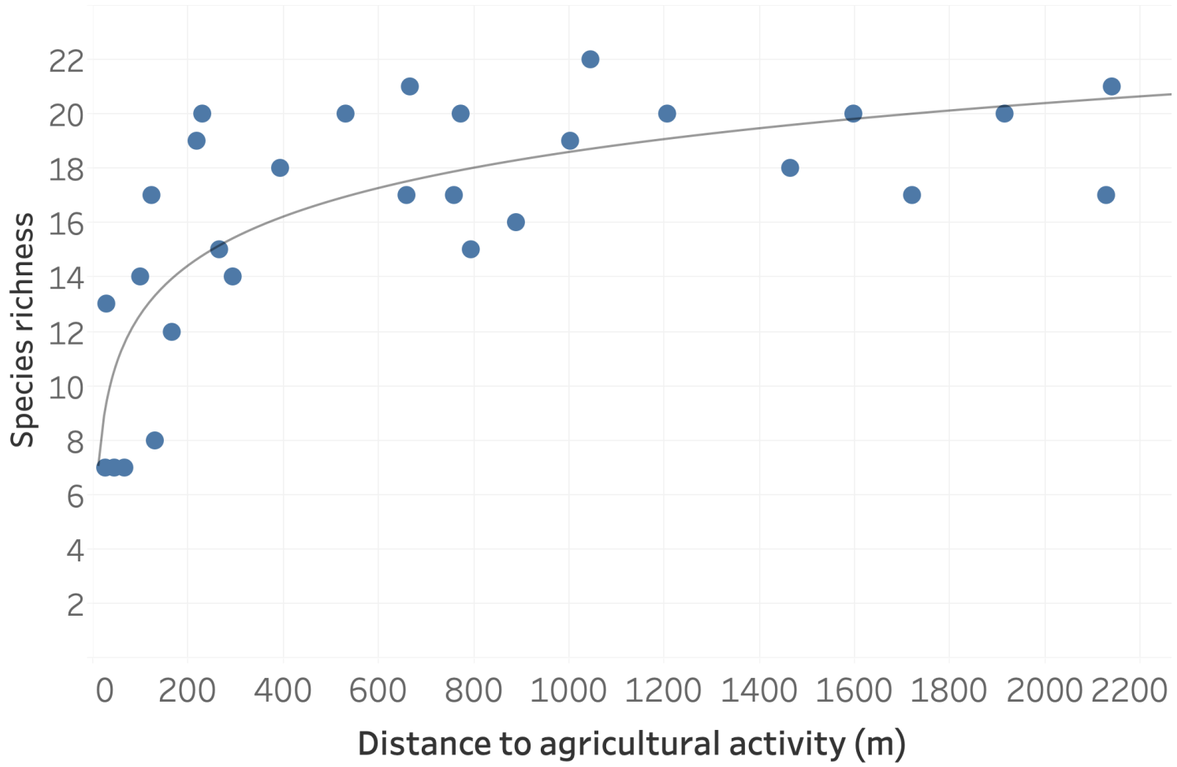
- $y = a * x + b$
- Regression: quantify how x causes y to change
- a = slope, b = intercept
- $richness = a * distance + b$
- $richness = 0.0038 * distance + 13.4$

Residuals and R^2 of linear model



- Goal is to minimize distance between observation and trend line
- The distance is called a residual
- Coefficient of determination R^2
- For linear model, R^2 = correlation coefficient squared
- Between `0` (worst) and `1` (best)
- $R^2 = 0.33$
- Explains $n\%$ of the variation

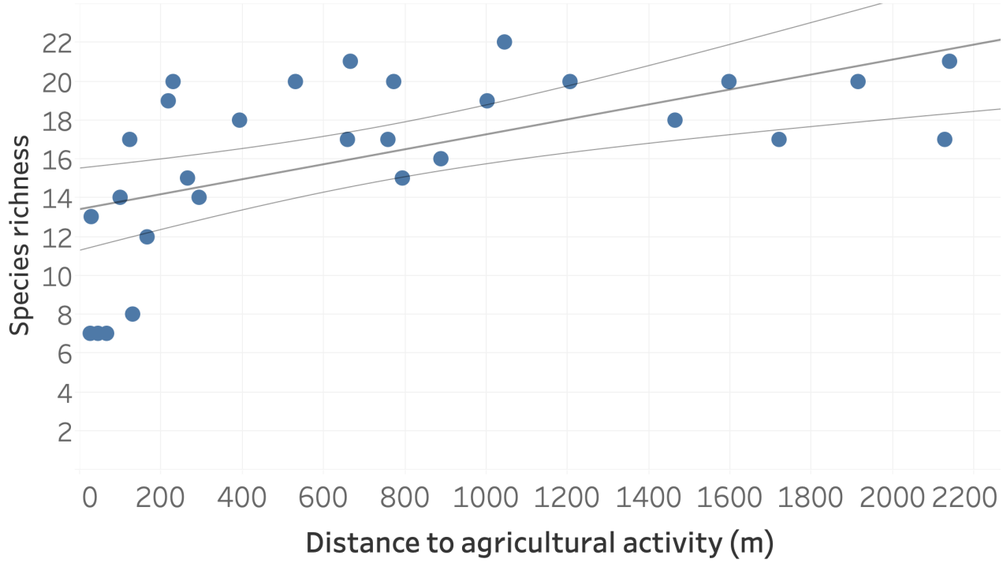
R^2 of logarithmic model



- $R^2 = 0.59$

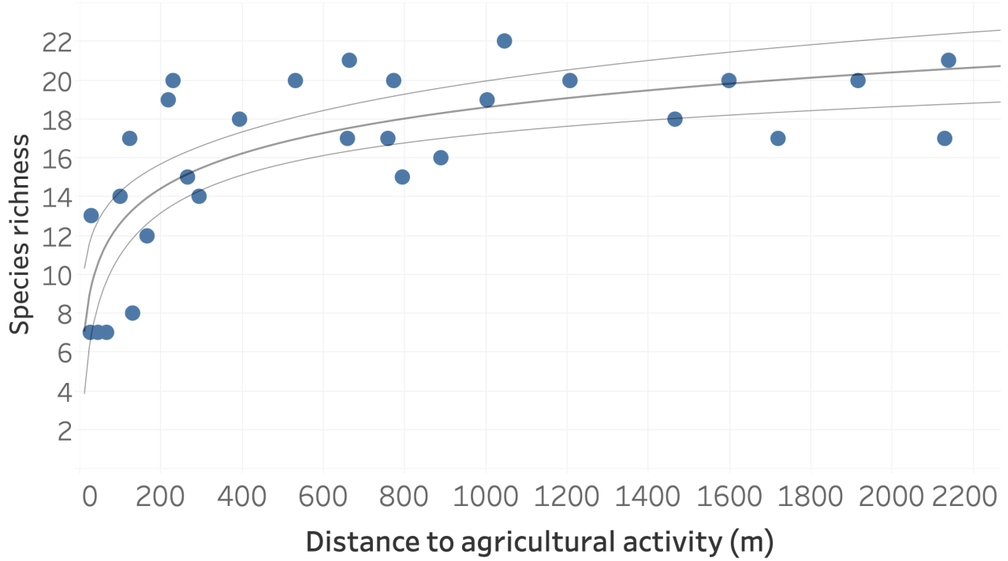
Residual standard error (RSE)

- Average difference between observed values and trend line
- Linear model: RSE = 3.69



- Confidence interval is wider at low and high distances

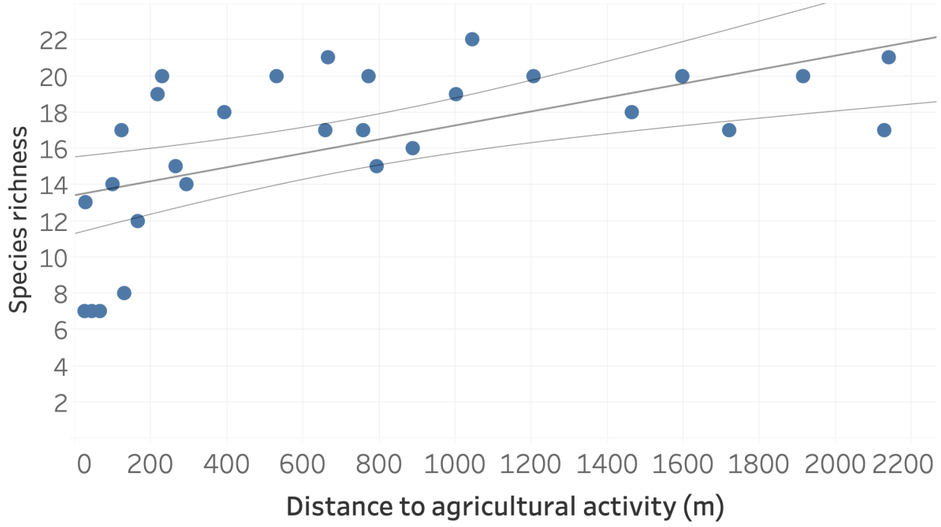
- Same unit as unit on y axis
- Logarithmic model: RSE = 2.91



- Confidence interval is only wider at very high distances

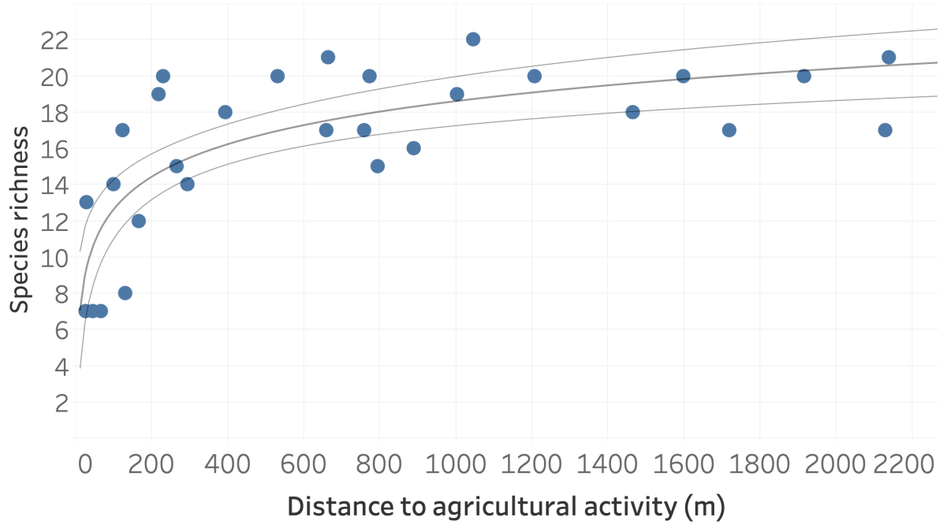
p-value

Linear model



- p-value = 0.001
- Chance of $\frac{1}{1000}$ there is no correlation
- p-value < 0.05: model is statistically significant, fits data well

Logarithmic model



- p-value = 0.0001
- Chance of $\frac{1}{10000}$ there is no correlation
- The lower the p-value, the better, but p-value isn't everything!

Let's practice!

STATISTICAL TECHNIQUES IN TABLEAU

Tableau: describing trend models

STATISTICAL TECHNIQUES IN TABLEAU

Maarten Van den Broeck
Content Developer at DataCamp

datacamp

Let's practice!

STATISTICAL TECHNIQUES IN TABLEAU