

DOI: 10.5846/stxb201402260333

周继华, 来利明, 郑元润. 一种模型模拟结果的统计检验方法. 生态学报, 2015, 35(19): - .

Zhou J H, Lai L M, Zheng Y R. A statistical method for validation of ecological models. Acta Ecologica Sinica, 2015, 35(19): - .

一种模型模拟结果的统计检验方法

周继华^{1,2}, 来利明¹, 郑元润^{1,*}

1 中国科学院植物研究所, 北京 100093

2 中国科学院大学, 北京 100049

摘要: 模拟结果的准确性是衡量生态学模型是否成功的关键, 但采用统计学方法判别模型模拟结果与观察值相符程度的报道较少。根据两个直线回归方程能否合并为一个方程的统计学检验方法, 提出了通过检验观察值与模拟值直线回归方程和 1:1 直线方程截距与斜率是否相同, 进而在统计显著水平上判断生态学模型模拟值与观察值一致性的统计学检验方法。数据检验表明, 此方法可以较好解决判断生态学模型模拟结果准确性的问题。

关键词: 生态学模型; 模拟精度; 统计学检验

A statistical method for validation of ecological models

ZHOU Jihua^{1,2}, LAI Liming¹, ZHENG Yuanrun^{1,*}

1 Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

2 University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Statistical validation has rarely been conducted to examine the consistency between observed data and the output of ecological models, although such validation is critical for determining the goodness-of-fit of such models. Based on the statistical principle of whether two regression models can be combined, a statistical method is proposed for validating ecological models. In the method, a linear regression is fit to observed (x-axis) and modeled (y-axis) data, then tested for significant differences in slope and intercept from the line $y = x$. A case study shows that the method could be used to validate ecological models.

Key Words: Ecological model; Model validation; Statistical test

采用数学模型模拟自然及社会中发生的各种过程是学科发展成熟的重要标志, 只有对这些过程有较为清晰的认识时才能采用模型进行模拟。随着研究的不断深入, 在包括生态学在内的自然与社会科学中开展了大量经典的模型研究工作^[1, 2]。但关于模型模拟结果与观察值相符程度的检验, 也即模型的验证一直是一个非常薄弱的环节^[3]。模型能否在统计显著水平上准确模拟所研究的自然现象一直是一个难题^[4, 5]。目前的生态学文献中对模拟值与观察值相符程度进行统计检验的研究较少。一般的方法是作出模拟值与观察值的散点图显示模拟结果的优劣, 并对观察值与模拟值进行回归分析^[1, 3, 4]; 有些通过分析模拟值在 1:1 直线两侧的分布情况, 判断模型模拟结果的准确性^[3, 6], 再进一步则是比较观察值与模拟值差值绝对值的大小, 绝对值越小, 模拟值与观察值相符越好, 模型的模拟结果越好^[6]。近年来一些研究采用均方根误差、绝对误差等统计量判断模型模拟结果的优劣^[7]。但上述方法的共同缺点是不能从统计学上解释模拟值与观察值的相符程

基金项目: 国家自然科学基金项目 (91225302)

收稿日期: 2014-02-26; 网络出版日期: 2014-12-04

* 通讯作者 Corresponding author. E-mail: zhengyr@ibcas.ac.cn

度,也即不能在统计意义上说明模型模拟结果的准确性,这无疑会增加使用模型进行预测与管理的不确定性,甚至会导致产生错误的结论。因此,在统计意义上判别模拟结果是否与观察值相符是一个亟待解决的问题。本文基于统计学检验两种回归直线是否来自同一总体,能否合并的原理提出了检验模型模拟结果与观察值是否在统计学上一致的方法,为生态学乃至所有模型的检验提供了一种简单易行的方法。

1 方法

统计学上,可以检验两种直线是否来自同一总体,即,通过检验直线回归方程(1)、(2)之间的斜率 b_1 , b_2 和截距 a_1 , a_2 是否存在显著差异,判断它们是否来自一个总体^[8]。

$$\hat{Y} = a_1 + b_1 X \quad (1)$$

$$\hat{Y} = a_2 + b_2 X \quad (2)$$

根据上述原理,可以检验模拟数据与观察数据之间的回归方程 $\hat{Y} = a + bX$ 与由观察数据构成的 1:1 直线 $Y = X$ 中 a 与 0 及 b 与 1 之间是否存在显著差异。如果不存在显著差异,可以认为模拟值与观察值相比具有统计显著水平上的一致性,否则尽管观察值与模拟值之间存在显著的相关关系,模拟结果仍不够理想。

检验回归系数 b_1 和 b_2 有无显著性差异的方法为^[8]:

假设, $H_0: \beta_1 - \beta_2 = 0$

备择假设, $\beta_1 - \beta_2 \neq 0$

$$\text{检验统计量: } t_b = \frac{b_1 - b_2}{\sqrt{MS_e \left(\frac{1}{S_{X_1X_1}} + \frac{1}{S_{X_2X_2}} \right)}}, \quad df = n_1 - 2 + n_2 - 2$$

其中: b_1 , b_2 分别为两条回归直线的截距, n_1 , n_2 分别为用于建立直线回归方程(1)和(2)的观察数据的样本含量。

$S_{X_1X_1}$ 与 $S_{X_2X_2}$ 分别为回归直线方程(1)、(2)中 X 的校正平方和,以直线回归方程(1)为例, X 的校正平方和 S_{XX} 可以表示为:

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$$

其中: x_i 与 $\bar{x}$ 分别为用于建立直线回归方程的 x 的第 i 个值及 x 的算术平均数。

$$MS_e = \frac{(n_1 - 2)MS_{e_1} + (n_2 - 2)MS_{e_2}}{(n_1 - 2) + (n_2 - 2)}$$

MS_{e_1} 与 MS_{e_2} 分别为回归直线方程(1)、(2)的剩余均方,以直线回归方程(1)为例, MS_{e_1} 可以表示为:

$$MS_{e_1} = \frac{S_{Y_1Y_1} - b_1 S_{X_1Y_1}}{n_1 - 2}$$

其中, $S_{X_1Y_1}$ 为直线回归方程(1)中观察值 XY 的校正交叉乘积和,可以表示为:

$$S_{X_1Y_1} = \sum_{i=1}^{n_1} (x_i - \bar{x})(y_i - \bar{y})$$

其中: y_i 与 $\bar{y}$ 分别为用于建立直线回归方程(1)的 y 的第 i 个值及 y 的算术平均数。

$S_{Y_1Y_1}$ 为回归直线方程(1)中关于 Y 的总校正平方和,可以表示为:

$$S_{Y_1Y_1} = \sum_{i=1}^n (y_i - \bar{y})^2$$

通过上述计算,可以求得 t_b , 以 $n_1 - 2 + n_2 - 2$ 为自由度查 t 的双侧分布表,如果计算所得 t_b 值小于查得 t 值,则接受零假设,认为 $b_1 = b_2$, 否则认为 b_1 与 b_2 不同。

检验回归系数 a_1 和 a_2 有无显著性差异的方法为^[8]:

假设, $H_0: \alpha_1 - \alpha_2 = 0$
备择假设, $\alpha_1 - \alpha_2 \neq 0$

检验统计量: $t_a = \frac{a_1 - a_2}{\sqrt{MS_e(\frac{1}{n_1} + \frac{\bar{x}_1^2}{S_{X_1X_1} + S_{X_2X_2}} + \frac{1}{n_2} + \frac{\bar{x}_2^2}{S_{X_1X_1} + S_{X_2X_2}})}} , df = n_1 - 2 + n_2 - 2$

通过上述计算,可以求得 t_a ,以 n_1-2+n_2-2 为自由度查 t 的双侧分布表,如果计算所得 t_b 值小于查得 t 值,则接受零假设,认为 $a_1 = a_2$,否则认为 a_1 与 a_2 不同。

为说明此方法的具体使用过程,以虚拟的植被净第一性生产力观察值与模拟值为例说明此方法的具体计算过程,分别采用两个模型的模拟值与观察值进行分析(表 1)。尽管表 1 与图 1 有些重复,但为了便于理解此方法,二者均保留在文中。

2 结果

查 t 分布表得 $t_{10, 0.05(双侧)} = 2.228$ 。对于回归直线 2, t 值均大于 2.228,模型 2 的模拟值在统计意义上与观察值显著不同,而对于回归直线 1, t 值均小于 2.228,模型 1 的模拟值在统计意义上与观察值不存在显著区别(表 2)。也即模型 1 可以准确模拟观察值,而模型 2 模拟效果较差。

表 1 植被净第一性生产力观察值与模拟值 ($t\text{ hm}^{-2}\text{ a}^{-1}$)
Table 1 Observed and modeled vegetation net primary productivity ($t\text{ hm}^{-2}\text{ a}^{-1}$)

观察值 Observed value	模型 1 模拟值 Values by model 1	模型 2 模拟值 Values by model 2	观察值 Observed value	模型 1 模拟值 Values by model 1	模型 2 模拟值 Values by model 2
1.5	1.57	2.5	1.71	1.89	2.71
1.93	2.05	2.93	2.23	2.09	3.23
2.47	2.34	3.47	2.89	3.22	3.89
3.52	3.92	5.12			

表 2 模拟值与观察值回归直线与 1:1 直线的统计学参数
Table 2 Parameters of regression lines and 1:1 line

统计量 Statistic parameters	回归直线 1 Regression line 1	回归直线 2 Regression line 2	1:1 直线 1:1 line	统计量 Statistic parameters	回归直线 1 Regression line 1	回归直线 2 Regression line 2	1:1 直线 1:1 line
N	7	7	7	$\bar{x}$	2.32	2.32	2.32
$\bar{y}$	2.44	3.41	2.32	S_{XX}	2.99	2.99	2.99
S_{YY}	4.14	4.74	2.99	S_{XY}	3.44	3.71	2.99
b	1.15	1.24	1.00	a	-0.23	0.53	0.00
MS_e	0.04	0.03	0.00	t_b	1.32	2.52	
t_a	-1.14	3.14					

3 讨论

在一些生态学模型中,仅比较模拟值与观察值随时间的变化,通过变化趋势是否一致,确定模型模拟结果的准确性,有时在某一时间点,观察值与模拟值差别较大^[1, 3]。有的模型比较不同模型模拟值在观察值 1:1 直线两侧的分布,但当多个模型的模拟值分布趋势一致时,难以判断模型的模拟能力^[4]。有的模型比较观察值与模拟值回归直线和观察值 1:1 直线的接近程度,但仍属于定性比较^[6]。比较不同模型模拟结果的均方根误差、绝对误差也可定量比较不同模型的模拟能力^[7]。由于本文主要为方法介绍,采用的两组虚拟模拟数据差别较大,因而,通过上述方法也可以判断模型 1 的模拟效果要好于模型 2,但不能说明模拟值与观察值的相

符程度是否达到统计显著水平,也即模拟结果是否在统计水平上可信。

其次,由图 1 可见,回归直线 1 与 2 的相关系数均达到统计显著水平,有时错误地认为两个模型的模拟值均与观察值相符较好,但采用上述方法^[6]也可发现回归直线 2 偏离了 1:1 直线,而回归直线 1 与 1:1 直线较为接近。通过本文提出的检验方法可以简单而有效地检验模型对观察值的模拟是否达到了统计显著水平,能够更加客观地评估模型的优劣,建议在生态学模型验证时采用这一方法。但由于生态过程太过复杂,目前的研究深度仍很有限,许多生态学模型的模拟结果仍与观察值有较大差异,采用本文的方法评估模型模拟结果的准确性可能会得出许多模型不能真实模拟自然过程的结果,有待于对生态过程更为深入的理解,进而开发模拟能力更强的模型。

参考文献 (References):

- [1] Burns T P, Rose K A, Brenkert K A. Quantifying direct and indirect effects of perturbations using model ecosystems. *Ecological Modelling*, 2014 (published on line). DOI: 10.1016/j.ecolmodel.2013.12.017.
- [2] Black K, Davis P, Lynch P, Jones M, McGettigan M, Osborne B. Long-term trends in solar irradiance in Ireland and their potential effects on gross primary productivity. *Agricultural and Forest Meteorology*, 2006, 141(2/4): 118-132.
- [3] Bullied W J, Flerchinger G N, Bullock P R, Van Acker R C. Process-based modeling of temperature and water profiles in the seedling recruitment zone: Part I. Model validation. *Agricultural and Forest Meteorology*, 2014, 188: 89-103.
- [4] Tayfur G, Zucco G, Brocca L, Moramarco T. Coupling soil moisture and precipitation observations for predicting hourly runoff at small catchment scale. *Journal of Hydrology*, 2014, 510: 363-371.
- [5] Snell R S, Huth A, Nabel J E M S, Bocedi G, Travis J M J, Gravel D, Bugmann H, Gutiérrez A G, Hickler T, Higgins S I, Reineking B, Scherstjanoi M, Zurbriegen N, Lischke H. Using dynamic vegetation models to simulate plant range shifts. *Ecography*, 2014, 37: 1-14.
- [6] Qiu G Y, Yano T, Momii K. An improved methodology to measure evaporation from bare soil based on comparison of surface temperature with a dry soil surface. *Journal of Hydrology*, 1998, 210: 93-105.
- [7] Zhu L H, Zhao X C, Lai L M, Wang J J, Jiang L H, Ding J Z, Liu N X, Yu Y J, Li J S, Xiao N W, Zheng Y R, Rimmington G M. Soil TPH concentration estimation using vegetation indices in an oil polluted area of eastern China. *PLoS ONE*, 2013, 8(1): e54028.
- [8] 杜荣骞. 生物统计学. 北京: 高等教育出版社, 海德堡: 施普林格出版社, 1999.

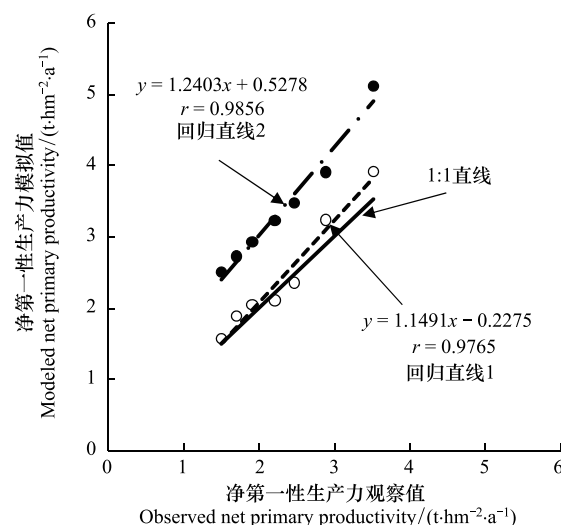


图 1 回归直线与 1:1 直线的关系

Fig.1 Relation between 1:1 line and regression lines