Mahalanobis distances for ecological niche modelling and outlier detection: implications of sample size, error, and bias for selecting and parameterising a multivariate location and scatter method

Thomas R. Etherington[1]

[1]**Manaaki Whenua — Landcare Research, PO Box 69040, Lincoln 7640, New Zealand**

Corresponding author:
Thomas R. Etherington[1]

Email address: EtheringtonT@landcareresearch.co.nz

ABSTRACT

The Mahalanobis distance is a statistical technique that has been used in statistics and data science for data classification and outlier detection, and in ecology to quantify species-environment relationships in habitat and ecological niche models. Mahalanobis distances are based on the location and scatter of a multivariate normal distribution, and can measure how distant any point in space is from the centre of this kind of distribution. Three different methods for calculating the multivariate location and scatter are commonly used: the sample mean and variance-covariance, the minimum covariance determinant, and the minimum volume ellipsoid. The minimum covariance determinant and minimum volume ellipsoid were developed to be robust to outliers by minimising the multivariate location and scatter for a subset of the full sample, with the proportion of the full sample forming the subset being controlled by a user-defined parameter. This outlier robustness means the minimum covariance determinant and the minimum volume ellipsoid are highly relevant for ecological niche analyses, which are usually based on natural history observations that are likely to contain errors. However, natural history observations will also contain extreme bias, to which the minimum covariance determinant and the minimum volume ellipsoid will also be sensitive. To provide guidance for selecting and parameterising a multivariate location and scatter method, a series of virtual ecological niche modelling experiments were conducted to demonstrate the performance of each multivariate location and scatter method under different levels of sample size, errors, and bias. The results show that there is no optimal modelling approach, and that choices need to be made based on the individual data and question. The sample mean and variance-covariance method will perform best on very small sample sizes if the data are free of error and bias. At larger sample sizes the minimum covariance determinant and minimum volume ellipsoid methods perform as well or better, but only if they are appropriately parameterised. Modellers who are more concerned about the prevalence of errors should retain a smaller proportion of the full data set, while modellers more concerned about the prevalence of bias should retain a larger proportion of the full data set. I conclude that Mahalanobis distances are a useful niche modelling technique, but only for questions relating to the fundamental niche of a species where the assumption of multivariate normality is reasonable. Users of the minimum covariance determinant and minimum volume ellipsoid methods must also clearly report their parameterisations so that the results can be interpreted correctly.

INTRODUCTION

The Mahalanobis distance (Mahalanobis, 1936) is a statistical technique that can be used to measure how distant a point is from the centre of a multivariate normal distribution. Mahalanobis distances are commonly applied to problems such as classifying data into groups and determining differences between groups (Manly, 2005). Mahalanobis distances have also been used to quantify species-environment relationships through habitat and ecological niche models (Dettmers et al., 2002; Johnson and Gillingham,

2005; Tsoar et al., 2007; Etherington et al., 2009). In this context Mahalanobis distances are classified as a presence-only technique because they do not require species absence or background environmental data (Peterson et al., 2011) and simply require a data matrix

$$\mathbf{A} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,n} \\ x_{3,1} & x_{3,2} & x_{3,3} & \dots & x_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & x_{m,3} & \dots & x_{m,n} \end{bmatrix}$$

that has m rows of species occurrences for which n columns of environmental variables have been obtained. The multivariate location and scatter of the data are defined by an n-dimensional vector $\hat{\mu}$ containing the sample means for each column of variables, and a sample variance-covariance matrix $\hat{\Sigma}$ of dimensions $n \times n$ that contains variances for each column along the main diagonal and pair-wise column covariances values elsewhere (Manly, 2005).

The Mahalanobis distance

$$D^2(\mathbf{x}) = (\mathbf{x} - \hat{\mu})^{\mathrm{T}} \hat{\Sigma}^{-1} (\mathbf{x} - \hat{\mu}) \tag{1}$$

can then be calculated for any vector $\mathbf{x} = [x_1, x_2, x_3, \dots, x_n]$ that represents a position in environmental space as defined by the n environmental variables. As D^2 is essentially the sum of n independent standard normal variables, the D^2 values from a multivariate normal population will follow a chi-squared distribution with degrees of freedom equal to the number of dimensions n (Manly, 2005). This means that a chi-squared cumulative distribution function $F_{\chi_n^2}(x)$ can be used to convert D^2 into a probability $P(\chi_n^2 \le D^2)$ that indicates if a location in environmental space has a D^2 that is greater than would be expected by chance (Etherington, 2019). For example, when applied to $m = 20$ hypothetical points in $n = 2$ dimensions, the calculated $P(\chi_n^2 \le D^2)$ values follow a characteristic elliptical pattern centred at the mean of each environmental variable, with $P(\chi_n^2 \le D^2)$ increasing from the centre outwards in a manner that accounts for the variability within and correlation between each environmental variable (Figure 1).

The elliptical form of Mahalanobis distances fits well with the theoretical concept of the fundamental niche, which Hutchinson (1957) p. 416 defined as "an n-dimensional hypervolume . . . which corresponds to a state of the environment which would permit the species . . . to exist indefinitely". Hutchinson (1957) used a rectangular model to define environment limits of the fundamental niche, but also stated that "If the variables are independent in their action on the species we may regard this area as the rectangle . . . but failing such independence the area will exist whatever the shape of its sides". So any convex shape, which includes the elliptical shape of $P(\chi_n^2 \le D^2)$, would be an appropriate model of the fundamental niche. Indeed, we see the use of ellipses alongside other convex shapes in later development of the niche concept (Hutchinson, 1978). In this context $\hat{\mu}$ represents the optimal environmental conditions at the centre of the fundamental niche, and $\hat{\Sigma}$ represents both the range of and interaction between environmental conditions within the fundamental niche.

While $P(\chi_n^2 \le D^2)$ is usually inverted to $P(\chi_n^2 > D^2)$ for use in ecological niche modelling to estimate the probability of an environmental location being within a fundamental niche (Etherington, 2019), $P(\chi_n^2 \le D^2)$ is also commonly used in statistics and data science to detect outliers (Aggarwal, 2017). However, when applied to detect outliers, the use of sample means and variance-covariance to estimate D^2 can be problematic because these measures of multivariate location and scatter are sensitive to outliers (Figure 1). So for outlier detection, D^2 can be calculated with different methods for defining the multivariate location and scatter of data, such as the minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE), which are much more insensitive to outliers (Rousseeuw, 1985).

The MCD approach estimates multivariate location $\hat{\mu}_{\mathrm{MCD}}$ and scatter $\hat{\Sigma}_{\mathrm{MCD}}$ from a subset numbering h data points that has the smallest variance-covariance matrix determinant (Hubert and Debruyne, 2010). The MVE approach is similar to the MCD in that it works with a subset of size h data points, but the MVE estimates multivariate location $\hat{\mu}_{\mathrm{MVE}}$ and scatter $\hat{\Sigma}_{\mathrm{MVE}}$ from the ellipsoid of minimal volume that encapsulates the h data points (Van Aelst and Rousseeuw, 2009). D^2 can then be calculated using either the MCD measures of multivariate location and scatter
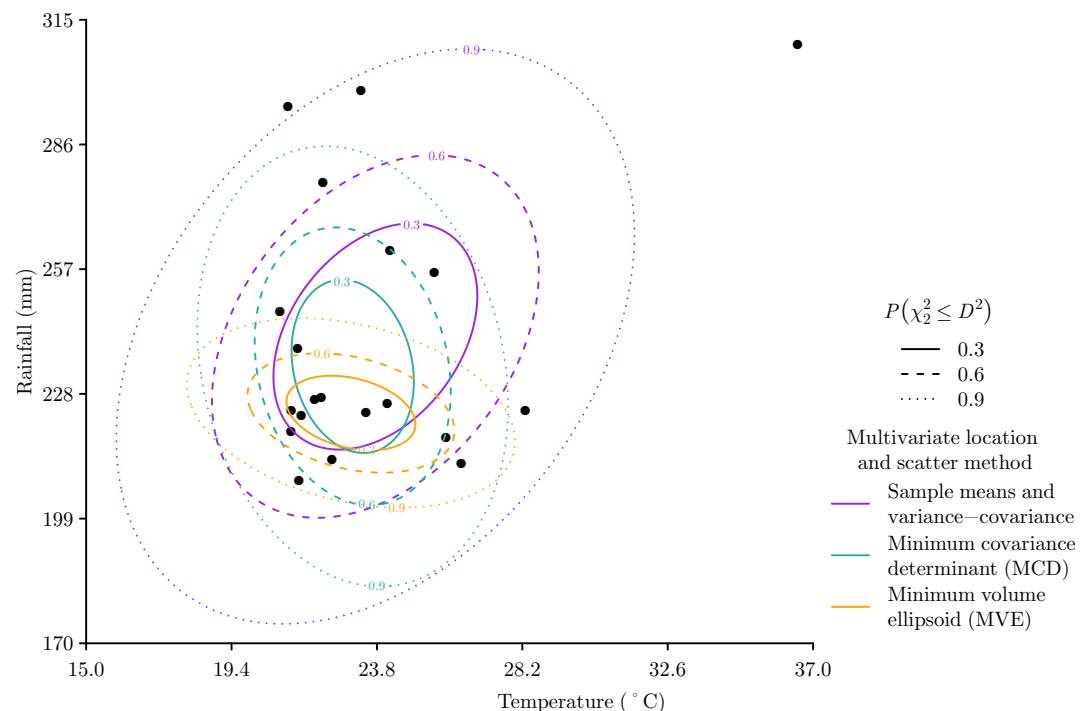
Figure 1. Hypothetical two-dimensional example of Mahalanobis distance D^2 with three different methods of defining the multivariate location and scatter of the data. For each method the ellipses show contours of probability $P(\chi_n^2 \leq D^2)$ that indicate if a location in environmental space has a D^2 that is greater than would be expected by chance. The sensitivity of the sample mean and variance-covariance matrix method to outlying data can be seen, which contrasts with both the minimum covariance determinant and minimum volume ellipsoid methods, which both focus on where data is concentrated.

$$D^2(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_{\text{MCD}})^{\text{T}} \hat{\Sigma}_{\text{MCD}}^{-1} (\mathbf{x} - \hat{\mu}_{\text{MCD}}) \tag{2}$$

89   or the MVE measures of multivariate location and scatter

$$D^2(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_{\text{MVE}})^{\text{T}} \hat{\Sigma}_{\text{MVE}}^{-1} (\mathbf{x} - \hat{\mu}_{\text{MVE}}) \tag{3}$$

90   for any vector $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_n]$ that represents a position in n-dimensional environmental space.
91   The utility of the MCD and MVE methods for defining the multivariate location and scatter of
92   data is highly relevant for ecological niche modelling based on digitally mobilised data through data-
93   sharing networks such as the Global Biodiversity Information Facility (Edwards et al., 2000). These
94   networks are reliant on natural history observation data that are likely to contain errors such as taxonomic
95   misidentification or incorrect and imprecise georeferencing (Graham et al., 2004), which can result in
96   species occurrences that are outliers in environmental space. Returning to our hypothetical example
97   (Figure 1), we can see that both the MCD and MVE methods ignore the apparent outlier and produce
98   measures of multivariate location and scatter that are focussed on where the data are more concentrated.
99   Given the robustness of MCD and MVE methods to outliers resulting from errors in natural history
100   observation data, it is perhaps no surprise that both methods have been adopted recently for ecological
101   niche modelling (Norris et al., 2006; Liu et al., 2018; Soberón et al., 2018; Yañez-Arenas et al., 2018;
102   Qiao et al., 2019; Altamiranda-Saavedra et al., 2020; Osorio-Olvera et al., 2020; Castaño-Quintero et al.,
103   2020).
104   On the other hand, natural history observation data will also contain sampling bias (Graham et al.,
105   2004) that can be extreme and result in greater amounts of data for more charismatic species, more

accessible places, more developed countries, and more recent times (Meyer et al., 2016). If unaccounted for this could skew models results. Therefore, returning to our hypothetical example (Figure 1), we may also have a situation in which the apparent outlier only appears to be an outlier due to sampling bias at environments with lower temperatures and rainfall. If this were the situation then the MCD and MVE methods would be providing a poorer estimate of multivariate location and scatter because they are both focussing on the data bias, whereas the method based on the sample mean and covariance provides a better estimate that relates to all the data points.

Unfortunately, the differing effects of errors and bias means there is unlikely to be a best method for estimating the multivariate location and scatter of data in all situations, but this is consistent with ecological niche modelling more generally (Qiao et al., 2015) . Therefore, this paper uses a virtual ecology approach (Zurell et al., 2010) to simulate a series of ecological niche modelling experiments to understand when different methods of defining the multivariate location and scatter of the data are more appropriate as sample size, errors, and bias vary.

MATERIALS & METHODS

The virtual species used in the experiments is the Antipodean opaleye dragon, which is imagined to live in the mountain valleys of New Zealand (Scamander, 2001). I have defined the species' fundamental niche in terms of population growth rates (Maguire, 1973) measured as the finite rate of increase λ_F of a population in a two-dimensional environmental space of temperature and rainfall. The fundamental niche is defined by three parameters: $\lambda_{\max}$, the maximum finite rate of increase at the fundamental niche optimum; μ, a 2×1 column vector of means that gives the optimal temperature and rainfall condition; and Σ, a 2×2 variance-covariance matrix that determines the size and orientation of the fundamental niche in each dimension. With

$$\lambda_{\max} = 2.5, \mu = \begin{bmatrix} 7.5 \\ 1800 \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} 2 & -950 \\ -950 & 800000 \end{bmatrix}$$

the fundamental niche finite rate of increase $\lambda_F(\mathbf{x})$ of a virtual species for any vector $\mathbf{x}$ of environmental space coordinates is then calculated as

$$\lambda_F(\mathbf{x}) = \lambda_{\max} \times e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)} \tag{4}$$

that results in an elliptically shaped fundamental niche (Figure 2a).

Generating samples of species occurrences begins with an idealised virtual sampling of the niche space. An initial set of sampling locations $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i\}$ consisted of a series of locations in environmental space. S was randomly generated from a two-dimensional normal distribution $S \sim \mathcal{N}_2(\mu, \Sigma)$ with the same vector of means μ and variance-covariance matrix Σ as defined the fundamental niche. Then the probability that a sampled location $\mathbf{x}_i$ resulted in the species being both present and detected $P_d(\mathbf{x}_i)$ was described as a logistic function

$$P_d(\mathbf{x}_i) = \frac{1}{1 + e^{-10(\lambda_F(\mathbf{x}_i)-0.5)}} \tag{5}$$

which was parameterised so that $P_d(\mathbf{x}_i)$ increases as $\lambda_F(\mathbf{x}_i)$ increases, but with $P_d(\mathbf{x}_i) \approx 0$ where $\lambda_F(\mathbf{x}_i) \approx 0$ because the virtual species population is unlikely to exist under these environmental conditions, and $P_d(\mathbf{x}_i) \approx 1$ where $\lambda_F(\mathbf{x}_i) \gtrsim 1$ because above this population growth rate the population should always be present.

Using this idealised sampling, a sample of $m = 100$ occurrence locations can be produced that are concentrated towards the centre of the fundamental niche, and for which $P(\chi_2^2 > D^2)$ estimates for all three methods of multivariate location and scatter align with the elliptical shape of the fundamental niche (Figure 2a). Under these idealised sampling conditions we can see that the actual λ_F values and the estimated niche probabilities $P(\chi_2^2 > D^2)$ for the occurrence samples are very highly correlated, and that all three methods of determining the multivariate location and scatter of the niche perform equally well (Figure 2b).
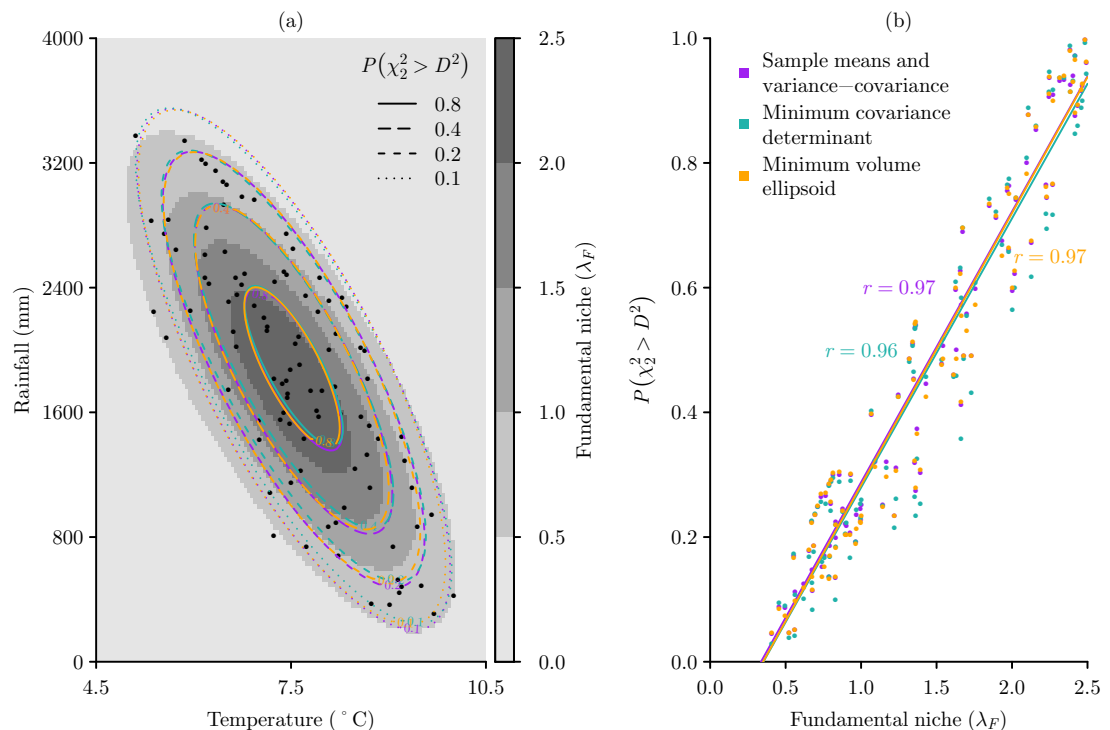
Figure 2. Modelling the fundamental niche λ_F of a virtual species with Mahalanobis distances D^2 based on three multivariate location and scatter methods. (a) Given an idealised sample of 100 occurrences across the virtual fundamental niche, the probabilities $P(\chi_2^2 > D^2)$ of the environmental space being within the niche are calculated for three multivariate location and scatter methods: the sample means and variance-covariance, the minimum covariance determinant, and the minimum volume ellipsoid. (b) With this idealised sample all three methods perform very well, producing very high correlations between the sample occurrences known λ_F and estimated $P(\chi_2^2 > D^2)$.

148   Of course this idealised virtual sampling is completely unrealistic, but does serve to demonstrate that it
149 is possible to produce useful models with enough accurate data. What is of interest is understanding under
150 what conditions of sample size, error, and bias the predictive ability of the various multivariate location
151 and scatter methods begin to break down. Therefore, to explore this, a series of virtual experiments were
152 conducted.
153   All experiments were done using `R` (R Core Team, 2019) with the `MASS` (Venables and Ripley, 2002),
154 `virtualNicheR` (Etherington and Omondiagbe, 2019), `fields` (Nychka et al., 2017), `raster`
155 (Hijmans, 2020), and `extrafont` (Chang, 2014) packages.

Sample size

157 It is common for many species to have as few as seven unique occurrence locations within the Global
158 Biodiversity Information Facility (Meyer et al., 2016). However, as the MCD approach needs $m \geq n \times 5$
159 (Hubert and Debruyne, 2010), and given the experiments are two-dimensional, the smallest value of m
160 that could be analysed is 10. Therefore, to explore the effects of sample size, the idealised sampling
161 approach was applied, but varying the sample size m from 10 to 160 in increments of 15. As m increases,
162 the idealised samples increasingly represent the elliptical shape of the fundamental niche (Figure 3a–c)
163   Regarding the choice of sample subset h used by the MCD and MVE methods, the standard choice
164 is $h = \lfloor (m+n+1)/2 \rfloor$ for both the MCD and MVE methods, because this produces the most robust
165 estimates (Hubert and Debruyne, 2010; Van Aelst and Rousseeuw, 2009). This standard choice uses just
166 over half the sample, so for sample sizes from 10 to 160 this would mean a standard choice of h from
167 6 to 81 as the sample size increases. Because h varies as a function of sample size, for experimental
168 consistency and to aid interpretation I specified the sample subset used by the MCD and MVE methods as
169 a proportion k of the sample size, such that $h = \lfloor k \times m \rfloor$. The standard choice of h was represented by

170 $k = 0.55$, and with $k = 0.75$ and $k = 0.95$ used to explore the effect of increasing k, and therefore h, on
171 the performance of the MCD and MVE methods.
172   Each sample of size m was replicated 500 times, and was applied to each multivariate location and
173 scatter method, with the MCD and MVE methods applied at the three different k values. The performance
174 of each multivariate location and scatter method was measured as the correlation between the actual λ_F
175 and estimated $P(\chi_2^2 > D^2)$ values.

Sample error

177 The virtual experiments to explore the effects of sample error followed the same process as the sample
178 size experiments except that the sample size was fixed at $m = 100$, and each sample was contaminated
179 with various levels of errors.
180   In generating errors I assumed that extreme errors can be identified using commonly used data
181 checking processes (Zizka et al., 2020), so errors were limited to locations within New Zealand that
182 comprise the core range of our virtual species. The climate space of New Zealand was described in terms
183 of mean annual temperature and annual precipitation climatologies for the period 1979-2013 with a 30 arc
184 second (around 1 km) grid resolution (Karger et al., 2017). An error was generated simply as a random
185 location within New Zealand, and the amount of error within each sample was varied from 0 % to 50 % in
186 increments of 5 %. As error increases, the samples with error reflect the fundamental niche less and the
187 New Zealand climate space more (Figure 3d–f).

Sample bias

189 The sample bias experiments followed the same process as the sample error experiments, except that
190 when selecting a random location within New Zealand the probability of the species being both present
191 and detected (Equation 5) was applied to limit bias to environments that are part of the fundamental niche.
192 The amount of bias within each sample was varied from 0 % to 100 % in increments of 10 %. As bias
193 increases, the biased samples become concentrated at the more commonly occurring climate space of
194 New Zealand that overlaps with the fundamental niche (Figure 3g–i).

RESULTS

196 There were obvious consistencies and trends amongst all the experiments. First, the results for the MCD
197 and MVE methods were very similar, making it hard to differentiate the performance of these two methods
198 (Figure 4, Figure 5, Figure 6). Second, in all cases, as k increases the MCD and MVE methods become
199 more similar to the sample means and variance-covariance method (Figure 4, Figure 5, Figure 6).
200   Regarding the effects of sample size, the sample means and variance-covariance method performed
201 better, but this difference only became notable when $m \lesssim 50$ and was less pronounced as k increased
202 (Figure 4). The fact that $m = 100$ gives good results regardless of the method is important to recognise,
203 because as we can be confident that any performance effects in the error and bias experiments that used
204 $m = 100$ will be a function of the imposed error or bias rather than the sample size.
205   Considering errors, in all cases the MCD and MVE methods performed better than the sample means
206 and variance-covariance method, though this difference was only evident at lower error levels for higher k
207 values (Figure 5). The pattern of response to bias was more subtle, with the sample means and variance-
208 covariance method performing better when bias was $\gtrsim 50$ %, but with this effect being very prominent
209 when $k = 0.55$, less obvious when $k = 0.75$, and no longer present when $k = 0.95$ (Figure 6).

DISCUSSION

Methodological differences

212 Errors and bias are inevitable in natural history data (Graham et al., 2004), and it is unlikely that an
213 error-free and unbiased set of data can be produced. Also, as the fundamental niche becomes less
214 similar to the sampling space, the potential effect of errors and bias should increase. This means that
215 ecological niche modellers must consider how to minimise the effects of error and bias in their analyses.
216 For those modellers using D^2 for niche modelling or outlier detection, the results from these virtual
217 ecology experiments demonstrate that MCD and MVE multivariate location and scatter methods provide
218 an opportunity to avoid the influence of errors, but this must be balanced against the influence of bias.
219 Therefore, modellers more concerned about the prevalence of errors should choose lower values for k,

while modellers more concerned about the prevalence of bias should choose higher values for k. This finding supports the advice of Qiao et al. (2015), who advise that there is no optimal modelling approach, and that choices need to be made based on the individual data and question.

As h is a free parameter that can vary $\lfloor (m+n+1)/2 \rfloor \leq h \leq m$ (Hubert and Debruyne, 2010), or in proportional terms $0.5 \lesssim k \leq 1$, it is critical that users of both MCD and MVE clearly report the value of h or k used. Specifying h or k is important for interpreting results, because when $h \to m$ or $k \to 1$ the MCD method becomes equivalent to the conventional sample means and variance-covariance method, and the MVE method produces ever larger ellipses that will eventually encapsulate all the data (Rousseeuw, 1985). Therefore, studies that do not specify the h or k parameter (Norris et al., 2006; Liu et al., 2018; Qiao et al., 2019) are not accurately reporting their methods. Also, given that the default value in software such as MASS (Venables and Ripley, 2002) is very close to $k = 0.55$, then based on the results here, if authors are not reporting a choice of h or k because they are relying on the default value, then they are potentially applying methods that do not perform well with the small sample sizes (Figure 4a) and high levels of bias (Figure 6a) that are prevalent in natural history data (Meyer et al., 2016). This supports the statement of Peterson et al. (2011) p. 113 that "it is generally poor practice to use default settings provided by software without justification, testing, and exploration of these values for a particular application", and hopefully the results presented here can provide some guidance for choosing the h value.

Those studies that have reported their choice of parameter have used values of $k = 0.95$ (Soberón et al., 2018; Altamiranda-Saavedra et al., 2020; Castaño-Quintero et al., 2020), $k = 0.975$ (Osorio-Olvera et al., 2020; Castaño-Quintero et al., 2020), and $k = 0.99$ (Yañez-Arenas et al., 2018). These seem to be sensible choices based on the virtual experiments conducted here, as in comparison to the sample means and variance-covariance method, when $k = 0.95$ the MCD and MVE methods will not be too negatively affected by small sample sizes (Figure 4c), are an improvement when errors $\lesssim 20\%$ (Figure 5c), and will not perform worse at any level of bias (Figure 6c).

In terms of choosing between MCD and MVE, while both were introduced simultaneously (Rousseeuw, 1985), MVE was initially more readily used due to its computational simplicity, but with the development of better algorithms MCD has been suggested as the preferred option due to its statistical efficiency (Rousseeuw and Van Driessen, 1999). However, the results of the experiments conducted here might indicate a slight preference for the MVE method as the MCD only performed slightly better in the error tests when the error percentage was above 30 % (Figure 5b), which is probably unrealistically large, while the MVE performed slightly better for all the bias tests (Figure 6). Ultimately I do not think there is much difference between the MCD and MVE methods, so the choice of either for ecological niche modelling is equally justifiable.

Reducing errors and bias in natural history observation data

The advantage of virtual experiments is that we can know the exact conditions of the data, but in reality the actual levels of error and bias remain unknown and can only be estimated based on experience with the data. This makes it hard to choose which multivariate location and scatter method is optimal in any given situation. However, the performance of all the multivariate location and scatter methods will improve with reduced errors and bias, so all ecological niche modellers should give this serious consideration.

Automated approaches can be tailored to rapidly detect likely errors within natural history observation data by checking for internal consistency of the meta-data and by comparison with complementary datasets (Zizka et al., 2020). In contrast, bias is harder to detect and correct (Graham et al., 2004), but is almost guaranteed to exist as this study has shown that even random sampling in geographical space leads to a biased sample of a niche in climatic space. This finding is supported by other research that also demonstrated that sampling bias is likely to become even worse as geographic sampling is further constrained to more accessible areas such as around roads (Albert et al., 2010) that is a common feature of natural history observation data (Reddy and Dávalos, 2003). This issue of bias is of particular importance for presence-only methods such as D^2 that are particularly attractive when working with natural history observation data that have no absences and that are sufficiently unstructured to reliably define the background environmental data (Etherington et al., 2009). However, while presence-only techniques have minimal data requirements, it becomes harder to detect and manage bias because the absence and background data can provide useful contextual information. When absence data are available, then presence-absence methods such as logistic regression may be expected to suffer less from bias, because biases in presence data can be balanced out by similar biases in absence data (Zadrozny, 2004).

274 Similarly, with presence-background methods, the background data can be created to have similar bias
275 to the presence data to minimise the effects of bias (Phillips et al., 2009). Effective bias reduction
276 options for presence-only methods include spatial filtering to reduce the intensity of the bias, either in
277 geographic (Boria et al., 2014) or more optimally in environmental space (Varela et al., 2014; Castellanos
278 et al., 2019). Spatial filtering has been shown to be more effective than background manipulation for
279 presence-background methods (Kramer-Schadt et al., 2013), and so it should be an effective bias reduction
280 technique for presence-only methods such as D^2, assuming the filtered sample sizes do not become
281 problematically small (Figure 4).

282 In summary, there are methods to reduce error and bias, but what level of errors and bias remain will
283 be unknown. Therefore, given bias is harder to detect than errors, and that there are reduced options
284 to control bias for presence-only models, I would suggest that ecological niche modellers err towards
285 multivariate location and scatter methods that are less sensitive to bias.

The assumption of normality

287 Regardless of the choice of multivariate location and scatter method used to calculate D^2, it is important to
288 consider if the fundamental niche can be reasonably approximated by the elliptical shapes resulting from
289 the underlying multi-dimensional normal distribution. Field studies of abundance or occurrence along
290 environmental gradients have shown some normally distributed species responses, but most responses,
291 while unimodal, are skewed, and some even show bimodal responses (Whittaker, 1952, 1956, 1960;
292 Terborgh, 1971; Austin, 1987). However, we need to recognise that it is ultimately impossible to truly
293 measure the fundamental niche in the real world, as biotic interactions mean that only the realised niche
294 can be measured, and realised niches may well take on very complex shapes that are quite different to the
295 fundamental niche (Austin and Smith, 1989; Blonder, 2016; Soberón and Peterson, 2020). In fact, the real
296 world situation is even more limited because the environmental space that can be sampled is actually a
297 complex interaction of the environments that currently exist, biotic interactions, and dispersal limitations
298 (Soberón and Peterson, 2005), and even the view of this limited environmental space is warped by the
299 sampling bias inherent in natural history observation data (Meyer et al., 2016). Given these complexities
300 of sampling from the real world, the fundamental niche can only really be measured through experimental
301 manipulations. However, there is very little of this experimental evidence (Soberón and Peterson, 2020),
302 so the expected shapes of fundamental niches remain unclear.

303 Ultimately, the inability to collect normally distributed data from the real world does not preclude the
304 use of D^2 as a fundamental niche model. Means and variances can be calculated from any distribution
305 of data, and the fact that D^2 models compare favourably with other modelling approaches (Dettmers
306 et al., 2002; Johnson and Gillingham, 2005; Tsoar et al., 2007) suggests the fundamental niche can be
307 approximated as elliptical in at least some settings. Even when it is not desirable to assume a fundamental
308 niche is normally distributed, D^2 can still be a used to eliminate outliers. This could support other
309 fundamental niche modelling methods, such as convex hulls (Pironon et al., 2019), that are not limited to
310 the assumption of normality but are sensitive to outliers (Blonder, 2018).

CONCLUSIONS

312 When using D^2 for ecological niche modelling and outlier detection, the performance of multivariate
313 location and scatter methods varies based on sample size, error, and bias. Comparison of the sample
314 means and variance-covariance, MCD, and MVE multivariate location and scatter methods provides
315 the clear conclusion that none of the methods, or individual parameterisations of any methods, can be
316 considered universally the best. Rather, any ecological niche modeller using these techniques needs to
317 think carefully about their data and objective to choose the method and parameterisation that are most
318 appropriate to their individual circumstances. For those modellers who wish to explore the potential of the
319 MCD and MVE methods, given these methods have been used widely in statistical analyses for some time,
320 these methods should be widely available in statistical software. However, modellers using the MCD and
321 MVE methods should carefully consider and clearly state the h or k parameter used in their analyses.

REFERENCES

323 Aggarwal, C. C. (2017). *Outlier Analysis*. Springer, Cham, Switzerland, 2nd edition edition.

324 Albert, C. H., Yoccoz, N. G., Edwards Jr, T. C., Graham, C. H., Zimmermann, N. E., and Thuiller, W.
325 (2010). Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography*,
326 33(6):1028–1037.
327 Altamiranda-Saavedra, M., Osorio-Olvera, L., Yáñez-Arenas, C., Marín-Ortiz, J. C., and Parra-Henao, G.
328 (2020). Geographic abundance patterns explained by niche centrality hypothesis in two Chagas disease
329 vectors in Latin America. *PLoS ONE*, 15(11):e0241710.
330 Austin, M. P. (1987). Models for analysis of species' response to environmental gradients. *Vegetatio*,
331 69(1-3):35–45.
332 Austin, M. P. and Smith, T. M. (1989). A new model for the continuum concept. *Vegetatio*, 83(1-2):35–47.
333 Blonder, B. (2016). Do hypervolumes have holes? *The American Naturalist*, 187(4):E93–E105.
334 Blonder, B. (2018). Hypervolume concepts in niche- and trait-based ecology. *Ecography*, 41(9):1441–
335 1455.
336 Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce
337 sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275:73–
338 77.
339 Castaño-Quintero, S., Escobar-Luján, J., Osorio-Olvera, L., Peterson, A. T., Chiappa-Carrara, X.,
340 Martínez-Meyer, E., and Yañez-Arenas, C. (2020). Supraspecific units in correlative niche mod-
341 eling improves the prediction of geographic potential of biological invasions. *PeerJ*, 8:e10454.
342 Castellanos, A. A., Huntley, J. W., Voelker, G., and Lawing, A. M. (2019). Environmental filtering
343 improves ecological niche models across multiple scales. *Methods in Ecology and Evolution*, 10(4):481–
344 492.
345 Chang, W. (2014). *extrafont: Tools for using fonts*. R package version 0.17.
346 Dettmers, R., Buehler, D. A., and Bartlett, J. B. (2002). *A test and comparison of wildlife-habitat modeling*
347 *techniques for predicting bird occurrence at a regional scale*, pages 607–616. Island Press, Washington.
348 Edwards, J. L., Lane, M. A., and Nielsen, E. S. (2000). Interoperability of biodiversity databases:
349 biodiversity information on every desktop. *Science*, 289(5488):2312–2314.
350 Etherington, T. R. (2019). Mahalanobis distances and ecological niche modelling: correcting a chi-squared
351 probability error. *PeerJ*, 7:e6678.
352 Etherington, T. R. and Omondiagbe, O. P. (2019). virtualNicheR: generating virtual fundamental and
353 realised niches for use in virtual ecology experiments. *Journal of Open Source Software*, 4(41):1661.
354 Etherington, T. R., Ward, A. I., Smith, G. C., Pietravalle, S., and Wilson, G. J. (2009). Using the
355 Mahalanobis distance statistic with unplanned presence-only survey data for biogeographical models of
356 species distribution and abundance: a case study of badger setts. *Journal of Biogeography*, 36(5):845–
357 853.
358 Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., and Peterson, A. T. (2004). New developments in
359 museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*,
360 19(9):497–503.
361 Hijmans, R. J. (2020). *raster: Geographic Data Analysis and Modeling*. R package version 3.3-7.
362 Hubert, M. and Debruyne, M. (2010). Minimum covariance determinant. *WIREs Computational Statistics*,
363 2(1):36–43.
364 Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*,
365 22:415–427.
366 Hutchinson, G. E. (1978). *An Introduction to Population Ecology*. Yale University Press, New Haven.
367 Johnson, C. J. and Gillingham, M. P. (2005). An evaluation of mapped species distribution models used
368 for conservation planning. *Environmental Conservation*, 32(2):117–128.
369 Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E.,
370 Linder, H. P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas.
371 *Scientific Data*, 4(1):170122.
372 Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M.,
373 Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W.,
374 Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima,
375 H., Duckworth, J. W., Breitenmoser-Wuersten, C., Belant, J. L., Hofer, H., and Wilting, A. (2013).
376 The importance of correcting for sampling bias in maxent species distribution models. *Diversity and*
377 *Distributions*, 19(11):1366–1379.
378 Liu, C. R., White, M., and Newell, G. (2018). Detecting outliers in species distribution data. *Journal of*

379   *Biogeography*, 45(1):164–176.

380   Maguire, B. (1973). Niche response structure and the analytical potentials of its relationship to the habitat.
381   *American Naturalist*, 107(954):213–246.

382   Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute*
383   *of Sciences of India*, 2(1):49–55.

384   Manly, B. F. J. (2005). *Multivariate Statistical Methods: a primer*. Chapman & Hall/CRC Press, Boca
385   Raton, 3rd edition.

386   Meyer, C., Weigelt, P., and Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global
387   plant occurrence information. *Ecology Letters*, 19(8):992–1006.

388   Norris, J. R., Jackson, S. T., and Betancourt, J. L. (2006). Classification tree and minimum-volume
389   ellipsoid analyses of the distribution of ponderosa pine in the western usa. *Journal of Biogeography*,
390   33(2):342–360.

391   Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). *fields: Tools for spatial data*. R package version
392   9.8-3.

393   Osorio-Olvera, L., Yañez-Arenas, C., Martínez-Meyer, E., and Peterson, A. T. (2020). Relationships
394   between population densities and niche-centroid distances in North American birds. *Ecology Letters*,
395   23(3):555–564.

396   Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínes-Meyer, E., Nakamura, M., and
397   Araújo, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press,
398   Princeton.

399   Phillips, S., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009).
400   Sample selection bias and presence-only distribution models: implications for background and pseudo-
401   absence data. *Ecological Applications*, 19(1):181–197.

402   Pironon, S., Etherington, T. R., Borrell, J. S., Kühn, N., Macias-Fauria, M., Ondo, I., Tovar, C., Wilkin, P.,
403   and Willis, K. J. (2019). Potential adaptive strategies for 29 sub-Saharan crops under future climate
404   change. *Nature Climate Change*, 9(10):758–763.

405   Qiao, H., Feng, X., Escobar, L. E., Peterson, A. T., Soberón, J., Zhu, G. P., and Papeş, M. (2019). An
406   evaluation of transferability of ecological niche models. *Ecography*, 42(3):521–534.

407   Qiao, H., Soberón, J., and Peterson, A. T. (2015). No silver bullets in correlative ecological niche
408   modelling: insights from testing among many potential algorithms for niche estimation. *Methods in*
409   *Ecology and Evolution*, 6(10):1126–1136.

410   R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for
411   Statistical Computing, Vienna, Austria.

412   Reddy, S. and Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation
413   priorities in Africa. *Journal of Biogeography*, 30(11):1719–1727.

414   Rousseeuw, P. J. (1985). *Multivariate estimation with high breakdown point*, pages 283–297. Reidel
415   Publishing Company, Dordrecht.

416   Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant
417   estimator. *Technometrics*, 41(3):212–223.

418   Scamander, N. (2001). *Fantastic beasts and where to find them*. Scholastic Press, New York.

419   Soberón, J. and Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and
420   species' distributional areas. *Biodiversity Informatics*, 2:1–10.

421   Soberón, J. and Peterson, A. T. (2020). What is the shape of the fundamental grinnellian niche? *Theoretical*
422   *Ecology*, 13(1):105–115.

423   Soberón, J., Peterson, A. T., and Osorio-Olvera, L. (2018). A comment on "Species are not most abundant
424   in the centre of their geographic range or climatic niche". *Rethinking Ecology*, 3:13–18.

425   Terborgh, J. (1971). Distribution on environmental gradients: theory and a preliminary interpretation of
426   distributional patterns in the avifauna of the Cordillera Vilcabamba, Peru. *Ecology*, 52(1):23–40.

427   Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., and Kadmon, R. (2007). A comparative evaluation of
428   presence-only methods for modelling species distribution. *Diversity and Distributions*, 13(4):397–405.

429   Van Aelst, S. and Rousseeuw, P. (2009). Minimum volume ellipsoid. *WIREs Computational Statistics*,
430   1(1):71–82.

431   Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters
432   reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*,
433   37(11):1084–1091.

434 Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer, New York, 4th
435   edition.
436 Whittaker, R. H. (1952). A study of summer foliage insect communities in the Great Smoky Mountains.
437   *Ecological Monographs*, 22(1):1–44.
438 Whittaker, R. H. (1956). Vegetation of the Great Smoky Mountains. *Ecological Monographs*, 26(1):1–80.
439 Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological*
440   *Monographs*, 30(3):279–338.
441 Yañez-Arenas, C., Rioja-Nieto, R., Martín, G. A., Dzul-Manzanilla, F., Chiappa-Carrara, X., Buenfil-Ávila,
442   A., Manrique-Saide, P., Correa-Morales, F., Díaz-Quiñónez, J. A., Pérez-Rentería, C., Ordoñez-Álvarez,
443   J., Vazquez-Prokopec, G., and Huerta, H. (2018). Characterizing environmental suitability of *Aedes*
444   *albopictus* (Diptera: Culicidae) in Mexico based on regional and global niche models. *Journal of*
445   *Medical Entomology*, 55(1):69–77.
446 Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of*
447   *the Twenty-First International Conference on Machine Learning*, ICML '04, page 114, New York, NY,
448   USA. Association for Computing Machinery.
449 Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J. F. R.,
450   Colli-Silva, M., Fantinati, M. R., Fernandes, M. F., Ferreira-Araújo, T., Gondim Lambert Moreira,
451   F., Santos, N. M. C., Santos, T. A. B., dos Santos-Costa, R. C., Serrano, F. C., Alves da Silva, A. P.,
452   de Souza Soares, A., Cavalcante de Souza, P. G., Calisto Tomaz, E., Vale, V. F., Vieira, T. L., and
453   Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *PeerJ*, 8:e9916.
454 Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J.,
455   Reineking, B., Schröder, B., and Grimm, V. (2010). The virtual ecologist approach: simulating data
456   and observers. *Oikos*, 119(4):622–635.
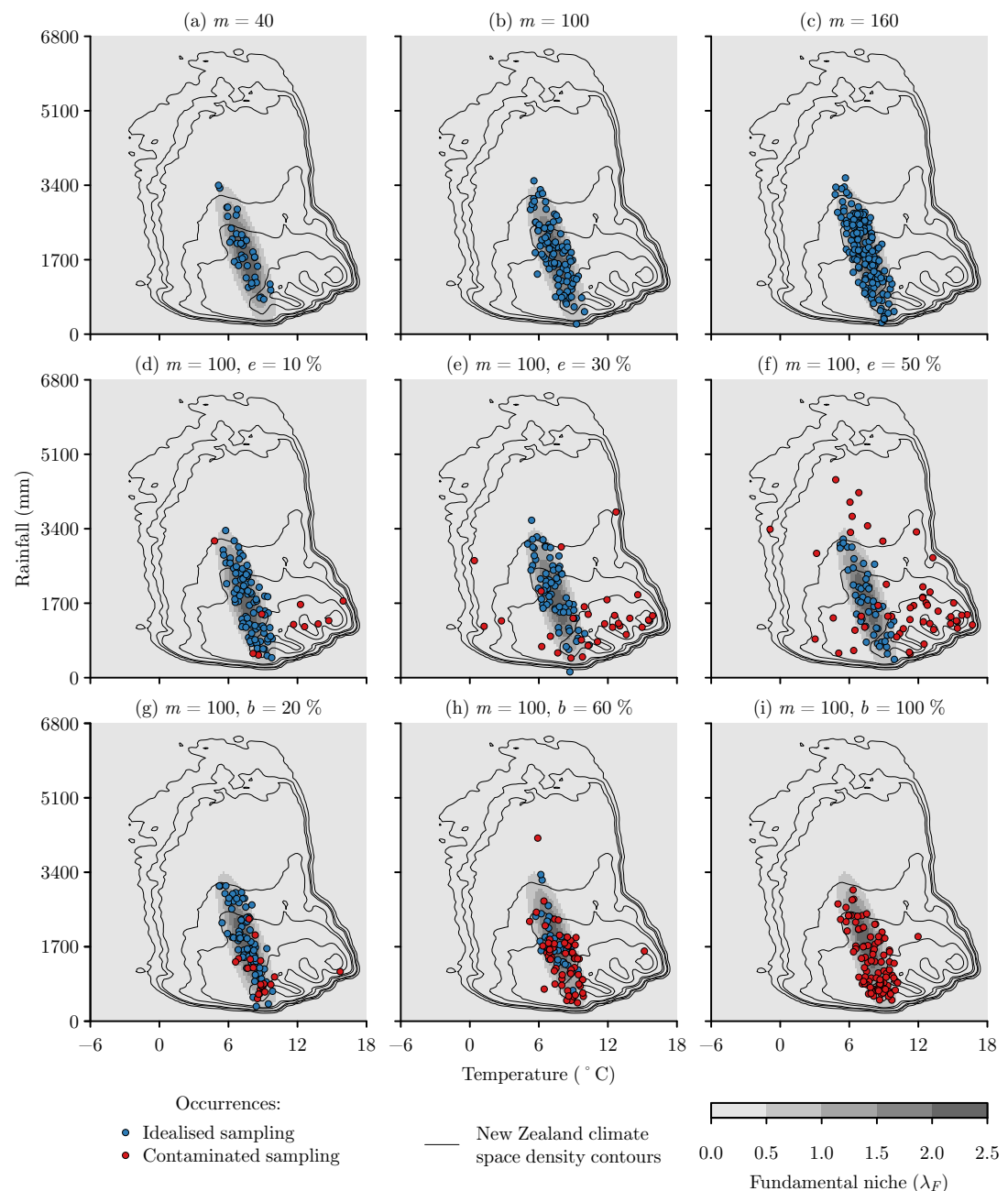
Figure 3. Examples of virtual species occurrence samples formed by idealised and contaminated sampling at varying levels of sample size (m), error (e), and bias (b). Idealised samples of (a) $m = 40$, (b) $m = 100$, and (c) $m = 160$ follow the elliptical shape of the fundamental niche with a greater intensity of occurrences towards the centre of the niche. Contaminated sampling was based upon the climate space of New Zealand, with increasing errors of (d) $e = 10\%$, (e) $e = 30\%$, and (f) $e = 50\%$ producing samples that increasingly represent the climate space rather than the niche, and increasing bias of (g) $b = 20\%$, (h) $b = 60\%$, and (i) $b = 100\%$ producing samples that increasingly represent the overlap between the climate space and the niche.
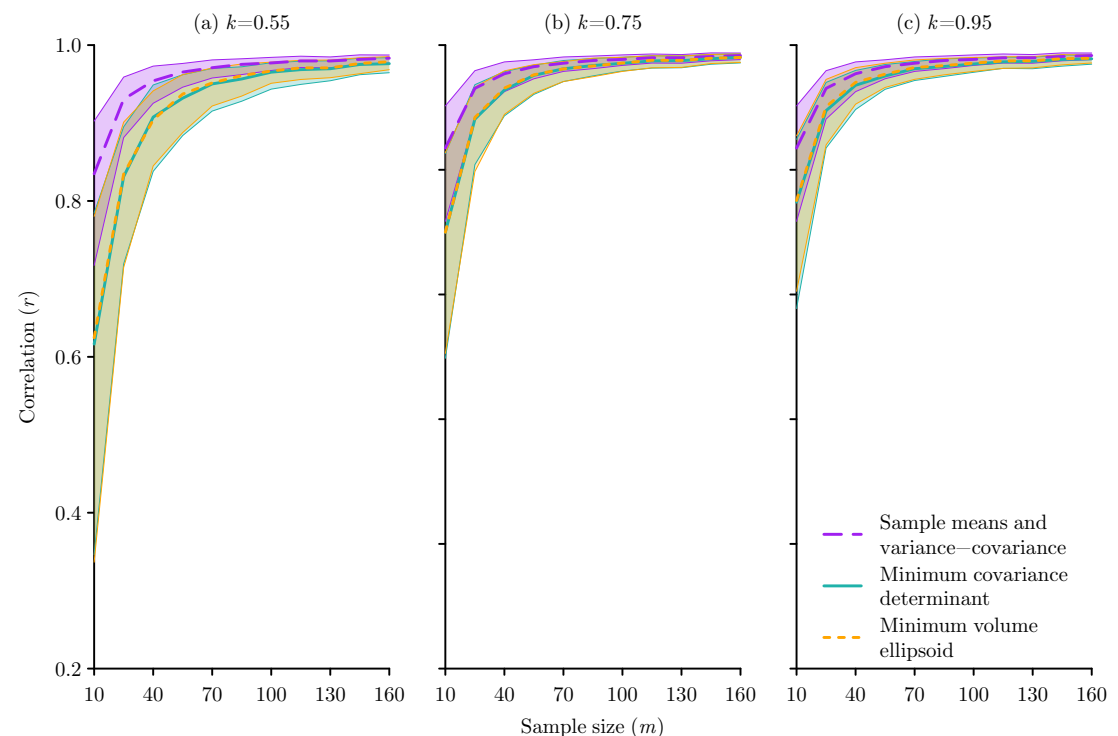
Figure 4. The effect of species occurrence sample size on the performance of Mahalanobis distance niche models based on three different multivariate location and scatter methods: sample means and variance-covariance, minimum covariance determinant, and minimum volume ellipsoid. The median and inter-quartile range of the correlation between the known niche value and the Mahalanobis distance probability for the occurrence sample from 500 replications are plotted for each method. The proportion k of occurrences used for the minimum covariance determinant and the minimum volume ellipsoid were set at (a) $k = 0.55$, (b) $k = 0.75$, and (c) $k = 0.95$.
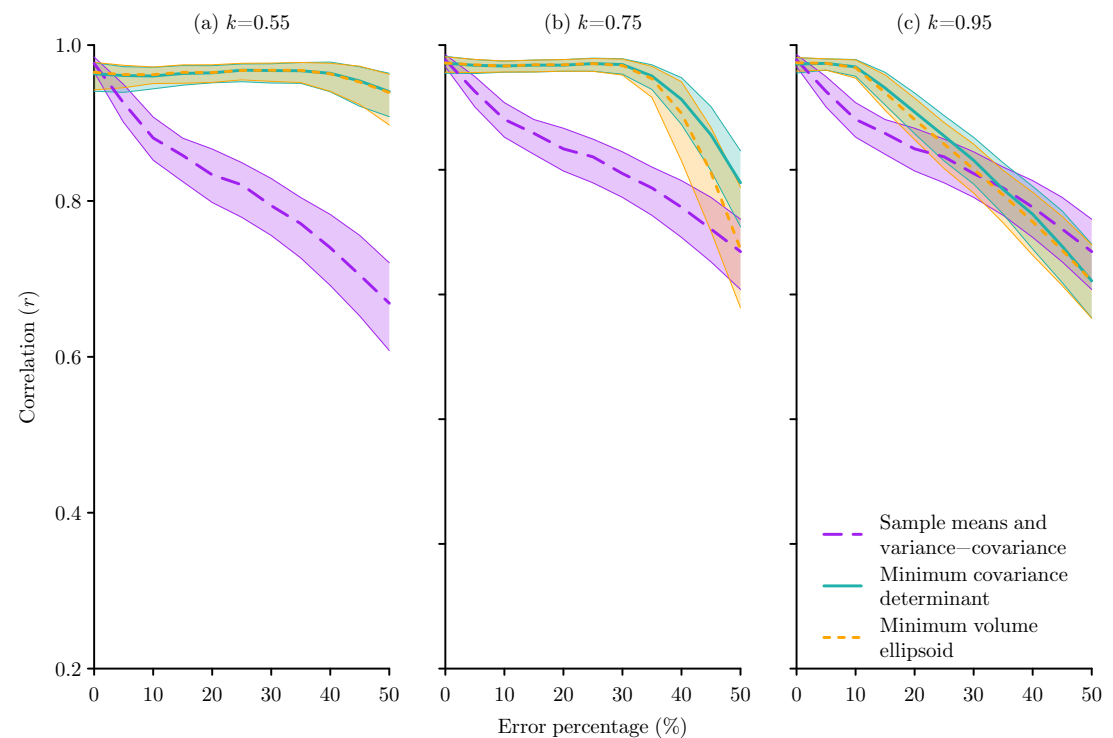
Figure 5. The effect of species occurrence sample errors on the performance of Mahalanobis distance niche models based on three different multivariate location and scatter methods: sample means and variance-covariance, minimum covariance determinant, and minimum volume ellipsoid. The median and inter-quartile range of the correlation between the known niche value and the Mahalanobis distance probability for the occurrence sample from 500 replications are plotted for each method. The proportion k of occurrences used for the minimum covariance determinant and the minimum volume ellipsoid were set at (a) $k = 0.55$, (b) $k = 0.75$, and (c) $k = 0.95$.
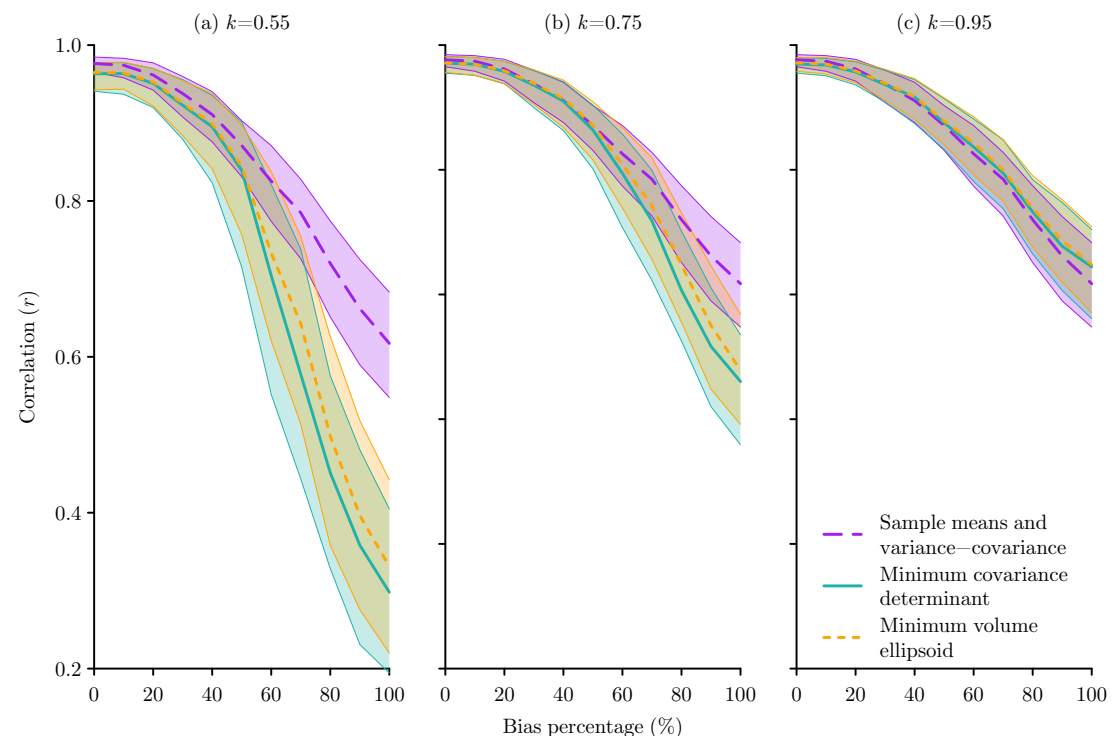
Figure 6. The effect of species occurrence sample bias on the performance of Mahalanobis distance niche models based on three different multivariate location and scatter methods: sample means and variance-covariance, minimum covariance determinant, and minimum volume ellipsoid. The median and inter-quartile range of the correlation between the known niche value and the Mahalanobis distance probability for the occurrence sample from 500 replications are plotted for each method. The proportion k of occurrences used for the minimum covariance determinant and the minimum volume ellipsoid were set at (a) $k = 0.55$, (b) $k = 0.75$, and (c) $k = 0.95$.