Understanding and Diagnosing Estimability Issues in Ecology with an Emphasis on Occupancy Models

A Thesis

Presented in Fulfillment of the Requirements for the

Degree of Master of Science with a Major in Statistical Science

College of Graduate Studies

University of Idaho

by

Amanda L. Bowe

Major Professor: Brian Dennis, Ph.D.

Committee Members: Timothy R. Johnson, Ph.D., Ryan Long, Ph.D.

Department Administrator: Chris Williams, Ph.D.

April 2016

AUTHORIZATION TO SUBMIT THESIS

This thesis of Amanda L. Bowe, submitted for the degree of Master of Science with a Major in Statistical Science and titled "Understanding and Diagnosing Estimability Issues in Ecology with an Emphasis on Occupancy Models," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____Date:_____

Brian Dennis, Ph.D.

Committee Members: _____Date:_____

Timothy R. Johnson, Ph.D.

_____Date:_____

Ryan Long, Ph.D.

Department

Administrator: _____Date:_____

Chris Williams, Ph.D.

ABSTRACT

Occupancy modeling is becoming increasingly popular in wildlife management as a method of monitoring trends in wildlife populations. One of the primary motivations for the use of occupancy modeling is the ability to make inferences about large landscape patches with a reduced number of surveys. However, this increased versatility comes at the risk of l. Previous research (Hubbard 2014) has explored the presence of inherent identifiability issues in occupancy models, little work has been done on the estimability the key parameters of these models: detection probability (p) and occupancy probability (psi).

Using maximum likelihood estimation, a combination of bootstrapped profile likelihoods, data simulation and the data cloning techniques of Lele et al. 2010 were used to diagnose estimability issues across a spectrum of parameter values for p and psi, sites and surveys. Preliminary results suggest estimability issues are present at smaller sample sizes or fewer repeat surveys as either p or psi approaches the boundaries (0 or 1). The potential use of Bayesian methods to mitigate these issues is still under exploration.

ACKNOWLEDGEMENTS

I greatly appreciate the help and support of the entire University of Idaho Statistics Department, especially my labmates: Cara Leatherman and Brenda Hanley. I would also like to acknowledge Dr. Darryl MacKenzie and Dr. Larissa Bailey for wonderful occupancy modeling workshops they hosted and for taking the time to discuss my thesis research.

DEDICATION

This thesis is dedicated to my roommate, Elizabeth Ehrsam for ensuring I continued to eat throughout the writing process, to Dr. Ann Abbott for allowing me to inhabit her spare bedroom to finish writing and to E. Stein for his constant companionship and support;

He kept me sane.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Occupancy modeling has rapidly become one of the most widely used statistical techniques in wildlife ecology and management. The appeal of occupancy modeling lies in its applicability to the study of more challenging species, those that are cryptic, elusive, or sparsely distributed on the landscape. Rather than attempting to estimate abundance or habitat preferences of individuals in a population, occupancy modeling seeks to make inferences regarding the proportion of a landscape occupied or used by the species. Habitat preferences can be further explored through the inclusion of site specific covariates and comparison of occupied versus unoccupied sites (Ball et al., 2005; Saracco et al., 2011).

In a typical occupancy study, each member of a group of randomly selected sites is surveyed multiple times throughout the course of a season for the presence of one or more target species. Sites can be any subdivision of the landscape either user-defined (grid cells, transects) or natural divisions of the landscape (habitat patches, water bodies). Repeated surveys of each site are necessary to obtain estimates of detectability (Mackenzie et al., 2003).

As the popularity of occupancy modeling grows, so does the complexity of the models being used. Occupancy models have been expanded to encompass multi-season models, multi-species and species co-occurrence models, multi-state models (breeding/non-breeding, age structure, etc; (Nichols et al., 2007; Mackenzie et al., 2009; Martin et al., 2009) and even combinations of these models (Jensen and Vokoun, 2013). Although a wide variety of work

has been done on the construction and analysis of more complex models, little work has been done regarding the limits of statistical applicability of this family of models.

One such topic is the statistical estimability of the model parameters. Hubbard (2014) conducted an exploration of analytical estimability issues in the form of parameter redundancy for the basic occupancy model. Lele et al. (2012) developed a method of using covariates to overcome the estimability problems of a single visit occupancy model. However, we can find no work regarding the estimability of the most common form of occupancy model: repeated visits to randomly selected sites. The lack of information about estimability for the occupancy model is concerning because the ability to make reliable inferences for any model depends on all relevant parameters being uniquely identifiable (Lele et al. 2010). Ideally, estimability of parameters should be confirmed for any model. Ecologists attempting to fit increasingly realistic models to data should be aware that estimability problems typically become more prevalent as complexity of a model increases.

In this paper, we explore the estimability of the simple occupancy model that forms the foundation for many of the more complex occupancy models. We use a recently developed technique called data cloning to evaluate estimability for this model. Data cloning takes advantages of the computational capabilities of a Bayesian Markov chain Monte Carlo (MCMC) to simulate maximum likelihood estimates (MLEs) from a posterior distribution. Data cloning is a particularly useful technique when the model involves complex likelihoods. Maximizing a complex likelihood analytically can be difficult or impractical.

ESTIMABILITY IN ECOLOGICAL MODELS

Types of Estimability Problems:

In general, estimability issues occur when multiple combinations of parameters can yield the same likelihood value. Estimability problems with a model can vary in both cause and severity and may manifest in different ways. Terminology for these issues varies between authors and disciplines and in some cases the same terminology can have different meanings. Here we have grouped estimability problems into two types distinguished by the underlying cause of the problem.

Type I:

A Type I estimability problem is often referred to as non-identifiability (Catchpole and Morgan, 1997), intrinsic non-estimability (Viallefont et al., 1998; McCullagh and Nelder, 1989), structural estimability (Raue et al., 2007) or parameter redundancy (Hubbard, 2014; Gimenez et al., 2004). A Type I estimability problem occurs when a model is built such that two or more parameters appear only as a combination. Such parameters are referred to as non-separable and can only be estimated as a unit. A simple example of this problem is a hierarchical model where X is distributed normally with mean μ and variance $\sigma^2 + \tau^2$. In this case, value of $\sigma^2 + \tau^2$ can be estimated from the data, but no unique MLE can be obtained for σ^2 or τ^2 individually. This form of estimability problem is the most severe because it is inherent in the model. Even with unlimited amounts of available data, it would

remain impossible to obtain MLEs of these parameters (Hubbard, 2014). A Type I estimability problem can only be addressed by reformulating the model.

A simple occupancy model with no repeat surveys is an example of a model with a Type I estimability problem. Without the additional information provided by a repeat survey, there is no way to distinguish between the true absence of the species at a site and the presence, but non-detection, of the species. Lele et al. (2012) developed an alternative solution to this problem using covariates to provide the additional information needed to determine absence vs. non-detection. However, occupancy models with repeated surveys remain the gold standard in ecology.

Type II:

A Type II estimability problem is referred to as nonestimable (Lele et al., 2010 a), inestimable (Campbell and Lele, 2014) extrinsically non-identifiable (Viallefont et al., 1998; McCullagh and Nelder, 1989), practical estimability (Raue et al., 2007) or nearly non-identifiable (Dennis et al., 2006). Unlike a Type I estimability problem, a Type II estimability problem is attributable to underlying issues in the data, not the model. With this type of estimability problem, all parameters can be estimated under the chosen model given a hypothetically infinite amount of data. However, the realized data contains insufficient information to separate the influence of an individual parameter from the others.

A Type II estimability problem can be solved by increasing the sample size, but the need for a larger sample is often not known prior to sampling. An unusual or unrepresentative

sample, the relationship between parameters and the actual parameter values can all cause estimability issues to occur, even with a seemingly sufficient sample size.

For an occupancy model, an observed set of data might have come from a population with high occupancy, but low detectability or moderate occupancy and moderate detectability. Without enough repeat surveys to tease apart the effects of both parameters, there is no way to know which combination of parameter values best approximates the true state of the population. A computer algorithm may still obtain a single MLE, but if the data are insufficient, the resulting estimate will have a high estimated variance.

Diagnosing Estimability Problems:

Estimability problems are commonly diagnosed by maximizing the likelihood across a range of fixed parameter values creating a profile likelihood plot. If the profile likelihood plot has a flat or ridge-like shape, it indicates an estimability problem is occurring. A perfectly flat profile-likelihood could indicate a Type I estimability problem, while a profile-likelihood that is locally flat or slightly jagged would indicated a Type II estimability problem. A contour plot is a combination of profile likelihoods used to visualize the joint likelihood of 3 parameters. With more than 3 parameters, a profile plot cannot be used to visualize the joint likelihood of parameters. A simulation study can be used to identify the cause of the problem. In a simulation study, multiple datasets are generated from the model across a spectrum of parameter values and sample sizes. These simulated datasets are then independently analyzed for estimability. If some of the simulated datasets are fully-estimable, the estimability problem observed is caused by some property of the specific data

used and a Type II estimability problem is occurring. However, if none of the simulated datasets are fully-estimable, the problem lies in the structure of the model itself and a Type I estimability problem should be suspected.

Other techniques for diagnosing estimability include: the symbolic methods used by Catchpole and Morgan (1997) to confirm Type I estimability, the Hessian matrix method (Gimenez et al., 2004) which can identify estimability problems using the condition number, but only when there is a clearly defined likelihood function and data cloning (Lele et al., 2007), which can identify both Type I and Type II estimability problems separately and does not require a closed-form likelihood.

USE OF DATA CLONING TO DIAGNOSE ESTIMABILITY

Simple Occupancy Model:

We investigated estimability for the simplest model for site occupancy with imperfect detection. The simplest model is for a single species over a single sampling season for which occupancy is assumed to be closed and the probability of detecting the species was constant across all sites and surveys. Sites were assumed to have been randomly selected from the landscape with no preference towards sites with known presence and surveys were assumed to be independent. In practice, the simple occupancy models is often expanded to include multiple species and/or seasons and a variety of covariates of interest. Any estimability problems in the underlying base model potentially have wide ranging consequences. The long-run goal of assessing estimability of occupancy models is best commenced at the base, where any lack of statistical performance will have to be corrected before incorporating realistic complexity.

Data are collected as a series of 1s and 0s indicating presence or non-detection respectively for each site at each survey period (x_{ij}). Occupancy data are typically compiled into a table of detection histories for each site as shown in Table 1.

	Survey		
Site	1	2	3
1	0	0	0
2	0	1	0
3	0	0	1
4	1	1	1
5	1	0	0

Table 1: Example of an occupancy detection history with sites (i) as rows and surveys (j) as columns. A 1 indicates the species was observed at that site during the survey. A 0 indicates the species was not observed.

The base occupancy model assumes a constant detection probability, so the order of detections and non-detections becomes irrelevant. Data for the base occupancy model can be condensed into two variables:

$x_{i.} =$ the total number of detections at the i^{th} site (summed across all j surveys)

$m_i =$ the number of surveys conducted the i^{th} site

where *i* ranges from 1 to *n* (the total number of sites surveyed)

A simple model of species occupancy with imperfect detection seeks to estimate two parameters:

- ψ: the probability that a randomly selected site is occupied,
- p: the probability of a species being detected at a site given that the site is occupied.

For any site (i), the occupancy is a Bernoulli random variable where site (i) is occupied with probability ψ or unoccupied with probability $1 - \psi$. Within an occupied site, detection for each survey $(j = 1\ to\ m_i)$, is also a Bernoulli random variable where the species is detected on survey (j) with probability p and not detected with probability $1 - p$. Combining the two yields the joint probability of observing a detection history where $X_i = x_{i.}$ at the i^{th} site over m_i surveys.

When the species is detected at least once $(x_{i.} > 0)$, the probability distribution is given by

$$\Pr(X_{i.} = x_{i.}) = \psi \binom{m_i}{x_{i.}} p^{x_{i.}} (1-p)^{(m_i - x_{i.})}$$

(Eq. 1)

For sites with no detections, ($x_{i.} = 0$) there exists two possibilities: either the species was absent (with probability $1 - \psi$) or the species was present and not detected on any of the surveys yielding a two part equation.

$$\Pr(X_{i.} = 0) = (1 - \psi) + \psi \binom{m_i}{x_{i.}} p^{x_{i.}} (1 - p)^{(m_i - x_{i.})}$$

(Eq. 2)

Using an indicator variable, θ_i, where $\theta_i = 1$ if the species was detected at least once at site i; else $\theta_i = 0$, allows us to combine Eq. 1 and Eq. 2 into a single equation.

$$\Pr(X_{i.} = x_{i.}) = (1 - \psi)(1 - \theta_i) + \psi \binom{m_i}{x_{i.}} p^{x_{i.}} (1 - p)^{(m_i - x_{i.})} \theta_i$$

(Eq. 3)

If the sites where selected at random, they are assumed to be independent and probability of observing the complete set of xi values (written as vector $\underline{x}$) across all sites, is simply the product of the probability for each individual site.

$$\Pr(X = \underline{x} \mid \psi, p, m_i) = \prod_{i=1}^{n} \left[(1 - \psi)(1 - \theta_i) + \psi \binom{m_i}{x_{i.}} p^{x_{i\cdot}} (1 - p)^{(m_i - x_{i\cdot})} \theta_i \right]$$

(Eq. 4)

This equation represents the probability of obtaining the observed data given values of ψ, p and m_i. It can also be considered as the likelihood of any estimates of ψ and p given the observed data

$$\mathrm{L}(\psi, p|m_i, \theta_i, X = \underline{x}) = \prod_{i=1}^{n} \left[(1-\psi)(1-\theta_i) + \psi \binom{m_i}{x_i} p^{x_i}(1-p)^{(m_i - x_i)}\theta_i\right]$$

(Eq. 5)

The likelihood given in Eq. 5 is used to calculate maximum likelihood estimates for ψ and p and is the foundation for the Markov chain Monte Carlo (MCMC) simulation used to diagnose estimability. The value of the combinatorial constant $\binom{m_i}{x_i}$ does not affect maximum likelihood parameter estimation. The combinatorial constant equals 1 in the likelihood terms where $x_i = 0$, and it is just a multiplicative constant that does not contain the parameters in the likelihood terms where $x_i > 0$.

Data Cloning Process:

Data cloning is a method of computing maximum likelihood estimates (MLEs) for complex models in which calculating the likelihood function requires high dimensional integration (Lele et al. 2007). Hierarchical models for random effects and latent variables are examples of models with likelihood functions that are difficult if not impossible to calculate. Data cloning utilizes a standard Bayesian framework using MCMC to generate a posterior distribution from prior distributions for each parameter and the joint likelihood of the data given parameters as shown in Eq.6

$$Pr(X = \underline{x} \mid \psi, p, m_i)\ g(p, \psi)\ \propto\ Pr\left(\psi, p | m_i, \theta_i, X = \underline{x}\right).$$

where $g(p, \psi)$ is the prior distribution,

$Pr(X = \underline{x} \mid \psi, p, m_i)$ is the joint likelihood of the data,

and $Pr\left(\psi, p | m_i, \theta_i, X = \underline{x}\right)$ is the posterior probability

(Eq. 6)

However, instead of just using the likelihood of the original data, one creates *k* copies of the original data (termed clones) and uses the likelihood of the cloned data. One can think of these clones as independent repetitions of the same experiment, which by random chance obtained the same results. According to Lele et al.'s (2007) modification of Walker's Theorem (Walker, 1969), as the number of clones increases, the data will overwhelm the priors and the means of the posterior distributions for the parameter estimates obtained through MCMC will converge to the MLEs.

The first step towards assessing estimability using data cloning is to set up a Bayesian MCMC simulation for the model in an appropriate software. For this study, we used the open source software, WinBUGS (Lunn et al. 2000). However, WinBUGS is only programmed for a few common distributions, so it was necessary to create a custom distribution corresponding to the desired likelihood function. The customized distribution was accomplished using what the WinBUGS manual refers to as the "zero-trick." The "zero-trick" uses as the primary data a vector of all zeros which are purported to arise from a Poisson distribution with parameter λ. An observation of zero from a Poisson(λ) distribution will have a likelihood equivalent to $e^{-\lambda}$. Therefore, we set λ_i equal to the negative log of our desired likelihood function for the i^{th} observation (Spiegelhalter et al.,

1999). The likelihood for one observation under the occupancy model is given by (Eq. 3). Numerical problems that can arise from calculating the binomial term in (Eq. 3) were avoided by taking its logarithm, summing the log terms, and then taking the antilogarithm of the sum, resulting in the equation below.

$$\lambda_i = -\log(e^{(\log(\psi)+x_i \log(p)+(m_i-x_i)\log(1-p))} + (1-\theta_i)(1-\psi))$$

(Eq. 6)

In (Eq. 6) we have omitted the combinatorial constant in the binomial distribution because its calculation is unnecessary. The observed data, summarized into x_i, m_i and θ_i for each site, are then read into WinBUGS and an MCMC was performed.

Simulation Study:

We used program R (R Core Team, 2015) to generate simulated occupancy datasets for each parameter combination of ψ and p in increments of 0.1. Each dataset consisted of presence ($x_{ij} = 1$) or non-detection ($x_{ij} = 0$) data for 25 hypothetical sites with 5 surveys per site. We used the data-cloning procedure detailed above to obtain summary statistics for each simulated dataset for $k = 1$, 5, 10, 25, and 50 clones of the data. For the MCMC analysis, we used a burn-in period of 10,000 followed by an additional 70,000 updates. The MCMC approximation was checked for convergence using trace and history plots. Summary statistics were obtained for ψ and p including MLEs, SEs and the correlation between ψ and p. The standard errors and correlations were combined in a variance-covariance matrix for each parameter combination and number of clones. We plotted the largest eigenvalue of the variance-covariance matrix against the number of clones to create a diagnostic plot for each

parameter combination as described in Lele et al. (2010). The eigenvalues for these plots were standardized to equal 1 when $k = 1$.

Code for simulating data from a simple occupancy model in R can be found in Appendix A. Code to calculate MLEs, SEs, and correlations for the simple occupancy model using WinBUGS is in Appendix B. Eigenvalue diagnostic plots can be generated from MLEs, SEs, and correlations using the R code found in Appendix C.

RESULTS

When there are no estimability problems occurring, the eigenvalue diagnostic plots will show an exponentially decreasing trend, converging to zero at a rate of approximate $1 / k$ (Figure 1) . See Lele et al. (2010) for proofs. All simulated datasets where the values of both ψ and p used in simulation were >0.3 had no evidence of an estimability problem as shown by the eigenvalue plots located in Appendix D.

When $p = 0.1$ and $\psi < 0.8$, the diagnostic plots tended to show a decreasing, yet concave down trend (Figure 2) or a decrease at a rate much slower than $1 / k$ (Figure 3). These plots still converge to zero indicating both ψ and p are estimable. However, the shape and speed of decrease suggests that the parameters are nearly non-estimable at these parameter values.

When either ψ or $p < 0.3$, the diagnostic plots show more signs of near estimability problems. Additional simulations (not shown) suggest that ψ and p in this range can often be estimated, but that the risk of drawing individual data sets with estimability problems is large (Figure 4).

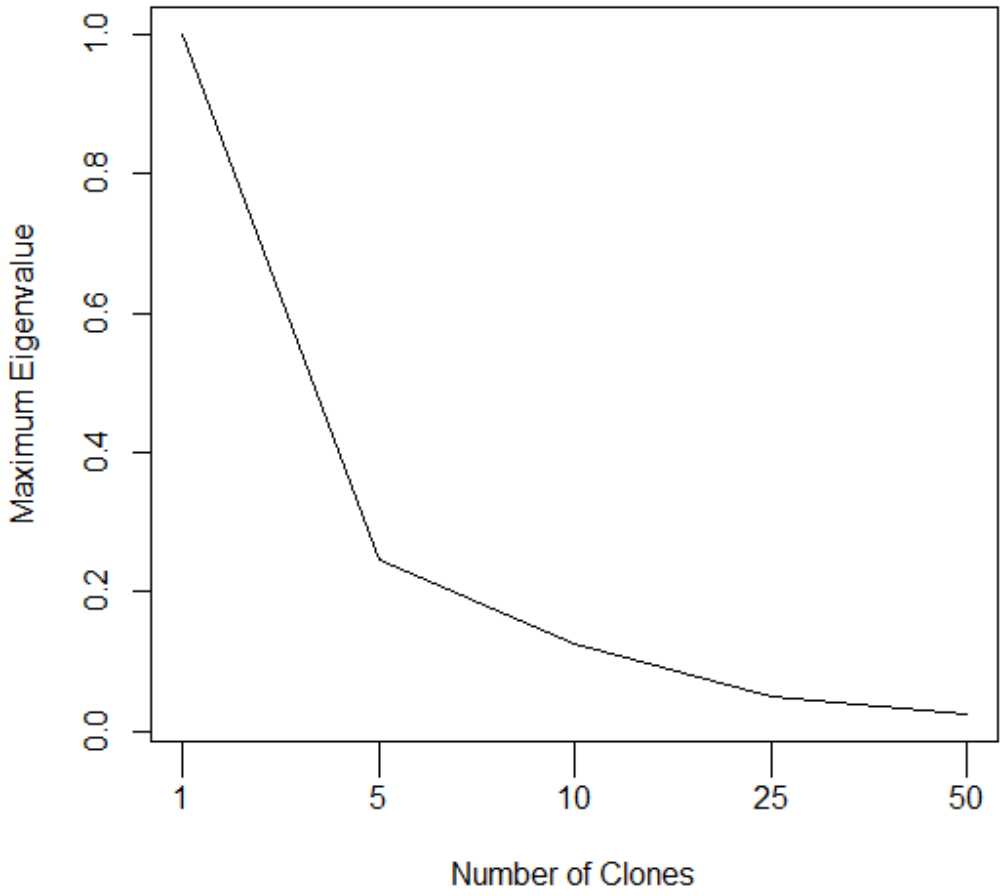

Figure 1: Eigenvalue diagnostic plot for $\psi = 0.8$ and $p = 0.4$ showing an exponentially decreasing trend converging to zero at a rate of approximate $(1 / k)$ indicating no estimability problems.

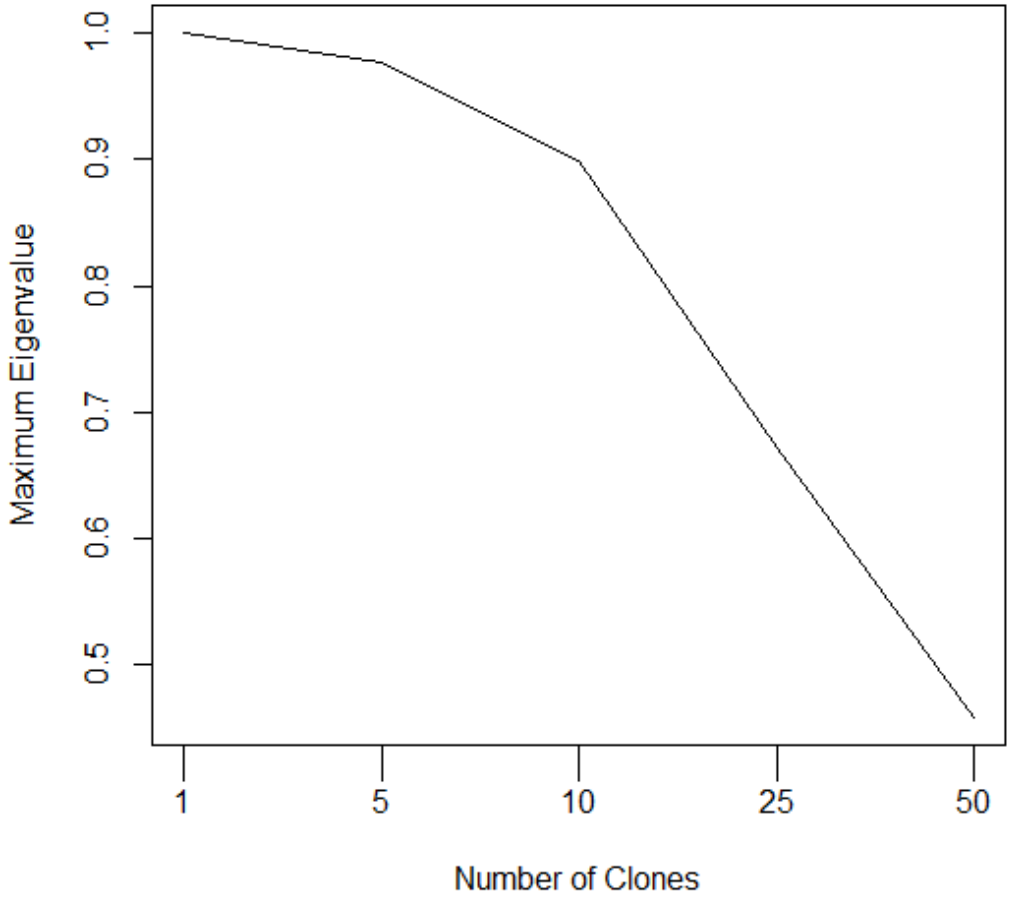

Figure 2: Eigenvalue diagnostic plots for $\psi = 0.1$ and $p = 0.4$ showing a concave down decreasing trend indicating near non-estimability of parameters.

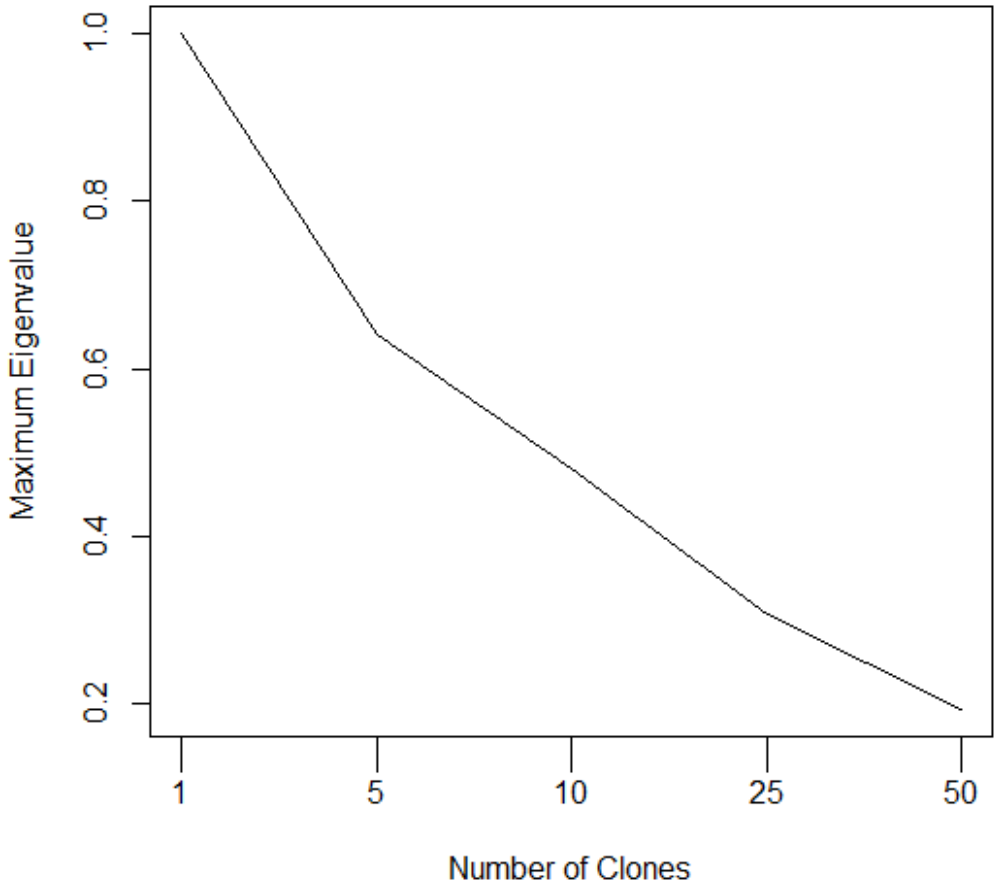

Figure 3: Eigenvalue diagnostic plots for $\psi = 0.7$ and $p = 0.1$ showing a concave down decreasing trend indicating near non-estimability of parameters.

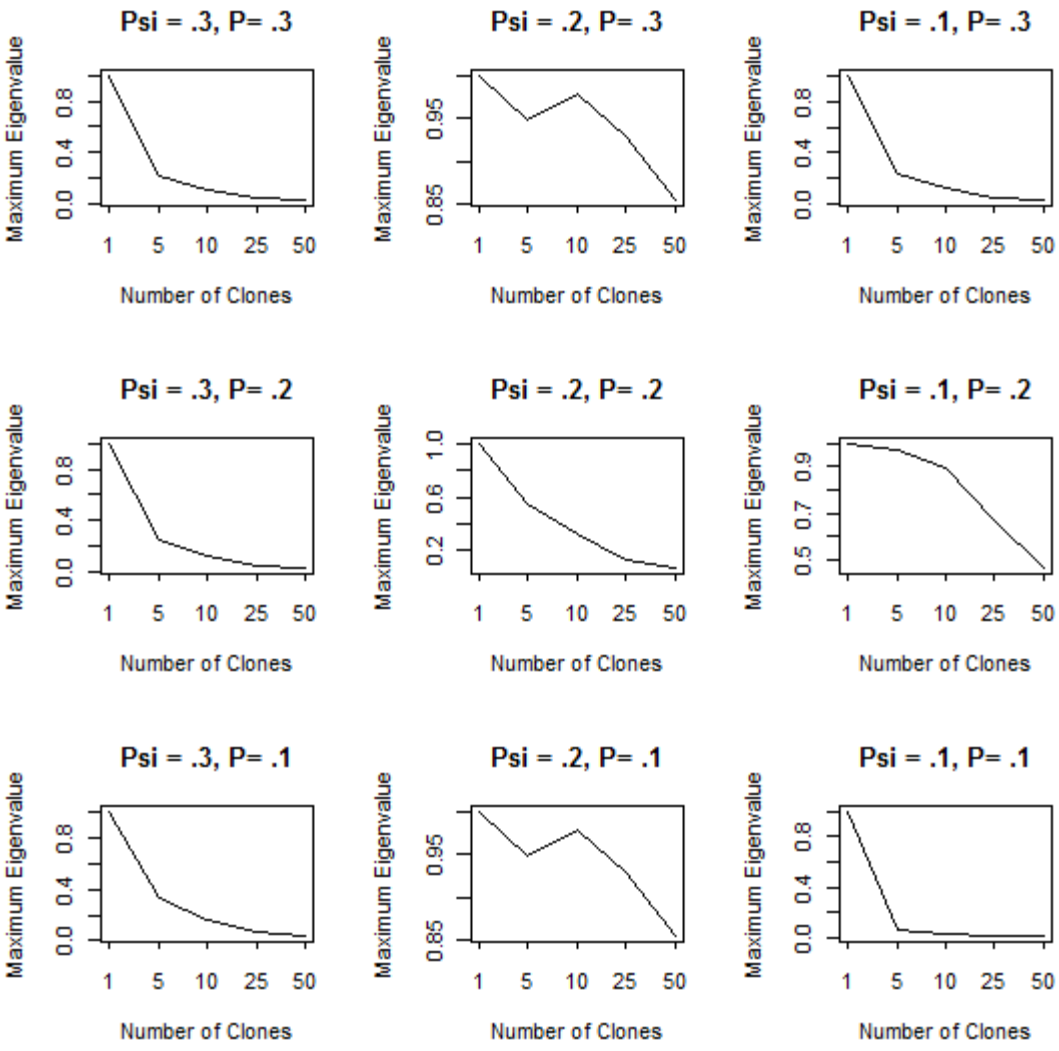

Figure 4: Eigenvalue diagnostic plots for ψ and p values between 0.1 and 0.3 showing abnormal trends indicating near non-estimability of parameters.

DISCUSSION

Importance of Assessing Estimability:

Failure to diagnose or account for estimability can lead to improper or misleading inferences about parameters. An estimate for a parameter can be obtained if a Type II estimability problem is occurring, but the estimate may not be valid or may have an impractically large variance. Ecological modeling studies are often intended to inform resource management decisions. Making decisions on an invalid parameter estimate due to an undiagnosed estimability problem could have serious consequences for management.

Estimability should be checked for any model regardless of its complexity. Although more complex models, especially hierarchical models, can be more prone to estimability issues, simple models can have estimability problems as well. In some cases, some parameters may be inestimable in a simple model but become entirely estimable when additional parameters are added (Gimenez et al., 2004; Lele et al., 2012).

Estimability of a Simple Occupancy Model:

Type I estimability problems in an occupancy model typically occur when the model is constructed in such a way that the parameters can only be estimated as a single unit ($\psi * p$), making it impossible to separate the effects of detectability and occupancy. Type I estimability problems can be eliminated through the use of repeat surveys (Hubbard, 2014)

or the methods of Lele et al. (2012) for dealing with detection error in occupancy studies with only one observation per site.

Most middle range parameter combinations for ψ and p produced no Type II estimability problems under the sample sizes simulated. However, several parameter combinations near the parameter range boundaries were only weakly estimable. When detectability was low ($p \leq 0.1$) decreasing the number of sites sampled or decreasing the number of surveys per site for populations with these characteristics could cause estimability problems to appear. When the target species is very difficult to detect or believed to be sparsely distributed throughout the study area, extra care should be taken to ensure parameter estimates are reliable. For species which are not exceedingly rare or cryptic, a sample size of 25 with 5 repeat surveys per site should be sufficient to estimate both ψ and p. In general, we recommend increasing the number of sampled sites and/or number of repeat surveys when detectability and/or occupancy are assumed to be low (< 0.3).

Advantages and Disadvantages of Data Cloning:

Data cloning is a versatile technique which can be used to assess estimability for almost any model. Data cloning works by using MCMC on the likelihood of a cloned dataset to provide the MLEs and standard errors for the original data. Data cloning can be performed using readily-available, free software such as WinBUGS, JAGS or OpenBUGS. Using the "zero trick" to coerce the model to a Poisson distribution, almost any model can be programmed into the modeling software. In fact, if the analysis is being conducted in a Bayesian framework, data cloning can be used to check estimability simply by duplicating the dataset

(Lele et al., 2010). The only difference is in computational power and time necessary to run the model, which can be significant for large datasets at large numbers of clones (k).

One limitation of data cloning is that it only yields the MLEs, not the actually value of the likelihood function at the maximum. The maximum likelihood value is used in computing profile likelihood confidence intervals, conducting hypothesis tests (comparing full vs. reduce model) and performing model selection using information criteria such as AIC and BIC (Ponciano et al., 2009). The standard errors obtained using data cloning can be used to generate Wald confidence intervals for the parameter estimates (Ponciano et al., 2009). However, Wald intervals often have incorrect coverage and are generally considered inferior to other techniques for developing confidence intervals such as profile likelihood-based intervals (Meeker and Escobar, 1995). Ponciano et al. (2009) developed a means of calculating the likelihood ratio using data cloning, which can be used in place of the maximized likelihood to construct profile likelihood-based confidence intervals, conduct likelihood ratio hypothesis tests or perform model selection in a frequentist framework. It is important to note that data cloning only uses the information contained in the original dataset. It does not increase the sample size or compensate for sparsity of data in any way (Lele et al. 2007).

Estimability in a Bayesian Framework:

Another growing trend in ecological research is the use of Bayesian methods to model complex ecological systems. The Bayesian approach is especially popular in uses of hierarchical models in ecology including in occupancy models. In a hierarchical occupancy model, parameters such as ψ, p, habitat covariates, etc can each be nested within the

hierarchy and each is represented with its own prior distribution. However, hierarchical models experience problems with estimability. Estimability problems in a Bayesian framework often result in the excessive influence of the prior in the posterior distribution. Alternatively, if an uninformative prior is used, the model can drift towards the bounds of the parameter (Gelfand and Sahu, 1999). In extreme cases (Type I estimability issues), the posterior distribution for the inestimable parameter simply returns the prior distribution.

Computationally, the analysis of a hierarchical model in a Bayesian framework is equivalent to the data cloning analysis for the original data ($k = 1$). Once the model is programmed into WinBUGS or any software using MCMC, data cloning is simply a matter of copying the data and repeating the analysis. The downside to this method is an increase in time and possible computational power necessary to run the analysis with larger number of clones. However, once the MCMC results are obtained, it is relatively simple to generate the eigenvalue plots and assess estimability using the methods of Lele et al. (2010).

The additional time and effort spent using data cloning to diagnose estimability problems can help ensure that results of fitting a realistic model to ecological data are reliable.

LITERATURE CITED

Ball, Lianne C., Paul F. Doherty Jr, and Matthew W. McDonald. "An occupancy modeling approach to evaluating a Palm Springs ground squirrel habitat model." *Journal of Wildlife Management* 69.3 (2005): 894-904.

Campbell, David, and Subhash Lele. "An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems." *Computational Statistics & Data Analysis* 70 (2014): 257-267.

Catchpole, Edward A., and Byron JT Morgan. "Detecting parameter redundancy." *Biometrika* 84.1 (1997): 187-196.

Gelfand, Alan E., and Sujit K. Sahu. "Identifiability, improper priors, and Gibbs sampling for generalized linear models." *Journal of the American Statistical Association* 94.445 (1999): 247-253.

Gimenez, Olivier, Ann Viallefont, Edward A. Catchpole, Réni Choquet, Byron J.T. Morgan. "Methods for investigating parameter redundancy." *Animal Biodiversity and Conservation* 27.1 (2004): 561-572.

Hubbard, Ben Arthur. *Parameter Redundancy with Applications in Statistical Ecology*. Diss. University of Kent, 2014.

Jensen, Timothy, and Jason C. Vokoun. "Using multistate occupancy estimation to model habitat use in difficult-to-sample watersheds: bridle shiner in a low-gradient swampy stream." *Canadian Journal of Fisheries and Aquatic Sciences* 70.10 (2013): 1429-1437
.

Lele, Subhash R., Brian Dennis, and Frithjof Lutscher. "Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods." *Ecology letters* 10.7 (2007): 551-563.

Lele, Subhash R., Khurram Nadeem, and Byron Schmuland. "Estimability and likelihood inference for generalized linear mixed models using data cloning." *Journal of the American Statistical Association* 105 (2010): 1617-1625.

Lele, Subhash R., Monica Moreno, and Erin Bayne. "Dealing with detection error in site occupancy surveys: what can we do with a single survey?." *Journal of Plant Ecology* 5.1 (2012): 22-31.

Lunn, David J., Andrew Thomas, Nicky Best, and David Spiegelhalter. "WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility." *Statistics and computing* 10.4 (2000): 325-337.

MacKenzie, Darryl I., James D. Nichols, James E. Hines, Melinda G. Knutson, and Alan B. Franklin. "Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly." *Ecology* 84.8 (2003): 2200-2207.

MacKenzie, Darryl I., James D. Nichols, Mark E. Seamans, and R. J. Gutiérrez.. "Modeling species occurrence dynamics with multiple states and imperfect detection." *Ecology* 90.3 (2009): 823-835.

Martin, Julien, Carol L. McIntyre, James E. Hines, James D. Nichols, Joel A. Schmutz, and Maggie C. MacCluskie. "Dynamic multistate site occupancy models to evaluate hypotheses relevant to conservation of Golden Eagles in Denali National Park, Alaska." *Biological Conservation* 142.11 (2009): 2726-2731.

McCullagh, Peter, and John A. Nelder. *Generalized linear models*. Vol. 37. CRC press, 1989.

Meeker, William Q., and Luis A. Escobar. "Teaching about approximate confidence regions based on maximum likelihood estimation." *The American Statistician* 49.1 (1995): 48-53.

Nichols, James D., James E. Hines, Darryl I. Mackenzie, Mark E. Seamans, and R. J. Gutierrez. "Occupancy estimation and modeling with multiple states and state uncertainty." *Ecology* 88.6 (2007): 1395-1400.

Ponciano, Jose Miguel, Mark L. Taper, Brian Dennis, and Subhash R. Lele. . "Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning." *Ecology* 90.2 (2009): 356-362.

R Core Team. "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. (2015). Version 3.1.3. http://www.R-project.org/.

Raue, Andreas, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood." *Bioinformatics* 25.15 (2009): 1923-1929.

Saracco, James F., Rodney B. Siegel, and Robert L. Wilkerson. "Occupancy modeling of Black-backed Woodpeckers on burned Sierra Nevada forests." *Ecosphere* 2.3 (2011): 1-17.

Spiegelhalter, D.J., A. Thomas, and N.G. Best. 1999. WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit.

Viallefont, Anne, Jean-Dominique Lebreton, Anne-Marie Reboulet, and Gerard Gory. "Parameter Identifiabîlîty and Model Selection in Capture-Recapture Models: A Numerical Approach." *Biometrical Journal* 40.3 (1998): 313-325.

Walker, A. M. "On the asymptotic behaviour of posterior distributions." *Journal of the Royal Statistical Society. Series B (Methodological)* (1969): 80-88.

SUPPLEMENTAL MATERIAL

Appendix A:

Simulation and Replication R Code:

```
#### R Program for Simulating and Replicating Data from a
####        Simple Single-Season Occupancy Model
####  A. Bowe - M.S. Candidate in Statistical Science: University of Idaho
####  Version 5.0 - April 11, 2016

#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#

#~#~#~#~#~#~# Create Function for Generating Simulated Datasets #~#~#~#~#~#~#

GenSimData = function(n,m,psi,p) {
      Sites=rbinom(n,1,psi)

      Det= matrix(NA, nrow = length(Sites), ncol = m)
            for(row in 1:n){Det[row,] = rbinom(m,1,p) }
      His=Sites*Det
       x=vector(length=n)
            for (i in 1:n){x[i] = sum(His[i,]) }
      M=rep(m,n)
      theta=ifelse (rowSums(His)>=1,1,0)

      Data=cbind(His,x,M,theta)
      WinData=cbind(x,M,theta)

      result=list(Data,WinData)
      return(result)
}

#~#~#~#~#~#~# Simulate data using the function #~#~#~#~#~#~#

n=25        # number of sites
m=5         # of surveys
psi=.7      # simulated value of psi
p=.9        # simulated value of p

Sim<-GenSimData(n,m,psi,p)

# Sim[1] is the first component of the result, this is the complete set of
data
#     simulated, including the detection histories
# Sim[2] is the simulated data in the format need to enter into WinBUGS


#~#~#~#~#~#~# Write Simulated Data to Text Files #~#~#~#~#~#~#
```

```
write.table(Sim[1],file="C:/Users/.../F_n25m5psi7p9k1_2.txt")
write.table(Sim[2],file="C:/Users/.../W_n25m5psi7p9k1_2.txt",row.names=FAL
SE,col.names=TRUE)


# File names are interpreted as:
#     F    = full data (Sim[1])
#     W    = data in WinBUGS format
#     n25  = simulated using samplel size n = 25
#     m5   = simulated number of surveys m = 5
#     psi7 = simulated using a psi = 0.7
#     p    = simulated using a p = .9
#     k    = number of replications of the data (clones)
#     _2   = 2nd dataset simulated from these parameters


#~#~#~#~#~# Generate Clones of Dataset #~#~#~#~#~#~#

# Read in dataset to be replicated
Data = read.table("C:/Users/.../W_n25m5psi7p9k1_2.txt",header=TRUE)

K5=rbind(Data,Data,Data,Data,Data)
K10=rbind(K5,K5)
K25=rbind(K5,K5,K5,K5,K5)
K50=rbind(K10,K10,K10,K10,K10)

write.table(K5,file="C:/Users/.../W_n25m5psi7p9k5_2.txt",row.names=FALSE,c
ol.names=TRUE)

write.table(K10,file="C:/Users/.../W_n25m5psi7p9k10_2.txt",row.names=FALSE
,col.names=TRUE)

write.table(K25,file="C:/Users/.../W_n25m5psi7p9k25_2.txt",row.names=FALSE
,col.names=TRUE)

write.table(K50,file="C:/Users/.../W_n25m5psi7p9k50_2.txt",row.names=FALSE
,col.names=TRUE)
```

Appendix B:

WinBUGS Code for Obtaining Estimates from the Simple Occupancy Model:

This section contains code for calculating MLEs, standard errors and correlations from the simple occupancy model. The model statement is written using the "zero trick" setting λ_i = to the log-likelihood of the data given in Eq. 5. Initial values for psi and p were set to 0.5 and 0.5 and a $Beta(2,2)$ prior was used for both. The sample data shown was simulated with parameters $n = 25, m = 5, \psi = 0.6$, and $p = 0.4$ and is shown with $k = 1$ clones.

```
model {
   for (i in 1:N) {
      zeros[i]<-0
      lambda[i]<--log(exp(log(psi)+x[i]*log(p)+(k[i]-x[i])*log(1-p))+
         (1-theta[i])*(1-psi))
   zeros[i]~dpois(lambda[i])
}
   psi~dbeta(2,2)
   p~dbeta(2,2)
}


DATA
list(N=25)
x[] k[] theta[]
0 5 0
0 5 0
0 5 0
3 5 1
2 5 1
0 5 0
1 5 1
2 5 1
3 5 1
0 5 0
0 5 0
0 5 0
3 5 1
1 5 1
2 5 1
3 5 1
0 5 0
0 5 0
2 5 1
3 5 1
0 5 0
1 5 1
2 5 1
2 5 1
0 5 0
END

INITS
list(p = .5, psi = .5)
```

Appendix C:

R Code for Creating Eigenvalue Diagnostic Plots:

```
#### R Program for Generating Eigenvalue Diagnostic Plots for Statistics
#     Obtain through Data Cloning
####  A. Bowe - M.S. Candidate in Statistical Science: University of Idaho
####  Version 6.0 - April 12, 2016

#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#~#
#~#~#


#~#~#~#~#~# Read in Table of Maximum Likelihood Estimates, Standard Errors
#           and Correlations obtained via MCMC

Data <- read.delim("~/R/Thesis/SimResults_15Feb2016.txt")
attach(Data)
head(Data)

#~#~#~#~#~# Create a function to generate the Var-Cov matrix and
#           calculate the Largest Eigenvalue of this matrix

# Intialize Vectors
n=length(P)
Vp=vector(length=n)
Vpsi=vector(length=n)
Cor=vector(length=n)
Cov=vector(length=n)
maxE=vector(length=n)

MaxE=function(pSE,psiSE,Corr){
  Vp=pSE^2
  Vpsi=psiSE^2
  Cov=pSE*psiSE*Corr
  VC.mat=matrix(c(Vp,Cov,Cov,Vpsi),ncol=2,nrow=2)
  VC.mat
  ee=eigen(VC.mat,symmetric=TRUE)
  VC.mat
  ee
  maxEigen=max(ee$values)
  return(maxEigen)
}

#~#~#~#~#~# Use function to Generate a Vector of Maximum Eigenvalues for
      #For every Parameter Combination at each Number of Clones (k)
MaxEigen = vector(length=n)

for (i in 1:n){
  MaxEigen[i] = MaxE(p_SE[i],psi_SE[i],Correlation[i])
}
MaxEigen
```

```
#~#~#~#~#~#  Sort and Standardize with each Parameter Combination

# Place in 5 column matrix; each row will equal one parameter value with
#     each level of cloning (1,5,10,25,50) as a column.

ME=matrix(MaxEigen,nrow=81,ncol=5,byrow=TRUE)

# Standardize this matrix by dividing by dividing by the maximum
Eigenvalue for
#     each parameter combination

SME=matrix(MaxEigen,nrow=81,ncol=5,byrow=TRUE)

for(i in 1:81){
SME[i,1] = ME[i,1]/max(ME[i,])
SME[i,2] = ME[i,2]/max(ME[i,])
SME[i,3] = ME[i,3]/max(ME[i,])
SME[i,4] = ME[i,4]/max(ME[i,])
SME[i,5] = ME[i,5]/max(ME[i,])
}
SME

#~#~#~#~#~#  Generate An Eigenvalue Diagnostic Plot #~#~#~#~#~#
K=c(1,5,10,25,50)

plot(SME[2,1:5],type="l", main="Sim2",xlab="Number of Clones",
ylab="Maximum Eigenvalue",xaxt="n")
axis(side = 1, at=c(1,2,3,4,5),labels = K, tck=-.05)
```

Appendix D:

Additional Eigenvalue Diagnostic Plots:

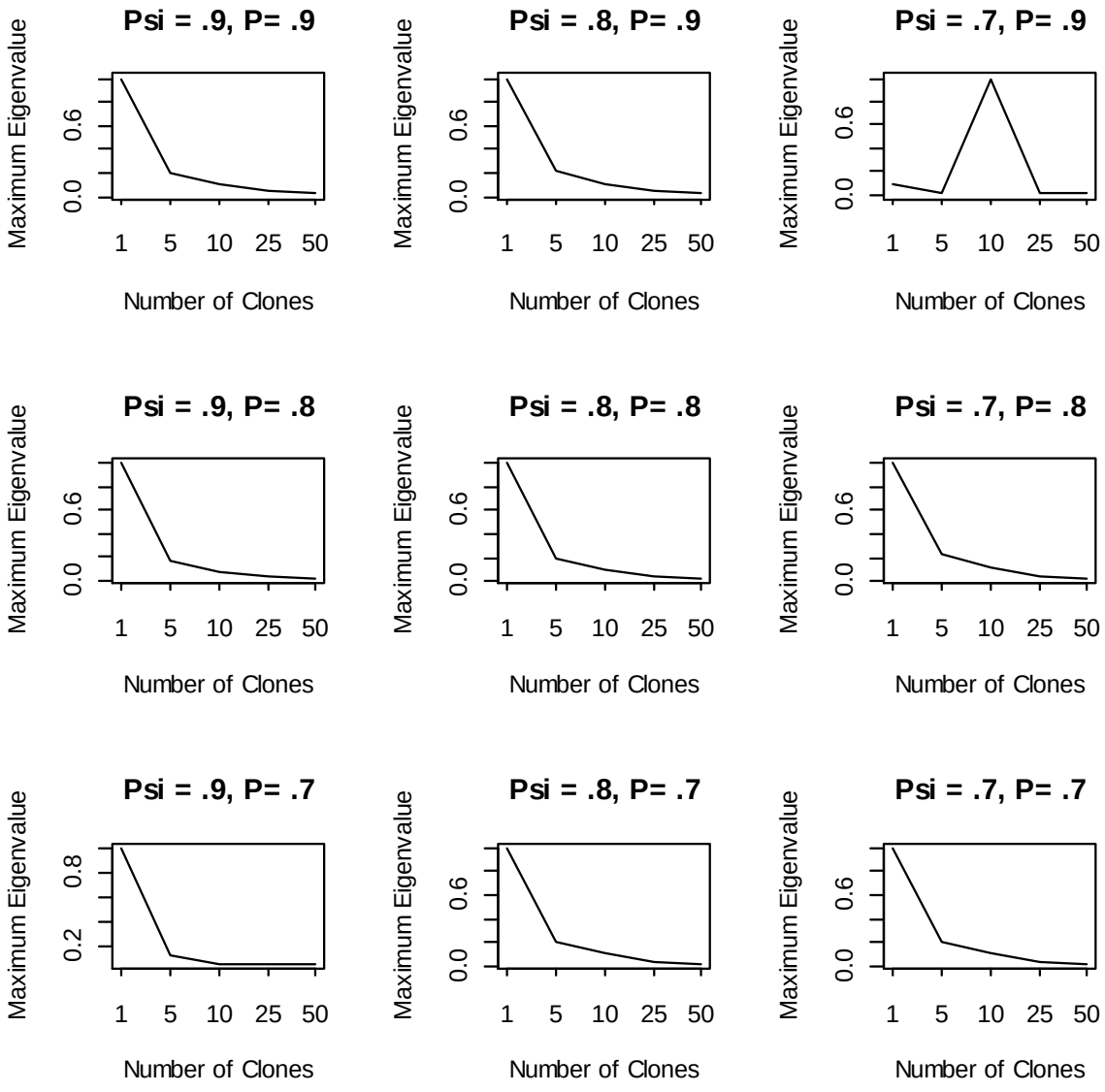

Figure 5: Eigenvalue diagnostic plots for ψ and p values between 0.7 and 0.9 showing trends indicating the estimability of parameters.

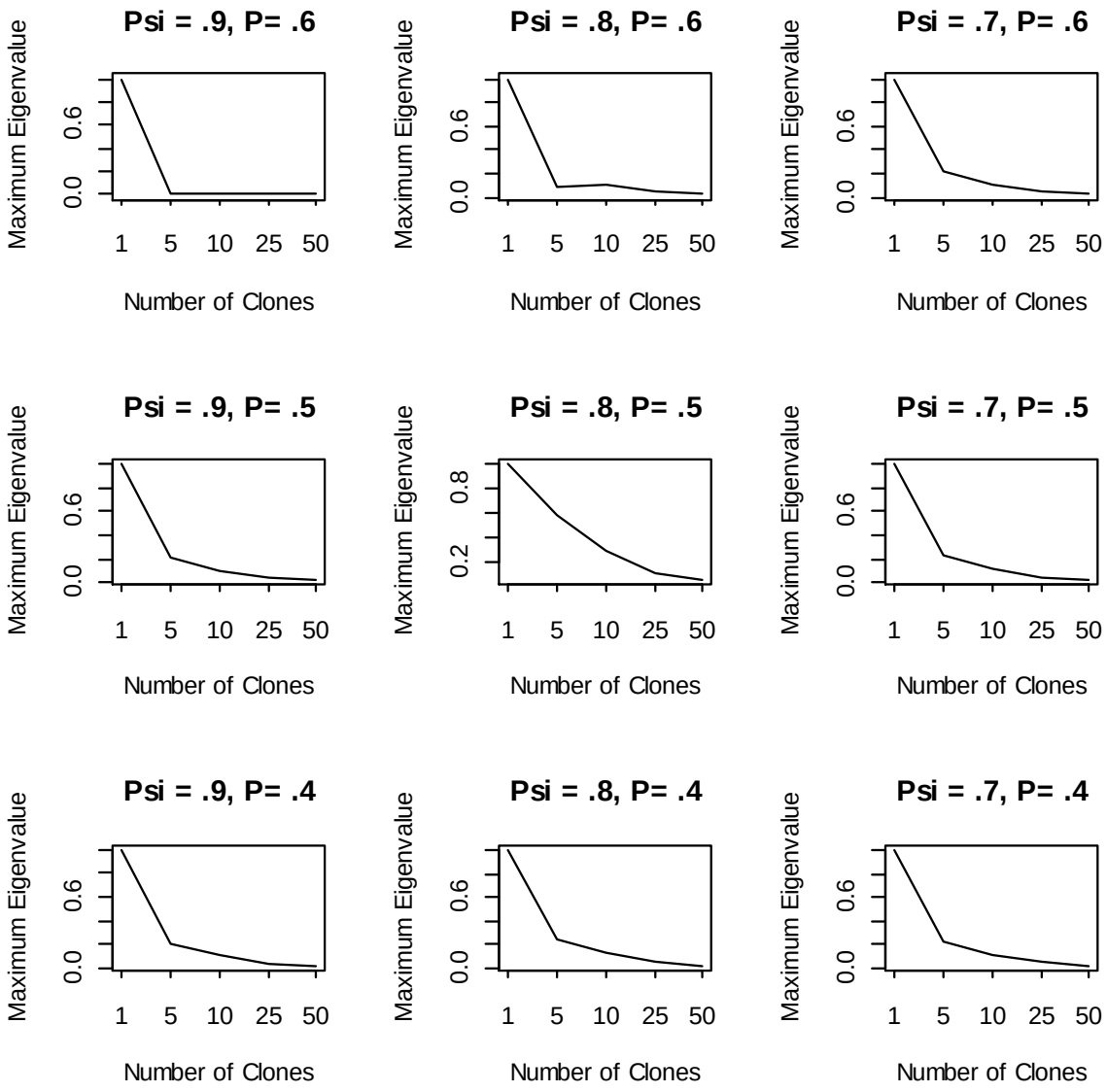

Figure 6: Eigenvalue diagnostic plots for ψ between 0.7 and 0.9 and p values between 0.4 and 0.6 showing trends indicating the estimability of parameters.

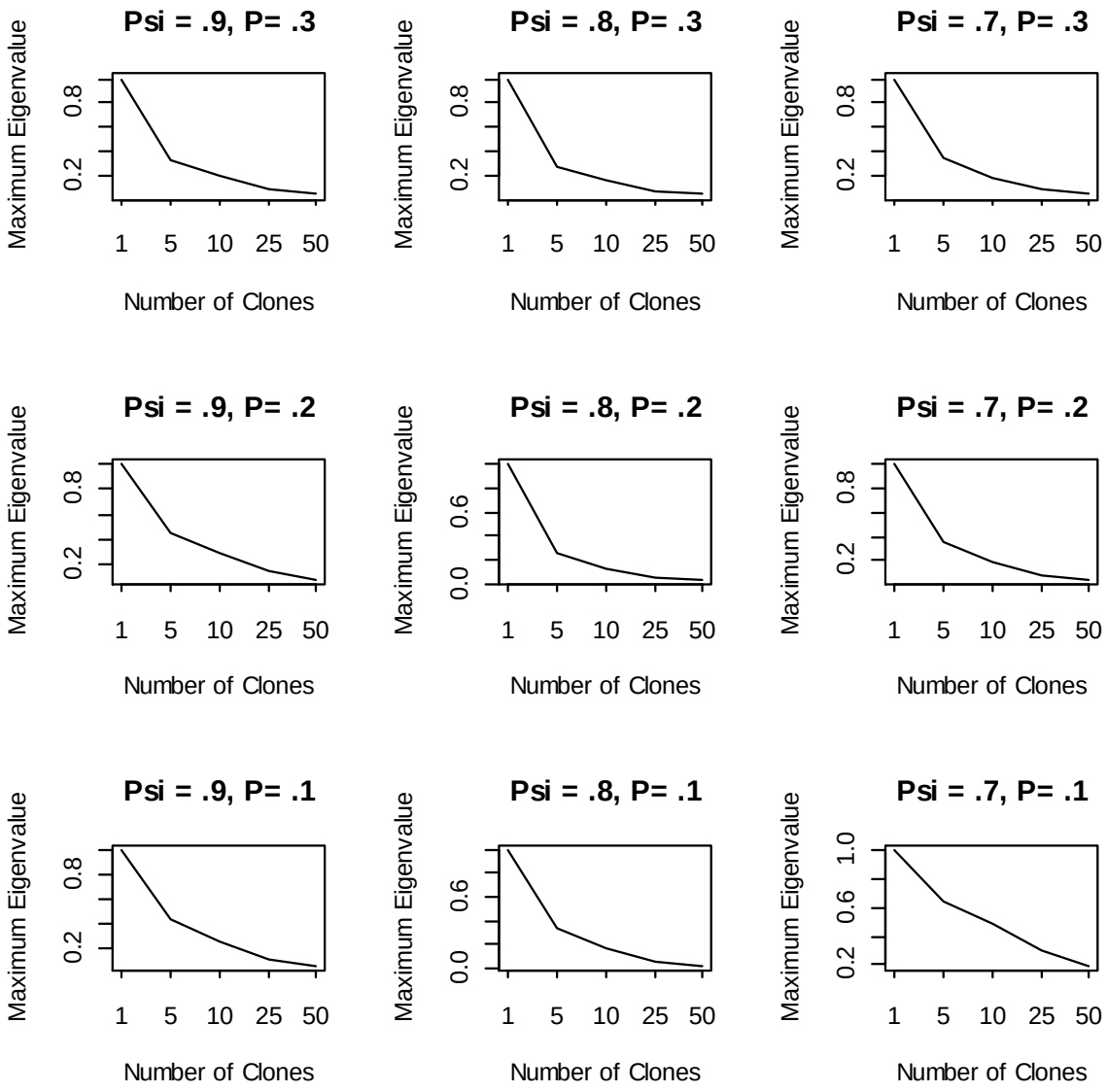

Figure 7: Eigenvalue diagnostic plots for ψ between 0.7 and 0.9 and p values between 0.1 and 0.3 showing trends indicating the estimability and near non-estimability of parameters.

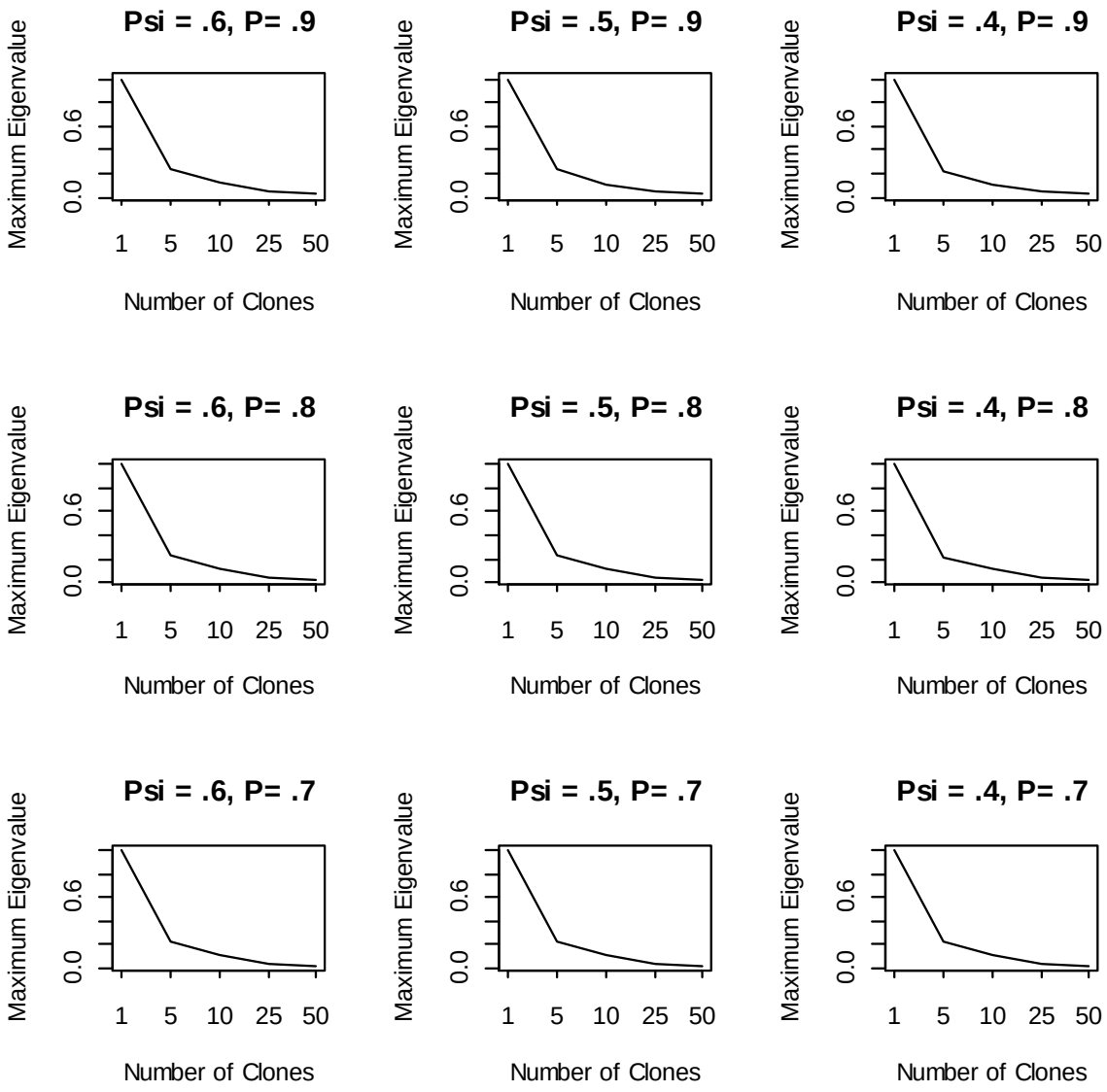

Figure 8: Eigenvalue diagnostic plots for ψ between 0.4 and 0.6 and p values between 0.7 and 0.9 showing trends indicating the estimability of parameters.

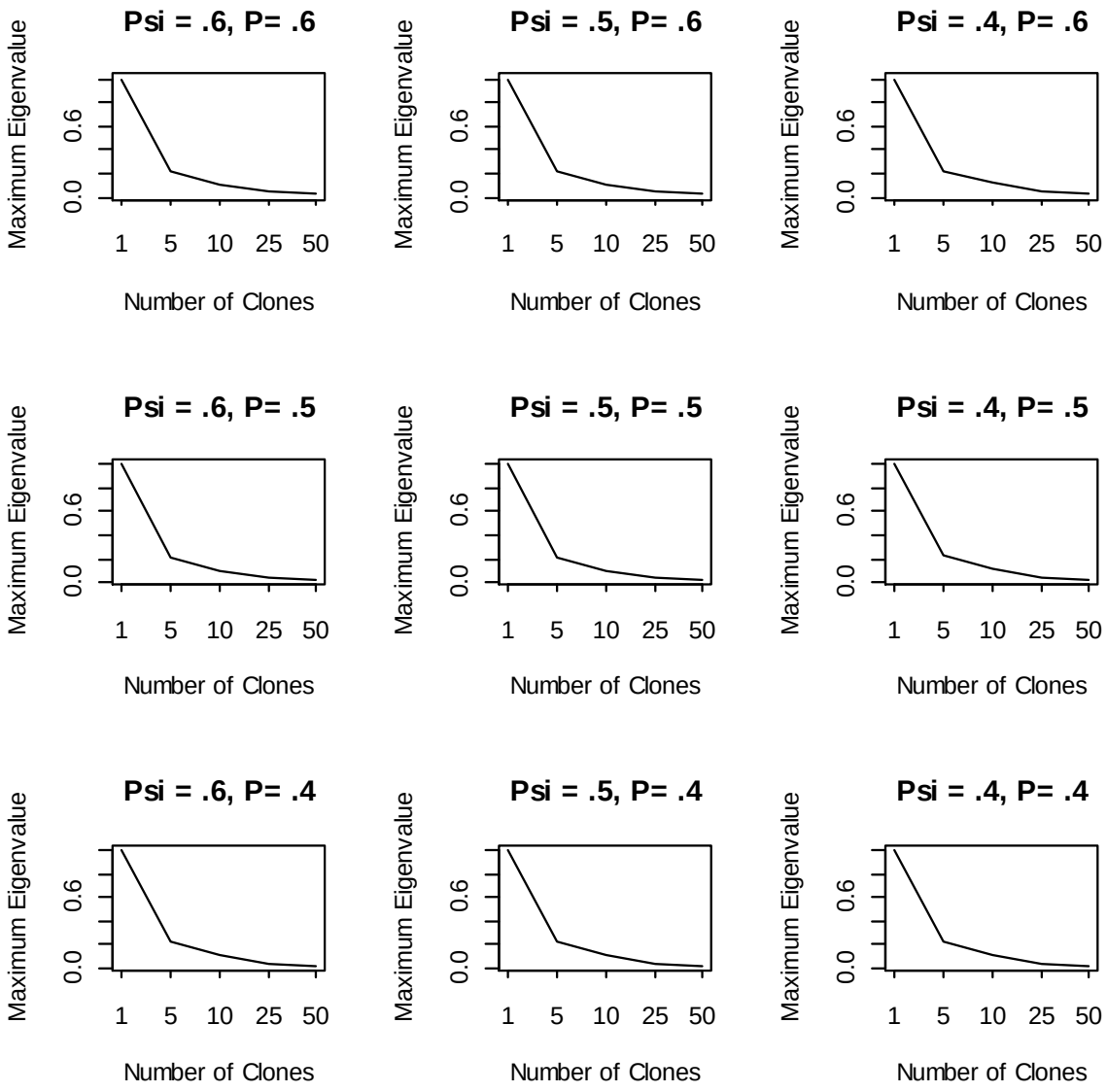

Figure 9: Eigenvalue diagnostic plots for ψ and p values between 0.4 and 0.6 showing trends indicating the estimability of parameters.

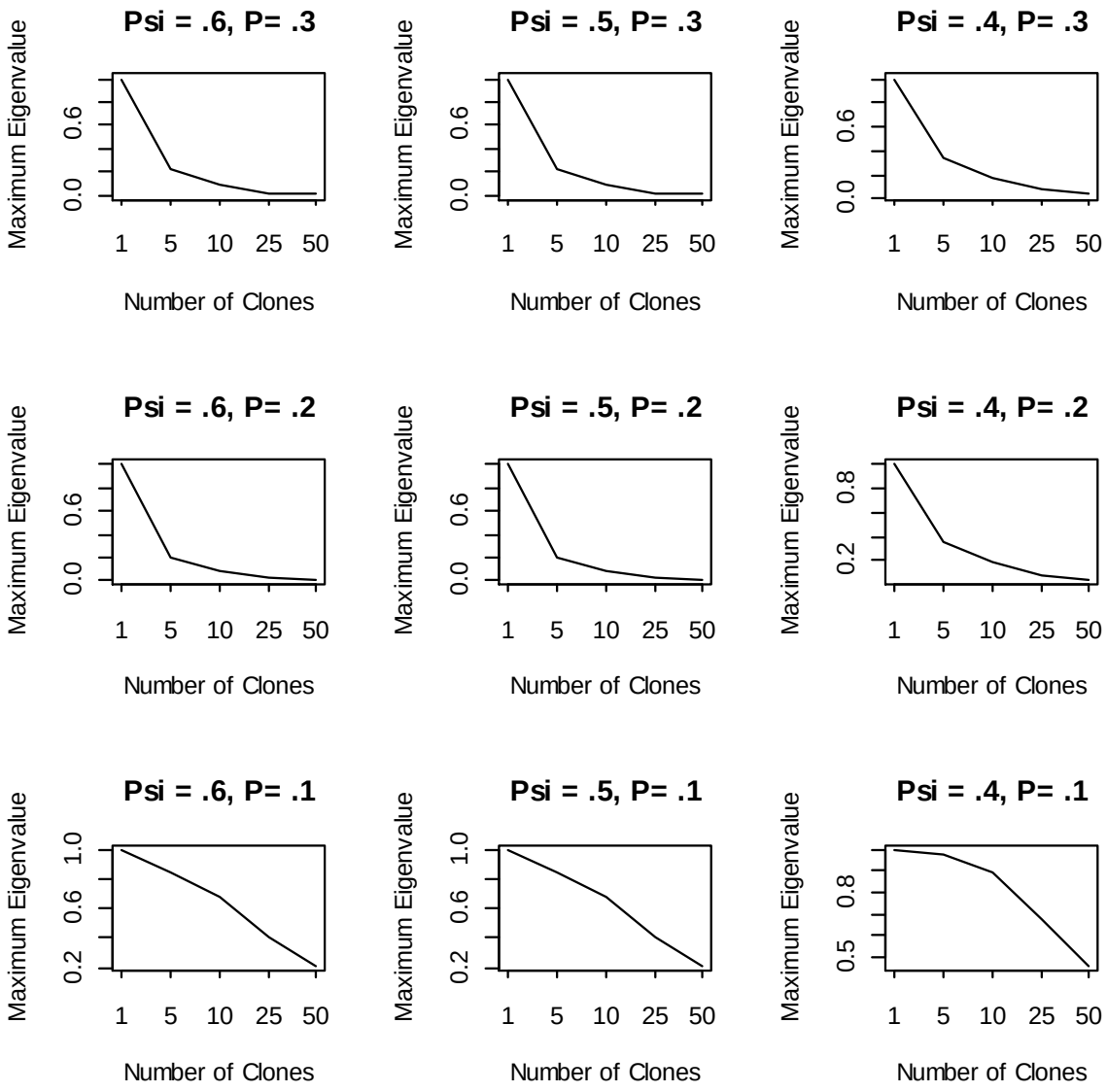

Figure 10: Eigenvalue diagnostic plots for ψ between 0.4 and 0.6 and p values between 0.1 and 0.3 showing trends indicating the estimability and near non-estimability of parameters.

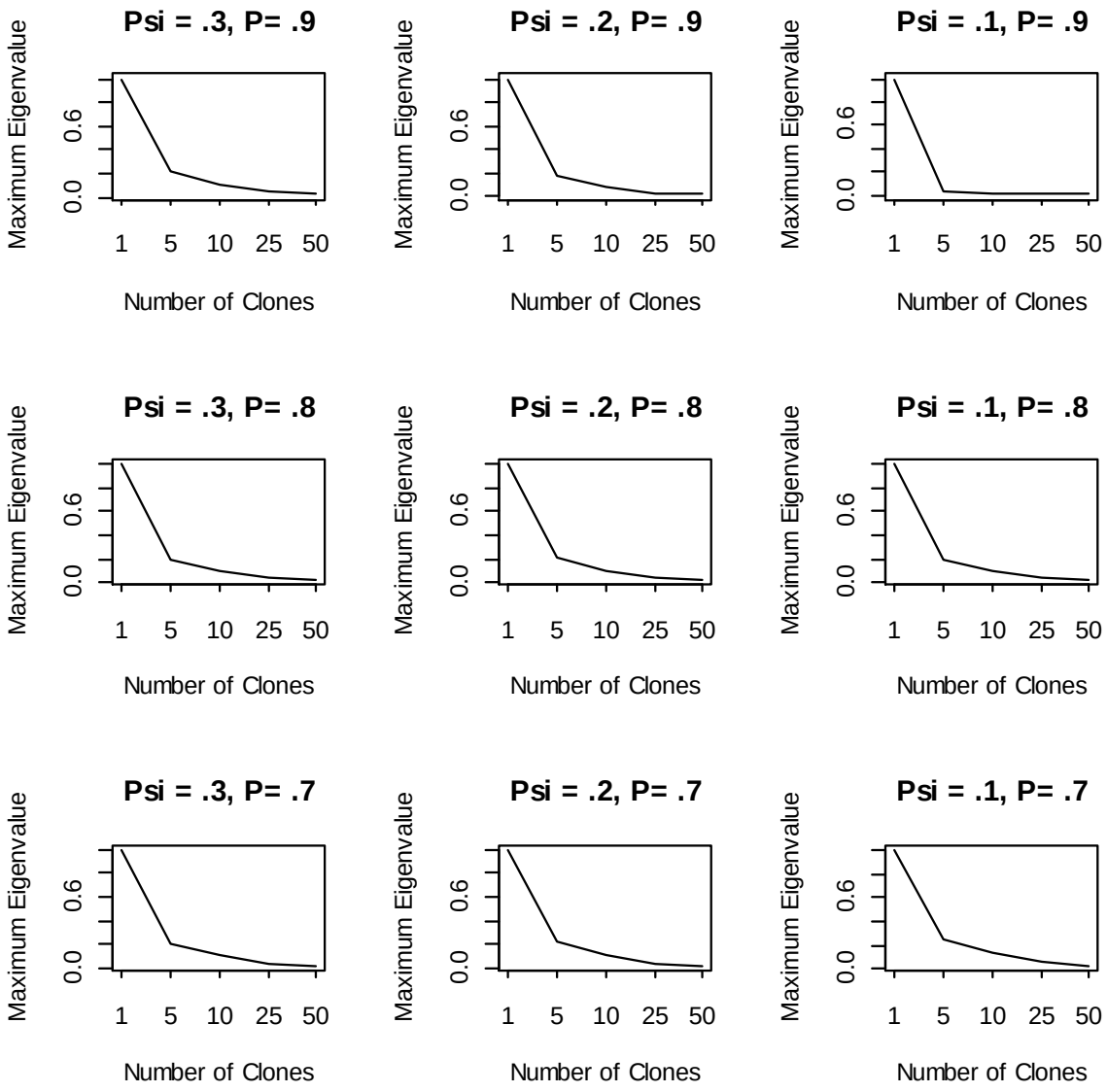

Figure 11: Eigenvalue diagnostic plots for ψ between 0.1 and 0.3 and p values between 0.7 and 0.9 showing trends indicating the estimability of parameters.

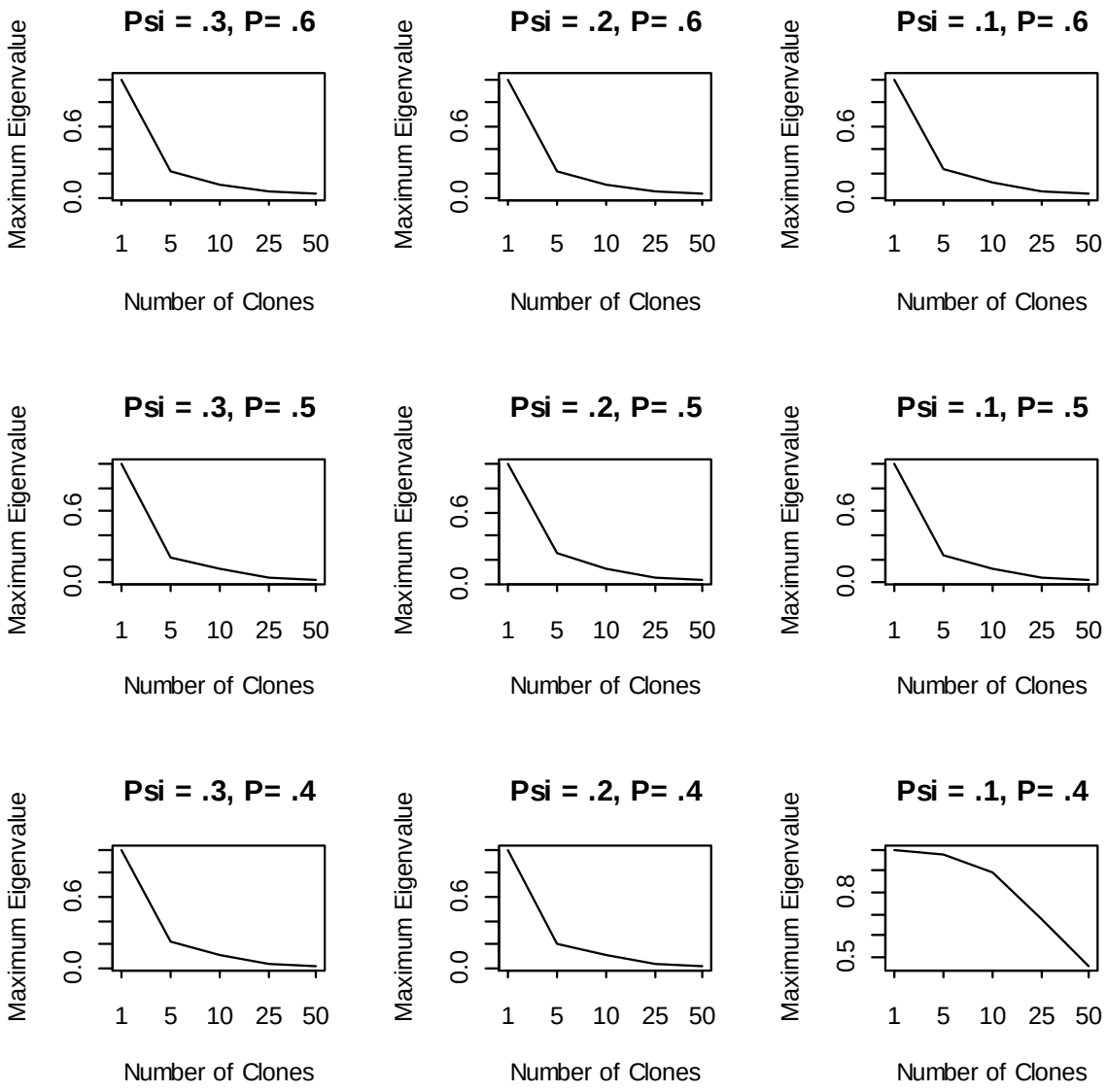

Figure 12: Eigenvalue diagnostic plots for ψ between 0.1 and 0.3 and p values between 0.4 and 0.6 showing trends indicating the estimability and near non-estimability of parameters.