Diagnostic Methods for Maxent Models in Ecology

by

Jeffrey T. Daniel

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics and Statistics

Guelph, Ontario, Canada

© Jeffrey T. Daniel, December, 2014
Abstract

Diagnostic Methods for Maxent Models in Ecology

Jeffrey T. Daniel
University of Guelph, 2014

Advisors:
Gary J. Umphrey, David W. Kribs

Understanding the geographic distributions of species is a fundamental problem in ecology. Many different statistical methods for modelling species distributions exist, but the most popular method is currently the machine learning algorithm Maxent. Despite its popularity, Maxent lacks the diagnostic tools available to more mature statistical models. In this thesis, we introduce leverage, influence, and residual methods for Maxent. We do so by applying recent results demonstrating the equivalence of Maxent and Poisson point process models. These results allow us to take methods from linear model theory and spatial statistics and adapt them to fit the Maxent framework. The result is a set of diagnostic methods for the critical evaluation of Maxent models. We illustrate these methods by applying them to Maxent models of the distributions of two ant species of the genus Trachymyrmex.
Acknowledgements

I would like to express my sincerest gratitude to my advisors, Dr. Gary Umphrey and Dr. David Kribs. I am truly thankful for all of their guidance and support during the production of this thesis and throughout my graduate studies.

I would also like to thank Dr. Ed Carter and Dr. Sanjeena Dang for taking time out of their busy schedules to serve on my examination committee.

Finally, I would like to thank Dr. Norman Johnson of The Ohio State University, whose *Pelecinus polyturator* dataset provided the initial inspiration for this thesis.
# Contents

1 Introduction 1

1.1 Species Distribution Modelling .......................... 1

1.2 Theoretical Basis of SDMs ............................... 2

1.3 Presence-Only Data .................................... 2

1.4 Modelling with Presence-Only Data ...................... 3

1.4.1 Maxent ............................................. 3

1.4.2 Maxent Model Evaluation ............................ 4

1.4.3 Maxent and the Inhomogeneous Poisson Process ...... 5

2 Models for Presence-Only Data 6

2.1 Maxent ................................................ 6

2.1.1 Motivation ....................................... 6

2.1.2 Model Definition .................................. 7

2.1.3 Model Interpretation ............................... 8

2.2 Inhomogeneous Poisson Process ......................... 10

2.2.1 Motivation ....................................... 10

2.2.2 Model Definition .................................. 10

2.2.3 Numerical Approximation and Poisson Regression ... 11

2.2.4 Model Interpretation ............................... 12

2.3 Equivalence of Maxent and the IPP Model ................ 13

3 Maxent Diagnostics 17

3.1 Leverage ............................................. 17

3.1.1 Classical Definition ............................... 17

3.1.2 Leverage for IPP Models ........................... 18
List of Tables

4.1 Bioclimatic variables used for Maxent models of *T. septentionalis* and
*T. turrifex*. ................................................................. 30
List of Figures

4.1 Maxent distribution of *T. septentrionalis*, logistic output.
4.2 Maxent distribution of *T. septentrionalis*, raw output.
4.3 Maxent distribution of *T. turrifex*, logistic output.
4.4 Maxent distribution of *T. turrifex*, raw output.
4.5 Leverage values for Maxent distribution of *T. septentrionalis*.
4.6 Leverage values for Maxent distribution of *T. turrifex*.
4.7 Influence values for Maxent distribution of *T. septentrionalis*.
4.8 Influence values for Maxent distribution of *T. turrifex*.
4.9 Maxent distribution of *T. septentrionalis* excluding Texas records.
4.10 Maxent distribution of *T. septentrionalis* excluding Georgia records.
4.11 Marginal influence of *T. septentrionalis* presence records by US state.
4.12 Smoothed residuals for Maxent distribution of *T. septentrionalis*.
4.13 Smoothed residuals for Maxent distribution of *T. turrifex*.
4.14 Cumulative residual plots for Maxent distribution of *T. septentrionalis*.
4.15 Cumulative residual plots for Maxent distribution of *T. turrifex*. 
Chapter 1

Introduction

1.1 Species Distribution Modelling

Understanding the relationship between a species and its environment has been identified as one of the great challenges in modern biology (Schwenk et al., 2009). Species distribution models (SDMs) offer a quantitative approach to identifying species-environment relationships, and how these relationships influence the geographic distributions of species. In species distribution modelling, the goal is to estimate the geographic range of a species using information about sites where a species has or has not been observed, along with the environmental conditions present at each site (Elith and Leathwick, 2009).

SDMs play a central role in many branches of the biological and environmental sciences. In biogeography, ecology, and evolutionary biology, SDMs are essential tools for answering fundamental questions about the geographic distributions of species (Miller, 2010). In conservation biology and wildlife management, SDMs provide crucial information for making informed management and policy decisions (Franklin, 2010). There is considerable interest in using SDMs to predict how climate change and other environmental shifts will affect species distributions (Bond et al., 2011). SDMs have also been used to identify regions at risk from invasive species (Barnes et al., 2014). This variety of applications had led to the development of many different SDM methods. However, most SDMs are motivated by the same underlying ecological theory.
1.2 Theoretical Basis of SDMs

The theoretical underpinning of SDMs is the concept of the ecological niche, which can be simply defined as “the ecological conditions that a species requires to maintain populations in a given region” (Peterson et al., 2011). Hutchinson (1957) gives a mathematical definition of the ecological niche. Suppose the distribution of a species is determined by a set of \( p \) environmental variables. These variables form a \( p \)-dimensional hyperspace, and the species’ niche is then “every point [in the \( p \)-dimensional hyperspace] which corresponds to a state of the environment which would permit the species to exist indefinitely” (Hutchinson, 1957). Thus the ecological niche corresponds to a hypervolume in \( p \)-dimensional environmental space; SDMs estimate this hypervolume, and then “project” it into geographic space to produce a mapped distribution (Elith and Leathwick, 2009). The \( p \) environmental variables are usually climatic in nature (i.e. temperature, precipitation). In such cases, the boundary of the species’ niche hypervolume is referred to as its bioclimatic envelope (Heikkinen et al., 2006).

Early SDMs, such as BIOCLIM (Busby, 1991), estimated the bioclimatic envelope directly by constructing a hyper-rectangle around observations in environmental space. Current SDMs take more sophisticated approaches to modelling species-environment relationships. More information regarding SDMs and their underlying ecological theory can be found in Elith and Leathwick (2009), Franklin (2010), and Peterson et al. (2011).

1.3 Presence-Only Data

Many different SDMs have been introduced, but selecting the right model for a particular application depends upon the available data. Ideally, species occurrence data is collected in a survey. For example, the study region may be divided into regular plots of land, and a random sample of plots selected. A surveyor then visits each sampled plot of land for a certain amount of time and records whether or not the species of interest is observed. If the species is observed, the species is recorded as present in that plot of land; otherwise, it is recorded as absent. Such designs produce presence-absence data, and the distribution of the species can be modelled.
using logistic regression and its many extensions (Hastie et al., 2009). These kinds of models are well-studied in the ecological literature, where they are often also referred to as resource selection models (in which a species “uses” the resources available in a plot) (Manly et al., 2002), or site occupancy models (in which a species “occupies” a plot) (MacKenzie et al., 2006).

Unfortunately, surveys are expensive and time-consuming, and they can be infeasible for rare, endangered, or difficult-to-detect species. For most species, all that is available is presence-only data, which consists only of records of where a species has been observed, with no information about where the species has not been observed (Pearce and Boyce, 2006). Sources of presence-only data include historical records held in museum and university collections, species lists and atlases, and radiotelemetry studies. Logistic regression and its extensions cannot be applied to presence-only data, requiring the use of alternative modelling methods. In this thesis, we consider two particular models for presence-only data: the Maxent model, and the inhomogeneous Poisson process model. While these two models have proved very popular and continue to hold much promise, how best to deal with presence-only data remains an open question in ecology and statistics (Hastie and Fithian, 2013).

1.4 Modelling with Presence-Only Data

1.4.1 Maxent

Maxent is by far the most popular presence-only SDM in use today: according to Google Scholar, a key Maxent paper (Phillips et al., 2006) has been cited over 3700 times as of October 2014. To model the distribution of a species, Maxent treats presence data as observations from a probability distribution defined over the study region, and estimates that distribution using the principle of maximum entropy (Phillips et al, 2006). Originally introduced in the context of statistical physics, the maximum entropy principle states that when fitting a probability distribution to data subject to a set of constraints, the best distribution is that which maximizes entropy (Jaynes, 1957). This approach is appealing because it identifies the distribution which uses all available information (the data and constraints) while making
the fewest assumptions. The resulting Maxent distribution is discrete, partitioning
the study region into a set of grid cells and assigning a probability to each cell,
where the probability associated with cell $i$ is the probability that, given an individ-
ual of the species was observed, that individual derived from cell $i$ (Phillips et al.,
2006). When total population size is known, these values can be used to estimate
the expected number of individuals occurring within each grid cell, which is the
species occurrence rate. However, total population size is almost never known; in
such cases, Maxent can only estimate relative occurrence rate (Fithian et al., 2013).
A full mathematical description of Maxent is given in Chapter 2.

Maxent was first introduced in 2004, and a software implementation was made
available in 2006. Maxent’s popularity is mostly due to an influential paper (Elith
et al., 2006) which compared the performance of several prominent SDMs, includ-
ing Maxent, well-established models such as generalized linear models (GLMs) and
generalized additive models (GAMs), simple bioclimatic envelope models such as
BIOCLIM, and more sophisticated presence-only models such as the Genetic Algo-
rithm for Rule-Set Production (GARP) (Stockwell and Peters, 1999). Maxent was
found to outperform GLMs, GAMs, BIOCLIM, and GARP with respect to pre-
dictive accuracy. Maxent was also hailed for its ability to fit complex functions of
environmental variables, which gave it a clear advantage over simpler SDMs. As a
result, Maxent quickly rose to become the most-used SDM; since 2006, BIOCLIM
and GARP have all but fallen out of use.

1.4.2 Maxent Model Evaluation

Despite Maxent’s popularity, it remains a “black box” model whose inner work-
ings are often poorly understood by practitioners (Elith et al., 2011). As a result,
Maxent lacks many of the inferential and diagnostic tools and methods that are
available to users of other, better-understood statistical models such as GLMs and
their extensions (Faraway, 2005). Additionally, the SDM community has been pre-
occupied with the development of performance metrics and methods for assessing
the accuracy of predictions (Elith and Leathwick, 2009). Examples include the re-
ceiver operator characteristic (ROC) curve, and the area under that curve (AUC),
which assess model accuracy by measuring classification error rates when applying
a model to test data (Hastie et al., 2009). Performance measures like AUC are ubiquitous in the SDM literature, and their advantages and disadvantages have been enthusiastically debated (Lobo et al., 2008). Discussion of other methods of Maxent model evaluation has been conspicuously absent. What is needed are powerful techniques for evaluating Maxent models on the basis of criteria other than predictive performance.

### 1.4.3 Maxent and the Inhomogeneous Poisson Process

Maxent may be the current SDM of choice for many practitioners, but an alternative modelling approach based on spatial statistics has rapidly been gaining ground. Spatial statistical models have been argued to hold several advantages over Maxent, including a more intuitive modelling framework and more easily interpretable results (Aarts et al., 2012). Recently, interest in spatial methods has increased dramatically following the publication of a number of papers demonstrating that Maxent is equivalent to a spatial statistical model called an inhomogeneous Poisson process (IPP) model (Aarts et al., 2012; Renner and Warton, 2013; Fithian et al., 2013). A mathematical description of the IPP model is given in Chapter 2.

There is a vast body of knowledge underlying the theory of point processes, especially concerning their relationship to linear model theory (Cressie, 1993). Recent research has been concerned with developing diagnostic methods for point process models, analogous to those for linear models (Baddeley et al., 2005, 2008, 2013). If we wish to develop diagnostic methods for Maxent models, it seems sensible to examine those that exist for the IPP model and exploit the IPP–Maxent equivalence to adapt those IPP methods for use with Maxent.

In Chapter 2 we describe the Maxent and IPP models and demonstrate their equivalence. In Chapter 3, we present diagnostic methods for Maxent models. We introduce measures of leverage and influence analogous to those available for linear models, and we also show how residuals for point processes can be adapted for use with Maxent. We then demonstrate these diagnostics by applying them to real data, evaluating Maxent distributions of two North American fungus-growing ants: *Trachymyrmex septentrionalis* and *Trachymyrmex turrifex*.
Chapter 2

Models for Presence-Only Data

2.1 Maxent

2.1.1 Motivation

The problem of modelling species distributions is essentially a problem of density estimation: given a set of sample points and a set of constraining variables, how can we find the best estimate of the unknown distribution? One solution, first proposed by Jaynes (1957), is known as the maximum entropy principle: the probability distribution subject to the given constraints which best describes the data is that with maximum entropy.

In statistics and information theory, the entropy of a probability distribution \( p(x) \) is defined as

\[
H(x) = - \sum_i p(x_i) \ln[p(x_i)].
\]  

(2.1)

Shannon (1948) describes entropy as “a measure of how much ‘choice’ is involved in the selection of an event”. In this context, the “event” is the presence of a species, the outcome of which is “selected” by the species distribution. The maximum entropy distribution is then that for which there are the most choices available – that is, the least constrained distribution. In this way, the maximum entropy principle leads to “the least biased estimate possible on the given information … [that is] maximally noncommittal with regard to missing information” (Jaynes, 1957).
2.1.2 Model Definition

Let $D$ be a geographic domain of area $|D|$ partitioned into $n$ discrete grid cells, each of equal size $|D|/n$. For each grid cell $d_i \in D$, define a binary random variable $y_i$ equal to 1 if the species is known to be present in $d_i$, and equal to 0 otherwise. There will be $m$ grid cells for which $y_i = 1$, and we assume that this set of presence cells $\{d_1, \ldots, d_m\}$ constitutes a random sample from some probability distribution $\pi(d)$. Our goal is to estimate the unknown distribution $\pi$.

Let $x_i$ be a $p$-length vector of environmental covariates defined for all $d_i \in D$. These covariates are generally referred to as features in the Maxent literature. They can be raw environmental variables, such as temperature and precipitation, or functions of these variables. The Maxent software accepts raw environmental data as input and can automatically compute features of six classes: linear, quadratic, product, threshold, hinge, and categorical (Phillips and Dudik, 2008). Additionally, any other function can be accommodated as a feature by computing it manually and having Maxent treat it as a linear feature (Merow et al., 2013).

To apply the maximum entropy principle, we must define constraints using the environmental features. Phillips et al. (2004) proposed constraining the distribution so that the expected values of the environmental features under the distribution $\pi$ match their sample means:

$$\hat{E}_\pi[x_i] = \frac{1}{m} \sum_{i:y_i=1} x_i. \quad (2.2)$$

Under these constraints, it can be shown (Della Pietra et al., 1997) that the resulting maximum entropy solution is a Gibbs distribution of the form

$$\pi(d_i) = \frac{e^{\beta' x_i}}{\sum_D e^{\beta' x_i}}, \quad (2.3)$$

where $\beta$ is a $p$-length vector of feature coefficients. Furthermore, it can also be shown, via convex duality, that the maximum entropy Gibbs distribution is equivalent to the maximum likelihood Gibbs distribution (Della Pietra et al., 1997). Therefore, the Maxent distribution can be found by maximizing the log-likelihood function of
the Gibbs distribution (Phillips et al., 2006):

$$
\ell(\beta) = \sum_{i:y_i=1} \log \left( \frac{e^{\beta x_i}}{\sum_{D} e^{\beta x_i}} \right).
$$

(2.4)

In the current software implementation of Maxent, the constraints are relaxed. Strict equality is not required, only that for each feature $x_{ij},$

$$
|E[x_{ij}] - \frac{1}{m} \sum_{i} x_{ij}| \leq r_j,
$$

(2.5)

where the $r_j$ are nonnegative constants chosen automatically by the software on the basis of empirical tuning studies (Phillips et al., 2006). This leads to a form of lasso regularization (Hastie et al., 2009) where the Maxent solution is found by optimizing a penalized Gibbs log-likelihood function:

$$
\ell(\beta) = \sum_{i:y_i=1} \log \left( \frac{e^{\beta x_i}}{\sum_{D} e^{\beta x_i}} \right) - \sum_{j=1}^{p} r_j |\beta_j|.
$$

(2.6)

This regularization scheme prevents overfitting by penalizing features with large coefficient values $\beta_j$. In fact, in models with a large number of features, most of the feature coefficients will be set to 0, resulting in a model that uses only the most significant features (Phillips et al., 2006). Simpler models are less likely to overfit to the data, and are preferred on the principle of parsimony.

More information about Maxent, including details of the fitting algorithms and the software implementation, can be found in Phillips et al. (2004), Phillips et al. (2006), Phillips and Dudik (2008), and Elith et al. (2011).

### 2.1.3 Model Interpretation

Maxent estimates $\pi(d)$, a probability distribution fitting the observed presence grid cell data. How exactly should $\pi(d)$ be interpreted? Consider the following idealized sampling scheme presented by Phillips and Dudik (2008). Suppose we select a grid cell $d_i \in D$ at random. The probability of selecting any cell is $Pr(d_i) = 1/n$. Under this scheme, $\pi(d_i)$ corresponds to $Pr(d_i|y_i = 1)$, the probability that a given presence record was derived from the chosen grid cell $d_i$. These probabilities constitute
Maxent’s raw output.

Maxent provides information about $Pr(d_i|y_i = 1)$, but the quantity of interest is usually $Pr(y_i = 1|d_i)$, the probability that the species is present in cell $d_i$. To find this probability, we apply Bayes' rule:

$$Pr(y_i = 1|d_i) = \frac{Pr(d_i|y_i = 1)Pr(y_i = 1)}{Pr(d_i)} \quad (2.7)$$

$$= \pi(d_i)Pr(y_i = 1)n. \quad (2.8)$$

We see that presence probability is proportional to $\pi(d_i)$. However, estimating $Pr(y_i = 1|d_i)$ requires information about $Pr(y_i = 1)$, the overall species prevalence. As demonstrated by Ward et al. (2009), accurate estimation of prevalence depends on the availability of both presence and absence data; when only presence data is available, $Pr(y_i = 1)$ is unidentifiable.

Thus, Maxent cannot directly estimate presence probability. At best, Maxent’s output values can be interpreted as relative probabilities. For example, if we have two cells $d_1$ and $d_2$, and the value of $\pi(d_1)$ is twice that of $\pi(d_2)$, then we can conclude that the presence probability of $d_1$ is also twice that of $d_2$. However, the absolute probabilities remain unknown.

Phillips and Dudik (2008) introduced a logistic output format for Maxent, transforming the raw model output $\pi$ to $e^{H\pi}/(1 + e^{H\pi})$, where $H$ is the model entropy. The logistic transformation approximates presence probability “assuming that the sampling design is such that typical presence localities have probability of presence of about 0.5.” (Phillips, 2008). Merow et al. (2013) notes that this assumption is often untenable in practice, and recommends avoiding logistic output. Regardless, logistic output is the default output setting for the Maxent software, and as such it is the most commonly-used output in the Maxent modelling literature.

The probabilistic interpretation of Maxent output depends upon the assumption that the sampling scheme described above is in fact true and presence data represent a random sample of geographic space. This assumption is often violated due to sampling bias and the haphazard collection of presence data. In such cases, Maxent output can still be interpreted qualitatively as an index of habitat suitability (Merow et al., 2013).
2.2 Inhomogeneous Poisson Process

2.2.1 Motivation

Diggle (1983) defines a *spatial point process* to be “any stochastic mechanism which generates a countable set of events... in the plane.” In the context of species distribution modelling, we consider the species presence locations to be the set of events, which occur within a bounded subset of the plane corresponding to the study region. The mechanism that determines the locations of events can be interpreted as the combination of environmental features that define the species’ bioclimatic envelope.

Point processes were first applied to species distribution modelling by Warton and Shepherd (2010). Their approach was motivated by perceived problems with the “pseudo-absence” approach to presence-only modelling. When no absence data is available for a species, it is common practice to take localities randomly sampled from the study region as “pseudo-absences” and use a presence-absence modelling method, such as logistic regression. Warton and Shepherd (2010) note several problems inherent in this approach, including issues of model interpretation. They proposed the use of point process models, which do not require pseudo-absences and have a natural and intuitive interpretation. Since then, many other authors have explored the use of point process models in distribution modelling, and especially its relationship to Maxent (Aarts et al., 2012; Fithian and Hastie, 2013; Renner and Warton, 2013).

2.2.2 Model Definition

Let \( D \subseteq \mathbb{R}^2 \) be a continuous geographic domain, and let \( Z = \{ z_i \} \) be a set of points corresponding to presence locations within the domain. An inhomogeneous Poisson process (IPP) is a model for such a set of points, and is defined by its *intensity function*

\[
\lambda : D \to [0, \infty)
\]  

(2.9)

where \( \lambda(z) \) gives the expected number of presence points \( z_i \) falling near \( z \). For any subregion \( A \subseteq D \)

\[
\Lambda(A) = \int_A \lambda(z)dz.
\]  

(2.10)
Finally, we assume that $\Lambda(D) < \infty$.

Following Diggle (1983), we note that IPPs have the following properties:

1. Let $N(A)$ be the number of presence points $z_i$ in the subregion $A$. Then $N(A)$ has a Poisson distribution with mean $\Lambda(A)$.

2. $N(A)$ and $N(B)$ are independent for disjoint subregions $A$ and $B$.

3. Conditional on the number of presence points $N(A)$, the presence points $z_i \in A$ form an independent random sample from the distribution with PDF $p(z) = \lambda(z)/\Lambda(A)$.

The IPP model is called “inhomogeneous” because the value of the intensity function $\lambda(z)$ depends on the value of one or more features (in a homogeneous Poisson process model, $\lambda$ would be constant). Let $x(z)$ be a $p$-length vector of environmental features defined for all points $z \in D$. Warton and Shepherd (2010) proposed modelling $\lambda(z)$ as a log-linear function of $x(z)$:

$$\lambda(z) = e^{\beta_0 + \beta'x(z)} \quad (2.11)$$

where $\beta_0$ is an intercept parameter and $\beta$ is a $p$-length vector of feature coefficients.

IPP models of this form are fit by maximizing the log-likelihood function (Cressie, 1993)

$$\ell(\beta_0, \beta) = \sum_Z (\beta_0 + \beta'x_i) - \int_D e^{\beta_0 + \beta'x(z)} dz - \log(|Z|!). \quad (2.12)$$

When there are a large number of features, IPPs can also be regularized to prevent overfitting, and recent software advances allow the implementation of Maxent-like lasso regularization for IPP models (Renner and Warton, 2013).

For more information about the IPP and other point process models, see Diggle (1983), Cressie (1993), or Gaetan and Guyon (2009).

### 2.2.3 Numerical Approximation and Poisson Regression

In practice, the IPP log-likelihood (2.12) usually cannot be evaluated analytically, and must be estimated numerically. Even if an analytic solution could be found, implementing the model on a computer would require some sort of discrete approximation. We present a numerical approximation method first proposed by Berman
and Turner (1992), and first applied in a species distribution modelling context by Warton and Shepherd (2010). First, partition the domain $D$ into $n$ discrete grid cells $d_1, \ldots, d_n$ of size $|D|/n$. Discretize the environmental features so that each grid cell $d_i$ has an associated feature vector $x_i$. Finally, discretize the presence points: let $y_i$ be a binary random variable equal to 1 if at least one presence point occurs in grid cell $d_i$, and equal to 0 otherwise; there will be $m$ grid cells for which $y_i = 1$. The approximate IPP log-likelihood is then

$$
\tilde{\ell}(\beta_0, \beta) = \sum_{i: y_i = 1} (\beta_0 + \beta' x_i) - \frac{|D|}{n} \sum_{i: y_i = 0} e^{\beta_0 + \beta' x_i} - \log(m!). \tag{2.13}
$$

This log-likelihood function is actually that for a weighted Poisson regression model of the form

$$
\mu(y|x) = \frac{|D|}{n} e^{\beta_0 + \beta' x} \tag{2.14}
$$

where each observation is weighted by the grid cell size. Indeed, the advantage of Berman and Turner’s method is that it allows IPP models to be fit using any software capable of fitting Poisson regression models.

The approximation of the continuous IPP model by the discrete Poisson model leads to some data loss due to discretization and incurs some bias. Baddeley et al. (2010) studied this bias and found that as $|D|/n \to 0$, $\tilde{\ell}(\beta_0, \beta) \to \ell(\beta_0, \beta)$; that is, as the grid cells become infinitely small (the so-called “fine-pixel limit”), the bias disappears and the Poisson log-likelihood converges to the IPP log-likelihood.

### 2.2.4 Model Interpretation

For every point $z \in D$, the IPP model estimates intensity $\lambda(z)$, the expected number of events occurring in the immediate vicinity of $z$. Additionally, for any region $A \subseteq D$, the IPP model estimates $\Lambda(A)$, the expected number of events occurring within the region $A$. In the context of species distribution modelling, the IPP intensity is interpreted as occurrence rate, the mean number of individuals of the species per unit area.

Occurrence rate is considered by many ecologists to be a more natural and intuitive measurement than presence probability (Fithian et al., 2013). Unlike presence probability, which depends crucially on the size of the grid cells for which it is de-
fined, occurrence rate is scale-independent, and “specifying the occurrence rate is equivalent to specifying occurrence probability simultaneously for all quadrat [grid cell] sizes” (Fithian et al., 2013).

To see why this is the case, recall IPP Property 3 given in section 2.2.2. We see that conditional on the total number of points $|Z|$, the set $Z$ forms an independent random sample from the probability distribution

$$p(z) = \frac{\lambda(z)}{\Lambda(D)} = \frac{e^{\beta_0+\beta'x(z)}}{\int_D e^{\beta_0+\beta'x(z)}dz}. \quad (2.15)$$

The value $p(z)$ may be interpreted as the probability of the species being present in the immediate vicinity of $z$, and this interpretation can be extended to give the presence probability for any subregion $A \subseteq D$.

Accurate estimation of $p(z)$ requires accurate estimation of $\Lambda(D)$, the overall population size within the study region, which is usually unknown. Examining equation (2.15), we can see that the role of the “intercept” parameter $\beta_0$ is actually to make sure that $p(z)$ integrates to 1. Thus, $\beta_0$ depends on the population size $\Lambda(D)$, and if the population size is unknown, $\beta_0$ cannot be accurately estimated. The result is that, as with Maxent, absolute presence probabilities cannot be estimated, and values of $p(z)$ must be interpreted as relative probabilities (Fithian et al., 2013).

The validity of the IPP model itself depends on the assumption that the presence locations $Z$ constitute a random sample of the species population, and due to sampling bias, this assumption is often violated (Fithian et al., 2013). In the worst cases, as with Maxent, IPP model output can be interpreted as a qualitative habitat suitability index.

Everything in this section applies equally to the Poisson GLM model. The only difference is the form of (2.15) is discretized:

$$p(d_i) = \frac{\lambda_i}{\Lambda(D)} = \frac{n!}{\sum_{i:y_i=0} e^{\beta_0+\beta'x_i}}. \quad (2.16)$$

### 2.3 Equivalence of Maxent and the IPP Model

So far in this chapter, we have discussed the Maxent and IPP models, and shown how the IPP reduces to a Poisson GLM when it is applied to discrete data. In this
section, we show that, despite their apparently very different formulations, Maxent and the IPP are in fact equivalent, and fit the same model.

To demonstrate this, we begin by considering an IPP model fitted to discrete data (i.e. a Poisson GLM). By deriving the maximum likelihood estimate (MLE) of $\beta_0$ and substituting it into the log-likelihood function, we can derive a Poisson GLM that has been partially maximized with respect to $\beta_0$. We will then show that the log-likelihood function of this new model is equivalent to that of the Maxent model, so that the two models have identical solutions for $\beta$. As we have already shown, the Poisson GLM is a discrete approximation to the continuous IPP. Since Maxent is equivalent to a partially-maximized Poisson GLM, Maxent must also be a discrete approximation to a partially-maximized IPP model. Thus, when applied to discrete data, Maxent and the partially-maximized IPP model are equivalent, and will yield the same solution.

Recall the log-likelihood function of the Poisson GLM, as given in equation (2.13). To find the MLE of $\beta_0$, we differentiate $\tilde{\ell}$ with respect to $\beta_0$ and set the result equal to zero:

$$0 = \frac{\partial}{\partial \beta_0} \left[ \sum_{i:y_i=1} (\beta_0 + \beta^\prime x_i) - \frac{|D|}{n} \sum_{i:y_i=0} e^{\beta_0 + \beta^\prime x_i} - \log(m!) \right]$$

$$0 = m - \frac{|D|}{n} \sum_{i:y_i=0} e^{\beta_0 + \beta^\prime x_i}$$

$$m = \frac{|D|}{n} \sum_{i:y_i=0} e^{\beta_0 + \beta^\prime x_i}. \quad (2.17)$$

Solving for $\beta_0$, we find that the MLE of $\beta_0$ is

$$\hat{\beta}_0 = \log(m) - \log \left( \frac{|D|}{n} \sum_{i:y_i=0} e^{\beta^\prime x_i} \right)$$

$$= \log \left( \frac{m}{\frac{|D|}{n} \sum_{i:y_i=0} e^{\beta^\prime x_i}} \right). \quad (2.18)$$

Substituting $\hat{\beta}_0$ into $\tilde{\ell}$, we obtain the partially-maximized Poisson GLM log-likelihood:

$$\tilde{\ell}^*(\beta) = \sum_{i:y_i=1} (\hat{\beta}_0 + \beta^\prime x_i) - \frac{|D|}{n} \sum_{i:y_i=0} e^{\hat{\beta}_0 + \beta^\prime x_i} - \log(m!)$$

14
\[
\sum_{i:y_i=1} \left[ \log(m) - \log \left( \sum_{i:y_i=0} \frac{|D|}{n} e^{\beta' x_i} \right) + \beta' x_i \right] - m - \log(m!) \\
= \sum_{i:y_i=1} \left[ \beta' x_i - \log \left( \sum_{i=1}^n e^{\beta' x_i} \right) \right] + C \\
= \sum_{i:y_i=1} \left[ \log(e^{\beta' x_i}) - \log \left( \sum_{i:y_i=0} e^{\beta' x_i} \right) \right] + C \\
= \sum_{i:y_i=1} \log \left( \frac{e^{\beta' x_i}}{\sum_{i:y_i=0} e^{\beta' x_i}} \right) + C. 
\] (2.19)

Compare (2.19) with the Maxent log-likelihood (2.4). We see that they differ only by a constant:
\[
\ell(\beta)_{\text{Maxent}} = \ell^*(\beta)_{\text{GLM}} - C. 
\] (2.20)

Since the constant \(C\) does not depend on \(\beta\), the maximum likelihood solutions for the two models are the same:
\[
\hat{\beta}_{\text{Maxent}} = \hat{\beta}_{\text{GLM}}. 
\] (2.21)

Therefore, for discrete data:
\[
\hat{\beta}_{\text{Maxent}} = \hat{\beta}_{\text{IPP}}. 
\] (2.22)

Furthermore, we can differentiate \(\tilde{\ell}^*\) with respect to \(\beta\) and set the result equal to zero:
\[
0 = \sum_{i:y_i=1} \frac{d}{d\beta} \left[ \beta' x_i - \log \left( \sum_{i:y_i=0} e^{\beta' x_i} \right) \right] \\
0 = \sum_{i:y_i=1} \left( x_i - \frac{\sum_{i:y_i=0} x_i e^{\beta' x_i}}{\sum_{i:y_i=0} e^{\beta' x_i}} \right) \\
0 = \sum_{i:y_i=1} x_i - m \frac{\sum_{i:y_i=0} x_i e^{\beta' x_i}}{\sum_{i:y_i=0} e^{\beta' x_i}} \\
\frac{1}{m} \sum_{i:y_i=1} x_i = \frac{\sum_{i:y_i=0} x_i e^{\beta' x_i}}{\sum_{i:y_i=0} e^{\beta' x_i}} \\
\frac{1}{m} \sum_{i:y_i=1} x_i = \tilde{E}_p[x]. 
\] (2.23)

This is the score equation for \(\tilde{\ell}^*(\beta)\), and it shows that the MLE of \(\beta\) is the \(\hat{\beta}\) that
satisfies the equation (2.23). This result is identical to equation (2.2), the Maxent model constraints. Thus, we see that Maxent and the discrete IPP employ the same fitting criterion: the log-likelihood is maximized by the coefficients $\hat{\beta}$ for which the expected values of the environmental features match their sample means.

Therefore, Maxent and the IPP model are equivalent for discrete data, and they fit the same feature coefficients $\hat{\beta}$. The IPP has an extra parameter $\beta_0$, but this serves only to scale the IPP model output; the relative magnitude of the Maxent and IPP model outputs are identical.
Chapter 3

Maxent Diagnostics

In the previous chapter we defined the Maxent and IPP models and showed their equivalence when fitted using discrete data. In this chapter, we exploit this relationship to derive measures of leverage and influence for Maxent, as well as methods for residual analysis of Maxent model fit.

Diagnostic methods for linear models are well-known, and discussion of such methods can be found in any regression textbook, as well as monographs by Cook and Weisberg (1982) and Atkinson (1985); for GLMs, the canonical reference is McCullagh and Nelder (1989). For IPPs and spatial processes in general, diagnostic methods analogous to those used for linear models have been derived only recently. The key references, to which we will refer often, are Baddeley et al. (2005) for residuals, and Baddeley et al. (2013) for leverage and influence measures.

For our Maxent derivations, we will take a consistent approach: we will first give the definition of the diagnostic for classical linear models, and then extend these definitions to apply to the IPP model. We then consider the form of the IPP diagnostic in the discrete case, and use this result as our definition of the Maxent diagnostic.

3.1 Leverage

3.1.1 Classical Definition

Consider a GLM of the form $E[Y] = g^{-1}(X\beta)$, where $Y$ is an $n$-length vector of responses, and $X$ is an $n \times p$ covariate matrix. Let $W$ be an $n \times n$ diagonal matrix
whose elements $w_{ii}$ are weights for each of the $n$ observations. Then the $n \times n$ leverage matrix (also called a “hat” matrix or a “projection” matrix) is defined

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}.$$  \hspace{1cm} (3.1)

The leverage of the $i$th observation is defined to be the diagonal entry $h_{ii}$ of the leverage matrix $H$ (McCullagh and Nelder, 1989).

An observation’s leverage may be interpreted as a measure of the distance between that observation and the $p$-dimensional mean of all other observations in covariate space. An observation lying at an extreme value of one or more covariates may “pull” the fitted model towards it, and cause the model to fit that observation much more closely than it would otherwise. In this way, leverage identifies outliers in covariate space, and indicates those observations that may have a disproportionate effect on model fit (McCullagh and Nelder, 1989).

Leverage values have the following properties:

1. $0 \leq h_{ii} \leq 1$.
2. $\sum_{i=1}^{n} h_{ii} = \text{tr}(H) = p$.

Thus, the average leverage value is expected to be $p/n$. A general rule of thumb is that observations for which $h_{ii} > 2p/n$ are unusual and should be examined more closely (Hoaglin and Welsch, 1978).

The weights $w_{ii}$ differ depending on the model type. For traditional linear models, $w_{ii} = 1$ and the leverage values depend only on the covariates. For GLMs, the weights depend on the particular model being fitted; for example, a Poisson GLM has $w_{ii} = \hat{\lambda}_i$ (McCullagh and Nelder, 1989).

### 3.1.2 Leverage for IPP Models

For the continuous IPP model, Baddeley et al. (2013) define, for all $z \in D$, a leverage function

$$h(z) = \lambda(z)x(z) \left[ \int_D x(z)x(z)'\lambda(z)dz \right]^{-1} x(z)' \hspace{1cm} (3.2)$$

which is analogous to the leverage values $h_{ii}$ defined in section 3.1.1 and has the same properties, specifically:
1. \( 0 \leq h(z) \leq 1 \).

2. \( \int_D h(z) dz = p \).

### 3.1.3 Leverage for Maxent Models

Baddeley et al. (2013) provide a heuristic derivation of their leverage function \( h(u) \) in which they partition \( D \) into \( n \) disjoint subsets, fit a Poisson GLM to the counts of events within each subset, and then observe the behaviour of the leverage matrix \( H \) as \( n \to \infty \) (the fine-pixel limit). If we partition \( D \) into \( n \) grid cells of equal size, and for each cell replace the event count with a binary variable indicating presence or absence, then the resulting Poisson GLM will have exactly the form of a Maxent model, as demonstrated in Section 2.3. The leverage matrix of a Maxent model is therefore the same as for a Poisson GLM.

**Definition.** Suppose we have a Maxent model fitted to \( n \) grid cells, such that for each cell \( d_i \) we have a vector of \( p \) environmental features \( x_i \), a binary response variable \( y_i \), and a fitted probability \( \pi_i \). Then the **leverage** of grid cell \( d_i \) is the entry \( h_{ii} \) of the \( n \times n \) matrix

\[
H = \Pi^{1/2} X (X'\Pi X)^{-1} X'\Pi^{1/2}
\]

where \( X \) is an \( n \times p \) covariate matrix and \( \Pi \) is an \( n \times n \) diagonal matrix with entries \( \Pi_{ii} = \pi_i \).

The leverage of a single grid cell can be computed as \( h_{ii} = \pi_i x_i (X'\Pi X)^{-1} x_i' \), making the Maxent leverage value \( h_{ii} \) a discrete analogue to the IPP leverage function \( h(z) \).

### 3.2 Influence

#### 3.2.1 Classical Definition

Consider again a GLM of the form \( \mathbb{E}[Y] = g^{-1}(X\beta) \). The **influence** of the \( i \)th observation is defined as:

\[
s_i = \frac{2}{p} \frac{L(\hat{\beta})}{L(\hat{\beta}_{(-i)})},
\]

(3.4)
where $L(\hat{\beta})$ is the likelihood of the fitted model, $L(\hat{\beta}_{(-i)})$ is the likelihood of a model fit excluding the $i$th observation, and $p$ is the number of model parameters. We can rewrite this definition in terms of model log-likelihoods, which makes for easier computation:

$$s_i = \ell(\hat{\beta}) - \ell(\hat{\beta}_{(-i)}) + \log(2/p),$$

where $\ell(\hat{\beta})$ is the log-likelihood of the fitted model, $\ell(\hat{\beta}_{(-i)})$ is the log-likelihood of a model excluding the $i$th observation, and $p$ is as before.

Influence is known as a “deletion diagnostic” (Cook and Weisberg, 1982), and its underlying idea is quite intuitive. Suppose we have a model fitted to a set of data. If we delete an observation from the dataset and refit the model, we expect the newly-fitted model to differ from the original model, and the magnitude of the difference depends on the observation that has been deleted. Some observations may contribute relatively little to the model, and their deletion changes the fitted model only slightly. Other observations may contribute substantially to the model, and their deletion will result in a fitted model very different from the original. Observations whose presence in the model has a major effect on the model’s outcome are said to be influential, and we can quantify this influence by measuring the difference in log-likelihood between the full model and that excluding the observation. The latter model will always have a log-likelihood smaller than the full model because it uses less data, so $s_i \geq 0$ with higher values indicating greater influence. Unlike leverage, there is no agreed-upon threshold above which observations should be considered unusually influential; influence is generally assessed visually by plotting $s_i$ for all observations (Atkinson, 1985).

Computing the influence of observations from the definition requires fitting an extra model for each observation, and when there are a large number of observations the process can become computationally expensive. An alternative method, due to Williams (1987), is to compute the matrix

$$S = \frac{1}{p}(y - \hat{y})X(X'WX)^{-1}X'(y - \hat{y})'$$

where $y$ and $\hat{y}$ are vectors of observed and fitted responses, $X$ an $n \times p$ data matrix, and $W$ an $n \times n$ diagonal weighting matrix. Then the diagonal elements $s_{ii}$ of $S$
approximate the influence values \( s_i \) of (3.4).

### 3.2.2 Influence for IPP Models

The definition of influence for point processes is the same as for GLMs, with one important extension: influence is now measured for a subregion of the domain. Suppose we have an IPP fitted to data \( Z \) within a domain \( D \). The influence of any subregion \( A \subset D \) is defined

\[
s(A) = \ell(\hat{\beta}) - \ell(\hat{\beta}_{(-A)}) + \log(2/p) \tag{3.7}
\]

where now \( \ell(\hat{\beta}) \) is the log-likelihood of the fitted model, \( \ell(\hat{\beta}_{(-A)}) \) is the log-likelihood of a model excluding any events in \( A \) (Baddeley et al., 2013).

Baddeley et al. (2013) also define an approximate influence measure analogous to the diagonal entries of the influence matrix \( S \) for GLMs. The influence of event \( z_i \) is

\[
s_i \approx \frac{1}{p} x(z_i) \left[ \int_D x(z)x(z)' \lambda(z) dz \right]^{-1} x(z_i)' \tag{3.8}
\]

### 3.2.3 Influence for Maxent Models

The general definition of influence applies to Maxent:

**Definition.** Suppose we have a Maxent model fitted to grid cell data. The influence of the \( i \)th presence cell is

\[
s_i = \ell(\hat{\beta}) - \ell(\hat{\beta}_{(-i)}) + \log(2/p) \tag{3.9}
\]

where \( \ell(\hat{\beta}) \) is the log-likelihood of the fitted model, \( \ell(\hat{\beta}_{(-i)}) \) is the log-likelihood of a model fit excluding data in the \( i \)th presence cell, and \( p \) is the number of model parameters.

Recall that for Maxent, the maximum likelihood and maximum entropy solutions are equivalent (Della Pietra et al., 1997). We can therefore redefine influence for Maxent models to be in terms of the entropy, rather than the log-likelihood:

\[
s_i = \frac{2}{p} \frac{H(\beta)}{H(\beta_{(-i)})} \tag{3.10}
\]
where $H(\beta)$ and $H(\beta_{-i})$ are the entropy of the full fitted model and the model excluding the data in the $i$th presence cell, respectively. This way of computing influence is easier to implement, as model entropy is given as part of Maxent’s output, while likelihood is not and would have to be calculated manually.

As with IPPs, we can expand the definition of Maxent influence to include sets of observations. Let $D_j \subseteq D$ be a set of presence grid cells. Then the cumulative influence of the cells $D_j$ is given by

$$s(D_j) = \frac{2}{p} \frac{H(\beta)}{H(\beta_{-j})}$$ (3.11)

where $H(\beta)$ is the entropy of the full model, and $H(\beta_{-j})$ is the entropy of the Maxent model fit excluding the data in the presence cells $D_j$.

We can justify computation of the influence matrix $S$ for the same reasons as the leverage matrix $H$. Thus, following the method of Williams (1987), the influence of the $i$th presence grid cell in a Maxent model can be approximated by computing

$$s_i \approx \frac{1}{p} (y_i - \hat{\pi}_i)x_i (X'\Pi X)^{-1}x'_i.$$ (3.12)

### 3.3 Residuals

#### 3.3.1 Classical Definition

In regression analysis, the residual of an observation is the difference between the observed response value and the fitted response value estimated by the model. Residuals represent estimates of the statistical error of the model, which is unobservable (Cook and Weisberg, 1982). As such, they play a key role in evaluating model fit. For a Poisson GLM, the raw residual can be written:

$$r_i = y_i - \hat{\lambda}_i$$ (3.13)

where $y_i$ is the observed response, and $\hat{\lambda}_i$ is the fitted model response.
Dividing the raw residual by the standard error of $y_i$ yields the **Pearson residual**:

$$ r_{P,i} = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}. \quad (3.14) $$

Although the unobservable statistical error of the model has constant variance, the observed raw residuals do not; specifically, the residual variance is generally larger for observations that lie farther from the mean in covariate space. Pearson residuals standardize the raw residuals relative to their variance and allow for more informative comparisons.

The sum of the squared Pearson residuals is called the **Pearson $\chi^2$ statistic**, and it follows a $\chi^2$ distribution with $n - p$ degrees of freedom, where $n$ is the number of observations and $p$ is the number of model parameters:

$$ \chi^2_P = \sum_{i=1}^{n} \left( \frac{y_i - \hat{\lambda}_i}{\lambda_i} \right)^2 \sim \chi^2_{n-p}. \quad (3.15) $$

The Pearson $\chi^2$ statistic can be used to formally test the goodness-of-fit of Poisson GLMs (McCullagh and Nelder, 1989).

### 3.3.2 Residuals for IPP Models

Baddeley et al. (2005) define residuals for IPP models using the same conceptual framework as for traditional residuals, but they extend the definition to apply to subregions of the domain. For any subregion $A \subseteq D$, the residual of $A$ is defined

$$ R(A) = n(Z \cap A) - \int_A \hat{\lambda}(z)dz \quad (3.16) $$

where $n(Z \cap A)$ is the number of points of $Z$ observed in $A$, and the second term is the modelled intensity of $A$. $R(A)$ is in fact a measure, taking on value 1 at each point $z_i \in Z \cap A$, and $-\lambda(z)$ at every other point $z \in A$.

Pearson residuals for IPP models can also be defined:

$$ R_{P}(A) = \sum_{z_i \in Z \cap A} \frac{1}{\sqrt{\hat{\lambda}(z_i)}} - \int_A \sqrt{\hat{\lambda}(z)}dz. \quad (3.17) $$

However, the Pearson $\chi^2$ statistic is no longer valid (Baddeley et al., 2005).
3.3.3 Residuals for Maxent Models

If we discretize an IPP model in the manner of Berman and Turner (1992), then IPP residuals are computed for each of the discrete grid cells. These residuals will then be equivalent to those for a Maxent model.

**Definition.** Suppose we fit a Maxent model to grid cell data. The **raw residual** of the $i$th grid cell is

$$r_i = y_i - \hat{\pi}_i$$  

where $y_i$ is a binary response variable equal to 1 if cell $i$ contains presence data, and 0 otherwise, and $\hat{\pi}_i$ is the estimated probability of presence in cell $i$ estimated by the model. The **Pearson residual** of the $i$th grid cell is

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i}}.$$  

For a more informative residual analysis, we can aggregate grid cells into larger regions, compute the Pearson residual for each region, and conduct Pearson’s $\chi^2$ test to evaluate the Maxent model’s goodness-of-fit. We can also apply the visualization methods of the following section.

3.3.4 Visualization of Maxent Residuals

For linear models and GLMs, residuals can be visually inspected by plotting them against a predictor variable. Theory dictates that if the fitted model is correct, the residuals should be distributed with mean 0 and a constant variance over the range of the predictor. Substantial deviation from this ideal implies that the fitted model is inappropriate. Similarly, we could map Maxent residuals and attempt to detect spatial trends that could reveal deficiencies in the model. However, a residual map is not very useful: since each nonpresence grid cell $j$ has raw residual $-\hat{\pi}_j$ and Pearson residual $-\sqrt{\hat{\pi}_j}$, such a map simply reproduces the Maxent distribution (or a transformation of it) negatively over all nonpresence grid cells, with positive masses occurring at presence grid cells.

A better way to visualize residuals is to smooth them. For IPPs, Baddeley et al.
(2005) introduce the smoothed residual field, defined as, for all \( z \in D \):

\[
s(z) = \frac{1}{\int_D k(z-v)dv} \left[ \sum_z k(z-z_i) - \int_D k(z-v)\lambda(v)dv \right]
\] (3.20)

where \( k \) is a smoothing kernel. Any smoothing kernel could be used, but Baddeley et al. recommend the Gaussian kernel

\[
k(z-v) = \exp \left\{ \frac{(z-v)^2}{2b} \right\}.
\] (3.21)

The parameter \( b \) is the bandwidth that controls the smoothness of \( s(z) \), and many different bandwidth selection methods exist; see Hastie et al. (2009). We adopt the method of Berman and Diggle (1989), which uses cross-validation to select the bandwidth that minimizes the model mean-square error.

Approximating the integral in (3.20) discretizes \( s(z) \) and allows us to apply it to Maxent. If the fitted model is adequate, we expect that \( s(z) \approx 0 \) over the whole domain, with minor random fluctuations. A clear spatial trend in values of \( s(z) \) suggests a model which is inadequate, at least in certain regions.

Another way to detect trends in Maxent residuals is through the use a cumulative residual plot, also due to Baddeley et al. (2005). Let \( x \) be an environmental feature or a spatial variable such as latitude and longitude. Then for each value \( x^* \), plot the cumulative residual \( \sum_i r_i \) for which \( x_i \leq x^* \). Like traditional residual plots, we expect the cumulative residuals to be distributed randomly about zero, and any systemic deviation suggests a misspecified model.

3.4 Ecological Interpretation of Diagnostics

Before continuing, it is worthwhile to consider how to interpret Maxent diagnostics in an ecological context. In addition to their standard roles in evaluating the fitted model, leverage and influence can be used to inform sampling effort for future studies.

Recall that leverage is a measure of the extremity of a observation’s covariates relative to the average covariate values of the entire dataset. For Maxent models, covariates take the form of environmental features, and thus grid cells with high leverage contain environmental conditions which differ substantially from the aver-
age environmental conditions of the study region. Maxent leverage values are also weighted by the fitted model responses, so that among grid cells with similar environments, those with a higher predicted response will be assigned higher leverage.

Thus, high leverage areas in Maxent models are those which exhibit unique environmental conditions relative to the study region, and yet have a high predicted presence. Of most interest would be those areas with high leverage and a very small number of presence records. The large response values indicate that the species is likely to occur within the area, and has simply not been recorded; the high leverage values indicate that any presence records found within the area may potentially have a substantial effect on the model. Therefore, if researchers are interested in improving a species’ predicted distribution and producing a more representative presence dataset, high-influence regions would be an excellent area to target sampling efforts.

There is a very important consideration when interpreting Maxent leverage values: the magnitude of a cell’s leverage depends upon the extent of the study region. Different study regions will exhibit different ranges of environmental conditions, and so will have different “average” conditions. For example, in a Maxent model fitted to the eastern United States, grid cells in southern Florida will have high leverage because of their tropical environment differs significantly from the study region’s average temperate environment. If however, the Maxent model were restricted to just the state of Florida, the leverage of the grid cells in southern Florida would be much smaller, as their environmental conditions are much closer to those of Florida in general.

Influence is a measure of how much effect an observation or group of observations have on the model fit. To interpret what this means for Maxent, recall the Hutchinsonian niche described in Chapter 1. A species’ distribution is a geographical realization of its bioclimatic envelope, which consists of the hypervolume of environmental space within which conditions are suitable for the species. In practice, this hypervolume is estimated by analyzing the positions of the observations in environmental space, and the boundaries of the estimated envelope will be primarily determined by those observations lying on the periphery of the distribution. In this way, observations towards the edges of the distribution will tend to be more influential than observations in the centre of the distribution. The more outlying
distributions are contributing more to the definition of the species’ bioclimatic envelope, and their removal may substantially alter the estimated boundaries of the species distribution.

Influential observations and subregions can serve as useful guides to further sampling efforts. Sampling in the vicinity of influential observations and within influential areas of the study region will produce more outlying presence records which can be used to better define the boundaries of the species’ bioclimatic envelope. A clearer definition can improve estimates of the species distribution, and better characterize the relationship between the species and its environment.
Chapter 4

Application to *Trachymyrmex* Data

Having defined leverage, influence, and residual diagnostics for Maxent, we now demonstrate these methods by applying them to models of the distributions of two ant species, *Trachymyrmex septentrionalis* and *Trachymyrmex turritex*.

4.1 Background

The fungus-growing ants (tribe Attini) are a diverse group of species that share the remarkable practice of cultivating fungi for their own consumption. These ants gather plant material and use it to culture and sustain fungal colonies (tribe Leucoceoprineae) in underground chambers within the ant nest. In this way, the ants provide the fungi with shelter, nourishment, and protection from pathogens and fungivores; in return, the ants use the fungi as their primary food source. This relationship is a classic example of a biological mutualism (Hölldobler and Wilson, 1990). Fungus-growing ants occur throughout the tropical and semitropical regions of North and South America. The most well-known fungus-growing ants are the leafcutter ants (genera *Acromyrmex*, *Atta*), whose colonies are among the largest and most complex of any ant (Weber, 1972), and who are economically significant agricultural pests (Hölldobler and Wilson, 1990).

*Trachymyrmex* is another notable genus of fungus-growing ants. Although the ant colonies of this genus are much smaller and less conspicuous than their leafcut-
ter relatives, *Trachymyrmex* is the most widely-distributed attine genus in North America, with nine species occurring within the United States. Most North American *Trachymyrmex* species are desert ants restricted to southern Arizona and New Mexico, but two species tolerate a broader range of environments (Rabeling et al., 2007). *Trachymyrmex septentrionalis* has a range extending from Texas to Florida, and northwards to Illinois, Ohio, and New York. *Trachymyrmex turrifex* occurs primarily in Texas, extending south into Mexico, north into Oklahoma, and east into Louisiana. Where their distributions overlap, these two species are easily distinguishable by their nest entrances: *T. septentrionalis* deposits excavated material in conspicuous crescent-shaped mounds round the opening, while *T. turrifex* constructs small earthen “turrets” over each entrance (Creighton, 1950).

The North American branch of *Trachymyrmex* was the subject of a recent taxonomic review by Rabeling et al. (2007). Included in the review were presence records for all nine North American *Trachymyrmex* species. To illustrate the Maxent diagnostics derived in Chapter 3, we use the data in Rabeling et al. (2007) to model the distributions of *T. septentrionalis* and *T. turrifex* and evaluate the results.

### 4.2 Methods

To model the bioclimatic envelope of *T. septentrionalis* and *T. turrifex*, we made use of 18 bioclimatic variables from the WorldClim dataset (Hijmans et al., 2005), which is freely available online. The names and descriptions of the variables used are given in Table 4.1. WorldClim variables are stored in raster format. A raster is a data structure consisting of a matrix of real values accompanied by a resolution, an extent, and a coordinate reference system (ESRI, 2008). Taken together, these latter three pieces of information define a rectangular grid over a region of the Earth’s surface. The resolution gives the size of each grid cell, the extent gives the size of the total grid, and the coordinate reference places the grid in the correct location. The elements of the matrix can be interpreted as the average values of the variable of interest within each grid cell. WorldClim variables have a resolution of 10′ latitude by 10′ longitude, which corresponds to a grid cell area of about 225 km² at mid-latitudes.
### Table 4.1: Bioclimatic variables used for Maxent models of *T. septentionalis* and *T. turrifex.*

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO1</td>
<td>Annual Mean Temperature (°C)</td>
</tr>
<tr>
<td>BIO2</td>
<td>Mean Diurnal Range (°C)</td>
</tr>
<tr>
<td>BIO3</td>
<td>Isothermality</td>
</tr>
<tr>
<td>BIO4</td>
<td>Temperature Seasonality</td>
</tr>
<tr>
<td>BIO5</td>
<td>Max Temperature of Warmest Month (°C)</td>
</tr>
<tr>
<td>BIO6</td>
<td>Min Temperature of Coldest Month (°C)</td>
</tr>
<tr>
<td>BIO8</td>
<td>Mean Temperature of Wettest Quarter (°C)</td>
</tr>
<tr>
<td>BIO9</td>
<td>Mean Temperature of Driest Quarter (°C)</td>
</tr>
<tr>
<td>BIO10</td>
<td>Mean Temperature of Warmest Quarter (°C)</td>
</tr>
<tr>
<td>BIO11</td>
<td>Mean Temperature of Coldest Quarter (°C)</td>
</tr>
<tr>
<td>BIO12</td>
<td>Annual Precipitation (mm)</td>
</tr>
<tr>
<td>BIO13</td>
<td>Precipitation of Wettest Month (mm)</td>
</tr>
<tr>
<td>BIO14</td>
<td>Precipitation of Driest Month (mm)</td>
</tr>
<tr>
<td>BIO15</td>
<td>Precipitation Seasonality</td>
</tr>
<tr>
<td>BIO16</td>
<td>Precipitation of Wettest Quarter (mm)</td>
</tr>
<tr>
<td>BIO17</td>
<td>Precipitation of Driest Quarter (mm)</td>
</tr>
<tr>
<td>BIO18</td>
<td>Precipitation of Warmest Quarter (mm)</td>
</tr>
<tr>
<td>BIO19</td>
<td>Precipitation of Coldest Quarter (mm)</td>
</tr>
</tbody>
</table>

Presence records for *T. septentionalis* and *T. turrifex* were obtained from Rabeling et al. (2007). Records were georeferenced using the GEOLocate software client (Tulane University Biodiversity Research Institute, 2010) and manually verified using Google Earth (Google Inc., 2013). The point presence locations were then discretized: in environmental raster grid cells containing at least one presence record, the point(s) were replaced by a presence point located at the centre of the grid cell. This discretization scheme matches that of Berman and Turner (1992).

Maxent models were fit using the *Trachymyrmex* data and environmental variables with the Maxent software package (Phillips and Dudik, 2008) via the dismo package (Hijmans and Elith, 2013) of the R statistical programming language (R Core Team, 2014). While Maxent has the capability to compute multiple feature types, only linear features (the raw environmental variables) were used in the models. Figures 4.1 and 4.2 give the resulting Maxent distribution for *T. septentionalis*; figures 4.3 and 4.4 give the Maxent distribution for *T. turrifex.* In all four figures, the black dots indicate presence grid cells. Maxent diagnostics were then computed using R. The relevant computer code for implementing the diagnostics is given in Appendix A.
Figure 4.1: Maxent distribution of *T. septentrionalis*, logistic output.

Figure 4.2: Maxent distribution of *T. septentrionalis*, raw output.
Figure 4.3: Maxent distribution of *T. turrifex*, logistic output.

Figure 4.4: Maxent distribution of *T. turrifex*, raw output.
The mapped Maxent distributions appear quite reasonable given the available data. One issue is that the \textit{T. septentrionalis} model is exhibiting probable sampling bias within Florida: the density of presence records in Florida may be due to greater prevalence of \textit{T. septentrionalis} in Florida, but could be a result of more intensive sampling within Florida than within other states. Dealing with sampling bias in Maxent bias is an area of ongoing research (Syfert et al., 2013). As these models are only being used for illustrative purposes, we do not attempt to correct for sampling bias.

4.3 Results

4.3.1 Leverage

Leverage for the \textit{T. septentrionalis} model is mapped in Figure 4.5. We see that leverage is highest in southern Florida, the Bahamas, and southern Louisiana. As we predicted, regions with climates that differ most significantly from the average temperate climate of the southeastern United States have been identified as regions of high leverage, with especially high leverage being attributed to regions with with large Maxent values. Florida has high leverage due to its tropical climate and the density of \textit{T. septentrionalis} observations there. The Bahamas and southern Louisiana also exhibit high leverage, but have no observations at all. The implication is that, based on the estimated bioclimatic envelope, \textit{T. septentrionalis} may occur in the Bahamas and southern Louisiana, and that presence records from those areas may have a significant effect on the fitted model.

Figure 4.6 maps leverage for the \textit{T. turrifex} model. High leverage values occur along the Texas Gulf Coast, whose subtropical maritime climate differs substantially from the more arid, continental conditions of inland Texas. Isolated pockets of high leverage can be seen in the southwest portion of the study region. These correspond to the peaks of the Sierra Madre mountain range. Areas of high elevation often exhibit conditions very different from the surrounding lowlands, and this result suggests that the Sierra Madre may be suitable for \textit{T. turrifex}. The Gulf Coast region and the Sierra Madre lack \textit{T. turrifex} observations, so the leverage map suggests that researchers focus \textit{T. turrifex} sampling efforts on those areas.
Figure 4.5: Leverage values for Maxent distribution of *T. septentrionalis*.

Figure 4.6: Leverage values for Maxent distribution of *T. turrifex*. 

34
Figure 4.7: Influence values for Maxent distribution of *T. septentrionalis*.

Figure 4.8: Influence values for Maxent distribution of *T. turrifex*. 
Figure 4.9: Maxent distribution of *T. septentrionalis* excluding Texas records.

Figure 4.10: Maxent distribution of *T. septentrionalis* excluding Georgia records.
4.3.2 Influence

The Maxent distribution of *T. septentrionalis* is given again in Figure 4.7, with circles indicating the influence of the presence grid cells. The most influential cell in the model is in the upper centre, and corresponds to two records of *T. septentrionalis* in southern Illinois. Other influential cells correspond to observations in New York and New Jersey, in the southern Appalachians, in northern Texas, and along the Florida coast.

Figure 4.8 shows the influence of presence grid cells in the *T. turrifex* model. The southernmost Mexican presence record is the most influential, with other influential records occurring in western Louisiana. The results for *T. septentrionalis* and *T. turrifex* are consistent with our expectation that influential observations are those which contribute most to defining the boundaries of the species’ bioclimatic envelope, and will tend to occur on the periphery of the species distribution.

We can also measure of the influence of subsets of observations. For example, we can partition the *T. septentrionalis* records by US state, and calculate the marginal influence of the records from each state. Figure 4.11 gives the results. We see that records in Texas exert the most influence on the distribution, while records in Georgia exert the least influence. We can visualize this influence by fitting and mapping two new Maxent distributions of *T. septentrionalis*: one excluding all records from Texas (Figure 4.9), and one excluding all records from Georgia (Figure 4.10). Comparing these distributions with that in Figure 4.1, we see that excluding Texas...
records dramatically alters reduces the predicted range of *T. septentrionalis* in the north and west. Conversely, excluding Georgia records barely changes the modelled distribution because nearly all information being contributed by the Georgia observations to the model has already been contributed by observations in neighbouring states. These results suggest that more *T. septentrionalis* records should be collected in Texas and New Jersey in order to improve estimates of *T. septentrionalis’* western and northern range. Similarly, the influence values in Figure 4.8 suggest that the estimated distribution of *T. turrifex* can be improved by targeting sampling efforts from Mexico and from western Louisiana.

### 4.3.3 Residuals

In section 3.3.4 we introduced ways of visualizing Maxent residuals to assess model fit, and we now illustrate these methods for the *Trachymyrmex* data. Figures 4.12 and 4.13 show smoothed residual fields for the fitted Maxent models of *T. septentrionalis* and *T. turrifex*, respectively. The residual fields were computed using functions available in the R package *spatstat* (Baddeley et al., 2005), and the kernel bandwidths for both fields were selected according to the method of Berman and Diggle (1989), which is also available in the *spatstat* package.

Examining Figures 4.12 and 4.13, we see no really obvious spatial trends in the residuals. Identifying a trend in the residuals from the field alone can be challenging, so we also plot cumulative residuals against latitude and longitude for each Maxent model. From these plots, we see clear systemic deviations in both models, indicating that both models have been misspecified.

Our Maxent models included eighteen bioclimatic covariates. What could we be missing? One environmental feature that we have neglected is soil type. Rabeling et al. (2007) note that *T. septentrionalis* is found in habitats “characterized by very sandy soils”, and that *T. turrifex* is abundant in “the black clay soils of central Texas and somewhat less abundant in the sandy soils of eastern Texas and western Louisiana”. Clearly, soil type is an important predictor of *Trachymyrmex* presence, and including soil type in our models would improve their performance and fit. However, species-environment relationships are extremely complex, and it may prove impossible to fully and correctly specify a Maxent distribution model.
Figure 4.12: Smoothed residuals for Maxent distribution of *T. septentrionalis*.

Figure 4.13: Smoothed residuals for Maxent distribution of *T. turrifex*. 
Figure 4.14: Cumulative residual plots for Maxent distribution of *T. septentrionalis*.

Figure 4.15: Cumulative residual plots for Maxent distribution of *T. turrifex*.
Chapter 5

Discussion

5.1 Summary

Species distribution modelling is a critical problem in biogeography, ecology, conservation biology, and many other fields. One of the most widely-used SDMs is Maxent, a presence-only modelling method which has been found to consistently outperform nearly all competing presence-only models. However, despite its popularity, the Maxent program is a “black box” which remains poorly understood by users. As such, Maxent lacks many of the model evaluation tools available for better-understood methods such logistic regression and its many extensions.

In this thesis, we have introduced measures of leverage and influence, and residuals for Maxent models. To do so, we have made use of recent research demonstrating the equivalence, via Poisson GLMs, of Maxent and the spatial IPP model: when fitted to discrete grid cell data, Maxent and the IPP yield identical solutions. We gave descriptions of Maxent and IPPs and showed their equivalence in Chapter 2. In Chapter 3, we derived diagnostic measures for Maxent by discretizing the diagnostic methods that have been defined for IPP models. In Chapter 4, we applied these diagnostic methods to Maxent models for the distribution of two ant species of the genus *Trachymyrmex*, and found the methods to be consistent with intuition and with ecological theory. We also proposed ways in which leverage and influence diagnostics could be used to assist sampling efforts by identifying areas where new sampling efforts will have the most effect on the modelled distribution.
5.2 Future Work

In this thesis, we focused on the diagnostics that are most commonly used for linear and generalized linear models. We did not consider more advanced diagnostics such as measures of deviance, or information-theoretic methods such as Akaike’s information criterion (AIC) (Faraway, 2005; Hastie et al., 2009). The extension of any of these methods to Maxent would be worthy of further study. We also did not consider diagnostic measures unique to spatial point process models, such as distance methods and Ripley’s $K$ function (Diggle, 1983; Cressie, 1993). Species distributions are spatial phenomena, and evaluating Maxent models (and other SDMs) using explicitly spatial techniques may yield insights beyond those of the methods considered in this thesis.

This thesis has made extensive use of results equivocating two previously independent SDMs. We believe that a major goal of future SDM research should be the ongoing unification of the SDM literature and the development of a cohesive framework for modelling with presence-only data. In the meantime, regardless of the methods used, we recommend that SDM users make greater efforts to critically evaluate their models using appropriate diagnostics. Doing so will result in more sensible models, stronger inferences, more accurate predictions, and better science.
Appendix A

Maxent Diagnostics in R

This appendix contains R code for implementing the diagnostic methods described in Chapter 3 of the thesis. All functions require the dismo and spatstat packages as prerequisites.

```r
library(dismo)
library(spatstat)
```

A.1 Leverage

# Function to compute leverage values for a fitted Maxent model.
#
# Input:
# - maxent.model : Maxent object representing fitted model
# - species : Species occurrence records, matrix or dataframe
# - predictors : Environmental data, raster object
#
# Output:
# - lev.df : Dataframe containing leverage value for each grid cell

leverage <- function(maxent.model, species, predictors){
  raw.values <- predict(maxent.model, predictors, 
                         args='outputformat=raw')
  df <- rasterToPoints(raw.values)
  df <- data.frame(df)
  coords <- df[,,-3]
  X <- extract(predictors, coords)

  # Further code to compute leverage values...
  # (Code not shown due to space constraints)
}
```
V <- diag(extract(raw.values, coords))
Z <- solve(t(X) %*% V %*% X)

leverage <- numeric(dim(coords)[1])
for(i in 1:dim(coords)[1]){
  leverage[i] <- raw.values[i,3] * t(X[i,]) %*% Z %*% X[i,]
}

lev.df <- data.frame(coords, leverage)
return(lev.df)

A.2 Influence

# Function for computing influence of individual observations in a
# fitted Maxent model.
#
# Inputs:
# - maxent.model : Maxent object representing fitted model
# - species : Species occurrence records, matrix or dataframe
# - predictors : Environmental data, raster object
# - arguments : Optional character vector of Maxent model
#   arguments (see Phillips, 2008)
#
# Outputs:
# - inf.df : Dataframe of species records and their
#   computed influence

influence <- function(maxent.model, species, predictors,
                      arguments=NULL){

  entropy <- maxent.model@results["Entropy",]
  ent <- numeric(nrow(species))
  for(i in 1:nrow(species)){
    data <- species[-i,]
    model <- maxent(predictors, data, args=arguments)
    ent[i] <- model@results["Entropy",]
  }

  Influence <- 2/8*log(entropy/ent)
  inf.df <- data.frame(species, Influence)
  return(inf.df)
}
# Function for computing influence of a subset of observations in a # fitted Maxent model.

# Inputs:
# - maxent.model : Maxent object representing fitted model
# - species : Species occurrence records, matrix or dataframe
# - subset : Subset of species occurrence records, numeric vector
# - predictors : Environmental data, raster object
# - arguments : Optional vector of Maxent model arguments (see Phillips, 2008)

# Outputs:
# - Influence : Computed influence of given subset

influence.subset <- function(maxent.model, species, subset, predictors, arguments=NULL){
  entropy <- maxent.model@results["Entropy",]

  data <- species[-subset,]
  model <- maxent(predictors, data, args=arguments)
  ent <- model@results["Entropy",]

  Influence <- 2/8*log(entropy/ent)
  return(Influence)
}

### A.3 Residuals

# Function to convert Maxent model to IPP model.

# Inputs:
# - species.data : Species occurrence records, matrix or dataframe
# - predictors : Environmental data, raster object

# Outputs:
# - ppm.model : Fitted IPP model

ipp <- function(species.data, predictors){

  # Function to convert raster to im
  raster.to.im <- function(raster){
    raster.mat <- as.matrix(raster)
    return(raster.mat)
  }

  ...
raster.mat <- apply(raster.mat, 2, rev)

raster.im <- im(raster.mat, 
    xcol = seq(extent(raster)@xmin, extent(raster)@xmax, length=dim(raster)[2]), 
    yrow = seq(extent(raster)@ymin, extent(raster)@ymax, length=dim(raster)[1]))

return(raster.im)
}

# Convert predictors from raster to im, place in covariates list
covariates <- NULL
for(i in 1:length(names(predictors))){
    covariates[[i]] <- raster.to.im(subset(predictors, i))
}
names(covariates) <- names(predictors)

# Create trend formula
trend <- "~"
for(i in 1:length(names(covariates))){
    trend <- paste(trend, paste(names(covariates)[i], "+", sep=""), sep="")
}
trend <- substr(trend, 1, nchar(trend)-1)
trend <- as.formula(trend)

# Rasterize species data
species.grid <- rasterize(species.data, predictors)
species.grid <- rasterToPoints(species.grid)[,-3]

# Define ppp object
W <- levelset(covariates[[1]], -100, ">")
species.ppp <- ppp(species.grid[,1], species.grid[,2], window=W)
print(bw.diggle(species.ppp))

# Define quadrature scheme
G <- gridcentres(W, nx=dim(predictors)[2], ny=dim(predictors)[1])
dummy.ppp <- ppp(G$x, G$y, window=W)
Q <- quadscheme(data=species.ppp, dummy=dummy.ppp)

# Fit point process model
ppm.model <- ppm(Q, trend, covariates=covariates)
return(ppm.model)
residuals <- function(maxent.model, species, predictors, type, bandwidth){
  temp.mod <- ipp(species.data, predictors)
  temp.res <- diagnose.ppm(temp.mod,
    which="smooth",
    type=type,
    sigma=bandwidth, plot.it=F)
  field <- temp.res$smooth$Z$v
  field <- apply(field, 2, rev)
  field <- raster(field)
  crs(field) <- crs(predictors)
  extent(field) <- extent(predictors)
  return(field)
}

cum.res <- function(maxent.model, species, predictors){
  maxent.ipp <- ipp(species, predictors)
diagnose <- diagnose.ppm(maxent.ipp)
x <- diagnose$xcumul$empirical$covariate
vx <- diagnose$xcumul$empirical$value
y <- diagnose$ycumul$empirical$covariate
vy <- diagnose$ycumul$empirical$value
par(mfrow=c(1,2))
plot(x, vx, type="l", xlab="Longitude", ylab="Cumulative Residual")
plot(x, vy, type="l", xlab="Latitude", ylab="Cumulative Residual")
Bibliography


