Data Analysis Report - Sparrow

Shangrong Chi

Introduction

This study investigates how variables such as female population density, nest location, and year of study influence song sparrows' offspring. The dataset comprises observations of data with researchers monitoring every 2-7 days during the mating season and tagged in 15 years.

The analysis focused on several key questions: How does female population density affect reproductive success? Do spatial factors (nest location) influence offspring outcomes? Furthermore, what role do other factors play in reproductive performance? Therefore, using statistics analysis and hierarchical models, breeding success is strongly influenced by population density and nest location, with fewer offspring produced when female population density is high and in eastern nesting locations. While annual environmental conditions substantially affect the outcomes, there is significant zero-inflation in reproduction patterns, suggesting that breeding failure occurs more frequently than the model predicts and that additional factors beyond density and location likely influence reproductive success.

Study Design

- Sample Size:

- Total observations: 742

- Observed over 19 years

- 356 unique birds identified by band numbers

- Each bird observed between 1-7 times (median: 2 observations per bird)

- Time Structure:

- Data collected in 15 of these 19 years

- Regular nest monitoring every 2-7 days during mating season

- Cohorts range from year 2 to 19 (101 missing values)

- Spatial Structure:

- Nest locations recorded using x-y coordinates

- X-coordinates range: 0.25-32.83

- Y-coordinates range: 1.375-5.250

- Only 5 missing spatial coordinates

- Other Measures:

- Female population size (fpop) ranges from 4 to 72 birds

- Age structure from 1-7 years old (101 missing values)

- Reproductive output (spf) as response variable ranges from 0-12 offspring

- Macro Level Variables:

- Year

- Population size(fpop)

- Micro Level Variables:

- Individual bird characteristics (age, cohort, and band)

- Nest location effects (x and y)

- Key Features:

-Complete and consistent sparrow monitoring

-Multi-generational tracking

-Precise nests locations

-Cohort, age, x, y and band columns have missing values (NA)

-Age, cohort and year are related data

-Reproductive success measured by number of fledglings (spf)

- Deal with missing values:

The initial dataset: cohort and age each had 13.6% missing values, while x, y and band ID had minimal missing data (less than 1%). Given the low proportion of missing values, we first removed missing x, y, or band data, reducing the sample size from 742 to 736. For the remaining missing data in cohort and age (100 observations each), we attempted several approaches including Generalized Linear Mixed Models (GLMMs) with multiple imputations (5 imputations) and maximum likelihood estimation (MLE). While the Negative Binomial GLMM showed reasonable overdispersion (ratio = 0.961), diagnostic checks revealed potential issues. Given that cohort, year, and age are mathematically related, and considering the challenges in imputing these values, we just retain only the year variable and drop cohort and age columns. Thus, these steps allowed us to maintain data integrity and properly deal with missing values.

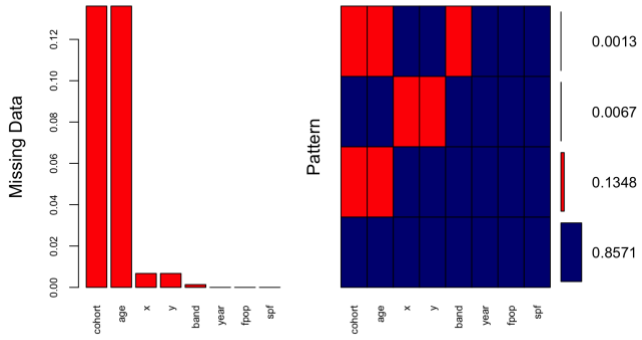

Figure 1: Missing Values Proportion for 'cohort', 'age', 'x', 'y' and 'band'

- Confounding Variables

The potential confounding variables revealed a complex pattern of relationships in this study. While the correlations between numeric predictors remained weak (all $|r| < 0.13$) and statistically significant for year-x: r=0.090, p=0.015; spf-x: r=-0.103, p=0.005; x-y: r=0.126, p<0.001, the inclusion of band ID as a random effect impacted the variance inflation factors. High VIF values ranging from 2.55 to 42,235.08 indicated significant multicollinearity when accounting for sparrows, suggesting that the band strongly influences the relationships among predictors. Besides, the dataset included 356 unique birds with an average of 2.07 observations per bird (range: 1-7) and showed substantial individual variation in reproductive success (SD = 2.079 offspring). Spatial effects on female population size were non-significant when controlling for individual sparrows (x: p=0.822; y: p=0.113), while time effects remained strong (year estimate = 3.41, p<2e-12). These patterns suggest that while the basic predictors show minimal direct confounding, the repeated observations introduce important dependencies. The substantial variation and the high VIF values support a mixed-effects modeling approach for this analysis.

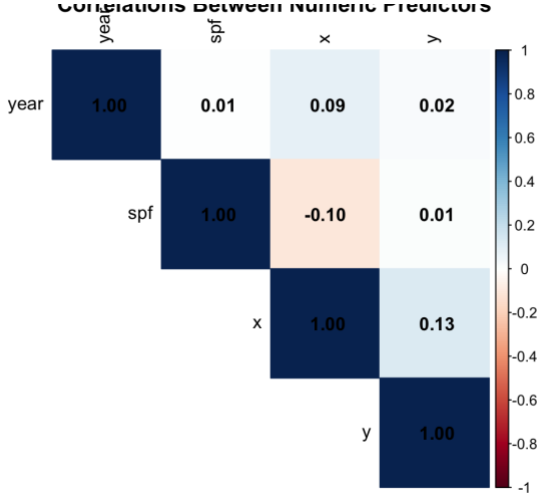

Figure 2: Correlations Between Numeric Predictors

Data description

The sparrow population data exhibits several important sources of heterogeneity across multiple scales. At the individual level, there is substantial variation among the 356 unique birds in the study, with each bird being observed between 1 and 7 times (mean = 2.07 observations per bird). Moreover, the female population size (fpop) shows large variation, ranging from 4 to 72 birds, with a median of 52.67 and large variation within birds (median SD = 8.08). The spatial distribution of nests, represented by scaled x and y coordinates, shows significant clustering effects, though their direct influence on population size is minimal (x: $p = 0.644$, y: $p = 0.871$ in fixed effects).

Moreover, time heterogeneity is also evident in the analysis, with a significant positive trend in population size over time (year_scaled estimate = 1.75, $p = 0.011$). Firstly, the mixed-effects model comparison includes random slopes for year effects ($\Delta AIC = 2.3$, $p = 0.041$), indicating that sparrow responds differently to time changes. Secondly, the random effect's structure reveals substantial individual variation (intercept SD = 6.73) and time variation (year slope SD = 5.80), with a moderate positive correlation between intercepts and slopes ($r = 0.447$). Finally, the model explains 36.7% of the total variance (conditional $R^2 = 0.367$), though fixed effects alone account for only 1.4% (marginal $R^2 = 0.014$), highlighting the importance of individual-level variation. To sum up, this heterogeneity is further supported by the significant deviation from normality in random effects (Shapiro-Wilk $p < 0.001$) and the relatively high intraclass correlation coefficient (ICC = 0.358), indicating strong individual effects on population dynamics.

In addition, among the 356 unique birds in the dataset, 190 sparrows (representing 569 out of 742 total observations) showed inconsistent nesting locations across years, indicating that contrary to the initial assumption. This finding challenges the original study description, and this discrepancy is important to consider and should be mentioned as a limitation or unexpected finding in the study conclusions.

More information revealed significant heterogeneity. The spatial distribution of birds showed varying densities across quadrants, with the highest concentration in the Central-North region (59 birds in Central-East) and lowest in the North (2-12 birds per quadrant). In addition, there is a significant year-to-year variation ($F = 14.881$, $p < 2.2e\text{-}16$) and density dependence was evident in the negative relationship between female population size and offsprings ($\beta = -0.046$, $p < 4.85e\text{-}15$). This complex pattern of variation suggests that while population-level effects are important, individual behavioral differences play a crucial role in reproductive success.

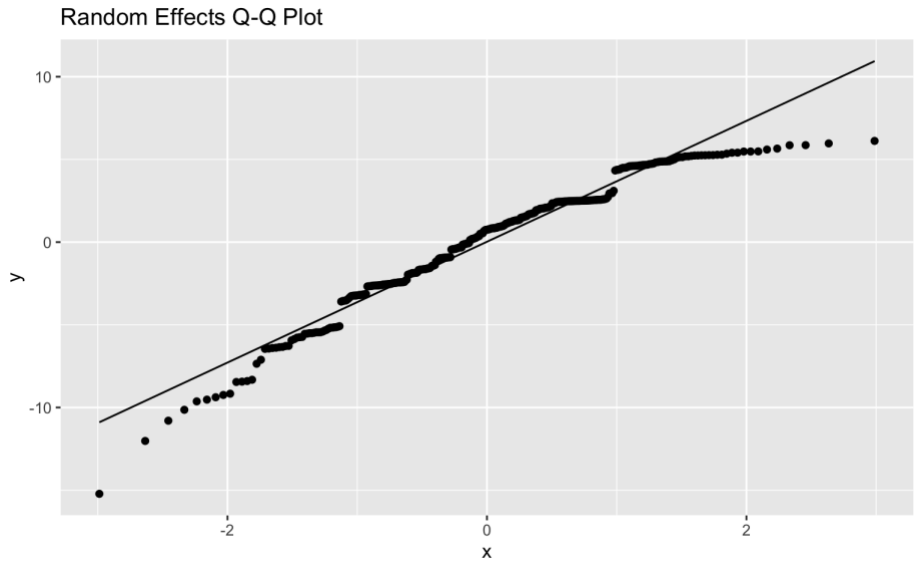

Figure 3: Random Effects Q-Q Plot: largely follows a normal distribution but deviates slightly at the extreme values, particularly in the upper tail.

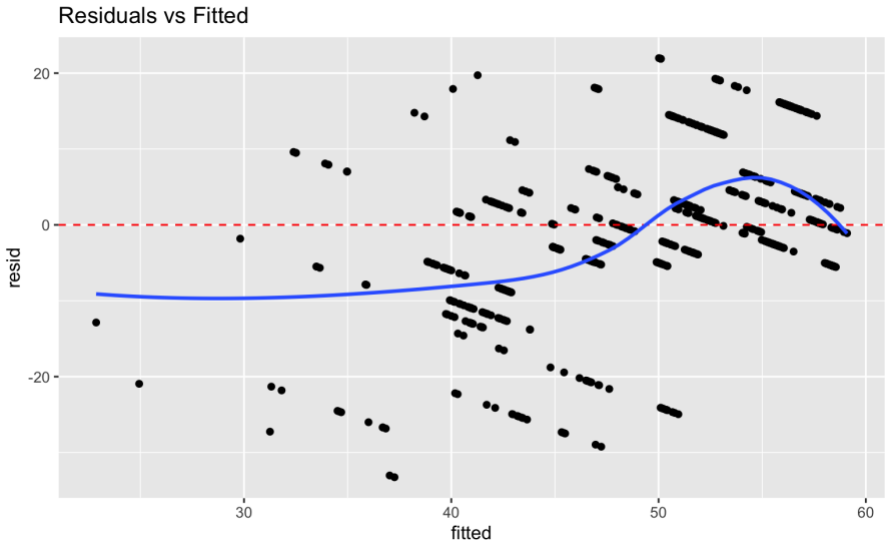

Figure 4: Residuals vs Fitted: a nonlinear pattern in residuals across fitted values, with systematic deviations suggesting potential model fit issues.

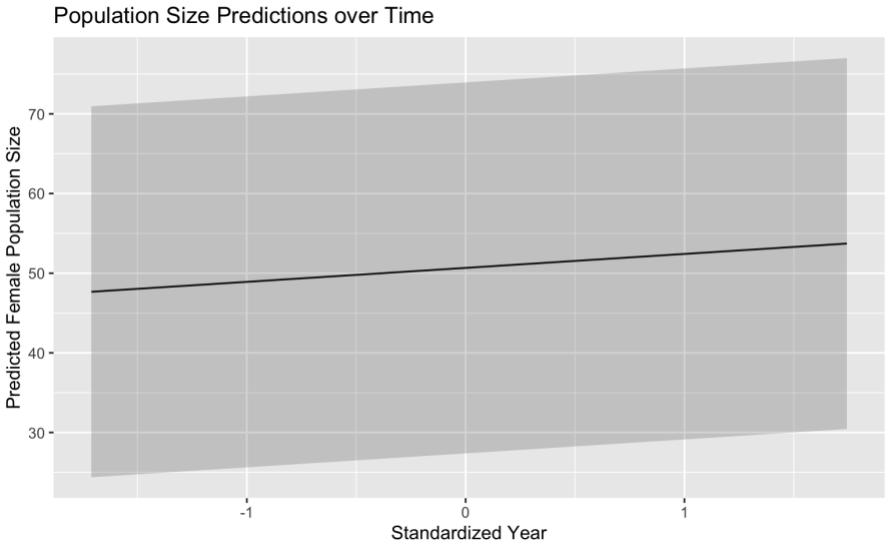

Figure 5: Population Size Predictions over Time: a slight upward trend in predicted female population size over time with widening confidence intervals, suggesting increasing uncertainty in longer-term predictions.

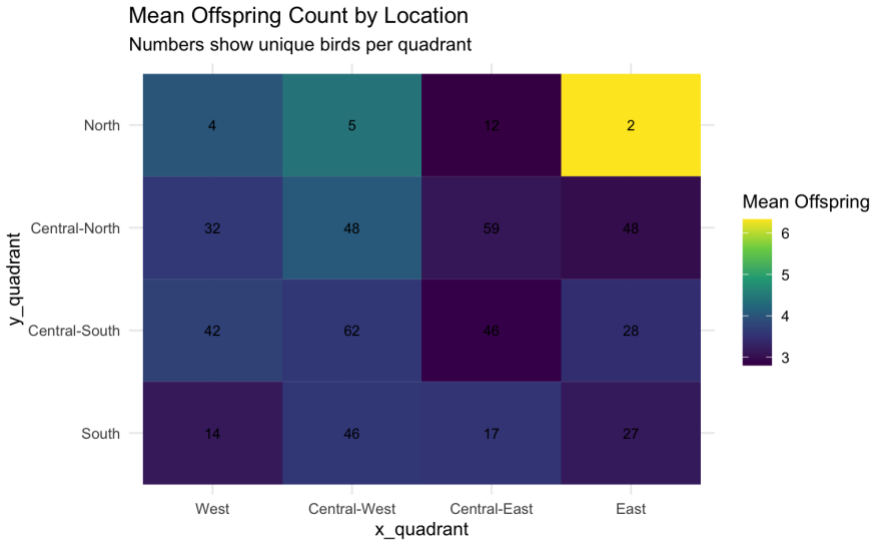

Figure 6: Mean Offspring Count by Location: the density of unique birds across different quadrants of the study area, with higher concentrations in the central regions.

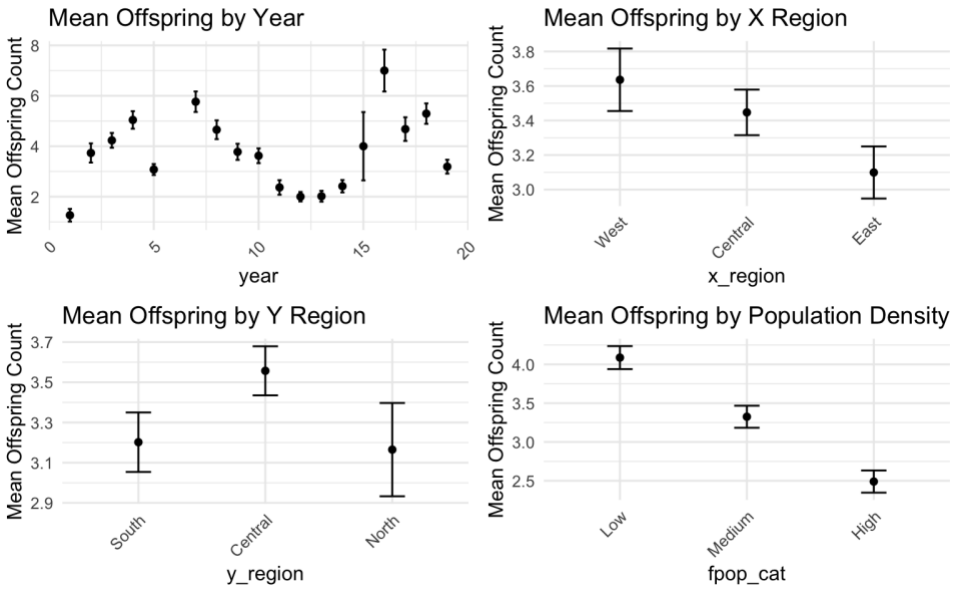

Figure 7: (a). Mean Offspring by Year: offspring counts over 20 years, with notable peaks around years 15-17. (b). Mean Offspring by X Region: slightly decreasing mean offspring counts from west to east regions. (c). Mean Offspring by Y Region: highest mean offspring counts in the central Y region compared to north and south. (d). Mean Offspring by Population Density: inverse relationship between population density and mean offspring count.

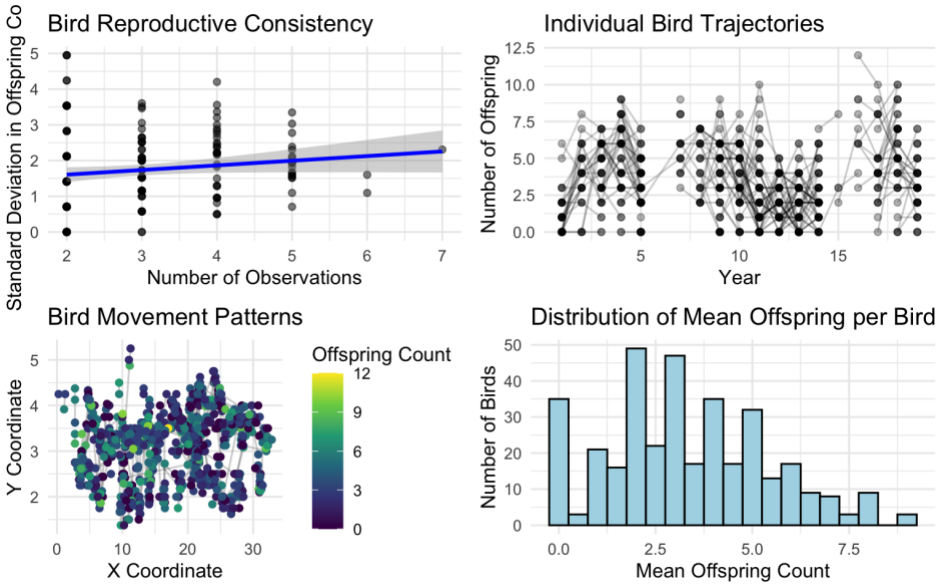

Figure 8: (a). Bird Reproductive Consistency: a slight increase in offspring count variation with more observations. (b). Individual Bird Trajectories: individual birds' offspring counts over time, showing high variability. (c). Bird Movement Patterns: bird locations with offspring count indicated by color intensity. (d). Distribution of Mean Offspring per Bird: the frequency distribution of mean offspring counts, with most birds having 2-4 offspring.

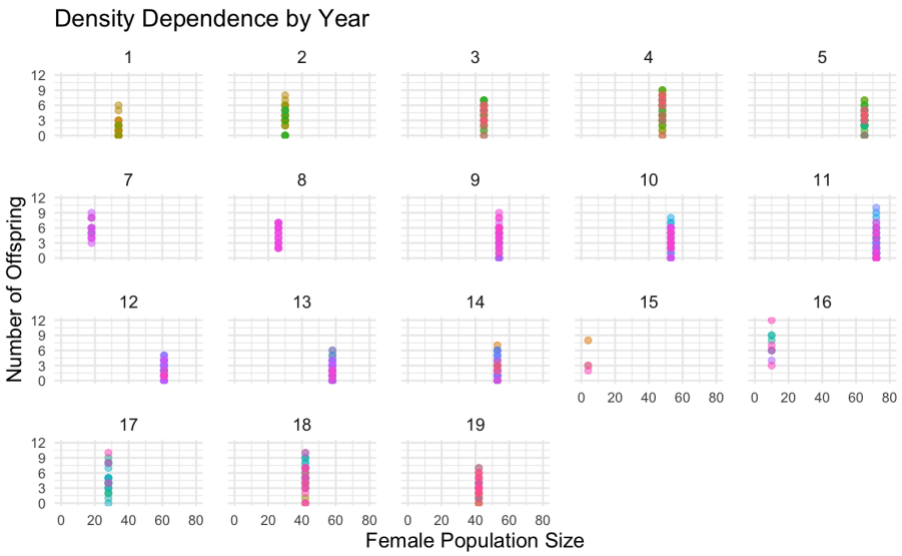

Figure 9: Density Dependence by Year: the relationship between female population size and number of offspring across different years.

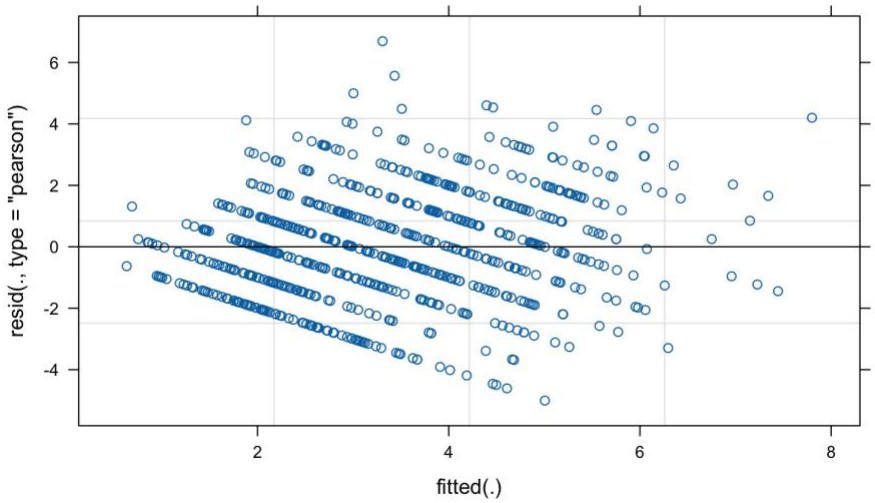

Figure 10: Model Residuals vs Fitted Values: a fan-shaped pattern in residuals, suggesting potential heteroscedasticity in the statistical model.

Model fitting and diagnostics

Firstly, we applied two models:

- m1: Basic random intercepts for year and bird ID.

- m2: Added random slope for population size by year.

- Thus, the more complex m2 model didn't significantly improve fit ($\chi^2 = 3.88$, $p = 0.144$), suggesting the simpler m1 model is adequate.

- Model Performance:

1. Female population size has a significant negative effect ($\beta = -0.70$, $p < 0.01$), indicating density dependence.

2. X-coordinate (east-west position) shows a significant negative effect ($\beta = -0.22$, $p < 0.01$).

3. Y-coordinate (north-south position) shows no significant effect ($\beta = 0.006$, $p = 0.94$).

4. Fixed effects explain about 10% of the variance (Marginal $R^2 = 0.098$).

5. The full model including random effects explains about 36% of the variance (Conditional $R^2 = 0.359$).

6. Moderate clustering by year (ICC = 0.156) and bird ID (ICC = 0.111).

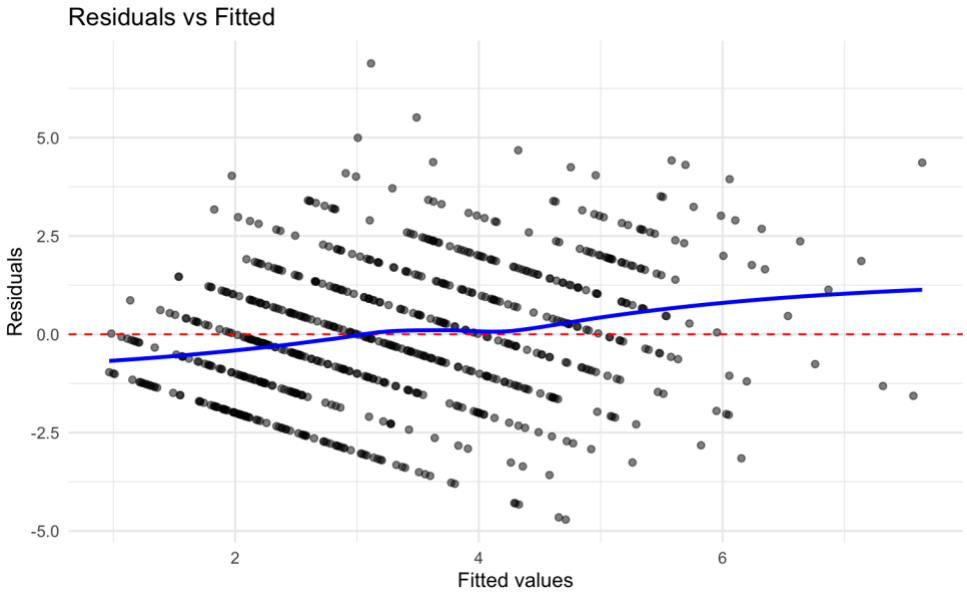

Figure 11: Assumption check: Residuals vs Fitted: distinct banding pattern suggesting potential Poisson-like count structure and slight upward trend in residuals indicates possible heteroscedasticity.

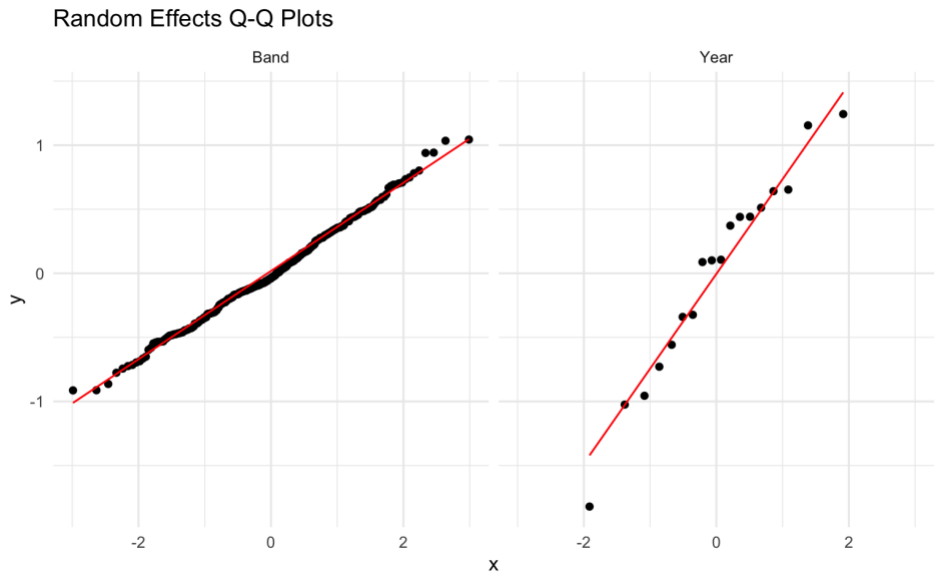

Figure 12: Assumption check: Random Effects Q-Q Plots: Both band and year random effects show normal distributions.

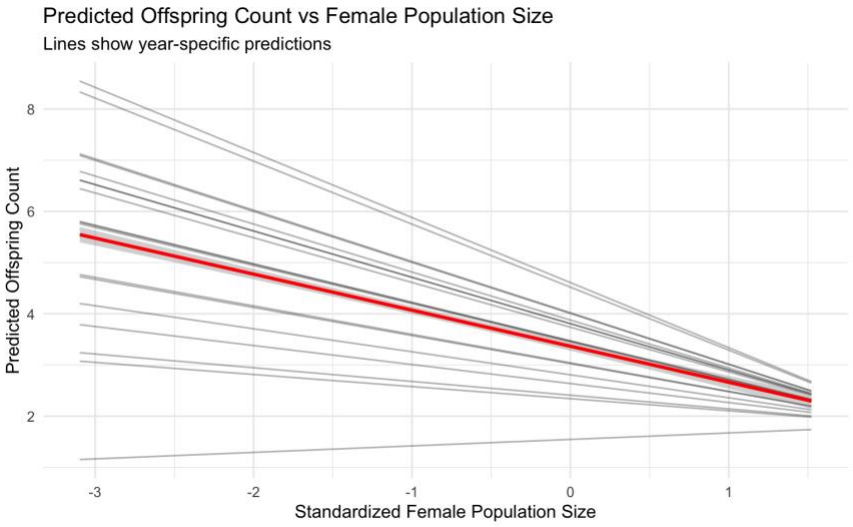

Figure 13: Assumption check: Model Predictions: consistent negative density dependence across years and year-specific slopes vary but maintain negative trend.

Due to the limitation for these 2 models, consider a generalized linear mixed model (GLMM) with Poisson or negative binomial distribution to select the best model.

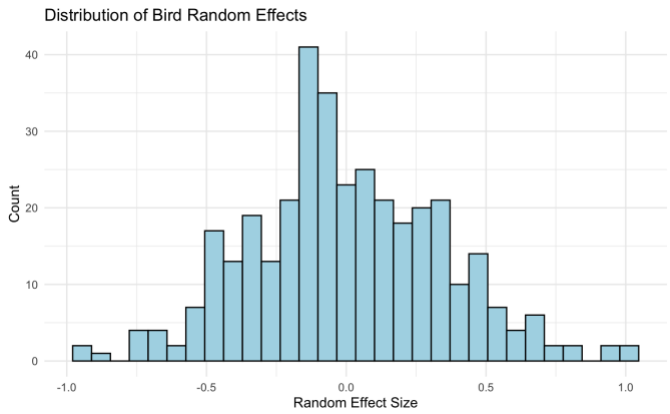

Figure 14: Distribution of Bird Random Effects: the distribution of random effect sizes for sparrow, centered near 0 with a slight negative skew and most values falling between -1 and 1.

Now, we added three more models:

- Negative Binomial GLMM (AIC: 3131.34).
- GAM with Spatial Smooth (AIC: 3150.36).
- Spatial Cluster Model (AIC: 3131.37).

- Improvements over the Previous Model:

1. Better handling of count data structure.

2. More appropriate error distribution and lower AIC value.

3. More stable parameter estimates.

Thus, the best model is Negative Binomial GLMM:

- Lowest AIC score (3131.34).

- More parsimonious (7 df vs 81.95 for GAM).

- Key Findings from Best Model:

1. Fixed Effects:

- Significant negative effect of population size ($\beta = -0.183$, $p < 0.01$).

- Significant east-west location effect ($\beta = -0.067$, $p < 0.01$).

- Non-significant north-south effect ($\beta = 0.003$, $p = 0.90$).

2. Random Effects:

- Moderate year-to-year variation ($\sigma = 0.318$).

- Lower individual bird variation (σ = 0.230).

- Relatively low ICCs (year: 0.088, bird: 0.046).

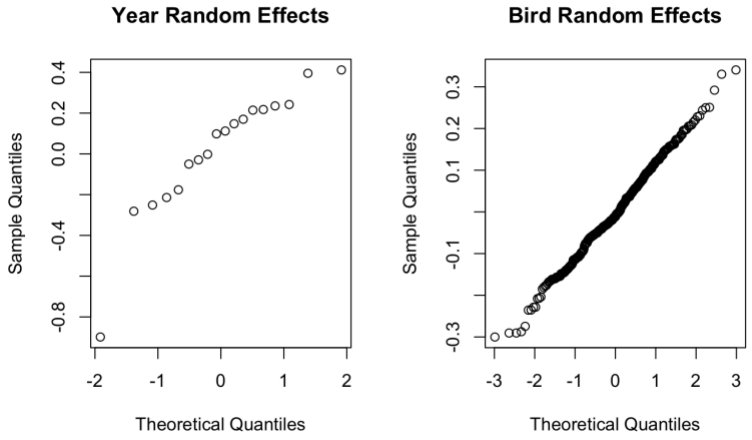

Figure 15: Random Effects Q-Q Plots: some deviation from normality (left) and normal distribution well (right).

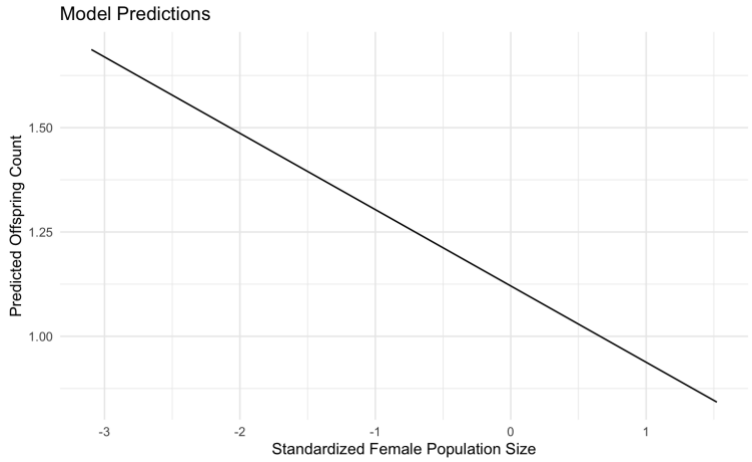

Figure 15: Model Predictions: Clear, appropriate and negative linear relationship between population size and offsprings.

However, there are also some limitations:

1. Relatively low ICCs suggest possible unmeasured important variables.

2. Some deviation from normality in year random effects.

3. Spatial effects might be oversimplified.

Model and data analysis interpretation

1. Female population size (fpop):

- Significant negative effect ($\beta = -0.183$, $p < 0.01$).

- For each standard deviation increase in female population size, the expected number of offspring decreases by about 16.7% ($\exp(-0.183) = 0.833$).

- Robust evidence for density-dependent reproduction despite model limitations.

2. Spatial Effects:

- East-West Location (x_scaled): Significant negative effect ($\beta = -0.067$, $p < 0.01$).

  * About 6.5% fewer offspring per standard deviation eastward.

  * Effect remains significant even after accounting for spatial autocorrelation.

- North-South Location (y_scaled): No significant effect ($\beta = 0.003$, $p = 0.90$).

3. Residual Diagnostics:

- Non-uniform residuals (Kolmogorov-Smirnov test, $p = 0.024$).

- No significant overdispersion (dispersion ratio = 0.942, $p = 0.864$).

- Significant zero-inflation detected (ratio = 1.85, $p < 0.001$).

4. Spatial Structure:

- No significant spatial autocorrelation in residuals (Mantel test, $p = 0.987$).

- Model adequately captures broad spatial patterns.

- Some fine-scale spatial structures may be missed.

5. Random Effects:

- Year variation (SD = 0.318, ICC = 0.088).

- Individual bird variation (SD = 0.230, ICC = 0.046).

6. Primary Conclusions:

- Strong evidence for negative density dependence.

- Effect is robust across different models.

- Clear east-west gradient and no significant north-south pattern in breeding success.

- Spatial independence of residuals supports reliability of spatial effects.

- Year variable stronger than individual effects.

- Zero-inflation suggests some sparrows consistently fail to reproduce.

7. Model Limitations:

- Zero-Inflation.

- Non-uniform Residuals.

- Unexplained Variation.

From the analysis above and the plots below, the study primarily demonstrates that competition from other females is the strongest predictor of reproductive success, while location plays a secondary role. However, the presented results do not reveal age, a question that will need to be solved in the future.

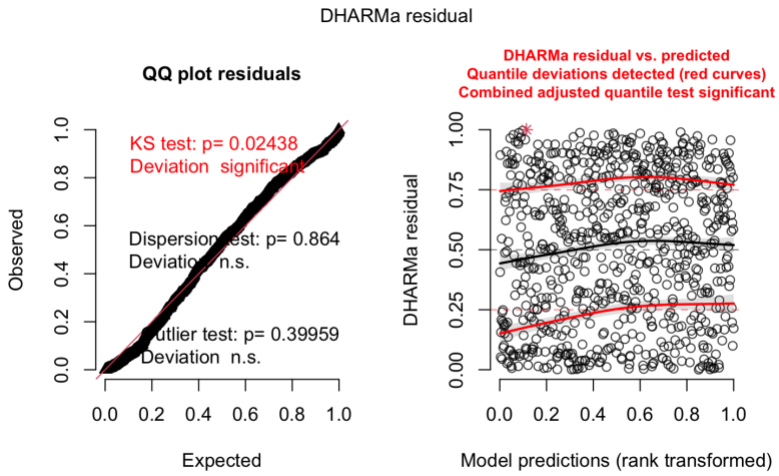

Figure 16: DHARMa Residual Plots: left panel shows a Q-Q plot of residuals against expected values with KS test results, while right panel displays residual values against model predictions with quantile curves.

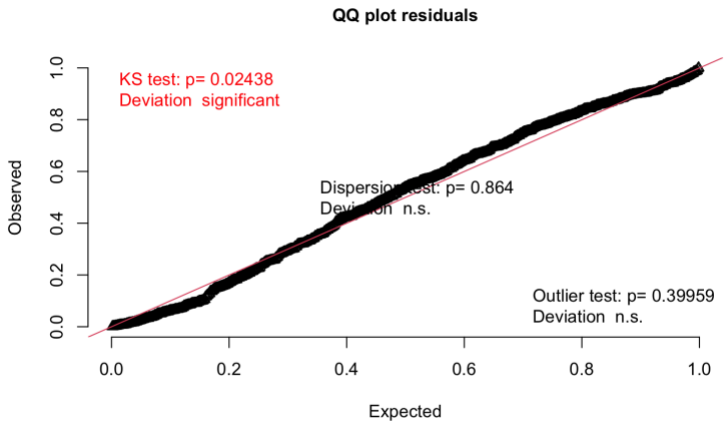

Figure 17: Q-Q Plot of Model Residuals: extended view of quantile-quantile plot comparing observed versus expected residual values, showing slight deviation from normality (KS test p=0.02438).

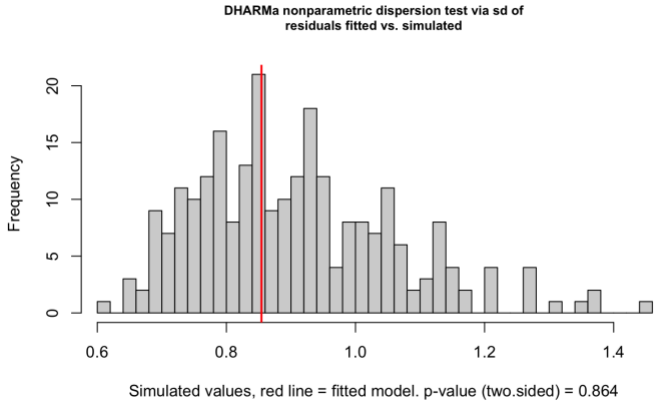

Figure 18: DHARMa Dispersion Test: distribution of simulated residual standard deviations compared to fitted model (red line), indicating no significant overdispersion (p=0.864).

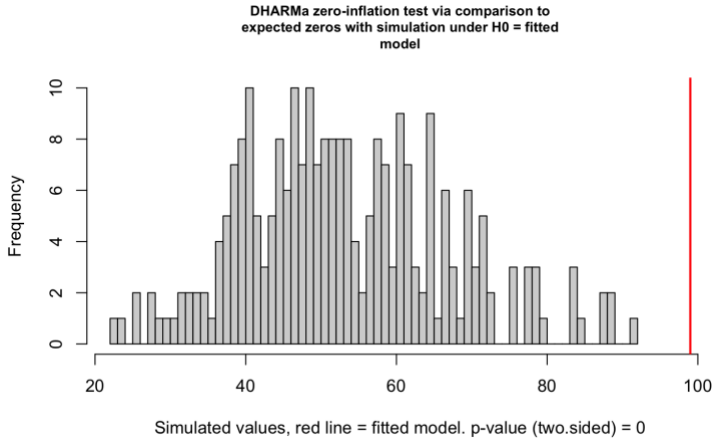

Figure 19: Zero-inflation Test Results: histogram comparing observed versus expected zero counts under the fitted model, showing significant zero-inflation (p < 0.001).

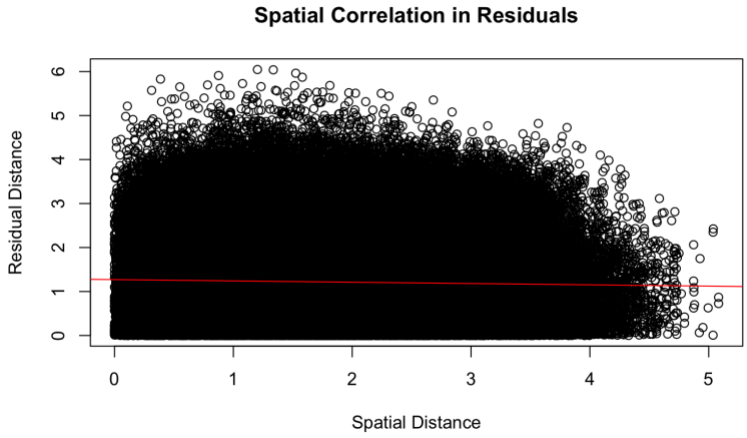

Figure 20: Spatial Correlation in Residuals: residual distances against spatial distances between observations, with a near-horizontal trend line indicating minimal spatial correlation.

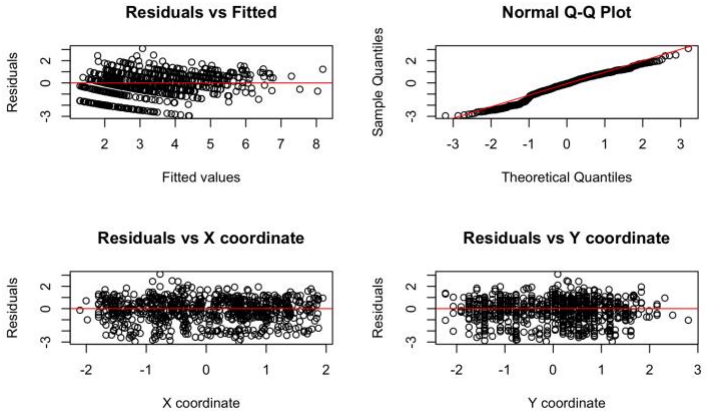

Figure 21: Model Diagnostic Plots: A four-panel diagnostic plot showing residuals versus fitted values (top left), normal Q-Q plot (top right), and residuals versus X and Y coordinates (bottom panels), used to assess model assumptions about residual patterns, normality, and spatial independence.

Conclusions

To sum up, this analysis provides strong evidence for female density population and habitat-specific effects on sparrow reproductive success. Based on the plots and tables, several key findings emerge:

1. The study reveals strong evidence for density-dependent effects (fpop) on reproduction, where a higher female population leads to reduced offspring, suggesting competition for resources or suitable nesting sites.

2. Location plays a significant role, with birds nesting in western areas producing more offspring than those in eastern areas, while north-south position shows no meaningful effect.
3. Substantial year-to-year suggests important environmental influences, while individual sparrow effects are relatively modest.

However, the model has several important limitations: zero-inflation indicates the model does not adequately capture the frequency of failed breeding attempts, and there is evidence of non-normal residuals despite using a negative binomial distribution. In addition, while spatial effects were included, the model may not capture fine-scale habitat variations. Therefore, future studies would benefit from incorporating more detailed habitat quality measurements, fewer missing values, explicit consideration of age, and investigation of weather.

Appendix

1. data analysis.Rmd: A comprehensive R-based analysis code of a song sparrow dataset, implementing statistical methods learned from course materials, R package documentation, and statistical programming resources.