Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error

Thomas R Etherington Corresp. 1

1 Manaaki Whenua – Landcare Research, Lincoln, New Zealand

Corresponding Author: Thomas R Etherington
Email address: etheringtont@landcareresearch.co.nz

The Mahalanobis distance is a statistical technique that can be used to measure how distant a point is from the centre of a multivariate normal distribution. By measuring Mahalanobis distances in environmental space ecologists have also used the technique to model: ecological niches, habitat suitability, species distributions, and resource selection functions. Unfortunately the original description of the Mahalanobis distance technique for ecological modelling contained an error describing how Mahalanobis distances could be converted into probabilities using a chi-squared distribution. This error has been repeated in the literature, and is present in popular modelling software. In the hope of correcting this error to maximise the potential application of the Mahalanobis distance technique within the ecological modelling community, I explain how Mahalanobis distances are calculated, and through a virtual ecology experiment demonstrate how to correctly produce probabilities and discuss the implications of the error for previous Mahalanobis distance studies.

Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error

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 can be used to measure how distant a point is from the centre of a multivariate normal distribution. By measuring Mahalanobis distances in environmental space ecologists have also used the technique to model: ecological niches, habitat suitability, species distributions, and resource selection functions. Unfortunately the original description of the Mahalanobis distance technique for ecological modelling contained an error describing how Mahalanobis distances could be converted into probabilities using a chi-squared distribution. This error has been repeated in the literature, and is present in popular modelling software. In the hope of correcting this error to maximise the potential application of the Mahalanobis distance technique within the ecological modelling community, I explain how Mahalanobis distances are calculated, and through a virtual ecology experiment demonstrate how to correctly produce probabilities and discuss the implications of the error for previous Mahalanobis distance studies.

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. Consider a data matrix $\mathbf{A}$ with m rows of observations and n columns of measured variables.

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

The Mahalanobis distance D^2 for each observation vector $\mathbf{x}_m = [x_{m1}, x_{m2}, x_{m3}, \dots, x_{mn}]$ is calculated as a function of an n-dimensional vector $\bar{\mathbf{x}}$ containing the means for each column of variables, and a variance-covariance matrix $\mathbf{S}$ of dimensions $n \times n$ that contains variances for each column along the main diagonal and pair-wise column covariances values elsewhere (Manly, 2005).

$$D^2(\mathbf{x}_m) = (\mathbf{x}_m - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_m - \bar{\mathbf{x}}) \tag{1}$$

When applied to $m = 50$ points in $n = 2$ dimensions, the calculated D^2 values follow a characteristic elliptical pattern with D^2 radiating out from the central location of the distribution (Figure 1a).

While D^2 can be calculated for any n-dimensions, the values of D^2 are not comparable when n varies, as D^2 increases as n increases (Figure 2a-c). However, the $\bar{\mathbf{x}}$ and $\mathbf{S}^{-1}$ used to calculate D^2 (Equation 1) transform the values from each n into independent standard normal distributions, with $\bar{\mathbf{x}}$ centering and $\mathbf{S}^{-1}$ scaling and rotating each variable distribution. This means that D^2 is essentially the sum of n independent

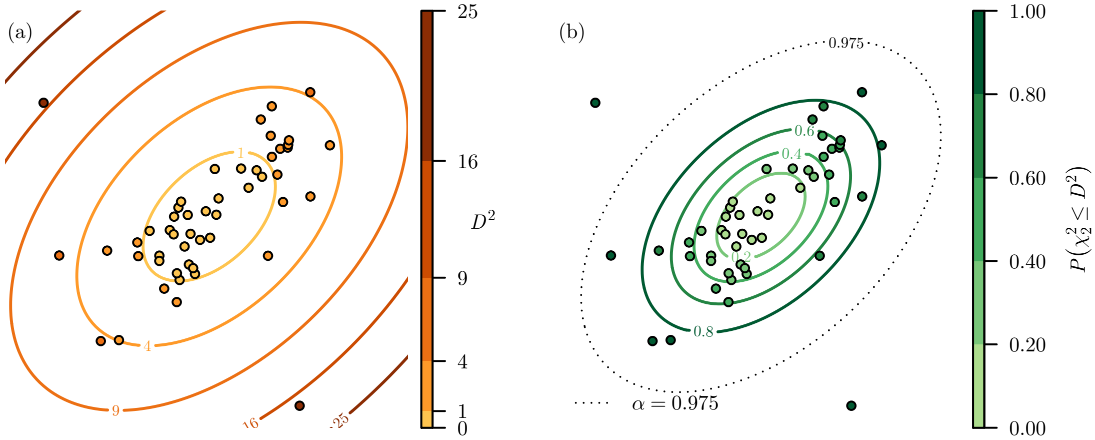

Figure 1. Two-dimensional example of Mahalanobis distance. (a) Given a set of points distributed in two-dimensional space, the Mahalanobis distances (D^2) for each point can be calculated. (b) The D^2 values can be transformed into probabilities using a chi-squared cumulative probability distribution. This highlights that there are two points points that have a very high probability of not belonging to the distribution and could be classified as outliers when $\alpha = 0.975$ as they are beyond that probability threshold.

standard normal variables, and as such follows a chi-squared distribution with degrees of freedom equal to the number of dimensions n (Manly, 2005).

For a chi-squared random variable with n degrees of freedom denoted as χ_n^2, the probability density function f of χ_n^2 when $x \geq 0$ is

$$f_{\chi_n^2}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} \tag{2}$$

where Γ is a gamma function (as the chi- squared distribution is actually a special case of the gamma distribution), and the associated cumulative distribution function is $F_{\chi_n^2}(x) = P(\chi_n^2 \leq x)$ (Johnson et al., 1994). When the variables in **A** are normally distributed, the association between D^2 and $f_{\chi_n^2}(x)$ can be clearly seen (Figure 2a-c).

By converting D^2 into probabilities using $F_{\chi_n^2}(x)$ we can put D^2 from any number of dimensions on a common 0-1 scale that indicates the probability $P(\chi_n^2 \leq D^2)$ that a location has a D^2 that is greater than that would be expected by chance (Figure 2d-f). By specifying a significance level α this process is commonly used as an outlier detection method as it is: parameter free, computational efficient, accounts for collinearity between variable dimensions, and is scale independent (Aggarwal, 2017). Returning to our earlier example, having transformed the values from D^2 to $P(\chi_n^2 \leq D^2)$ we can see that there are two points that are very likely to be outliers, and would be classified as outliers with $\alpha = 0.975$ (Figure 1b).

The potential of D^2 for use in habitat modelling was first identified by Clark et al. (1993) and then discussed further in the context of niche modelling by Farber and Kadmon (2003) — which is how I will continue the discussion. The premise here is that given a data matrix **A** of m species observations for which various environmental variables n are measured, D^2 can be used as a measure of niche suitability from an optimum location in environmental space. Having defined a niche in this way, by measuring the D^2 for each location on a landscape a map of niche suitability can then be produced. The key advantages of using D^2 over other methods are that the D^2 method needs only presence information, and so does not require either absences or a background definition, and that independence of explanatory variables is not required (Clark et al., 1993; Farber and Kadmon, 2003). Studies that have compared Mahalanobis distance to other modelling approaches have also shown that while the optimum method tends to vary by the species in question, the Mahalanobis distance approach performs well against a variety of other presence-only, presence-background, and presence-absence modelling approaches (Dettmers et al., 2002; Johnson and Gillingham, 2005; Tsoar et al., 2007).

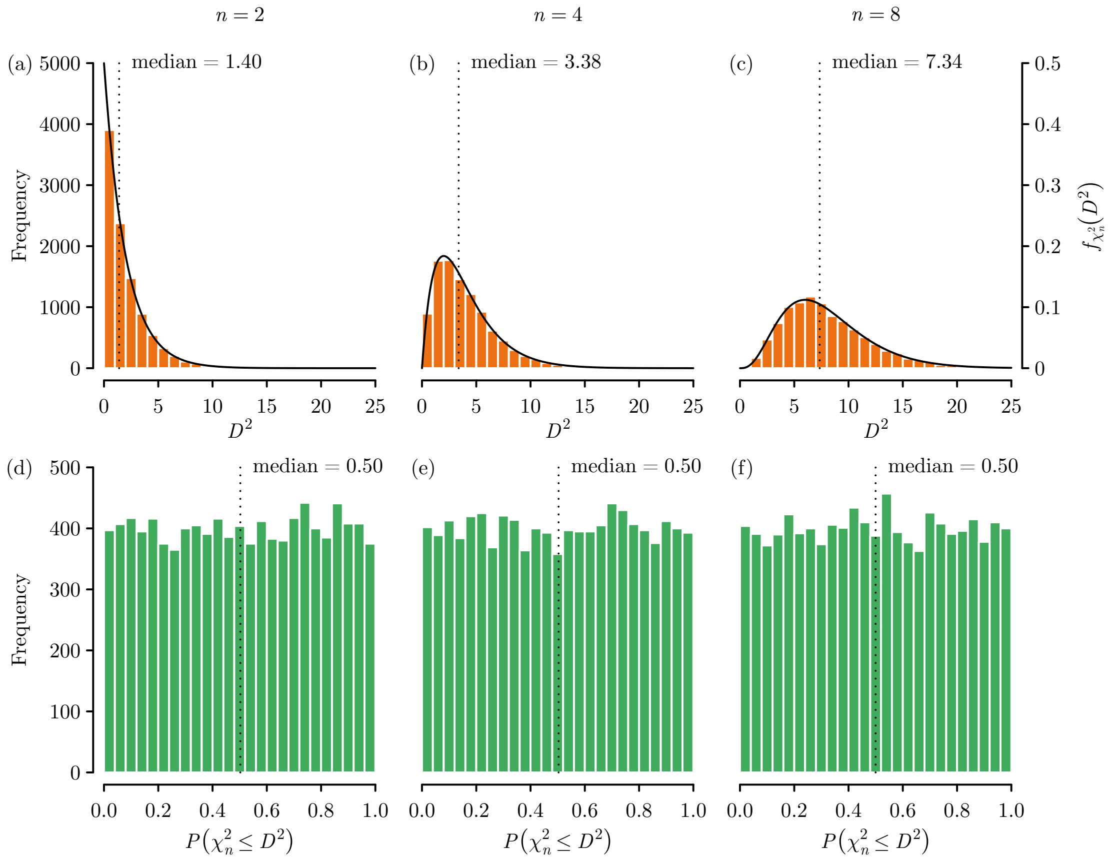

Figure 2. Mahalanobis distance (D^2) dimensionality effects using data randomly generated from independent standard normal distributions. We can see that the values of D^2 grow following a chi-squared distribution as a function of the number of dimensions (a) $n = 2$, (b) $n = 4$, and (c) $n = 8$. By using a chi-squared cumulative probability distribution the D^2 values can be put on a common scale, such that all values range 0-1 and so that statistics such as the median are consistent across dimensions. The near uniform distribution of the probability values is expected given the underlying data were randomly generated from independent standard normal distributions.

While introducing D^2 into the ecological modelling domain Clark et al. (1993) also highlighted that D^2 will follow a chi-squared distribution, and therefore the potential to convert the distances into probabilities. However, as ecological niche models aim to describe probability of belonging to the niche, rather than use the chi-squared cumulative distribution function $F_{\chi^2_n}(x) = P(\chi^2_n \leq x)$ that has 0 at the optimum, we use the inverse chi-squared cumulative distribution function $F^{-1}_{\chi^2_n}(x) = P(\chi^2_n > x)$ that has 1 at the optimum. This means that the probabilities $P(\chi^2_n > D^2)$ indicate locations with a D^2 that is less than that would be expected by chance, and hence are more likely to be within the niche.

Unfortunately, when describing the use of a chi-squared distribution to convert D^2 into probabilities, Clark et al. (1993, p.522) state that “Assuming multivariate normality, Mahalanobis distances are approximately distributed as Chi-square with $n-1$ degrees of freedom, where n equals the number of habitat characters.”, but this is incorrect. D^2 values follow a chi-squared distribution with degrees of freedom equal to n (Manly, 2005) as has already been clearly shown (Figure 2). This error has been repeatedly described in the literature (Knick and Rotenberry, 1998; Farber and Kadmon, 2003; Hellgren et al., 2007), and has even permeated into software such as the R package adehabitat (Calenge, 2006).

To demonstrate this error and to examine its implications, I present an experiment based on a virtual ecology approach (Zurell et al., 2010) – which allows us to examine methodologies in a controlled system uncomplicated by the uncertainties of the real world!

MATERIALS AND METHODS

The virtual ecology experiment began by defining, and therefore knowing truthfully, the fundamental niche of the imaginary species *Mimbulus mimbletonia* (Rowling, 2003). This fundamental niche N was defined using a multivariate normal distribution describing the niche in relation to two environmental variables of temperature and rainfall.

$$N_{(x,y)} = e^{-\frac{1}{2}((x,y)-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}((x,y)-\boldsymbol{\mu})}, \text{where } \boldsymbol{\mu} = \begin{bmatrix} 25 & 100 \end{bmatrix} \boldsymbol{\Sigma} = \begin{bmatrix} 9 & 60 \\ 60 & 625 \end{bmatrix} \tag{3}$$

A sample of observations was then created by randomly sampling the niche space ranging 15–35°C of temperature and 0–200 mm of rainfall. At each randomly selected location the species was considered detected using a probability equal to the fundamental niche. This process was continued until a sample of 200 observations was generated.

The values of temperature and rainfall at each of the 200 sampled locations were then used to estimate the fundamental niche using Mahalanobis distances. The D^2 was calculated for each of the sampled locations, and the resulting distribution of D^2 was compared against a range of inverse cumulative chi-squared distribution functions based on differing degrees of freedom.

Finally, the D^2 values for the samples were converted into probabilities of belonging to the fundamental niche, and these predictions were compared against the known fundamental niche values for the locations. Probabilities were calculated using chi-squared distributions with n and $n-1$ degrees of freedom to examine any differences.

RESULTS

The fundamental niche defined using the multivariate normal distribution (Equation 3) produced an elliptically shape niche with positive correlation between rainfall and temperature (Figure 3a). The random sampling resulted in a set of 200 samples that followed this elliptical niche pattern, with a greater concentration of samples towards the centre of the niche (Figure 3b).

The D^2 values calculated on the basis of these samples also followed an elliptical pattern (Figure 3c), and when the calculated D^2 values for each sample were plotted against the actual fundamental niche value for each sample, a trend that clearly follows the inverse cumulative chi-squared distribution when $n = 2$ can be seen (Figure 3d). This is as we would expect as in this example the fundamental niche is based on the two environmental variables of temperature and rainfall.

When the D^2 values were converted to probabilities using $n = 2$ degrees of freedom we see a near-perfect linear fit between the estimated niche suitability and the actual known niche suitability (Figure 3e). This proves quite clearly that to get a truthful estimate of our known fundamental niche, the probabilities need to be based on an inverse cumulative chi-squared distribution with n degrees of freedom. In contrast, when the D^2 values are converted to probabilities using $n-1$ degrees of freedom we see a badly-fitting curvilinear relationship that underestimates niche suitability (Figure 3f).

DISCUSSION

The results of the virtual ecology experiment (Figure 3) clearly shows the erroneous under prediction of niche suitability when using a chi-squared distribution with $n-1$ degrees of freedom. Fortunately most previous D^2 studies simply rescaled the D^2 into quantiles or ranks of increasing suitability (Knick and Dyer, 1997; Knick and Rotenberry, 1998; Johnson and Gillingham, 2005; Hellgren et al., 2007; Etherington et al., 2009) or binary classifications based on a threshold (Farber and Kadmon, 2003; Thatcher et al., 2006; Tsoar et al., 2007) and so are not affected by this problem. Also, although I could not find any examples of this, those studies that did create chi-square probabilities using $n-1$ degrees of freedom, but then converted to categories based on quantiles or a predictive threshold would not have impacted the conclusions of the study as while the shape of the relationship becomes curved with $n-1$, the trend is still one of monotonic increase, therefore the quantiles would be the same. However, those studies that have created chi-squared probabilities with $n-1$ as some form of suitability index (Clark et al., 1993) will have underestimated the niche suitability.

CONCLUSION

Given that D^2 values are unitless and unbounded and are not directly comparable for different dimensions, I would argue that anyone using the Mahalanobis distance method should present their results as chi-squared probabilities. This will put Mahalanobis distance models on a 0-1 scale that enables models based on differing numbers of n to be directly comparable, and is consistent with most other types of ecological niche models that also use a 0-1 scale. As such, it will be very important that the chi-square probabilities are calculated correctly and hopefully the methodological description and experimental evidence presented here will enable that to be achieved.

REFERENCES

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

Calenge, C. (2006). The package “adehabitat” for the r software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197(3-4):516–519.

Clark, J. D., Dunn, J. E., and Smith, K. G. (1993). A multivariate model of female black bear habitat use for a geographic information system. *Journal of Wildlife Management*, 57(3):519–526.

Dettmers, R., Buehler, D. A., and Bartlett, J. B. (2002). *A test and comparison of wildlife-habitat modeling techniques for predicting bird occurrence at a regional scale*, pages 607–616. Island Press, Washington.

Etherington, T. R., Ward, A. I., Smith, G. C., Pietravalle, S., and Wilson, G. J. (2009). Using the mahalanobis distance statistic with unplanned presence-only survey data for biogeographical models of species distribution and abundance: a case study of badger setts. *Journal of Biogeography*, 36(5):845–853.

Farber, O. and Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the mahalanobis distance. *Ecological Modelling*, 160:115–130.

Hellgren, E. C., Bales, S. L., Gregory, M. S., Leslie, D. M., and Clark, J. D. (2007). Testing a mahalanobis distance model of black bear habitat use in the ouachita mountains of oklahoma. *Journal of Wildlife Management*, 71(3):924–928.

Johnson, C. J. and Gillingham, M. P. (2005). An evaluation of mapped species distribution models used for conservation planning. *Environmental Conservation*, 32(2):117–128.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Probability Distributions: volume 1*. John Wiley & Sons, New York, 2nd edition.

Knick, S. T. and Dyer, D. L. (1997). Distribution of black-tailed jackrabbit habitat determined by gis in southwestern idaho. *Journal of Wildlife Management*, 61(1):75–85.

Knick, S. T. and Rotenberry, J. T. (1998). Limitations to mapping habitat use areas in changing landscapes using the mahalanobis distance statistic. *Journal of Agricultural, Biological, and Environmental Statistics*, 3(3):311–322.

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

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

Rowling, J. K. (2003). *Harry Potter and the Order of the Phoenix*. Bloomsbury, London.

Thatcher, C. A., Van Manen, F. T., and Clark, J. D. (2006). Identifying suitable sites for florida panther reintroduction. *Journal of Wildlife Management*, 70(3):752–763.

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

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., and Grimm, V. (2010). The virtual ecologist approach: simulating data and observers. *Oikos*, 119(4):622–635.

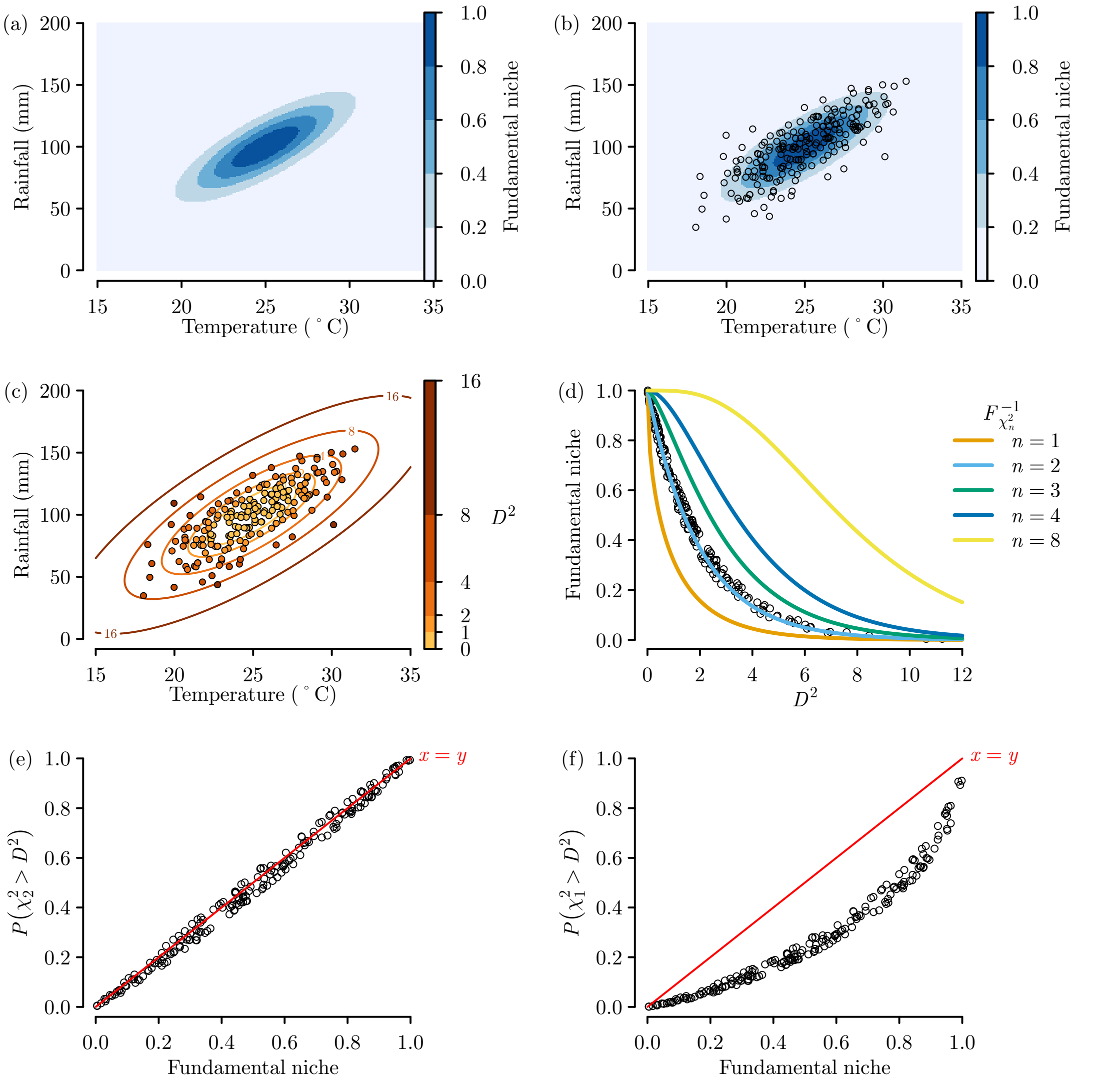

Figure 3. Virtual ecology experiment to examine the importance of chi-squared distribution degrees of freedom. (a) The virtual fundamental niche with (b) random sampling across niche space and (c) resulting D^2 calculations. (d) The trend of D^2 compared against the inverse chi-squared cumulative probability distributions with differing degrees of freedom n. The trend for the sampling locations between the actual fundamental niche values and the niche estimates from D^2 values converted to probabilities via an inverse cumulative probability distribution with degrees of freedom equal to (e) $n = 2$ and (f) $n = 1$ (or $n = 2 - 1$).