A Hierarchical Model Analyzing Springbok Data

Zach White & Jialiang Mao

STA 723 Case Studies

February 14, 2017

Overview

Introduction and EDA

Model

Analysis

Objectives

- Monitoring programs aimed to track long-term changes in population size are crucial for applied ecological studies.

Objectives

- Monitoring programs aimed to track long-term changes in population size are crucial for applied ecological studies.
- We seek to understand trends in the springbok population.

Objectives

- Monitoring programs aimed to track long-term changes in population size are crucial for applied ecological studies.
- We seek to understand trends in the springbok population.
  - Are there systematic trends from year to year in the springbok population?
  - What about from site to site?

Objectives

- Monitoring programs aimed to track long-term changes in population size are crucial for applied ecological studies.
- We seek to understand trends in the springbok population.
  - Are there systematic trends from year to year in the springbok population?
  - What about from site to site?
- We want to understand the abundance of springbok.
  - Can we estimate the overall abundance of springbok antelope based on this data?

Objectives

- Monitoring programs aimed to track long-term changes in population size are crucial for applied ecological studies.
- We seek to understand trends in the springbok population.
  - Are there systematic trends from year to year in the springbok population?
  - What about from site to site?
- We want to understand the abundance of springbok.
  - Can we estimate the overall abundance of springbok antelope based on this data?
- Are there any covariates that affect the springbok populations living habits?

Data

- The data come from aerial surveys

Data

- The data come from aerial surveys
- The variables of interest

Data

- The data come from aerial surveys
- The variables of interest
  - `site`: the site of each watering hole. Out of 25 watering holes studied, we have data from sites 12-21, 23, and 24.

Data

- The data come from aerial surveys
- The variables of interest
  - `site`: the site of each watering hole. Out of 25 watering holes studied, we have data from sites 12-21, 23, and 24.
  - year: the year the observation was recorded. We have data from 1990-2012

Data

- The data come from aerial surveys
- The variables of interest
  - `site`: the site of each watering hole. Out of 25 watering holes studied, we have data from sites 12-21, 23, and 24.
  - year: the year the observation was recorded. We have data from 1990-2012
  - `date`: The week of year the observation was recorded.

Data

- The data come from aerial surveys
- The variables of interest
  - `site`: the site of each watering hole. Out of 25 watering holes studied, we have data from sites 12-21, 23, and 24.
  - year: the year the observation was recorded. We have data from 1990-2012
  - `date`: The week of year the observation was recorded.
  - `hour`: The number of hours past noon from the recorded date of the observation

Data

- The data come from aerial surveys
- The variables of interest
  - `site`: the site of each watering hole. Out of 25 watering holes studied, we have data from sites 12-21, 23, and 24.
  - year: the year the observation was recorded. We have data from 1990-2012
  - `date`: The week of year the observation was recorded.
  - `hour`: The number of hours past noon from the recorded date of the observation
  - `counts`: The number of Springbok Antelope at the given watering hole.

Data

- The data come from aerial surveys
- The variables of interest
  - `site`: the site of each watering hole. Out of 25 watering holes studied, we have data from sites 12-21, 23, and 24.
  - year: the year the observation was recorded. We have data from 1990-2012
  - `date`: The week of year the observation was recorded.
  - `hour`: The number of hours past noon from the recorded date of the observation
  - `counts`: The number of Springbok Antelope at the given watering hole.

Data

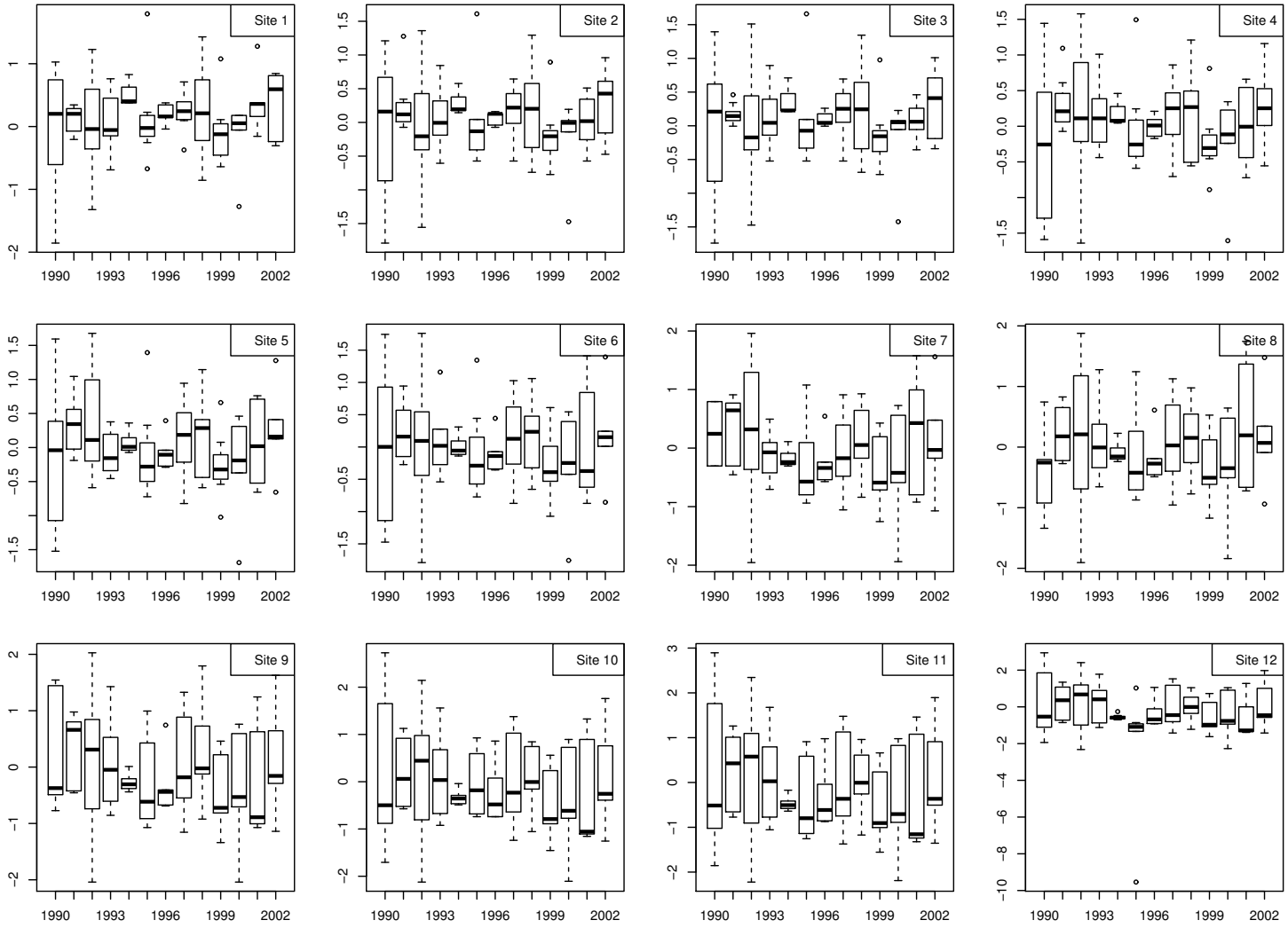

Figure 1: HourFromNoon by site

Data

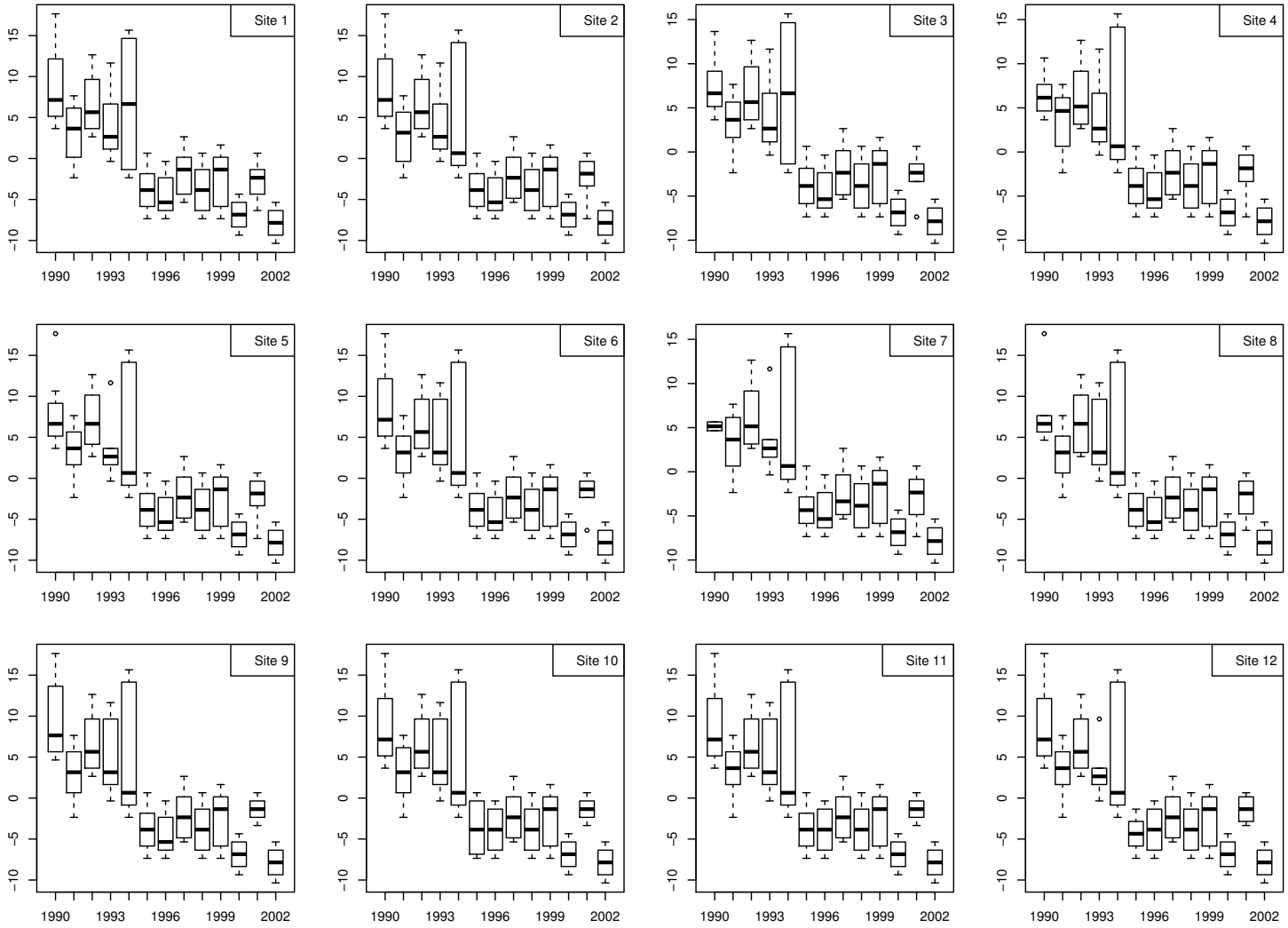

Figure 2: Date by site

Data

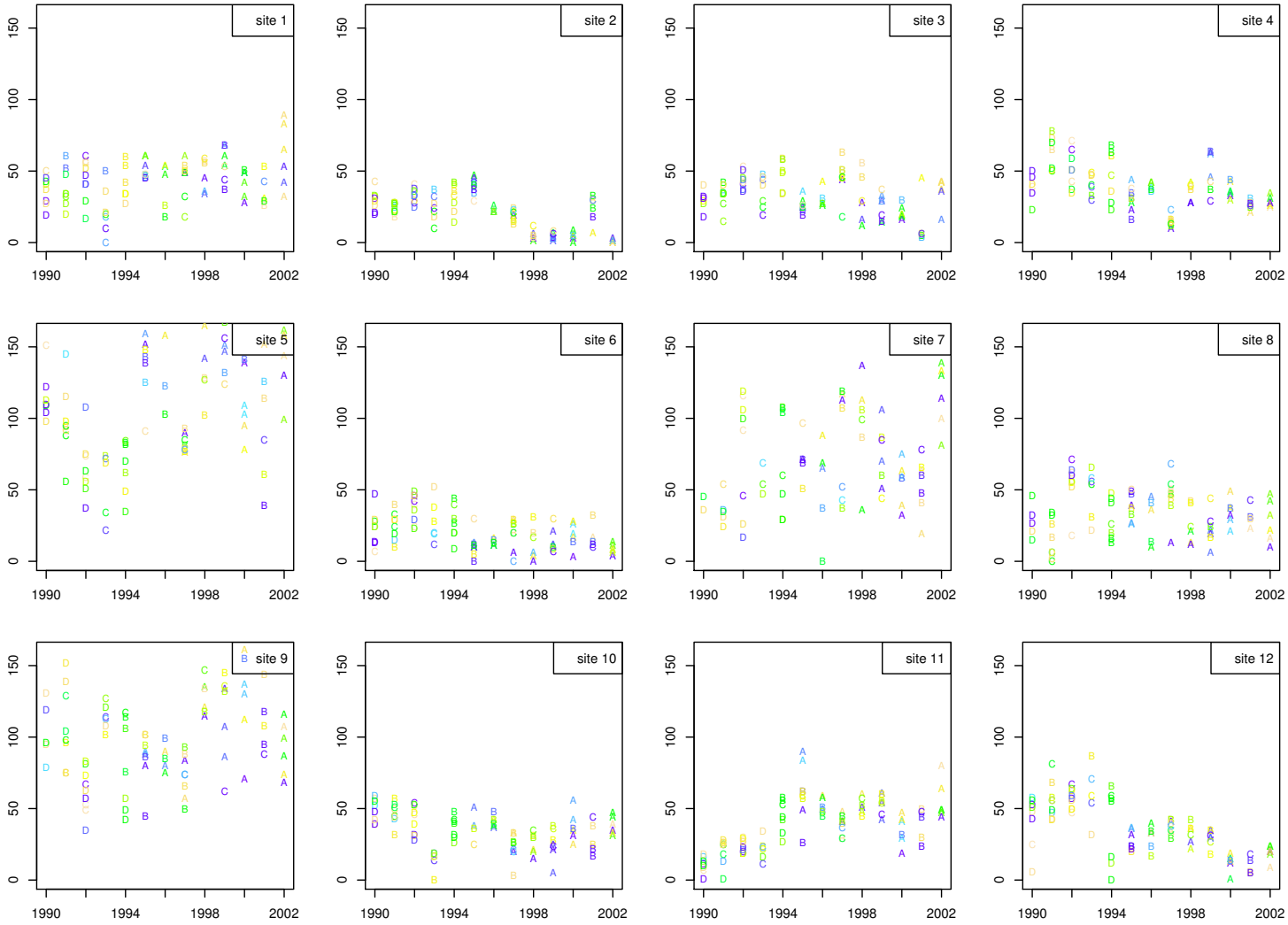

Figure 3: Data by site.

EDA

- Clearly, we should include both site and year in the model.
- For data in a given site of a given year, the larger the mean, the larger the variance.
- No clear trend - may need site-year interaction.
- Whether include `DATE` and `HourFromNoon` is not clear.
- The sampling scheme changed around 1995.

DATE	14-19	20-24	25-28	28-42
Group	A	B	C	D

Table 1: Quartiles of DATE.

EDA

Let Y_{ijt}, d_{ijt}, h_{ijt} denote the j-th `counts`, `date` and `HourFromNoon` at site i in year t, respectively.

$$Y_{ijt} \sim \text{Poi}(\mu_{ijt})$$

$$\begin{aligned}
\text{Model 1}: &\quad \log(\mu_{ijt}) = \mu + \alpha_i + \beta_t \\
\text{Model 2}: &\quad \log(\mu_{ijt}) = \mu + \alpha_i + \beta_t + \nu_{it} \\
\text{Model 3}: &\quad \log(\mu_{ijt}) = \mu + \alpha_i + \beta_t + \nu_{it} + \gamma d_{ijt} + \delta h_{ijt}
\end{aligned}$$

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	1026	10380.63			
2	894	5287.14	132	5093.50	0.0000
3	892	4901.69	2	385.44	0.0000

Table 2: ANOVA.

Model based on the sampling scheme

Let Y_{ijt}, d_{ijt}, h_{ijt} denote the j-th `counts`, `date` and `HourFromNoon` at site i in year t, respectively.

$$
\begin{aligned}
N_{it} &\sim \mathrm{Poi}(\mu_{it}) \\
Y_{ijt} &\sim \mathrm{Binom}(N_{it}, p_{ijt}) \\
g(p_{ijt}) &= \mu + \alpha_i + \beta_t + \gamma d_{ijt} + \delta h_{ijt}
\end{aligned}
$$

where g is some link function.

- The inference is cumbersome. (need to marginalize out N_{it}.)
- With limited data points in each site each year, it is hard to quantify the uncertainty.

Bayesian hierarchical log linear model

$$
\begin{aligned}
Y_{ijt} &\sim \mathrm{Poi}(\mu_{ijt}) \\
\log(\mu_{ijt}) &= \mu + \alpha_i + \beta_t + \gamma d_{ijt} + \delta h_{ijt} \\
\alpha_j &\overset{\text{i.i.d.}}{\sim} N(\mu_\alpha, \sigma^2_\alpha) \\
\beta_t &\overset{\text{i.i.d.}}{\sim} N(\mu_\beta, \sigma^2_\beta)
\end{aligned}
$$

let $\alpha_1 = \beta_1 = 0$ for identifiability. Use flat priors for μ, γ, δ, $\mu_\alpha, \mu_\beta, \sigma^2_\alpha, \sigma^2_\beta$.

Model with interaction

$$
\begin{aligned}
Y_{ijt} &\sim \mathrm{Poi}(\mu_{ijt}) \\
\log(\mu_{ijt}) &= \mu + \alpha_i + \beta_t + \nu_{it} + \gamma d_{ijt} + \delta h_{ijt} \\
\alpha_j &\overset{\text{i.i.d.}}{\sim} N(\mu_\alpha, \sigma_\alpha^2) \\
\beta_t &\overset{\text{i.i.d.}}{\sim} N(\mu_\beta, \sigma_\beta^2) \\
\nu_{it} &\overset{\text{i.i.d.}}{\sim} N(\mu_\nu, \sigma_\nu^2)
\end{aligned}
$$

let $\alpha_1 = \beta_1 = 0$ for identifiability. Use flat priors for μ, γ, δ, $\mu_\alpha, \mu_\beta, \mu_\nu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\nu^2$.

MCMC diagnostic

We run the chain for $N = 20000$, with a burn-in $B = 5000$.

- Traceplots.
- Geweke.

Overdispersion?

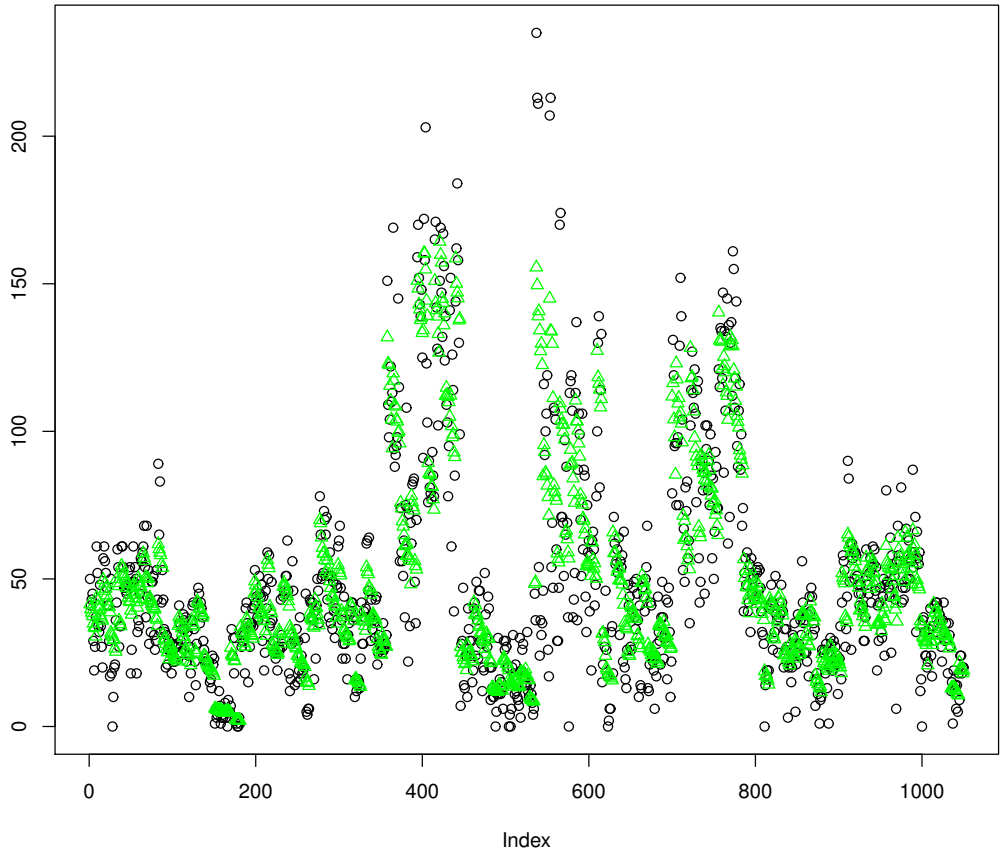

Figure 4: Data vs estimated.

Residuals

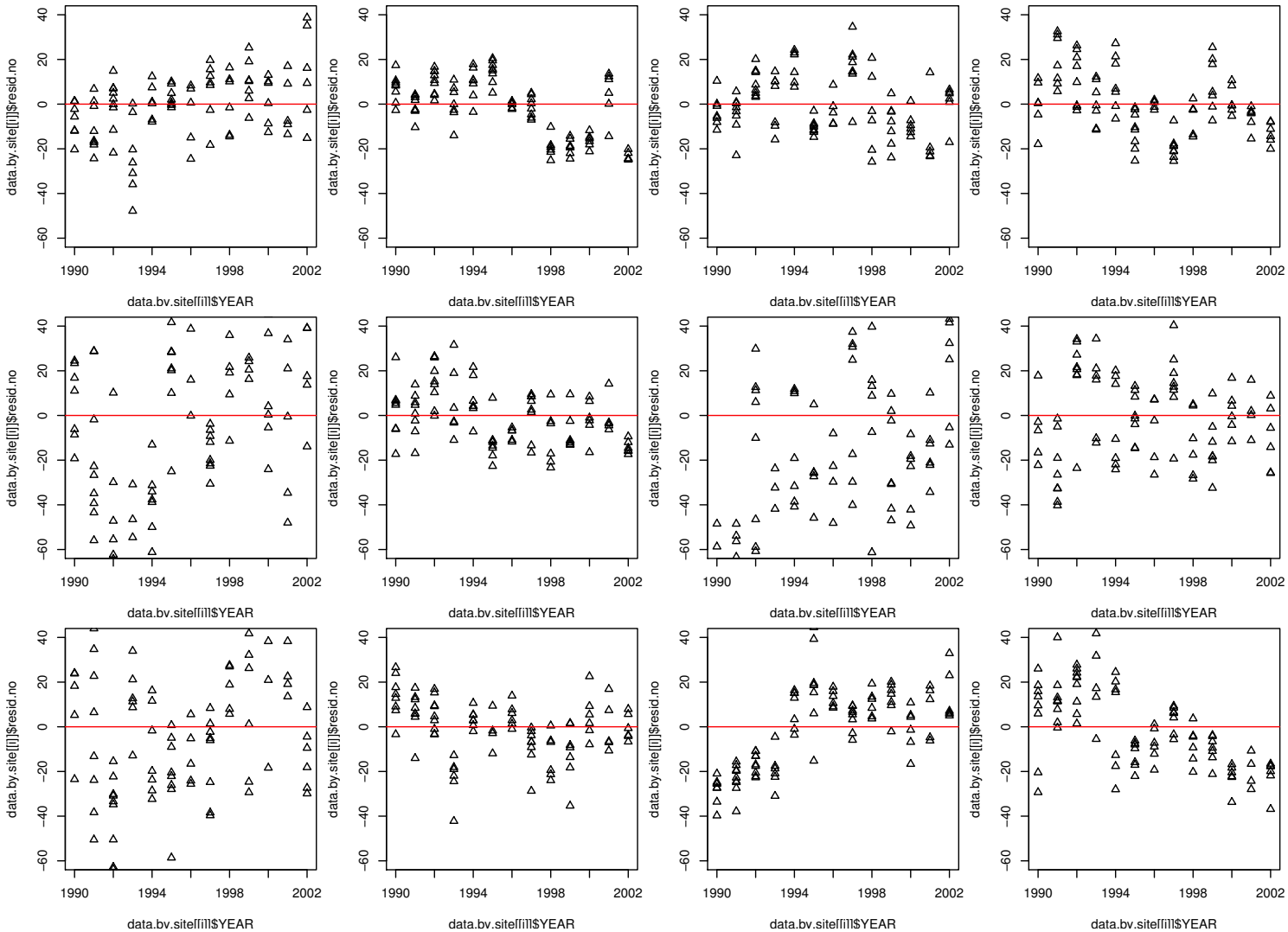

Figure 5: Residuals (without interaction)

Residuals

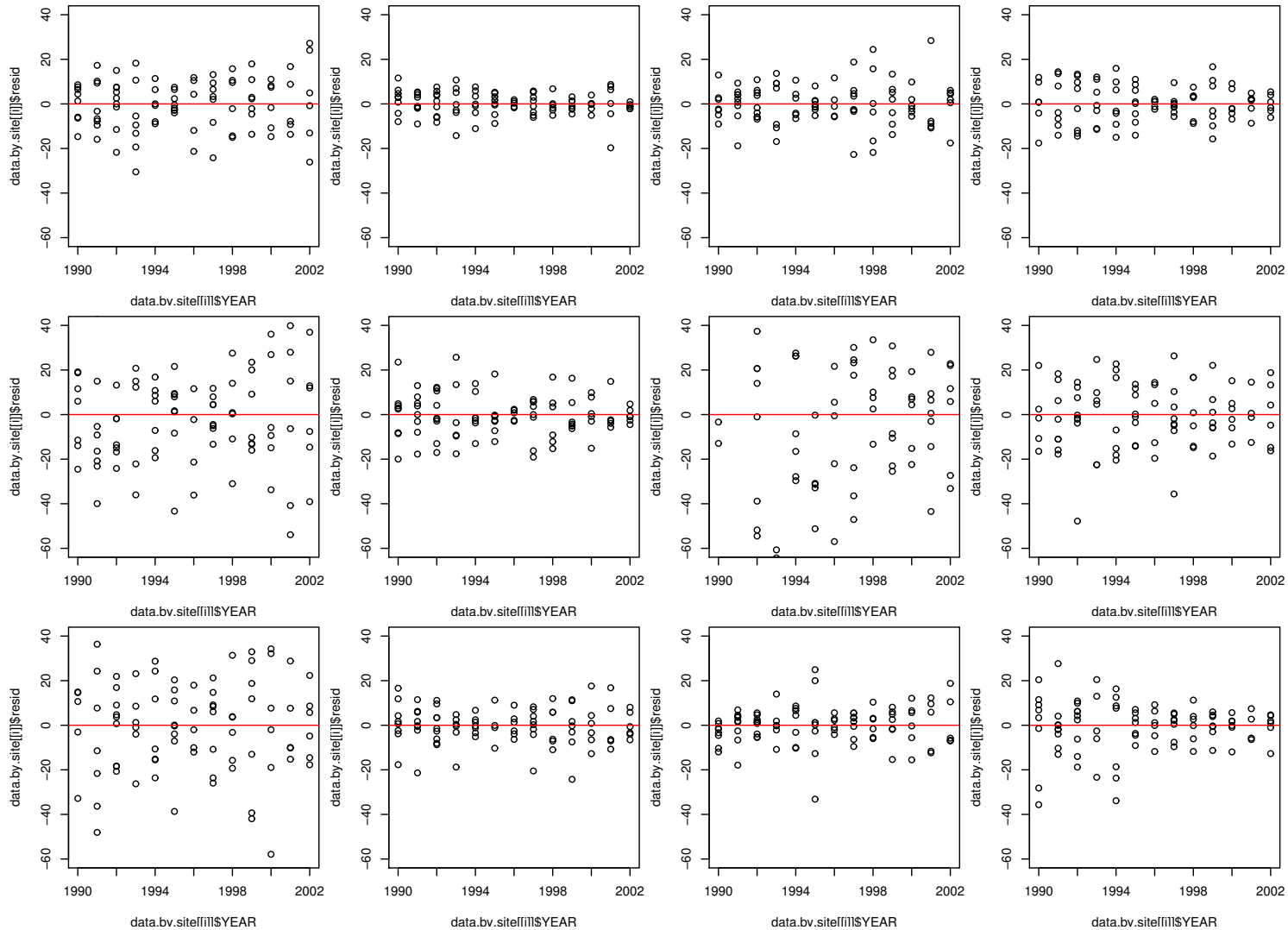

Figure 6: Residuals (with interaction)

Posterior summaries

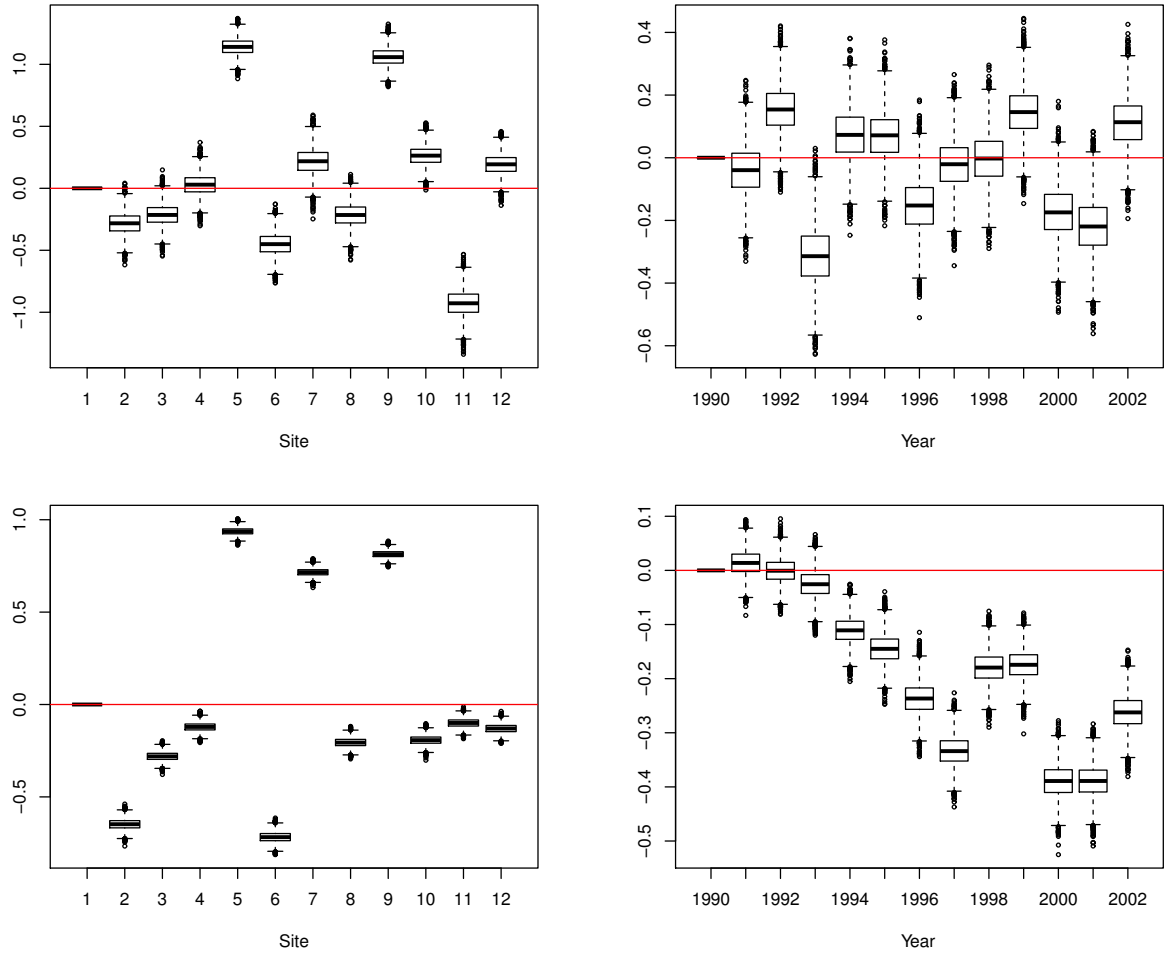

Figure 7: Posterior samples. (Up: with interaction; Bottom: without interaction.)

Posterior summaries

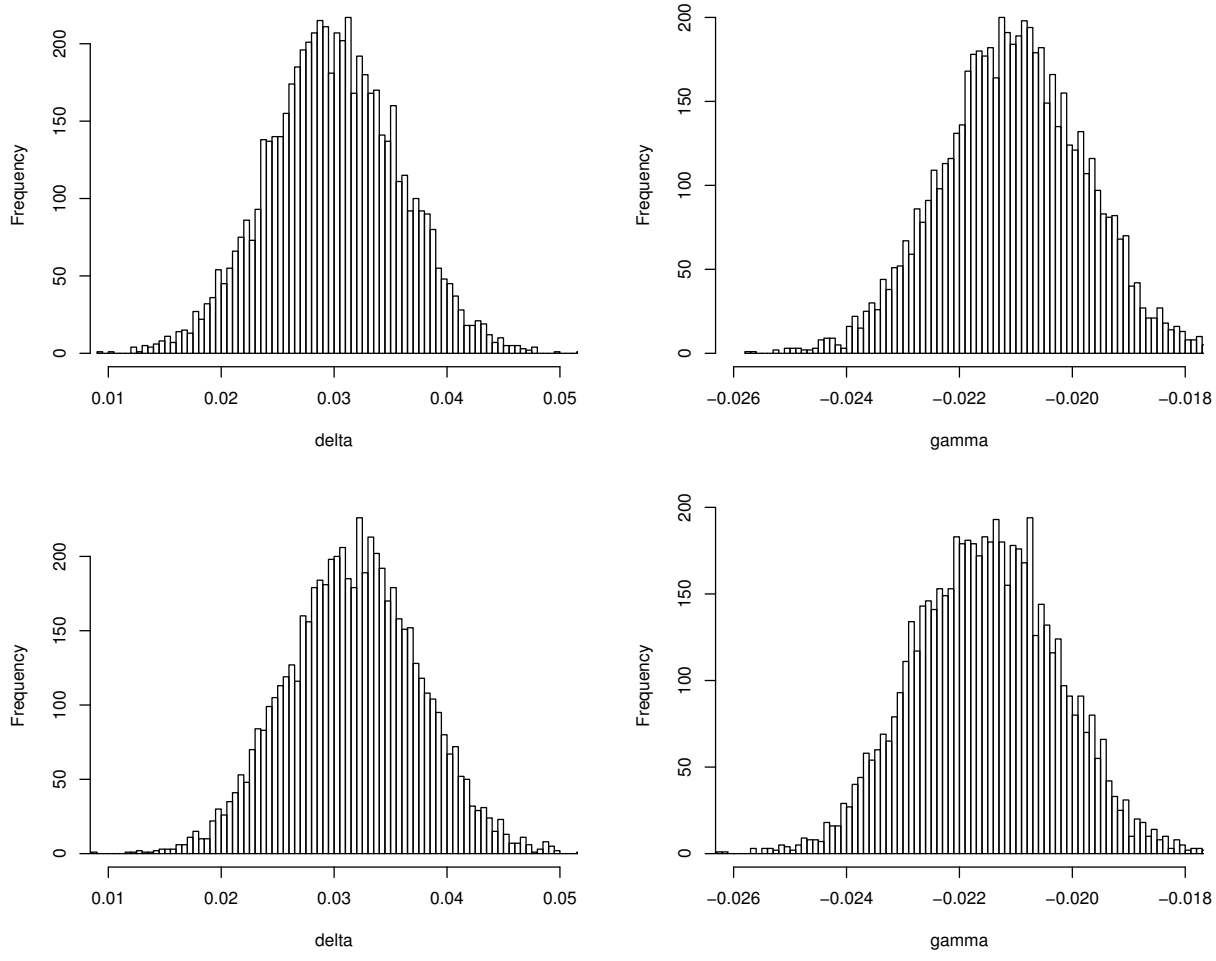

Figure 8: Posterior samples. (Up: with interaction; Bottom: without interaction.)

Model Fitting

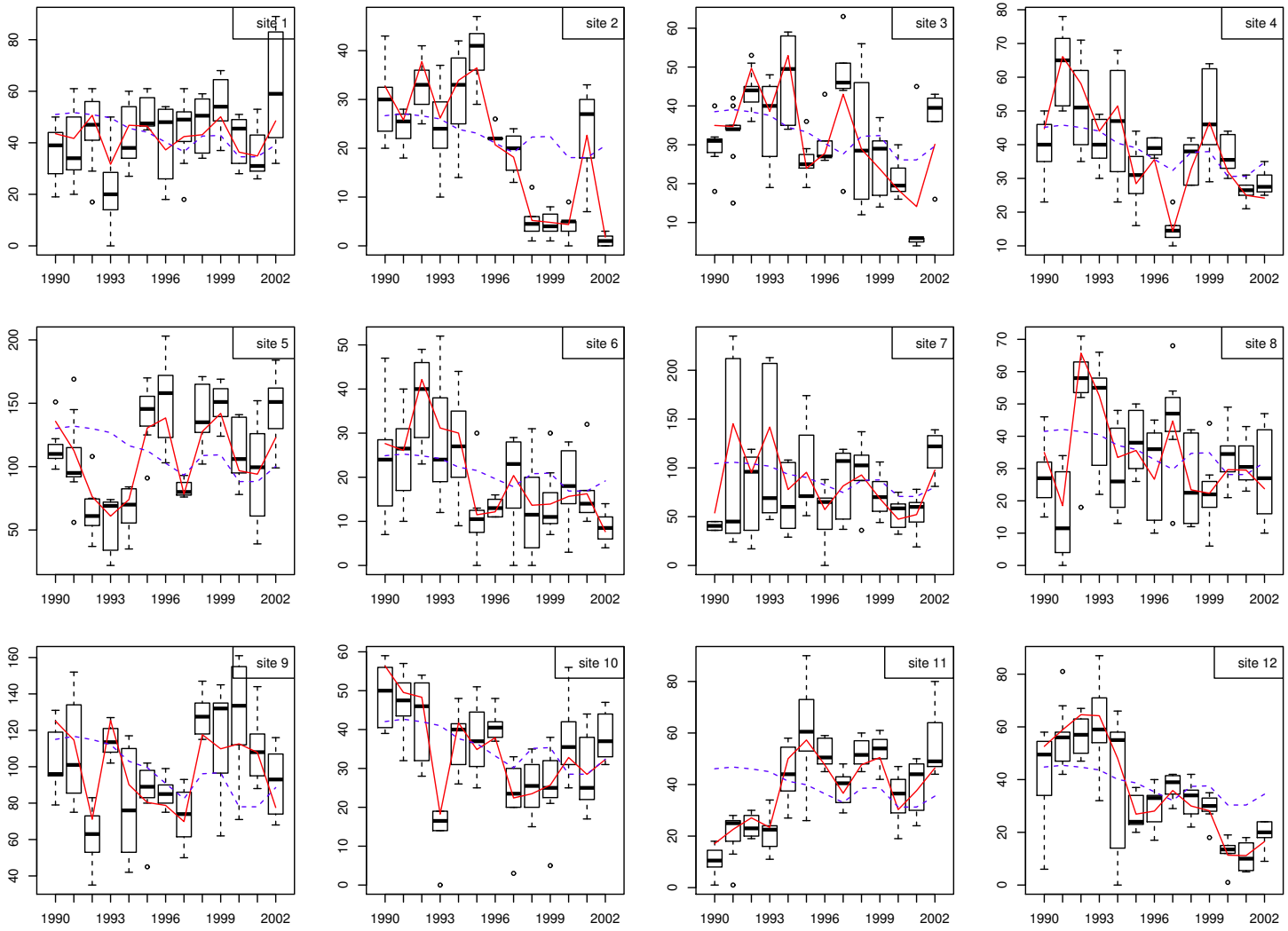

Figure 9: Posterior means.

Time Series

- No clear trend. (We have site-year interactions in our model.)
- Two time-relevant covariates. Different sampling schedule each year. How to define the time unit?

Conclusion

- Are there systematic trends from year to year in the springbok population?
  - Without the interaction, there does seem to be a semi-consistent negative trend.
  - With the interaction, there may be a cyclical trend.
- What about from site to site?
  - Under both models, sites 5 and 9 seem to be associated with higher springbok levels.
- Can we estimate the overall abundance of springbok antelope based on this data?
  - Under our model, this is not possible. However, if we were to use the other possible model we discuss, it could be possible.
- Are there any covariates that affect the springbok populations living habits?
  - Under both models, `date` and `hour` both have an impact.