Estimating Heterogeneity in Mosquito-Bird Contact Using Approximate Bayesian Computation

Senior Thesis Project: Biology 4F06

Jordyn Walton

McMaster University
2018/04/30

Contents

1 Abstract

The vector-host dynamics in West Nile Virus (WNV) are influenced by extreme heterogeneous contact between mosquitoes and several host bird species. The heterogeneous transmission that results can increase the spread rate of the epidemic and may have an effect on the total epidemic size [7],[9]. Here, employing approximate Bayesian computation (ABC), we attempt to quantify and model the heterogeneous contact between mosquitoes and individual birds within a species in numerical simulations. Better information about heterogeneous contact can improve the modelling of WNV and may be generalizable to other vector-borne diseases. A sampling design is presented that can provide occupancy information about bird populations using DNA sequenced from mosquito blood meals. First, a simple model, which assumes mosquito-feeding probabilities are homogeneous, is investigated using ABC. ABC was successful here, achieving precision and accuracy in estimating its parameter (bird population size); it was not found to be robust when the homogeneity assumption was violated. Next, a model using beta-distributed mosquito-feeding probabilities was investigated. Here, ABC was employed to approximate the joint density of the beta models population-size parameter, and a variance-related parameter, θ. The ABC density for bird population achieved improved accuracy compared to ABC under the homogeneous feeding assumption, but was still unreliable and unable to estimate populations greater than 35. This point estimates otherwise had moderate bias and MSE and credible intervals, which would in most cases cover the true value more than 75% of the time. Using ABC to estimate θ was not successful. Point estimates from ABC were biased and remained almost unchanged from the mean of the prior distribution . Lastly, the posterior density of the heterogeneity, here defined as the coefficient of variation (CV) in mosquito-feeding probabilities was approximated. Due to the unsuccessful estimation of θ, an alternate low-dimensional summary statistic was explored: the moments of the occupancy spectrum. Both procedures were tested based on simulations with heterogeneity where bird population size was kept consistent and in a small range (20-25 birds). Using the moments as a summary statistic was found to improve the performance of the procedure. As a result, it became more accurate and precise in estimating CV with smaller bias and MSE. Still, ABC using moments showed notable bias and MSE when estimating the more extreme values in the prior distribution. A possible explanation was suggested: the prior overwhelming the data. The ABC method with the beta model using moments of the occupancy spectrum as a summary statistic may be a viable estimator of heterogeneity. However, further testing is required to improve the method for extreme values in the prior and determine what leads to overestimating/underestimating the heterogeneity in these cases.

2 Introduction

Extremely heterogeneous contact between mosquitoes and their host animals dominate the transmission of West Nile Virus and other vector-borne disease. The resulting heterogeneous rates of disease transmission can ultimately increase the spread rate of the epidemic and may have an effect on the total epidemic size ([9]). Better information about heterogeneous contact that occurs within species may be able to provide insight into modelling WNV transmission and possibly other vector-borne disease.

The pattern of contact between mosquitoes and birds has been found to be heterogeneous due to numerous factors. Kelly found that the pattern of contact between blood-sucking insects and their host animals is extremely heterogeneous and non-random: most hosts will be bitten relatively infrequently while a subset will consistently be heavily sampled [6]. The attraction of blood-sucking insects to humans and other animals has also been reported to vary with several demographic parameters, including pregnancy [8], as well as age, body size, sex, disease status, skin colour and blood type [6]. Heterogeneous contact can also result from both the host and insects behaviour [6].

The contact between mosquitoes and different bird species has been previously explored by Kilpatrick et Al. [7]. When mosquitoes feed on the blood of animals, they are able to conveniently keep a DNA record of the individuals on which they feed in the form of blood meals. Using the DNA sequence data in mosquito blood meals, Kilpatrick was able to gain the species information about the birds that were sampled (see Kilpatrick et al. 2006,[7] for a description of blood meal genotyping methods). What they found was that although WNV is found in over 300 species of birds, the American Robin, a single, relatively uncommon species was responsible for the majority of West Niles infectious mosquitoes. Employing the same technique as Kilpatrick could also provide information about the individual birds being bitten within a species and could potentially reveal information about heterogeneity that occurs within a single speciesor intra-specific heterogeneity. If DNA was sequenced across several mosquito blood meals, an occupancy spectrum of the birds bitten could be an informative tool in revealing heterogeneity within mosquito-bird contact.

The occupancy spectrum is a commonly used statistic when estimating populations that describes the number of individuals that were represented r times in a sample, for $r = 0, 1, 2, .$ For instance, abundance data where only one individual was caught several times would indicate a much smaller population than data where 50 individuals were each caught only once. The occupancy spectrum is a widely used tool connected to several disciplines such as cryptography (see Good-Turing for the Repeat Rate, [5]) and community ecology.

The occupancy spectrum is also expected to change in the presence of heterogeneous contact rates between birds and mosquitoes. Heterogeneous contact rates mean that some birds will be bitten more often than other birds. In an extreme example, in two populations known to be of equal size, the population experiencing heterogeneous contact could have an occupancy spectrum where all mosquitoes bite the same bird. In contrast, a population with homogeneous contact may have a few mosquitoes bite the same birds and all others biting a distinct bird.

To represent this in a simple model, each bird was associated with a probability of being bitten and detected in a mosquito blood meal. If mosquito-bird contact was truly homogeneous, all birds in a population would have an equal probability of being bitten by a mosquito. Birds that experience heterogeneous contact would be associated with probabilities that have a larger coefficient of variation.

Because of the variation in representation of birds in the occupancy spectrum, heterogeneity can present a problem when trying to estimate the bird population size or the amount of heterogeneity itself. Due to practical limitations, not all individuals that are present will be detected. As a result, counts of population can are more likely to be inaccurate as many population estimators are based on the assumption that all individuals have an equal probability of being caught. This is analogous to the homogeneous case where all birds have an equal probability of being fed on. These estimators produce estimates that are negatively biased when there is actually heterogeneity present in capture probabilities, meaning they underestimate the actual population size([2]). In response, models without the equal-catchability restriction have been proposed such as Chaos lower bound estimator ([2]), Dorazio and Royles estimator [4] which model individual capture probabilities using a beta distribution, and Coull and Agrestis estimator which models capture probabilities as the logit of a normal distribution [3].

Similarly, the amount of heterogeneity may not be realistically estimable if the size of the population is truly unknown. The occupancy spectrum may appear similar in cases where the population is relatively small to the case where the population is larger but experiences highly heterogeneous contact rates with

mosquitoes. For simplicity, we will largely consider the case where either the amount of heterogeneity or the population size is treated as a known variable or known to constrict to a relatively small range of values. Trying to estimate both variables at the same time is a much more complex situation where variables can introduce confounding effects.

Before a method can be taken to practice, it is important to explore more theoretical considerations. Using approximate Bayesian computation, we will be able to approximate the density of bird population size and heterogeneity. Approximate Bayesian computation (ABC), a recent statistical technique, has been developed to infer parameters from complicated scenarios where a likelihood function cant be derived analytically or is too computationally costly to evaluate. The use of approximate Bayesian computation is ideal in the case of bird population and heterogeneity since several factors concerning mosquito-bird contact are unknown and an exact likelihood function has not be determined (see Beaumont, 2010 for more description [1]). ABC involves comparing the summary statistic of real data to the summary statistic of simulated data. ABC will also perform better in the case when the summary statistic for the data is low dimensional. Here there may be a trade-off since the occupancy spectrum is higher dimensional when more mosquitoes are caught in a sample. An alternative low dimensional summary statistic is also used, the moments of the occupancy spectrum.

Purpose

The purpose of this thesis project is to explore heterogeneity in mosquito-bird contact. The objectives are to 1) evaluate ABC in estimating population size in a simple model with the homogeneous probability assumption given heterogeneous and homogeneous mosquito feeding patterns; 2) evaluate ABC in estimating population size, B, when mosquito feeding probabilities are generated by a beta distribution with parameters (B and θ); 3) attempt to quantify a heterogeneity metric, defined as the coefficient of variation in mosquito-feeding probabilities 1, with ABC using the occupancy spectrum; 4) attempt to quantify the coefficient of variation instead using the moments of the occupancy spectrum as a summary statistic

$$h(p_i) = \frac{\sigma}{\mu} = \sqrt{\frac{B-1}{(\theta+1)}} \tag{1}$$

3 Methods

The following methods rest on the assumption that an occupancy spectrum can be obtained by DNA sequencing mosquito blood meals.

Mosquito-bird contact was simulated when feeding probabilities (p_i) were homogeneousequaland when they were heterogeneousgenerated by a beta distribution parameterized by B and θ ($B = \alpha/(\alpha+\beta), \theta = \alpha+\beta$) where B is the bird population size and θ is a parameter inversely proportional to variance. The beta distribution has expected value, $E(p_i) = 1/B$, and variance given by $Var(p_i) = (B-1)/((B^2(\theta+1)))$. A metric of heterogeneity, h is defined as the coefficient of variation of mosquito-feeding probabilities which can be approximated by the coefficient of variation in a beta distribution:

$$CV = \frac{\sigma}{\mu} = \sqrt{\frac{B-1}{\theta+1}}$$

Each occupancy spectrum was generated by treating feeding as a multinomial model like M balls falling into B bins, where M is the number of mosquitoes and B is the number of birds.

We will also be exploring more optimistic scenarios where data is high volume i.e. large amounts of mosquitoes are collected.

3.1 Approximate Bayesian Computation

All simulations were performed in R (v3.2.2).

The first rejection-ABC procedure was used to compute the parameter B, the population size under the homogeneous probability assumption,

Given an abundance or incidence data observed for a population, y, the following is repeated for 1000 trials

1. Draw $B_i \sim \pi(B)$
2. Simulate an observation, $x_i \sim p(x|B_i)$
3. Reject B_i if $\rho(S(x_i), S(y) > \epsilon$

where $\pi(B)$ is the prior distribution of B, $S(\cdot)$ is the occupancy spectrum of the abundance data and ρ is the standard Euclidean distance between summary statistics for simulated and observed mosquito-bird contact and ϵ is a cut-off value that is predefined.

The remaining B_i that havent been rejected are samples from the posterior distribution of the parameter value that gave rise to the observation y. The resulting distribution can also be visualized using a density plot.

The general rejection procedure was applied across two cases, when the homogeneous probability assumption was true, meaning observation y was simulated using homogeneous feeding probability and another where the assumption was violated, meaning observation y was generated using beta distributed feeding probabilities with parameters being the true value of B and $\theta = 10$.

A cut-off value for ABC was defined using the following procedure. The mean value of the prior distribution was taken as a parameter. One realization of an occupancy spectrum was then simulated as observation y given the prior mean. Several realizations of an occupancy spectrum were then simulated repeatedly, $\{x_i, i = 1, 2, 3, \}$ with the same parameter, the prior mean. The distance measure $\rho(S(x_i), S(y))$ was then computed for each observations summary statistic $S(x_i)$ with observed summary statistic $S(y)$. Once the distances were each collected, a cut-off value was chosen based on the 10th percentile of all of the distances.

Once an approximate posterior distribution was generated, several statistics were gathered. The mean of the posterior distribution was used as a point estimate of the relevant parameter. A credible interval was also obtained from the 2.5 percentile and the 97.5 percentile. The relative bias was collected as well as a mean squared error. The credible interval width as well as its coverageif the true value was within the credible intervalwas also collected.

Modified ABC procedure for B and θ estimation using Occupancy Spectrum

The rejection procedure with the homogeneous assumption was modified to account for heterogeneous biting probabilities by estimating the joint distribution of B and θ.

The modified ABC-rejection procedure for heterogeneous biting probabilities is given by

1. Draw $B_i \sim \pi(B)$ and $\theta_i \sim \pi(\theta)$
2. Generate mosquito-feeding probabilities by taking B_i samples from a beta distribution (B_i, θ_i)
3. Simulate an observation, $x_i \sim p(x|B_i, \theta_i)$
4. Reject B_i and θ_i if $\rho(S(x_i), S(y) > \epsilon$

where $\pi(\theta)$ is the prior distribution of θ.

Modified ABC procedure for B and θ estimation using Moments

This above procedure was again modified, to replace the summary statistic with the moments of the occupancy spectrum, calculated as

$$\eta_n = \sum_{i=1}^{M} i^n \frac{S_i}{\sum_{i=1}^{M} S_i} \tag{2}$$

where M is the number of mosquitoes.

3.2 Choosing Priors

From data provided in Bolker et al., (2017) the population size of American robins at a site in Foggy Bottom or Baltimore was thought to be in the range of 10-30 birds. The prior distribution of B, which is the population size of a bird species, was chosen with this in mind. As a result, a negative binomial with size 6 and probability 0.2 was chosen in most cases. Here, the resulting sample represents the number of trials.

The joint prior distribution was assumed to be the product of the marginal prior distribution of B and θ. The prior marginal distributions were chosen differently depending on whether we were assessing the efficacy of estimating B or of estimating θ. When estimating B, the prior marginal distribution of θ was chosen to be thinner and spread over a smaller range in order to isolate the effect of B on the occupancy spectrum. A gamma distribution with shape 40 and scale 0.5 was chosen. When estimating θ, the prior marginal distribution of θ was chosen to be a gamma distribution with shape 5 and scale 2 (see prior A in figure 8). The prior distribution of B was also modified to be thinner and over a smaller range so again a negative binomial distribution was chosen, but with altered parameters, size 20 and probability 0.95 (see prior B in 8). See figure 8 in the appendix for these marginal prior distributions.

4 Results and Discussion

4.1 Estimating population size

The first aim of this project was to assess the efficacy of ABC in estimating population sizes given different mosquito feeding patterns.

First the ABC procedure that estimated B under the homogeneous probability assumption was investigated. We wanted to test how the procedure performs when the homogeneous probability assumption is true versus when the assumption is violated. Figure 1 shows an approximate posterior generated for B using an observed occupancy spectrum generated from A) homogeneous feeding probabilities ($\sigma/\mu = 0$) and B) from heterogeneous feeding probabilities ($\sigma/\mu > 0$).

In A) we see that the estimator performs well as the posterior includes the true value. However, in B) the estimator is negatively biased. This is a common issue with population estimators with the homogeneous probability assumption [2]. The posterior in (A) is also thinner and more precise.When assessing how ABC performs, it is also important to note how similar the distribution of the posterior is to the prior. If the prior distribution remains unchanged it may indicate that the summary statistic chosen for ABC does not reveal any information about the parameter we are trying to estimate. However, this is not the case, as the distribution of the posterior in A) and B) appears to be different than the distribution of the prior.

Next, we tested the modified ABC-rejection procedure that uses heterogeneous feeding probabilities for how well it performs when they are, in fact, heterogeneous. The marginal posterior distributions were computed for θ and B from the joint posterior and are used to provide point estimates and credible intervals.

Relative bias, relative MSE, coverage and credible interval width were collected over 100 simulations for various combinations of M and B. These are shown in Figure 2 for the parameter, B. The prior for θ used in this computation was centered over the smaller range of 5-18 (provided in 8) in order to isolate the effect of changing B on the occupancy spectrum as a summary statistic. In A), many estimations are shown to have a negative bias, meaning bird populations were underestimated, especially when the bird population was large to begin with ($B = 55, 45, 35$). When the bird population size is smaller (B= 15), the bias is positive and the population was overestimated. Also as M increases, bias seems to decrease in the case when $B = 15$ and $B = 25$ and plateaus in other cases. In B) the relative MSE has a similar trend and seems to decrease for the case $B = 15$ and $B = 25$. In other cases, the MSE plateaus. The MSE is largest (0.2) when the bird population is largest, in the case $B = 55$. In C), coverage is high in several cases, sometimes greater than 95 %, but lower in the case when $B = 55$ where the true value is in the credible interval approximately 25% of the time. In D), credible interval width is shown to increase with B and decrease with larger M in the cases where B is small ($B = 15, 25, 35$). In E), mean estimates are in the range of 22-30 when M is small ($M = 25$) and then are decreasing in the direction of the true value of B with M increasing. When B is larger, the mean estimate seems to plateau and not reach the true value. In both the cases when $B = 45$ and $B = 55$, the mean estimate seems to stop approaching the true value once it reaches 30.

In the heterogeneous/heterogeneous case, we see that ABC estimates are sometimes unreliable when M is very small and that estimates will be similar no matter the true value, staying in the range of $15 - 35$. The bias does seem to decrease as M increases, but often in cases where the true value of B is close to the mean of the prior distribution ($B = 15, 25$, prior mean = 20). Bias is expected to keep decreasing as M increases, but plateaus in several cases. This could reflect an issue with the ABC summary statistic. As M increases, the summary statistic becomes higher dimensional, and this can cause issues in the ABC procedure because it relies on taking the Euclidean distance between summary statistics. This could also indicate a need to more rigorously investigate how the ABC procedure is affected by the prior chosen for B and θ. The ABC procedure could be tested with a weaker prior. The coverage is also greater than 95% in several situations. This is an example of over-coverage, which is a common issue with ABC since the use of a prior allows for much more variation in the posterior than with conventional confidence intervals. The ABC procedure accounting for heterogeneous feeding seems more desirable because it has higher coverage than the procedure with the homogeneous feeding assumption, where the true value was completely out of the range of the posterior. Still further testing would be required

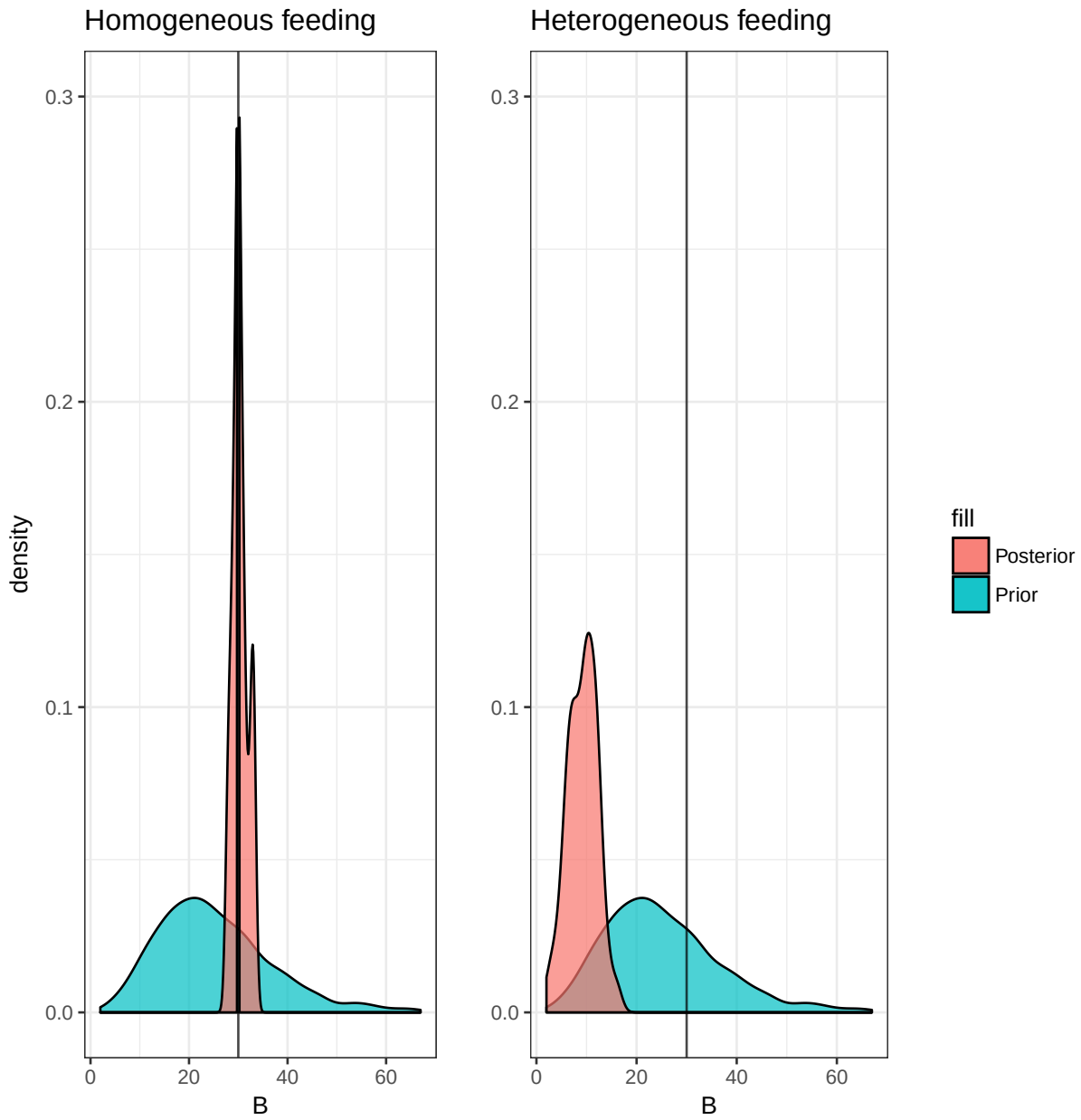

Figure 1: Prior (Posterior approximated by ABC with homogeneous probability assumption. True value of B drawn on black line (True B = 30). A) The homogeneous probability assumption is true. ($\hat{B} = 30.36$) B) Biting probabilities are heterogeneous. ($\hat{B} = 9.19$)

4.2 Quantifying heterogeneity

Lastly, since heterogeneity is important for population estimates and in studying host-vector contact but is not a readily available or enumerable quantity in practice, we want to be able to estimate it.

Figure 3 shows the model testing statistics for the heterogeneity parameter, θ. These results were obtained using this prior for B and θ (refer here to the Methods and to figure 8 in the appendix). From A) we see that θ was overestimated when it was small, when $\theta = 2, 5$. Here θ was overestimated to be 2 or 5 times its actual size. As θ increased this bias approached 0 and became negative. In B), the MSE is shown to be very large for $\theta = 2$, then to approach 0 when *theta* is larger. In C), coverage is high most of the time, greater than 95%, but much lower when $\theta = 20$. Coverage is 0 when $\theta = 2$. In D), credible interval width appears to approach 16 with larger M. In E) the estimates are shown all to be approximately 10, no matter the true value of θ being estimated. These coincide with the mean of the prior distribution (9.85)

Here, using the ABC-rejection procedure to estimate θ does not seem very promising. In 3 E), the estimates of θ dont seem to be sensitive to the true value of θ. The estimates also dont improve when M is large. The credible intervals do obtain high coverage, but not in the case of the relatively more extreme values (when $\theta = 2, 20$). Note coverage is 0 in the case when $\theta = 2$. This is due to the choice of prior distribution of θ because values of θ such as $\theta = 2$ have a lower density. Overall, gaining a credible interval or estimate of θ is sensitive to the choice of prior distribution. As M gets larger, estimates of θ seem to be dominated by the information held in the prior. Changing the prior in the ABC procedure could further test this hypothesis. A non-informative uniform prior may be beneficial. However, this could also indicate that the occupancy spectrum as a summary statistic does not reveal information about θ when M is larger. A possible explanation for this is the issue of occupancy spectrum tending to become higher dimensional when the number of *mosquitoes* is large.

Heterogeneity is further examined by computing the highest posterior density region of the joint density of B and θ and computing a posterior density of Heterogeneity (defined in 1) as a function of the joint density of B and θ.

A highest posterior density region of the joint distribution of B and θ is shown in figure 4. The posterior density region (shown in red) is much narrower over values of θ than that of the prior density (shown in black) which is much noisier and covers a smaller range of B. The true value ($B = 25, \theta = 10$) is in both the highest posterior density region of the prior joint density and that of the posterior joint density.

The prior for B used in this computation was centered over the smaller range of 20-27 (provided in 8) in order to isolate the effect of changing θ on the occupancy spectrum as a summary statistic. However keeping B relatively more consistent did not make the estimate of θ closer to the real value than the initial estimate given by the prior distribution. The credible interval outlined in red still contains the true values of B and θ, but has a large width in the θ direction, (~ 20) which makes it appear less precise.

As noted in 1, heterogeneity (σ/μ) can be computed as a function of B and θ. The prior and posterior densities of Heterogeneity were computed as a function of the joint densities of B and θ. This means that for each pair of observations (B, θ), the corresponding heterogeneity was computed using 1. A true value of heterogeneity was also calculated from the true values of B and θ. The density is provided in figure 5. In A) the posterior density appears to shift closer to the true value, which shows that estimating heterogeneity could be promising. However further testing is required.

Since several of the previous results seem to be affected by the summary statistic being too high dimensional when M is large, another summary statistic was also evaluated: the 1st, 2nd and 3rd moments of the occupancy spectrum. The posterior density was estimated using the moments as shown in 5B. Here, the posterior density shifts closer to the true value than the prior density. This could be promising, but again, further testing is required.

Relative bias, relative MSE, coverage, credible interval width and mean estimates were again collected over 100 simulations for various combinations of M and the coefficient of variation. First using the ABC procedure matching the whole occupancy spectrum. Then these statistics were collected using the ABC procedure matching the moments of the occupancy spectrum.

The performance statistics with the occupancy spectrum matching is shown in Figure 6. In A), the bias is shown to increase when M is large for the cases when $CV = 0.951, 1.09, 1.314$. Bias also increases with the true magnitude of the coefficient of variation. In B), the MSE seems to increase when the number of mosquitoes increases when the CV is smaller (0.951,1.09). Otherwise, the MSE plateaus as mosquitoes

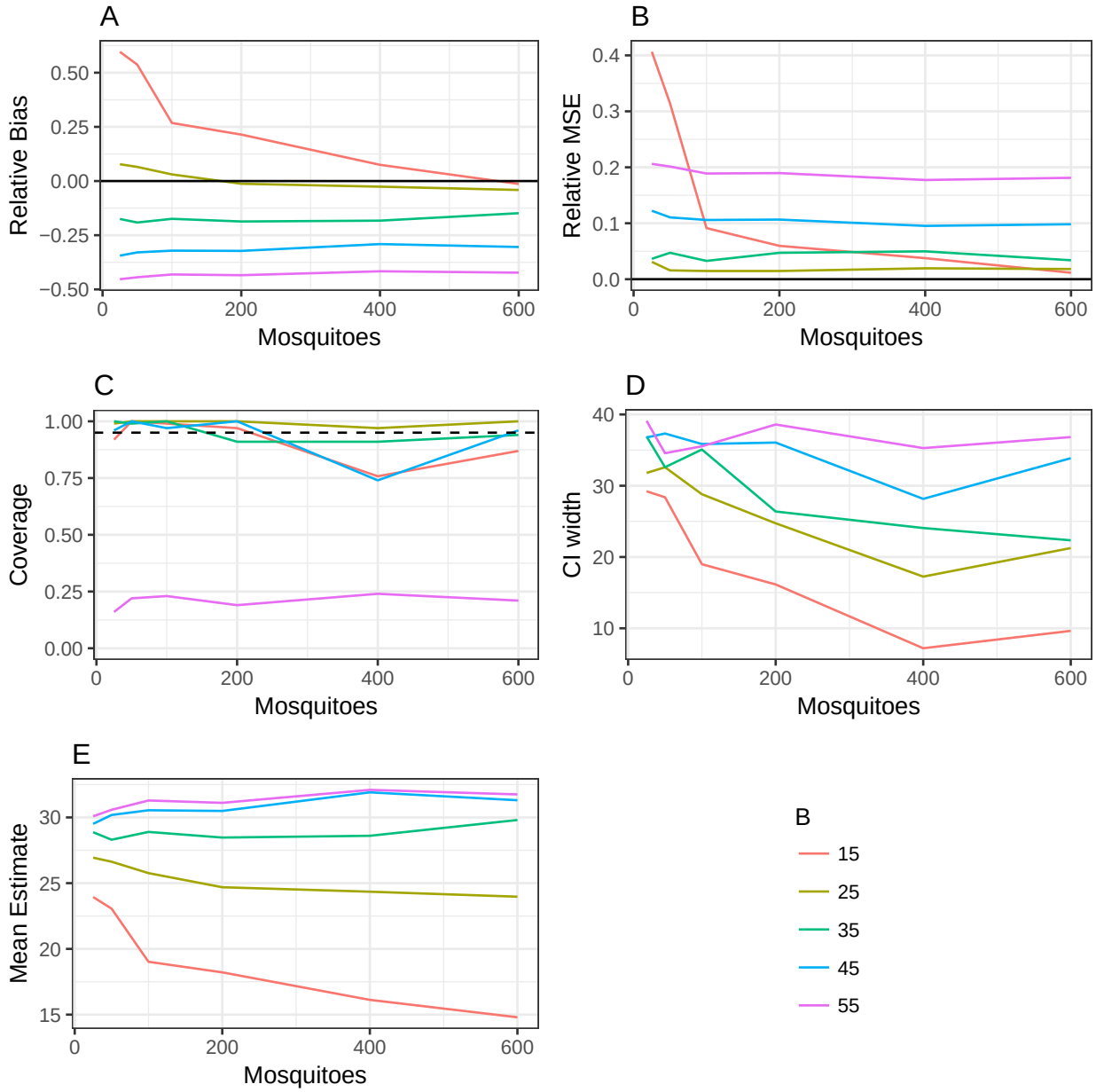

Figure 2: Estimation of bird population size with assumption of heterogeneous mosquito feeding probabilities. A) Bias, B) MSE and E) Mean estimate of $\hat{B}$, and C) Coverage and D) width of the corresponding credible interval generated from the approximated Posterior produced by ABC employed on occupancy spectrum modeled with beta distributed $(B, \theta = 10)$ capture probabilities.

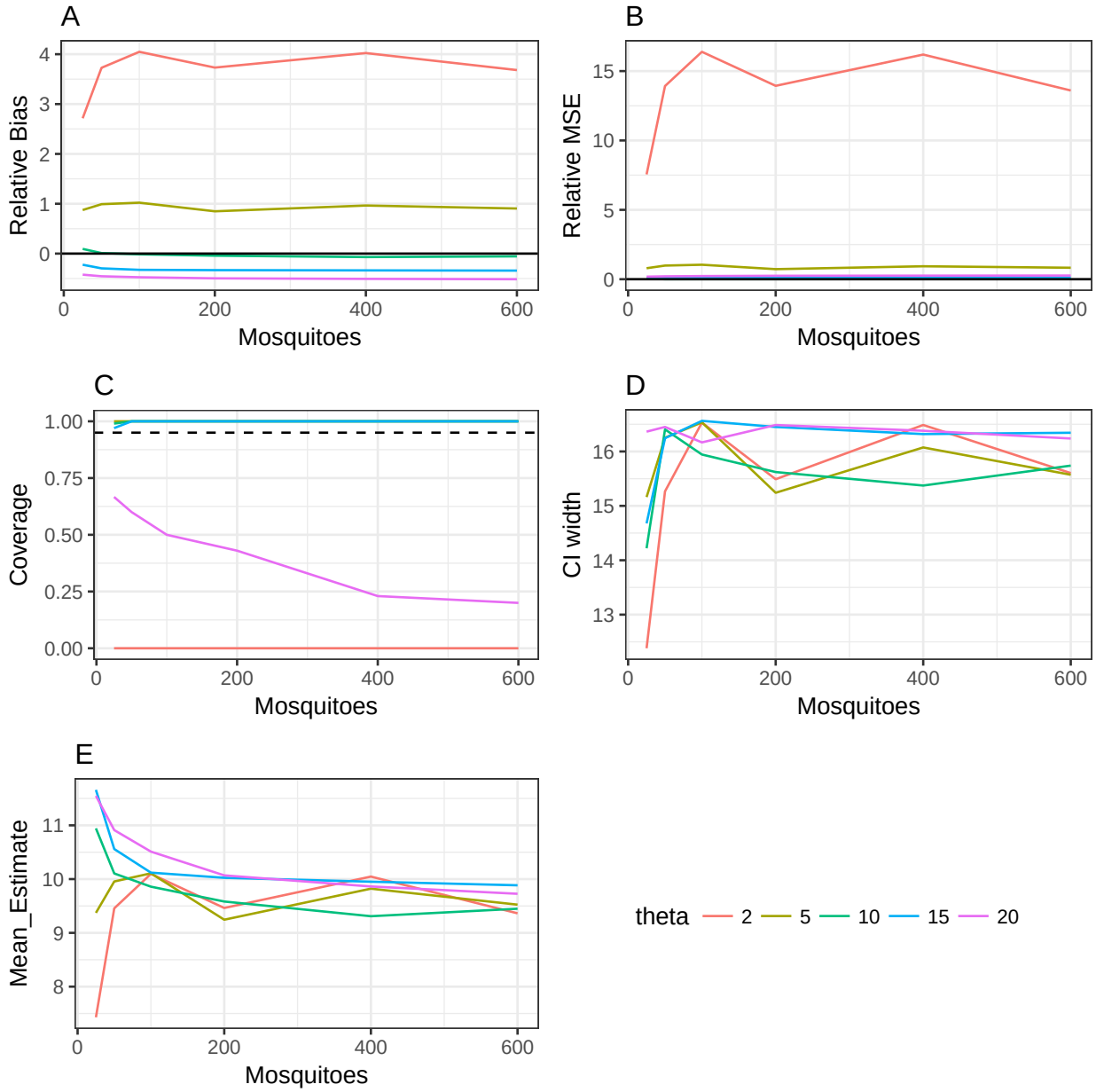

Figure 3: Estimation of θ, a paramter inversely proportional to variance in mosquito feeding probabilities. A) Bias, B) MSE and E) Mean estimate of $\hat{\theta}$, and C) coverage and D) width of the credible interval generated from the approximated Posterior produced by ABC employed on occupancy spectrum modeled with beta distributed $(B = 22, \theta)$ capture probabilities.

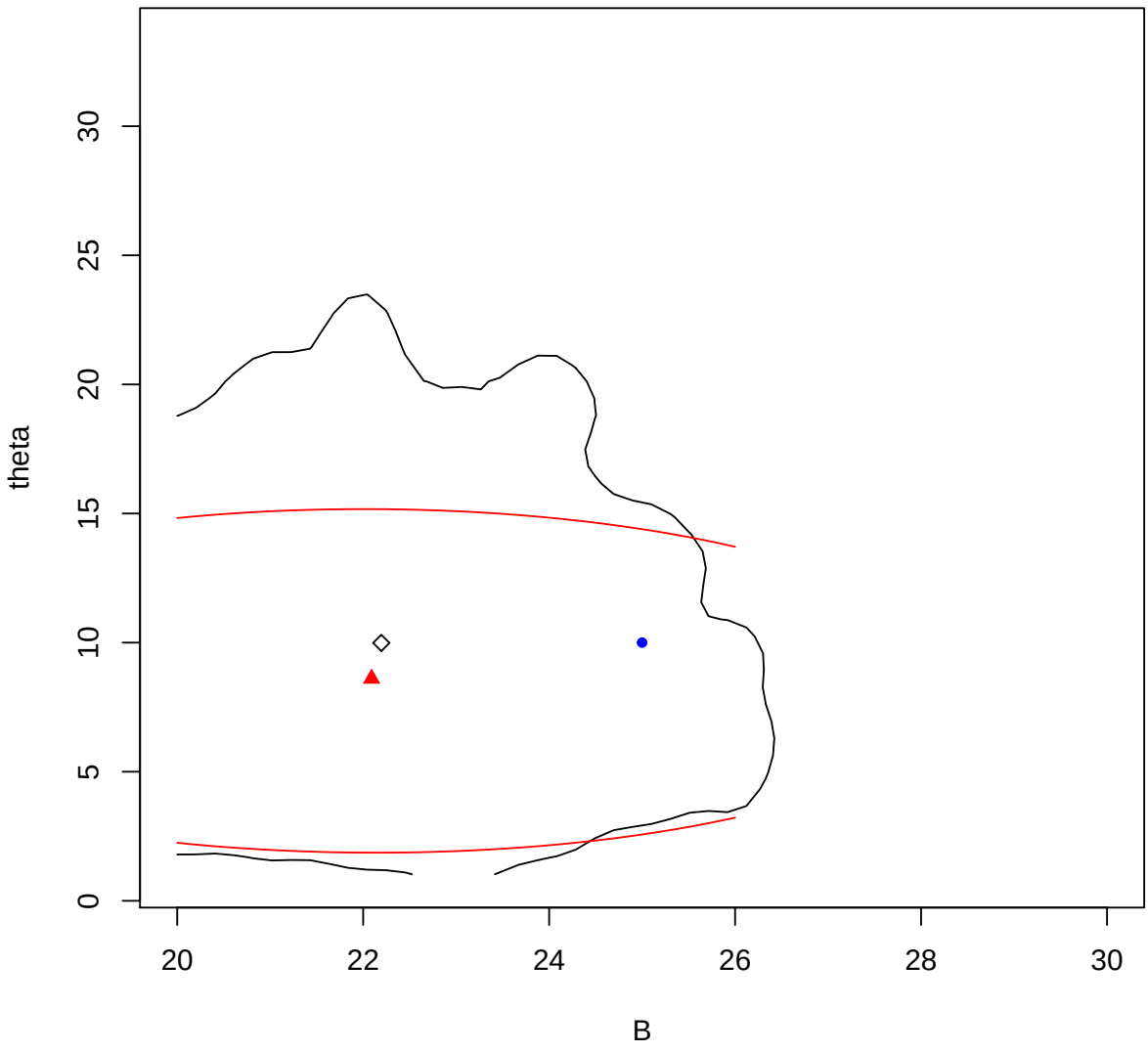

Figure 4: Highest posterior density region plot of the joint density of B and θ. The black corresponds to the prior's region and its mean ($\bar{B} = 21.047, \bar{\theta} = 10.142$), the red to that of the posterior and its mean ($\hat{B} = 22.12, \hat{\theta} = 8.148$). Blue corresponds to the true value ($B = 25, \theta = 10$).

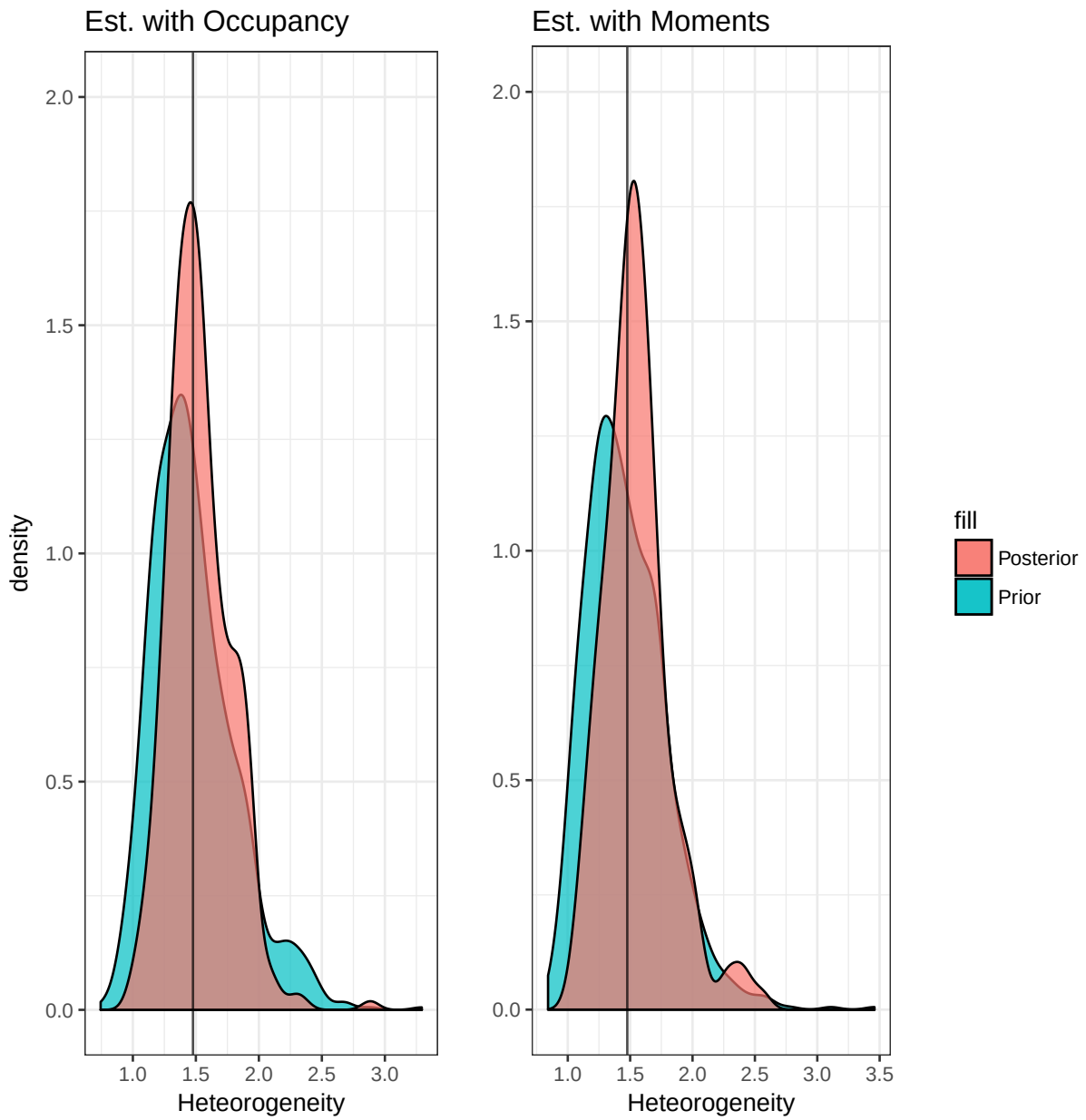

Figure 5: Estimation of posterior density by ABC and prior of heterogeneity metric where heterogeneity is the coefficient of variation of the beta distributed feeding probabilities. ($\hat{h} = 0.0202 > h = 0.0156$) Computed by 1

increase. The MSE is largest in the case when the coefficient of variation is smallest ($CV = 0.951$) where it reaches greater than 0.4. In C) the credible interval provided lower coverage in the cases when the true CV was more extreme ($CV = 0.951, 2.517$). In other cases, the credible interval is be greater than 0.75. In D), the credible interval width seems to increase, approaching approximately 1.2 1.4 in all cases. In E), the mean estimates increases, approaching 1.5 to 1.7 no matter the true magnitude of CV.

After further testing, the occupancy spectrum matching procedure does not seem as promising as Figure 5A may have indicated. The estimator gives inaccurate estimates that are insensitive to the true value of CV. As the number of mosquitoes increases, the mean estimate increases, no matter the true value of the CV. The cause of this pattern is unknown but it could be partly a result of the large dimensional summary statistic confusing the ABC procedure. As well as becoming more inaccurate, the estimate also becomes less precise since MSE and credible interval width increases with Mosquitoes. The credible intervals do obtain high coverage, but not in the case of the relatively more extreme values (when $CV = 0.951, 2.517$). This is likely the result of the prior where these values have a lower density and thus are not represented as often in the procedure. Overall the issues of estimating coefficient of variation are similar to the issues of estimating θ encountered previously.

The performance statistics with moment matching procedure are shown in Figure 7. In A), the bias appears to plateau in all cases. Bias appears to decrease with the true magnitude of CV. Absolute bias is consistently less than 0.3. In B) the MSE is largest when $CV = 0.951, 2.517, 1.09$. The MSE appears to be decreasing in the cases where $CV = 0.951, 1.09$. In other cases the MSE appears to stay relatively constant. MSE is consistently less than 0.10. In C) the coverage appears to be in the range of 80 to 95%. However it appears to be much lower in the cases where the true CV is the smallest (0.951) or the largest (2.517) of the values tested. In D), credible interval width may decrease or plateau with respect to mosquitoes, but is also consistently less than 1.0. In E), the actual estimates themselves appear to plateau as mosquitoes increase in number, but only two of the estimates approach their true values (1.78,1.314). The estimate of CV also increases as the true value of CV increases.

The moment matching procedure appears to be more promising than the occupancy matching procedure. Bias is in a smaller range. The MSE and CI width are smaller, meaning the procedure provides more accurate and precise estimates. The coverage is compatible. The mean estimates themselves are directly related to the true magnitude of the CV. However the procedure is still limited: the mean estimates seems to plateau and stop approaching the true value of CV. Values of CV which were estimated more accurately ($CV = 1.314, 1.78$) are closer to the mean of the prior distribution (1.432) than the other true values of CV. This suggests the prior may overwhelm the data in the estimation of CV. Here, further testing with a weaker prior is required.

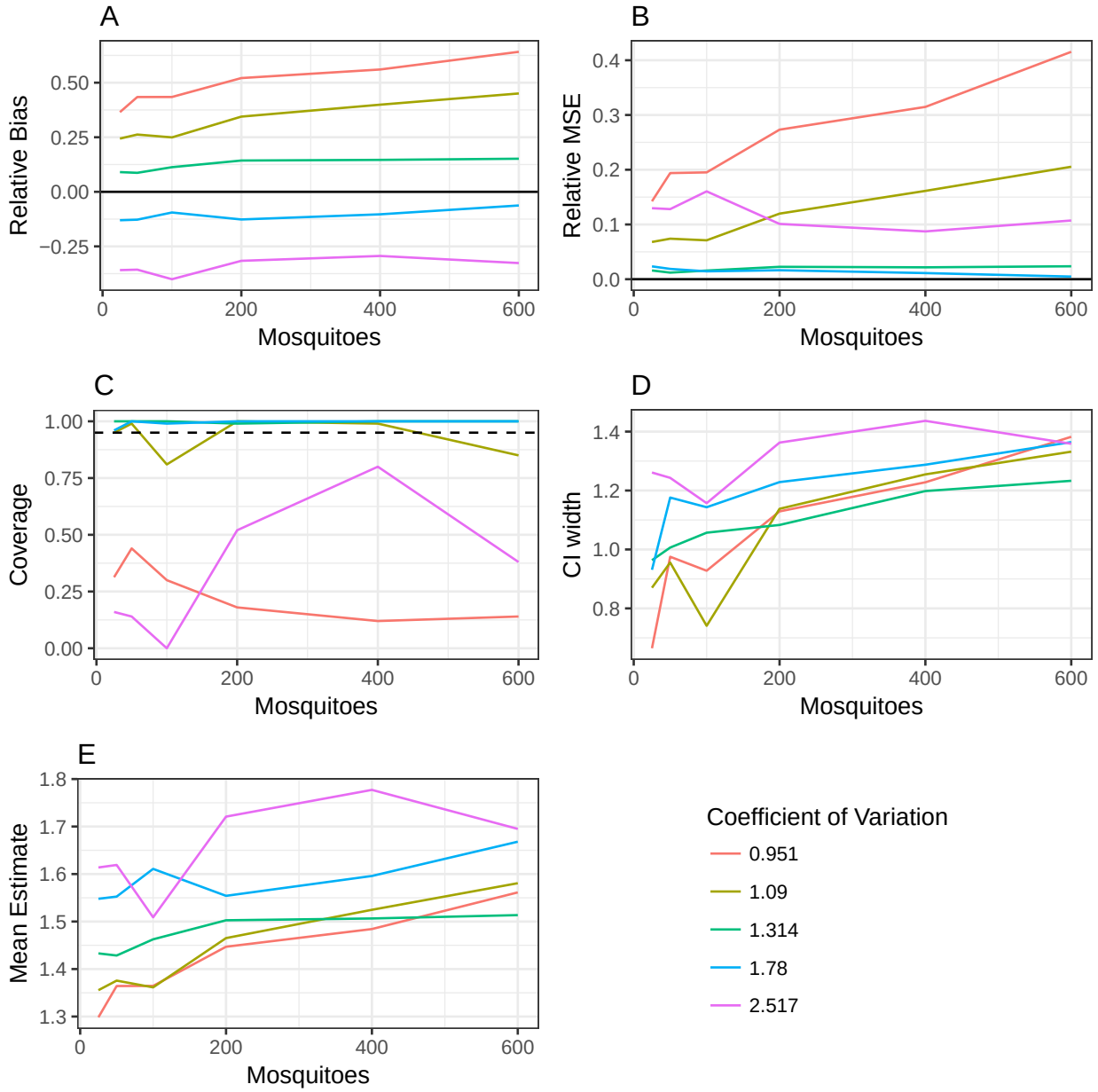

Figure 6: Estimation of coefficient of variation in mosquito feeding probabilities using occupancy spectrum. A)Bias, B) MSE, and E) mean estimate of CV and C) coverage and D) width of the credible interval generated from the approximated posterior produced by ABC employed on the occupancy spectrum modeled with beta distributed $(B = 22, \theta)$ capture probabilities.

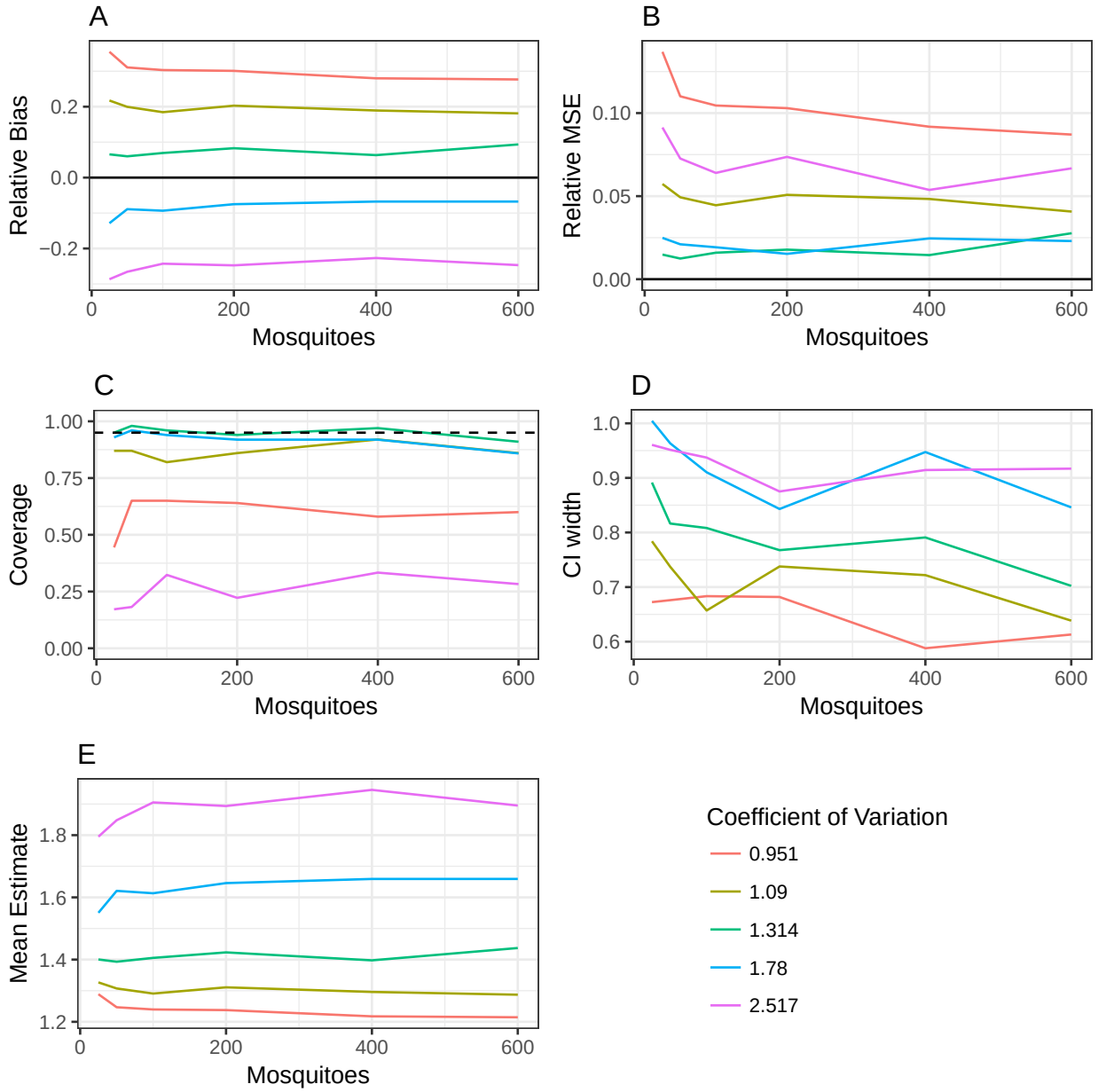

Figure 7: Estimation of coefficient of variation in mosquito feeding probabilities using moments of the occupancy spectrum. A)Bias, B) MSE, and E) mean estimate of CV and C) coverage and D) width of the credible interval generated from the approximated posterior produced by ABC employed on the moments of the occupancy spectrum modeled with beta distributed $(B = 22, \theta)$ capture probabilities.

5 Conclusion

First, a sampling design from Kilpatrick et al is presented that would provide occupancy information about bird populations using DNA sequenced from mosquito blood meals. Re-purposing Kilpatricks method, we wanted to test what the occupancy spectrum could reveal about heterogeneity in mosquito-feeding patterns within a bird species using numerical simulations in R.

A simple model, which assumes mosquito-feeding probabilities were homogeneous, was investigated using approximate Bayesian computation. The first aim was to assess the efficacy of this technique in estimating its only parameter, the bird population given an observed occupancy spectrum generated from different mosquito feeding patterns including homogeneous and heterogeneous feeding. Based on simulations of homogeneous mosquito feeding, ABC was able to reasonably estimate the population size. However, when the homogeneous feeding assumption was violated, ABC underestimated the population size.

Next, a model using beta-distributed mosquito-feeding probabilities (parameterized by $1/B = \alpha/(\alpha + \beta), \theta = \alpha + \beta$) was investigated. A heterogeneity metric was introduced and defined as the coefficient of variation in mosquito-feeding probabilities modeled by the beta distribution. The first aim of investigating this model was to test the efficacy of the ABC procedure in estimating the bird population size. The efficacy of this technique was also tested based on simulations of heterogeneous mosquito feeding. This estimator was able to achieve improved estimation compared to the ABC with the homogeneous feeding assumption, but was still unreliable, since the procedure would not predict population sizes greater than 35. Otherwise the procedure had moderate bias and MSE (often $|\text{bias}| > 0.25\%$, often $MSE < 0.4$, coverage $> 75\%$, but coverage $\approx 25\%$ when true $B = 55$).

The second aim of investigating the beta model was to test ABC in how well it could estimate θ. Overall, using ABC to estimate θ was not successful since point estimates were biased and mostly unchanged from the mean of the prior distribution. ($|\text{bias}| \approx 4, MSE > 10$, Coverage often $< 95\%$). The issue of estimating θ may be due to the use of a high dimensional summary statistic in the ABC procedure.

The last aim of investigating the beta model was to test how well ABC could estimate heterogeneity, which is a function of the B and θ estimates. Heterogeneity was estimated with both the occupancy spectrum and the moments of the occupancy spectrum as summary statistics in the ABC procedure. Replacing the occupancy spectrum with its moments provided a low dimensional summary statistic for ABC. Both procedures were tested based on simulations with heterogeneity, where the bird population was kept consistent in a small range (20-25). The ABC procedure using moments was able to produce more accurate and more precise estimates than the ABC procedure using the whole occupancy spectrum with less bias, less MSE and smaller credible intervals. Estimates were directly proportional to the true magnitude of coefficient of variation being estimated, but were more biased when estimating extreme values, further from the mean of the prior distribution (1.41). This may suggest that the prior is dominating the procedure, instead of the evidence provided by the moments themselves.

For ABC performance to improve, issues with the ABC rejections procedure should be further investigated. The ABC procedure using moments should be further explored. The efficacy of the procedure in estimating B and θ could provide more insight into the estimation of coefficient of variation. Further testing is also required for the choice of priors used in the ABC procedure and how this plays a role in estimating B and θ with both the occupancy matching and moment matching cases. Several priors should be tested to explore the effect of different choices on the estimates.

6 Appendix

```
## Error in as.data.frame(BThetaPrior):  object 'BThetaPrior' not found
## Error in plot_clone(plot):  object 'thetaPrior' not found
```

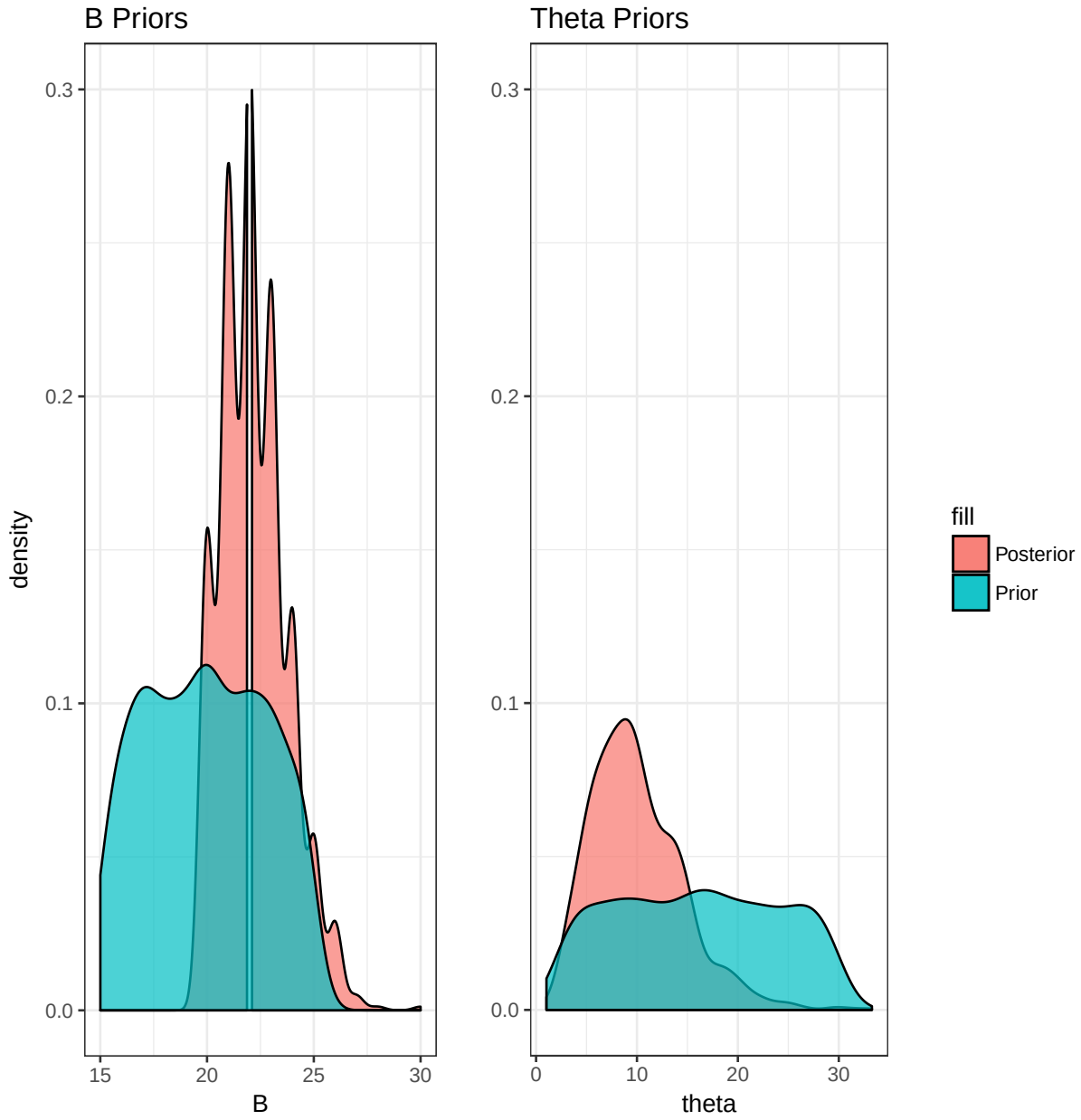

Figure 8: Priors for B and θ. Note: $E(B|\text{Prior A}) = 22, E(B|\text{Prior B}) = 20, E(\theta|\text{Prior A}) = 9.99, E(\theta|\text{Prior B}) = 15.96$.

References

[1] Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics 41*, 379–406.

[2] Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783–791.

[3] Coull, B. A. and A. Agresti (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics 55*(1), 294–301.

[4] Dorazio, R. M. and J. Andrew Royle (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics 59*(2), 351–364.

[5] Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika 66*(2), 393–396.

[6] Kelly, D. W. (2001). Why are some people bitten more than others? *Trends in parasitology 17*(12), 578–581.

[7] Kilpatrick, A. M., P. Daszak, M. J. Jones, P. P. Marra, and L. D. Kramer (2006). Host heterogeneity dominates west nile virus transmission. *Proceedings of the Royal Society of London B: Biological Sciences 273*(1599), 2327–2333.

[8] Lindsay, S., J. Ansell, C. Selman, V. Cox, K. Hamilton, and G. Walraven (2000). Effect of pregnancy on exposure to malaria mosquitoes. *The Lancet 355*(9219), 1972.

[9] Woolhouse, M. E., C. Dye, J.-F. Etard, T. Smith, J. Charlwood, G. Garnett, P. Hagan, J. Hii, P. Ndhlovu, R. Quinnell, et al. (1997). Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proceedings of the National Academy of Sciences 94*(1), 338–342.