# 昆虫区系多元相似性分析方法

申效诚[1,2],孙 浩[1],赵华东[1]
(1. 郑州大学生物工程系 郑州 450001;2. 河南省农科院植保所 郑州 450002)

**摘要**:由植物学领域首先提出的相似性概念已广泛应用于动植物区系地理研究以及生物学、生态学等诸多自然学科乃至社会科学领域。根据 Jacard 提出的二元相似性系数公式 $SI = C/(A + B - C)$ 和 Sprensen 提出的二元相似性系数公式 $SI = 2C/(A + B)$,分别推导出 2 个计算多元相似性系数的数学表达式,$SIJ_{ab\cdots n} = [(\sum H_{ij})2/n + (\sum H_{ijk})3/n + \cdots + H_{ab\cdots n}]/[\sum S_i - \sum H_{ij} - 2\sum H_{ijk} - \cdots - (n-1)H_{ab\cdots n}]$ 和 $SIS_{ab\cdots n} = [2(\sum H_{ij})2/n + (3\sum H_{ijk})3/n + \cdots + nH_{ab\cdots n}]/\sum S_i$,并用中国夜蛾广布种类在中国 7 个动物地理区的分布资料为例进行验证,从而可以直接从整体角度和宏观规模上简便、快捷地考量多个系统间的亲疏程度和相似关系。建议在以相似性为基础的聚类分析中,不必再先把 2 个系统合并成一个新系统后,再和第 3 个系统比较,而可直接计算多个系统的相似性系数,以避免由于合并带来的信息损失。还讨论了应该提高 Sprensen 公式 0.5 的显著性标准,以使同一组数据的两种计算结果趋向一致。

**关键词**:相似性;多元;公式;昆虫区系
**文章编号**:1000-0933(2008)02-0849-06 **中图分类号**:Q141,Q968 **文献标识码**:A

# A discussion about the method for multivariate similarity analysis of fauna

SHEN Xiao-Cheng[1,2], SUN Hao[1], ZHAO Hua-Dong[1]
1 *Department of Biological Enginering, Zhengzhou University, Zhengzhou* 450001, *China*
2 *Institute of Plant Protection, Henan Academy of Agricultural Science, Zhengzhou* 450002, *China*
***Acta Ecologica Sinica*, 2008, 28(2): 0849 ~ 0854.**

**Abstract**: The concept of similarity, originated from studies in botany, but has been widely applied in plant and animal geology, biology, ecology and even sociology. Based on the two bivariate analysis formulas $SI = C/(A + B - C)$ and $SI = 2C/(A + B)$, two expression equations, $SIJ_{ab\cdots n} = [(\sum H_{ij})2/n + (\sum H_{ijk})3/n + \cdots + H_{ab\cdots n}]/[\sum S_i - \sum H_{ij} - 2\sum H_{ijk} - \cdots - (n-1)H_{ab\cdots n}]$ and $SIS_{ab\cdots n} = [2(\sum H_{ij})2/n + (3\sum H_{ijk})3/n + \cdots + nH_{ab\cdots n}]/\sum S_i$ were developed to describe the coefficient of similarity for multivariate data. These two equations were validated using population distribution data of the common Noctuidae species from seven animal geological areas in China. By using these two equations, the species similarity and the relationship distance in the whole system overall can be efficiently measured. We suggest how to appropriately use these two equations to enhance the significant standard to 0.5 in Spensen's formula, so that a consistent result for the same set of data can be obtained from both formulas.

---

**基金项目**:河南省自然科学基金项目(411031700)
**收稿日期**:2006-11-14; **修订日期**:2007-04-29
**作者简介**:申效诚(1943 ~ ),男,河南民权人,硕士,教授,主要从事昆虫区系、昆虫生态研究. E-mail: shenxiaoc@126.com
**致谢**:中国科学院地理科学与资源研究所张镱锂研究员赠送部分资料,美国俄克拉荷马州立大学 K. Giles 博士和赵白鸽女士对写作给予帮助,特此致谢
**Foundation item**: The project was financially supported by the Natural Science Foundation of Henan Province (No. 411031700)
**Received date**: 2006-11-14; **Accepted date**: 2007-04-29
**Biography**: SHEN Xiao-Cheng, Professor, mainly engaged in insect fauna and insect ecology. E-mail: shenxiaoc@126.com

**Key Words**: similarity; multivariate; formula; fauna

由植物学领域首先提出的相似性概念,经过一个半世纪的发展,已广泛应用到多个学科和多个领域[1-10],科学家们设计了十几种乃至数十种相似性系数(similarity coefficient)的计算方法[11]。特别是 Jacard 提出的相似性系数公式 $SI = C/(A + B - C)$ [12]和 Sprensen 提出的公式 $SI = 2C/(A + B)$ [13],已成为两系统间诸多相似性计算中最基础、最常用、最直观的方法。但在实际运用中,当涉及到多个系统间的比较时,一般都是根据两两系统间的相似性系数的大小,逐个进行合并聚类,而没有一个直接计算多个系统间相似性系数的方法。本文试图提出多元相似性系数的计算公式,以供有兴趣者试用和讨论。

## 1 多元相似性的概念

在动植物区系研究中,常会出现多个地区区系特征的比较。多元相似性是多个系统间总体的相似性或相异性,它用多元相似性系数进行定量表达,多元相似性系数是要比较的系统间共有种类(common species)占所有种类的比例。超过 0.5,为显著水平。

和 Jacard 及 Sprensen 的二元相似性系数公式相比,要计算多个系统之间的相似性系数,主要是作为公式分子的共有种类数的确定,所有系统的共有种类,作用自然重要,而部分系统共有的种类,也应该在分子中占有一定分量,但权重应小于所有系统的共有种类,而且随着比较系统的增多,这些部分系统共有种类的层次也相应增多,其权重也应随之不同。

如比较华北、华中、西南 3 区的夜蛾科昆虫的属数,华北有 312 属,华中 435 属,西南 457 属,各属在这 3 区的分布型如表 1。

**表 1 华北、华中、西南 3 区夜蛾属的分布型**

**Table 1 The distribution pattern of genus of Noctuidae in North-China, Central-China and South-Weat regions**

| 分布型 Distribution pattern | | 属数 Genus | 华北区(NC) North China | 华中区(CC) Central China | 西南区(SW) South West |
|---|---|---|---|---|---|
| 单区型 Single region form | 华北区 NC | 40 | 40 | | |
| | 华中区 CC | 75 | | 75 | |
| | 西南区 SW | 114 | | | 114 |
| 二区型 Double regions form | 华北区 + 华中区 NC + CC | 37 | 37 | 37 | |
| | 华北区 + 西南区 NC + SW | 20 | 20 | | 20 |
| | 华中区 + 西南区 CC + SW | 108 | | 108 | 108 |
| 三区型 Three regions form | 华北区 + 华中区 + 西南区 NC + CC + SW | 215 | 215 | 215 | 215 |
| 合计 Total | | 609 | 312 | 435 | 457 |

按表 1 数据,使用 Jacard 的 $SIJ = C/(A + B - C)$ 公式,华北区($A$)和华中区($B$)的相似性系数达显著水平,华中区和西南区(C)也达到显著水平,华北区和西南区达不到显著水平:

$$SIJ_{AB} = (37 + 215)/(312 + 435 - 37 - 215) = 252/495 = 0.509^{*}$$

$$SIJ_{BC} = (108 + 215)/(435 + 457 - 108 - 215) = 323/569 = 0.568^{*}$$

$$SIJ_{AC} = (20 + 215)/(312 + 457 - 20 - 215) = 235/534 = 0.440$$

计算三者的相似性系数,直接用“属数”列的数据即可:

$$SIJ_{ABC} = [(37 + 20 + 108)2/3 + 215]/609 = 325/609 = 0.534^{*}$$

由此可以认为在华北、华中、西南 3 区中,夜蛾科昆虫在属级水平上存在显著的相似性。这 3 区互相毗连,有相同和相似的生态条件,具有显著的相似性,符合生物学逻辑,也符合统计学逻辑。

如果把相似性系数较大的华中和西南区合并聚类(cluster),再和华北区比较,其相似性系数为:

$$SIJ_{(BC)A} = (37 + 20 + 215)/609 = 0.447$$

显然,这样合并聚类,掩盖、改变了某些份量的作用。只有将所有能合并的方式都计算出来,再取其算术

平均数,才能得到 $SIJ_{ABC}$式的结果:

$$SIJ_{ABC} = (SIJ_{(AB)C} + SIJ_{(AC)B} + SIJ_{(BC)A})/3 = (0.563 + 0.591 + 0.447)/3 = 0.534^{*}$$

同样,使用 Sprensen 的 $SIS = 2C/(A+B)$,公式,也得到上面同样趋势的计算结果,而且合并后的信息损失量更大:

$$SIS_{AB} = 2(37 + 215)/(312 + 435) = 0.675^{*}$$

$$SIS_{BC} = 2(108 + 215)/(435 + 457) = 0.724^{*}$$

$$SIS_{AC} = 2(20 + 215)/(312 + 457) = 0.611^{*}$$

$$SIS_{ABC} = [2(37 + 20 + 108)2/3 + 3 \times 215]/(312 + 435 + 457) = 0.718^{*}$$

$$SIS_{(BC)A} = 2(37 + 20 + 215)/(312 + 559) = 0.625^{*}$$

所以,无论 $SIJ_{ABC}$式或 $SIS_{ABC}$式,都公平地、充分地显示了各个共有种类分量的作用,避免了因聚类合并而造成的信息偏差。

## 2　多元相似性系数公式

按 Jacard 的公式,在 $N$ 个系统间比较,其共有种类数有 $N-1$ 个层次:

每 2 个系统间的相同种类,其和为 $\sum H_{ij}$,其权重为 $2/n$

每 3 个系统间的相同种类,其和为 $\sum H_{ijk}$,其权重为 $3/n$

…

$N$ 个系统间的相同种类为 $H_{ab\cdots n}$, 其权重为 $n/n = 1$

$N$ 个系统之间相似性系数公式为:

$$SIJ_{ab\cdots n} = [(\sum H_{ij})2/n + (\sum H_{ijk})3/n + \cdots + H_{ab\cdots n}]/[\sum S_i - \sum H_{ij} - 2\sum H_{ijk} - \cdots - (n-1)H_{ab\cdots n}] \quad (\text{I})$$

式中,$SI$ 为相似性系数,$H$ 为共有种类数,$S$ 为某系统的总种类数。

按 Sprensen 的公式,$N$ 个系统之间相似性系数公式似乎更简单:

$$SIS_{ab\cdots n} = [2(\sum H_{ij})2/n + 3(\sum H_{ijk})3/n + \cdots + nH_{ab\cdots n}]/\sum S_i \quad (\text{II})$$

## 3　多元相似性分析的应用

根据本文第一作者的研究资料,中国有夜蛾 3751 种,广泛分布在中国 7 个动物地理区(zoogeographical region)内,任何两区之间的相似性都是不显著的。但夜蛾有 4 种区系成分,其中地跨古北、东洋两界的广布种类(Eurytopic species)共 85 种,它们理应广泛地分布在全国各个大区,但由于各地调查深入程度不同,目前分布状况如表 2。以西南区 South-West Region(SW)、华北区 North China Region(NC)、华中区 Central China Region(CC)最丰富,蒙新区 Mongolia-Xinjiang Region(MX)、东北区 North-East Region(NE)、青藏区 Qinghai-Xizang Region(QX)居中,华南区 South China Region(SC)最少。用 Jacard 公式进行二元比较分析(表 3),在 21 个相似性系数中,达到显著水平的有 10 个,青藏区和任何区都不密切,蒙新区、华南区分别和 2 个区达到显著水平,最高的华中区、华北区、西南区、东北区分别和 4 个区有显著相似性。按相似性大小进行聚类分析,华中、华北、西南、东北、青藏区先后进入显著性水平,蒙新区、华南区游离在外。

这些种类中,每两区之间共有种类共 11 种,每 3 区之间共有种类共 8 种,每 4 区共有为 10 种,5 区为 25 种,6 区为 11 种,全国分布 10 种。七大区间的相似性系数为:

$$SIJ_{1234567} = (11 \times 2/7 + 8 \times 3/7 + 10 \times 4/7 + 25 \times 5/7 + 11 \times 6/7 + 10)/85 = 49.57/85 = 0.5832^{*}$$

$$SIS_{1234567} = (11 \times 2 \times 2/7 + 8 \times 3 \times 3/7 + 10 \times 4 \times 4/7 + 25 \times 5 \times 5/7 + 11 \times 6 \times 6/7 + 10 \times 7)/(44 + 52 + 59 + 42 + 58 + 64 + 38) = 0.7150^{*}$$

表 2 中国夜蛾广布种类的分布型

Table 2 The distribution patterns of Eurytopic species of Noctuidae in China

| 分布型 Distribution patterns | 种类数 Species | 东北 NE | 蒙新 MX | 华北 NC | 青藏 QX | 华中 CC | 西南 SW | 华南 SC |
|---|---|---|---|---|---|---|---|---|
| 单区型 Single region form | | | | | | | | |
| 蒙新(MX) | 5 | | 5 | | | | | |
| 华北(NC) | 1 | | | 1 | | | | |
| 青藏(QX) | 2 | | | | 2 | | | |
| 西南(SW) | 2 | | | | | | 2 | |
| 双区型 Double regions form | | | | | | | | |
| 蒙新+青藏 MX+ QX | 3 | | 3 | | 3 | | | |
| 蒙新+西南 MX+ SW | 1 | | 1 | | | | 1 | |
| 华北+华中 NC+CC | 1 | | | 1 | | 1 | | |
| 青藏+西南 QX+SW | 3 | | | | 3 | | 3 | |
| 华中+西南 CC+ SW | 2 | | | | | 2 | 2 | |
| 西南+华南 SW+ SC | 1 | | | | | | 1 | 1 |
| 三区型 Three regions form | | | | | | | | |
| 东北+蒙新+青藏 NE+ MX+QX | 1 | 1 | 1 | | 1 | | | |
| 东北+华北+西南 NE+ NC+ SW | 1 | 1 | | 1 | | | 1 | |
| 蒙新+青藏+西南 MX+ QX+ SW | 1 | | 1 | | 1 | | 1 | |
| 华北+华中+西南 NC+ CC+ SW | 1 | | | 1 | | 1 | 1 | |
| 华北+华中+华南 NC+ CC+SC | 1 | | | 1 | | 1 | | 1 |
| 华中+西南+华南 CC+ SW+SC | 3 | | | | | 3 | 3 | 3 |
| 四区型 Four regions form | | | | | | | | |
| 东北+蒙新+华北+青藏 NE+ MX+ NC+ QX | 2 | 2 | 2 | 2 | 2 | | | |
| 东北+蒙新+华北+西南 NE+ MX+ NC+ SW | 1 | 1 | 1 | 1 | | | 1 | |
| 东北+华北+华中+华南 NE+ NC+ CC+SC | 1 | 1 | | 1 | | 1 | | 1 |
| 蒙新+华北+青藏+西南 MX+ NC+ QX+SW | 1 | | 1 | 1 | 1 | | 1 | |
| 蒙新+华北+华中+华南 MX+ NC+ CC+SC | 1 | | 1 | 1 | | 1 | | 1 |
| 蒙新+华中+西南+华南 MX+ CC+ SW+SC | 1 | | 1 | | | 1 | 1 | 1 |
| 华北+华中+西南+华南 NC+ CC+ SW+ SC | 2 | | | 2 | | 2 | 2 | 2 |
| 青藏+华中+西南+华南 QX+ CC+ SW+ SC | 1 | | | | 1 | 1 | 1 | 1 |
| 五区型 Five regions form | | | | | | | | |
| 东北+蒙新+华北+青藏+华中 NE+ MX+ NC+ QX+ CC | 2 | 2 | 2 | 2 | 2 | 2 | | |
| 东北+蒙新+华北+青藏+西南 NE+ MX+ NC+ QX+ SW | 2 | 2 | 2 | 2 | 2 | | 2 | |
| 东北+蒙新+华北+华中+西南 NE+ MX+ NC+ CC+ SW | 7 | 7 | 7 | 7 | | 7 | 7 | |
| 东北+华北+青藏+华中+西南 NE+ NC+ QX+ CC+ SW | 1 | 1 | | 1 | 1 | 1 | 1 | |
| 东北+华北+华中+西南+华南 NE+ NC+ CC+ SW+ SC | 5 | 5 | | 5 | | 5 | 5 | 5 |
| 蒙新+华北+青藏+华中+西南 MX+ NC+ QX+ CC+ SW | 2 | | 2 | 2 | 2 | 2 | 2 | |
| 蒙新+华北+华中+西南+华南 MX+ NC+ CC+ SW+ SC | 3 | | 3 | 3 | | 3 | 3 | 3 |
| 华北+青藏+华中+西南+华南 NC+ QX+ CC+ SW+ SC | 3 | | | 3 | 3 | 3 | 3 | 3 |
| 六区型 Six regions form | | | | | | | | |
| 除蒙新 Except MX | 2 | 2 | | 2 | 2 | 2 | 2 | 2 |
| 除青藏 Except QX | 3 | 3 | 3 | 3 | | 3 | 3 | 3 |
| 除西南 Except SW | 1 | 1 | 1 | 1 | 1 | 1 | | 1 |
| 除华南 Except SC | 5 | 5 | 5 | 5 | 5 | 5 | 5 | |
| 全国广布型 Nationwide distributed | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 合计 Total | 85 | 44 | 52 | 59 | 42 | 58 | 64 | 38 |

表 3　广布种类在各区之间的共有种类和相似性系数

Table 3　The common species and similarity coefficient between seven zoogeographical regions of Eurytopic species

| 动物地理区 Zoogeographical region | 东北区 NE 44 | 蒙新区 MX 52 | 华北区 NC 59 | 青藏区 QX 42 | 华中区 CC 58 | 西南区 SW 64 | 华南区 SC 38 |
|---|---|---|---|---|---|---|---|
| 1 东北区 NE 44 | | 34 | 43 | 26 | 37 | 37 | 22 |
| 2 蒙新区 MX 52 | 0.548 | | 40 | 30 | 35 | 37 | 19 |
| 3 华北区 NC 59 | 0.717 | 0.563 | | 31 | 51 | 49 | 32 |
| 4 青藏区 QX 42 | 0.433 | 0.469 | 0.443 | | 27 | 31 | 17 |
| 5 华中区 CC 58 | 0.569 | 0.467 | 0.773 | 0.370 | | 51 | 37 |
| 6 西南区 SW 64 | 0.521 | 0.468 | 0.662 | 0.413 | 0.718 | | 34 |
| 7 华南区 SC 38 | 0.367 | 0.268 | 0.492 | 0.270 | 0.627 | 0.500 | |

这说明 80 多种广布种类在七大区之间的分布是广泛且均匀的。这一方面说明广布种类固有的分布特性,另方面说明各区都进行了相当程度的区系调查,体现出了广泛分布的特点,但还存在着一定的相异性,说明有进一步深入调查的探索空间。

## 4　讨论

**4.1**　相似性系数已广泛应用于动植物区系地理研究以及生物学、生态学等诸多学科研究中。本文根据 Jacard 和 Sprensen 的二元相似性系数公式,推导出多元相似性系数计算式,不仅能从宏观或整体角度考量多系统间的亲疏程度和相似关系,而且简便、快捷。可在昆虫区系分析实践中进一步试用。

**4.2**　当进行多系统相似性比较时,原来由于没有多元相似形系数计算方法,只能按二元相似形系数的高低顺序进行聚类,逐个比较,从而付出了信息偏差的代价。建议采取“类而不聚”的办法,可以把最高相似度的系统划作一类,但不合并为一个新的系统,而是分别采用三元、四元、五元等相似性系数计算并分类,使所有共有种类的信息得到充分表达。

**4.3**　由于 Sprensen 公式比 Jacard 公式更强化共有种类的作用,对同一组数据的计算结果,*SIS* 要高于 *SIJ*。同样,本文提出的公式 II 的结果也要高于公式 I。两个公式采用同一个标准(0.5)来认定显著性,显然缺乏准确性。建议通过实践验证,提高 Sprensen 公式(或本文公式 II)的显著性标准,以使二者结果不必经过换算,就能得到较为一致的认定。

**Reference:**


[1] Huang X L, Feng L, Qiao G X. Similarity analysis and historical origin on aphid fauna in Taiwan and mainland of China. Acta Zootaxonomica Sinica, 2004,29(2):194－201.

[2] Lu J Q. Clustering analysis on the geographical distribution of glires in Henan Province. Chinese Journal of Ecology, 2000,19(4):43－45.

[3] Ding S Y, Lu X L. Comparison of plant flora of Funiu mountain and Jigong mountain natural reserves. Geographical Research, 2006,25(1):62－70.

[4] Ge Y, Yu M, Liu Q R. Pteridophyte flora in the area under the jurisdiction of Beijing. Acta Bot. Boreal-Occident Sin., 2006,26(8):1657－1662.

[5] Zuo J F, Fu D Z. Quantitative study on seed pland flora of China Ⅴ. Flotistic similarity. Journal of Tropical and subtropical Botany, 1996,4(3):18－25.

[6] Yan L H, Qi C J, Liu X X. The essential characteristics of vine flora in the subtropical zone of Central China. Journal of Central South Forestry University, 2006,26(4):36－41.

[7] Liu C M, Lian Z M. Quantitative classification and similarity coefficients of grasshopper community on the southern slope of Taibai mountain of Qinling. Journal of Northwest Forestry University, 2004,19(1):85－88.

[8] Tang Y. A survey of approaches for time series similarity analysis. Computer Engineering and Applications,2006,(1):20－26.

[9] Wang H C, Dong Z C, Liang Z M, *et al*. A study on the index system for storm-flood resemblance analysis. Journal of China Hydrology, 2006,

(2):4 －10.

[10] Zhang S G, Cui X Y. Analysis of similarity of Hang Seng Index to Shanghai Index. Operations Research and Management Science, 2006,(5):22 －29.

[11] Zhang Y L. The coefficient of similarity, a important parameter in studies on flora. Deographical Research, 1998, 17(4):429 －434.

[12] Jaccard P. Distribution de la flore alpine dans le Bassin des Dranses et dams quelque region vasines. Bull. Soc. Vaud. Sci. Nat., 1901, 37:241 －272.

[13] Sprensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons. Biol. Skr., 1948, 5(4):1 －34.

**参考文献:**

[1] 黄晓磊,冯磊,乔格侠. 台湾与大陆蚜虫区系的相似性分析和历史渊源. 动物分类学报,2004,29(2):194 ~ 201.

[2] 路纪琪. 河南省啮齿动物地理分布的聚类分析. 生态学杂志,2000,19(4):43 ~45.

[3] 丁圣彦, 卢训令. 伏牛山和鸡公山自然保护区植物区系比较. 地理研究,2006,25(1):62 ~70.

[4] 葛源, 于明, 刘全儒. 北京地区蕨类植物区系分析. 西北植物学报,2006,26(8):1657 ~1662.

[5] 左家哺, 傅德志. 中国种子植物区系定量化研究 Ⅴ. 区系相似性. 热带亚热带植物学报,1996,4(3):18 ~25.

[6] 颜立红, 祁承经, 刘小雄. 中国亚热带中部藤本植物区系的基本特点. 中南林学院学报,2006,26(4):36 ~41.

[7] 刘缠民, 廉振民. 太白山南坡蝗虫群落数量分类及相似性分析. 西北林学院学报,2004,19(1):85 ~88.

[8] 汤胤. 时间序列相似性分析方法研究. 计算机工程与应用,2006,(1):20 ~26.

[9] 王海潮, 董增川, 梁忠民, 等. 暴雨洪水相似性分析指标体系研究. 水文,2006,(2):4 ~10.

[10] 张曙光, 崔翔宇. 恒生指数和上证指数相似性分析. 运筹与管理,2006,(5):22 ~29.

[11] 张镱锂,1998. 植物区系地理研究中的重要参数——相似性系数. 地理研究,17(4):429 ~434.