生态学试验设计与解释中的常见问题

牛海山[1,2], 崔骁勇[1,2], 汪诗平[1,*], 王艳芬[2]

(1. 中国科学院西北高原生物研究所高原生物适应与进化重点实验室,青海省西宁市西关大街 59 号 810008;
2. 中国科学院研究生院,北京市石景山区玉泉路甲 19 号 100049)

摘要:一个“好”的试验,统计检验的显著差异可证明处理效应存在;而一个“坏”的试验,统计检验的显著差异本身并不能证明处理效应的真实存在。试验单位的不独立就可能使干扰因素偏倚地影响试验结果,是多种形式的伪重复的根本原因。如果不统筹地考虑试验方法和数据分析方法,环境要素的时空格局就可能被错误地当成处理效应。以案例分析的形式探讨了中国生态学试验设计与解释中常见的 3 个问题:(1)简单伪重复问题,这是 Hurlbert[2] 早已指出的伪重复情形中的一种,通常是把取样的重复当成了处理的重复;(2)把反复测量结果当成重复的问题,即对同一个对象的反应变量前后进行多次观测,却把这些观测值视为重复而进行统计检验所造成的问题;(3)混淆时空效应与处理效应的问题,由于取样方法(破坏性取样)或者研究对象(例如流动的水体)性质的特殊性等原因,数据中所体现出来的格局有可能由于时空效应而并非处理效应所造成。在这 3 种情况下,数据的产生方式与所用统计方法的前提相违背。

关键词:试验设计;生态学试验;简单伪重复;反复测量;误差估计;重复;时空变异

文章编号:1000-0933(2009)07-3901-10 **中图分类号**:Q143 **文献标识码**:A

Three common errors in Chinese ecological experimental design and interpretation

NIU Hai-Shan[1,2], CUI Xiao-Yong[1,2], WANG Shi-Ping[1,*], WANG Yan-Fen[2]

1 *Northwest Institute of Plateau Biology, CAS, Xining* 810008, *China*

2 *Graduate University of CAS, Beijing* 100049, *China*

***Acta Ecologica Sinica*, 2009, 29(7): 3901 ~ 3910.**

Abstract: The logic of experiments is discussed, and three common mistakes in designing and interpreting experiments among ecologists are analyzed. Researchers sometimes overlook logic problems, while paying too much attention to statistical techniques. A statistically significant difference can demonstrate treatment effect for a well-designed experiment, but it might not for a “bad” experiment. Any extraneous variables that change systematically along with the treatment threaten the internal validity of an experiment. The bias in some cases arises from interdependence between experimental units, which is the case with pseudo-replication, a term coined by S. Hurlbert in 1984. In this paper we review experiments published in one issue each of two of the most influential ecology journals, Journal of Plant Ecology and Acta Ecologica Sinica. Some cases are selected and analyzed to exemplify three mistakes which are commonly made by Chinese ecologists when designing an experiment or interpreting data. They are: (1) simple pseudo-replication, (2) confusing repeated measurements as replicates or temporal pseudo-replication in terms of Hurlbert, and (3) wrongly treating a spatial or temporal pattern as a treatment effect. Among the experiments reviewed, 17.9%－42.9% in Journal of Plant Ecology and 14.3%－42.9% in Acta Ecologica Sinica involved at least one type of the errors mentioned above.

基金项目:中国科学院西部行动计划资助项目(KZCX2-XB2-06-01);中国科学院百人计划择优支持项目

收稿日期:2008-03-26; **修订日期**:2008-09-08

* 通讯作者 Corresponding author. E-mail: wangship2008@yahoo.cn

Key Words: experimental ddesign; ecology experiment; simple pseudoreplication; repeated measurement; error estimation; replicate; spatial and temporal heterogeneity

生态学越来越成为一门试验科学,利用试验来检验假说和发现新问题在生态学研究中愈来愈普遍。虽然在试验完成之后才进行数据统计分析,但是往往在试验之前就要依据统计学的原则对试验单位的类型、数量、空间布局、重复数以及处理施加的顺序和方式等进行设计,以使得试验的逻辑结构和操作过程满足统计学的基本假设,并争取试验取得最大可能的统计功效(statistical power)①;这是现代试验设计的特征。现代试验设计由 Ronald A. Fisher 在洛桑农业试验站(Rothamsted Experimental Station)期间(20 世纪 20 ~ 30 年代)所开创,此后逐渐在工业生产质量控制等领域得到应用[1]。现代试验设计是以数理统计为基础的,但是有些时候试验者过分注重统计方法本身而忽视了试验的基本逻辑,设计出了即使统计检验显著也不能证明处理效应的试验。事实上,数理统计与试验设计中的基本逻辑不可能相互矛盾:凡是违反了逻辑的试验设计也一定违反了统计检验的基本前提。基本逻辑与统计检验的基本原则在道理上是相通的。

试验的目标是检验"处理(treatment)"对研究目标是否存在"效应(effects)";做法是,在研究目标的诸多影响因素之中有目的地改变处理因素,然后观测研究目标的反应变量如何改变[1]。然而,许多不受控的因素也可以造成反应变量的改变,不妨称之为干扰因素。因此,试验也是把处理因素的效应与干扰因素的效应区分开的过程。没有重复的试验无法区分处理因素和干扰因素的效应。譬如从研究目标中任选出两个试验单位(即可以单独施加不同处理的最小单位[2]),一个施加处理(T),另一个施加对照(CK),然后测定两者的某一指标(可以反映处理效应的变量),结果 T 与 CK 不同,于是说处理导致了 T 与 CK 不同。这样的试验没有说服力,因为除了处理因素不同而外 T 与 CK 接受到的其它因素影响也一定是不同的,在处理之前 T 与 CK 就不可能相同,所以 T 与 CK 的差异无法排它地解释为处理效应。在这种没有重复的设计中,如果试验者进行了统计检验就犯了简单伪重复的错误[2]。这种错误的通常情形是把每个试验单位内的若干抽样当成重复。如果试验有重复,但是(已知或未知的)干扰因素不是以随机的方式影响各处理,而是在受处理的试验单位(A_1, A_2, A_3, …, A_n)与对照的试验单位(B_1, B_2, B_3, …, B_m)之间还存在着(处理因素而外的)系统性的差异,这样的设计也一定是错误的。使干扰因素随机地影响各处理的前提是各试验单位必须在统计学上相互独立。把反复(repeated measurements, 即对同一个对象的前后多次测量)当成重复(replication)就极有可能造成 A_1, A_2, A_3, …, A_n与 B_1, B_2, B_3, …, B_m之间(处理因素而外的)系统性差异。这是另一种形式的伪重复设计[2,9]。此外,生态梯度无处不在,如果这种生态梯度又不是试验所要研究的因素,那么把 A_1, A_2, A_3, …, A_n 设置在生态梯度的一端而把 B_1, B_2, B_3, …, B_m设置在梯度的另一端也会造成试验结果的系统性偏差。在这 3 种情况下,即使对试验数据进行统计检验得到组间差异显著的结果,也无法证明处理效应的存在。这些朴素的逻辑容易理解,但是在实践中却由于各种各样的原因而没有得到遵循。

生态学试验中的"伪重复(Pseudoreplication)"是由美国生物学家 Stuart H. Hurlbert 最早发现并定义的,它是指"进行了推断统计以检验处理效果,而数据所源自的试验没有重复——尽管可能存在多个抽样(samples)——或者重复不独立"的情况[2]。后来他又对这个定义进行了限定,指出了两种没有重复却不构成伪重复的情形[3]。Hurlbert 分析了 1960 ~ 1980 年发表于生态学杂志上的 176 个实验,发现在使用统计检验的文章之中有 48% 存在"伪重复",1984 年文章发表当年 Hurlbert 即因为提出"伪重复"的问题而获得美国统计学会 George W. Snedecor 奖②,并且获得美国自然科学基金会(NSF)的资助以支持他进一步发现生态学试验设计与统计中的其它错误[4]③。2003 年在美国科学院第 140 次年会上他获得"科学评论"奖(NAS Award for

① 统计功效的定义为:如果处理效果真实存在,通过试验检验出这种效果(即使统计检验达到显著水平)的可能性;也即不犯第二类错误的概率($1-\beta$);国内也有翻译成统计效能

② http://www.amstat.org/awards/index.cfm?fuseaction=copss-past

③ http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=8509535

Scientific Reviewing)①。试验设计中的伪重复问题在美国生态学界引起高度关注,根据 Thomson ISI 的统计,Hurlbert 的这篇文章到目前已经被引用 2000 多次[5],早已成为 ISI 的引用经典(citation classic)[4];Google Scholar 显示被引用 2192 次②。

当然这篇文章也引起了广泛争论,尤其以大尺度试验研究者反弹强烈,但是这些反对意见中的大多数与其说是在否认伪重复问题的存在,不如说是在强调生态学试验中实现真重复的困难——即"现实性"与"真重复"的矛盾;他们认为,生态学上的许多过程只有在相应的(大)尺度试验下才能得到体现,而大尺度的试验往往难以实现重复[6,7]。事实上,Hurlbert 并非简单地把没有真重复定义为伪重复,而是指没有重复(或者重复不独立)却错误地使用需要重复的统计检验[2]。在所有反驳性的文章中,最值得一提的是 Oksanen 的[8];这一方面因为它传播较广,另一方面也因为根据这篇文章许多生态学家认为 Hurlbert 的"伪重复"观点终于被推翻了[5]。Oksanen 认为[8],试验是用于检验假说的,而一个值得去检验的假说应该具有足够小的先验概率,而检验这样的假说根本不需要重复,例如如果某一假说精确地预测了一次日食就足以验证其正确;这个观点以前也有人提出过[6]。但是这种观点似乎忽视了一点:凡是具有足够小的先验概率的假说,其内容描述与检验都不会基于统计模型。Oksanen 的另一个重要观点是[8],没有重复的试验降低了花费,那么在总经费一定的情况下就意味着可以在更多的地点上进行试验,如果这些试验结果都得到发表,就意味着为 meta-分析提供更多的材料以便在更一般层次上验证假说。Hurlbert[3]认为他夸大了 meta-分析的效力,因为实际的 meta-分析中并不是每个数据都能与其它来源数据匹配,因而总效率不及在具体试验中设置重复。

无论如何,对"伪重复"问题的讨论促进了生态学试验设计的发展,伪重复的试验明显减少;Heffner 等对 Hurlbert 统计的杂志在 1986 ~ 1990 年之间发表的论文进行了分析,发现试验伪重复的比率已经降低到了 18%[9]。而且这一比例一直在下降[10]。

相比之下,非英语世界的生态学家对于试验设计方法问题关注得非常不够。例如,有人对俄罗斯生态学家进行了调查[11],发现只有 2 个研究组知道伪重复问题并且阅读过 Hurlbert 的文章,而其他 30 个被调查的研究小组从来没有听说过这个概念,而且在 1987 ~ 2001 年所发表的文章中均没有人提到过这篇文章。与此类似,迄今为止以中文发表的生态学文章中只有最近的一篇文章以正确的方式引用过这篇文章[12]。虽然少数研究者认真对待伪重复的问题,但是更大多数的中国生态学家还没有关注过这件事。这种状况反映了对试验设计的一贯忽视。忽视试验设计轻则降低试验功效、浪费人力物力,重则得到虚假试验结果。

本文对 3 种常见试验设计问题在中国生态学研究中出现的情况进行了调查,并且以具体案例的形式加以分析。它们是:①Hurlbert[2]在 1984 年就已经提出的简单伪重复(Simple Pseudoreplication)问题;②把反复测量(repeated measurements)结果视为重复的问题;③混淆空间变异与处理效应的问题。其中,第②个问题实质上可归入 Hurlbert 所定义的时间伪重复(temporal pseudoreplication),但是第②种问题的严重程度较低,有时可以通过改变数据分析方法得到校正。第③个问题是逻辑错误,但是以前没有被具体地指出过。这 3 个问题都不能视为单纯的统计方法误用——如果不看整个试验的流程而只考察试验"数据",则看上去都满足所使用的统计方法的要求,因此它们应该被看成是由于数据生产与数据分析相互脱节而产生的问题。关于统计方法的误用,已有许多专门文章予以讨论[13,14]。

为了把问题表述得更明确,把认为有问题的文章单独列出来,并具体指出问题所在,以使讨论言之有物。这种针对具体问题的讨论可起到警醒和具体化的作用。

需要说明的是,许多生态学的野外观测活动应该归入"试验",因为这些观测活动为了突出某一因素的效应,会有意识地保持其它因素不变或控制其变动范围,而只保留目标因素变化,通常也设置对照。与控制试验的区别仅仅在于这种试验的处理是自然界施加的。

① http://www.nasonline.org/site/PageServer?pagename=AWARDS_scirev

② http://scholar.google.com 搜索引擎截至 2007 年 7 月显示被引次数为 2192 次

本文中所使用的案例并非使用随机抽样或排查的方法得到,因笔者从事生态学工作,大多研究中用到的都是国内重要的生态学期刊,为方便起见只选用《植物生态学报》2007 年 31 卷第 2 期和《生态学报》2007 年 27 卷第 5 期为例加以分析,以引起所有试验设计者及数据分析者的重视。

在伪重复文章的统计中(表 1)并没有将下述 4 种情况计入在内:(1)文章对试验设计的描述不清楚,以致无法复原试验设计,因而难以判断处理是否有重复的;(2)试验单元面积较大或者涉及景观结构的研究,即使没有处理重复也不计入;(3)虽然处理没有重复,但是数据处理的过程中没有使用方差分析或者 t-检验的;(4)在 Hurlbert 图 1 中的 B-3,B-4 的“伪重复”情形①,虽然在我国已报道的试验设计中也常见,但是本文中也不计入。情况(2)和(4)之所以没有被计入伪重复设计,是因为如果进行真正的处理重复将会使试验的工作量或者经费支出增加许多,因而属于可以被“原谅”的错误。

对本文中所述及的案例,所讨论的仅仅是其试验设计和数据解释方法,而不是评判其研究意义。

1 简单伪重复(Simple pseudoreplication)

简单伪重复是 Hurlbert 指出的若干种伪重复情形中的最简单的一种[2],但是在中国生态学试验中也是最常见的一种。譬如[例 1]:为了研究某种施肥对土壤微生物生物量的影响而设置了两个试验小区,其中一个施加施肥处理而另一块施加对照处理;一段时间以后分别在两个样地各取样 n 个;认为这是有 n 个重复的试验而对两个区的调查结果进行统计检验。这个例子就是一个简单伪重复的试验设计。

这个试验设计在基本逻辑上存在漏洞:任何两个试验小区——即使在不进行任何处理的情况下——土壤微生物生物量就一定是不同的,而且只要取样量(n)足够大这种差异就一定能达到统计上显著的水平[2]。如果任何两个小区之间本来就是显著不同的,又怎么能说是处理导致了这两个小区的差异呢?

下面来证明只要取样量(n)足够大,无论看上去多么“极其相似”的两个小区,其差异就一定能达到统计上显著的水平[2]。因为:整个样地土壤微生物生物量变异程度(可用标准差 s 表示)不会随取样数(n)多少而改变,两个试验小区土壤微生物生物量之差别(记为 D)也不会随取样数(n)改变而改变;统计上检验 D 是否达到显著水平的方法是看其是否在置信区间($-d, +d$)之中,即判断 D 是否大于 d,如果 $D \geq d$ 则显著、反之则不显著,d 是使两个试验小区达到统计检验上显著差异所需要的最小差别;d 与取样量(n)的平方根成反比,数量关系[15]为 $d \approx 2 \cdot s/\sqrt{n}$(如果置信度设为 0.95);也就是说,随着取样量 n 变大,d 越来越小,总会有一时刻使得 $D \geq d$。因此如果这个试验没有得出差异显著的结果,那一定是因为取样量(n)还不够大(尚未使得 $D \geq d$)。

如果试验得到施肥小区与对照小区在统计上有显著差异的结果,那的确可以说明这两个试验小区的土壤微生物生物量是不同的,但是这种证明毫无用处——不做任何试验就知道这一点。试验的真正目的在于证明是施肥造成了两个小区土壤微生物量的差异(更准确地说是要去推翻一个“施肥对于土壤微生物量无影响”的假设),但是如例 1 这样的试验永远也不能达到这个目标。

从统计学的角度,这个试验违反了统计检验的基本前提条件。方差分析(或者 t-检验以及回归分析)中要求各数据点(试验单位)相互独立,这是一个不容违反的强假设,而其它假设(例如方差齐性与正态性)却是可以轻微违反的弱假设[29]。为什么各数据点必须相互独立呢? 这是因为每一个数据点的大小都除了受到处理因素的可能影响而外,还会受到干扰因素(空间变异、取样误差以及测定误差等)的作用,这些干扰因素的影响在现实中是无法去除掉的,但是试验者可以让每一种处理方式(或者同种处理的不同水平)以随机的方式施加到每一个试验单位上去,这样干扰因素对处理(或处理的不同水平)的影响是随机的——而不是偏倚的,这样统计检验就可以正确地区分处理效应与随机干扰。而在[例 1]中,如果处理小区所在的地块(不妨称为 A 区)土壤微生物生物量的本底值就高于对照地块(不妨称为 B 区),那么即使不施肥,$A_1, A_2, A_3, \cdots, A_n$ 就显

① 举例说明这两种伪重复的情形:如果试验只有处理(T)和对照(CK)两个处理水平,那么处理(T)的所有试验单位在一个设备(例如气室、培养箱、温室)、而对照(CK)的所有试验单位在另一个设备接受处理就是 B-3 类型伪重复;或者虽然表面上相互独立,但是事实上处理(T)的各试验单位间通过某种渠道联系起来(例如共享同一套供水系统或者可通过廊道相互连通起来),而对照的试验单位间也通过另一渠道相互联系在一起,这就构成了 B-4 类型的伪重复。

著高于 $B_1,B_2,B_3,\cdots,B_n$,这就使得干扰因素偏倚地影响了处理因素。相互独立是不同处理(或者同处理的不同水平)得以随机地施加到试验单位的前提条件。[例 1]中,在试验开始之前如果确定 A_1 被施加的是施肥处理,则 $A_2,A_3,\cdots,A_n$ 也注定施加的是施肥处理而不是对照处理,这些点在处理施加之前就是相互联系在一起的,并不是独立的。

正确的设计如(但是不限于)下面的例子。[例 2]:可以把上例中的地块(Field)划分出 6 个小区(plots),以随机的方式将施肥和对照处理分别施加到其中的 3 小区(plots),经过一段时间后,每小区采集 m 个样本(Samples)用以估计各自小区的微生物生物量平均值;以 3 为重复数进行统计检验。这就是一个具有真重复的试验设计。在这个设计中,每一个小区都能够等可能地接受到施肥或者对照的处理,每一个小区接受何种处理与其它小区没有关系,因此每个小区是相互独立的试验单位;因为它们是相互独立的,处理才可能以随机的方式施加于其上;因为处理是随机施加的,干扰因素对施肥与对照处理的影响是随机地,所以在统计检验中处理效应与随机误差才能够被正确地估计。在这样的试验设计中,干扰因素也会发生作用,甚至会纯粹由于干扰因素作用而错误地得到差异显著的结果(即犯 α 类错误),但是这种可能性——从长久而言——不高于 5%。此例中每个小区(试验单位)之内的 m 个取样位点是"抽样"单位;多个抽样的平均值才能更充分地估计该小区土壤微生物生物量。

对比[例 1]与[例 2],可以发现[例 1]中的重复(即 $A_1,A_2,A_3,\cdots,A_n$ 和 $B_1,B_2,B_3,\cdots,B_n$)是处理之下的抽样(Sampling)的重复,而[例 2]中的重复首先是处理(即施肥处理与对照处理)本身被重复(各 3 次),其次是小区内抽样的重复。抽样重复的目的在于估计每一个处理小区的"真值"。每个处理小区的若干抽样得到统计检验中的一个数据点。抽样越多则抽样越精确,则每一个数据点越接近相应小区的真实值。处理重复的目的是为了估计处理效应和随机误差。[例 1]只有两个小区,只能汇总成两个数据,而试验者却把抽样的重复把当成了处理的重复进行统计检验[2]。如果把[例 1]中的抽样单位看成是试验单位,那么这些"试验单位"互相之间并不独立,$A_1,A_2,A_3,\cdots,A_n$ 互相关联、$B_1,B_2,B_3,\cdots,B_n$ 互相关联。Hurlbert 举了一个非常精彩的例子来说明简单伪重复的问题所在[2]:为了比较某湖湖底两个深度下(1m 和 10m)枫树叶分解速度,用尼龙网袋装满枫叶,分别放置 8 袋于 1m 等深线、8 袋于 10m 等深线,1 个月后测定质量改变;如果出于某种原因,试验者将 8 袋树叶全部放置于 1m 等深线的某一位点,而另 8 袋也全部放置于 10m 等深线的某一位点处,如果试验者把每个尼龙袋作为一个试验单位进行统计检验,那么就犯了简单伪重复的错误。干扰因素对于每一位点上的 8 个尼龙袋影响是相同的,而不是随机的,因为它们之间是高度关联的。

试验单位的不独立也是其它形式伪重复的根源。再例如[例 3]:如果想研究植被覆盖类型对于表土侵蚀率的影响,把一片草地、一片农田和一片林地各自分成了 3 个小区,在每个小区内布设观测;观测结果以每种植被覆盖类型(处理因素)3 个重复进行统计检验。这也是一种伪重复的试验设计。因为每一个植被覆盖类型下的 3 个小区互相之间并不相互独立,不能算是 3 个处理的重复。换言之,这个设计中每一个小区不能等可能地接受到处理(即草地、农田、林地),也就无法正确估计非处理因素对侵蚀率的影响(即随机误差项)。正如有人戏言:并不是把一条狗称量 100 次就获得了 100 个重复①。

需要说明的一点是:因为处理的重复是为了正确估计随机误差项,所以对例 1 和例 3 的试验数据进行分析时,如果不使用统计检验就不构成"伪重复"的错误[2]。

除了例 3 中的情况以外,还多种情形可能造成试验单位的不独立,因而不能正确估计随机误差项。例如为了研究植物对某种气体浓度增加的响应,设计了盆栽试验,每一盆是一个试验单位,增加该气体浓度处理与对照各 3 盆作为重复;但是可能出于节约开支的考虑,试验者可能会把接受增加浓度处理(或者对照)的 3 个盆都置于(共用)同一个气室,这种情况就会造成重复的不独立,因而也会构成伪重复的错误。这就是 Hurlbert 文图 1 中列出的 B-3 和 B-4 的情况[2]。这个问题在本次调查中也有发现[16,17];但是由于上述原因,

① http://www.math.unb.ca/~knight/BasicStat/cheating.htm

这种情况并没有计入表 1 里面的简单伪重复。

也有两种特殊情况,没有处理重复但是随机误差项可以得到正确估计,这就不能算是伪重复。这两种情况是 Hurlbert 后来补充的[3]。一种情况是基于回归分析的设计,残差平方和可用于估计随机误差的影响;另一种情况是多因素的析因试验设计(又称因子设计,factorial design)下,每一个处理组合中可能只有一个试验单位,方差分析中可用互作平方和作为随机误差平方和。在本次调查中,没有遇到这两种情况。

下面是本次调查中发现的简单伪重复的案例:

[案例 1][18]试图比较 4 个林分类型凋落物量以及组成的差异,但是每个林分类型只选取了一个样地作为"代表"。虽然每一个样地中"按一定距离间隔随机设置 34 个凋落物收集框",这只是取样的重复,林分类型这个处理并没有被重复,因此这是一个典型的伪重复。如果确实需要对这 4 种林分类型的凋落物进行互相比较,那么林分类型作为处理因素就必须要被重复。而每一个样地内可以降低取样量(即低于 34),这样可以在总取样量不增加(从而也不增加成本)的情况下避免伪重复的错误。

[案例 2][19]观测了 5 种林分土壤呼吸及自养与异养呼吸全年的动态及其对土壤温度、湿度的敏感性。每一个林分类型设置一个 20m × 20m 的样地,每个样地内随机设置 5 组断根样方,10 组非断根样方。该项研究中,虽然每林分的样地内有若干断根与非断根样方,但是林分这个因素是没有重复的,所以用这个样地观测到的参数作为林分类型的估计是不妥当的,对这些参数进行不同林分之间的统计差异检验也仅是立地差异的统计检验,而不能代表林分之间的差异。这是一个工作量非常大的观测试验,但是在同样的工作量下,作者可以减小每一林分的取样重复而增加林分的重复,从而避免伪重复的问题。

[案例 3][20]为了比较不同土地利用方式下土壤呼吸及对温度的敏感性而分别在林地、草地与轮作旱地 3 种类型的土地上布设多个观测点。但是每一个类型的代表性样地只有一个,即处理没有重复;虽然每一样地内可能测定了很多点,这只是试验单位内的重采样——因为属性的空间变异,必须多点采样才能估计一个试验单位相应属性的"真值"。在文章的数据分析部分,作者进行了不同土地利用类型的通量的 F 检验,因此在通量比较这一点存在伪重复问题。

我国许多研究单位在野外台站设有长期样地,由于各种原因这些样地虽然面积较大但是没有重复,在这些样地上进行的观测性试验都要避免简单伪重复问题。

2 把反复测量结果视为重复

在生态学的试验中经常有这样的情况,即对同一个试验单位(一株植物、一个动物或者一个试验小区)在时间序列上反复(repeated)抽样并进行测定,时间间隔或者以天、或者以年为单位。这种试验设计下获得的同一个试验单位系列观测值,不能视为重复(replication)。因为如上(简单伪重复部分)所述,重复是用来估计随机误差的,它要求重复所对应的试验单位相互独立,而对同一个试验单位不同时间的观测值相互不独立,所以把同一对象多次测定的结果当作重复进行 t-检验、ANOVA、相关分析或者回归分析都是错误的。Pearson 相关分析和回归分析都假设(变量内)观测值之间相互独立,除非有证据表明反应变量在观测间隔长度下已经足以摆脱自相关,否则把反复测量得到的观测值视为独立的观测值进行这两种分析违背了它们的前提假设。而对于这一点,中外生态学家都经常忽视。如果把反复当成了重复,那么由于"重复"的不独立,也就构成了一种伪重复[9]。

这种错误有时可以通过采用适合的方法重新分析数据得到纠正,其中的一些情况仍可用 ANOVA 以变通方式进行分析,或者利用专门的统计算法。目前比较流行的统计软件中也开始陆续加入专门处理反复测量数据的功能模块。事实上,反复测量设计(repeated measures design)①是一类花样繁多、并且非常高效的设计方法,也有相应的数据处理方式。研究者最好在试验部署之前就按照反复测量设计的要求设计试验,而不是已经拿到数据之后才寻找合适的数据处理方法,因为反复测量设计的处理施加顺序、取样频率等有一些特殊的

① 许多时候"repeated measures"被翻译成"重复测量",笔者认为译成"反复测量"才能够突出这种设计方法的特点。Repeated measures 是对同一个受试对象在时间序列上进行的若干次观测,这些观测值之间是相互关联的;而"重复(Replicates)"则要求统计上相互独立,这是它被用来估计随机误差的前提条件。正是因为不符合重复的这一前提,所以才出现了"反复测量"这么一种特殊的试验设计和数据处理方法。

要求,需要在试验之前就确定下来。

下面是把同一个处理对象反复测量结果视为重复的例子:

[案例 4][21]这篇文章是为了研究土壤水分条件对于菖蒲萌发和幼苗生长的影响,对菖蒲进行了 6 个水分梯度的试验,每处理 3 个重复,在 3 个阶段,观测了一些形态指标,整个试验的设计是正确的;但是文中表 4 组内自由度为48,是由 6 ×(3 ×3 -1)得来,说明作者数据处理时把 3 个阶段的数据合并起来进行单向方差分析,这就不正确了。因为组内的数据应该是相互独立的,而 3 个时间段得到的数据相互不独立。如果作者把每个阶段视为一个区组(block),套用区组设计的 ANOVA 方法就更好。

[案例 1][18]在 4 种林分类型的样地中布置凋落物观测,连续观测了 3a。从该文表 3 中可以看出,文章是把每一年的物凋落物观测值的作为一个重复进行方差分析和多重比较的。对同一个样地在不同年份进行的观测实际上构成反复而不是重复,因此这个试验犯了本文所讨论的第②类错误(即把反复测量结果当成重复)。

3　混淆时空格局与处理效应

所研究的目标因素对反应变量的改变称为处理效应。而对于绝大多数生态学对象而言,即使没有处理效应,反应变量也存在巨大的时空变异(可称为时空效应),因此如何把反应变量的时空变异与处理效应区分开是生态学试验的重要目标。如上所述,伪重复的试验注定无法分开这两种效应。

但是即使采用重复并满足随机的要求,在许多时候——特别是反应变量在时间或者空间上呈现某种规则性格局的时候——也不足以保证试验能够正确区分处理效应与时空效应。随机化可能会产生这样的试验方案:处理(或者处理组合)的梯度与某个干扰因素的梯度相重合。在这种情况下,该干扰因素所造成的随机误差是带有方向性的(即偏倚)。因此,生态学试验还要遵循另一个原则,即处理(或者处理组合)的梯度方向不能与某个干扰因素的梯度方向相重合——但是可以正交(例如采用区组设计)。

这一原则在一些教科书中并不等价地表述为:试验单位各方面条件(除处理因素而外)要尽可能一致。条件一致,一方面有使干扰因素均衡(公平)地、而不是偏倚地作用于各处理(或处理组合)的意思,另一方面也有使干扰因素造成的随机误差尽可能小的涵义。后者对于生态学试验(特别是观测性试验)往往是可欲而不可求的,它虽然可以使处理效应容易显现出来,但是并不是试验的必要条件;而前者才是保证试验结果正确的必要条件,否则就无法区分干扰因素的效应与处理因素的效应。

例如想研究一种植物的 A、B、C 3 种生态型在气孔导度上有何种差异,就不能在早晨测定物种 A 的 8 个试验单位、中午测物种 B 的 8 个试验单位、晚上测物种 C 的 8 个试验单位,因为植物的气孔导度在一天内存在一个规则性(单峰或者双峰型)的变化,用这样的试验方法就根本无法把时间变异与生态型效应区分开来。

有些时候干扰因素对反应变量影响的时空格局是已知的,那么如果随机化产生的试验方案恰好使处理水平沿着这种空间梯度方向,就要推翻这个方案重新产生一个[22],或者采用区组设计等。而有些时候这种格局是未知的,这种情况下就要考虑试验单位的分散布置,即同一处理(组合)下作为重复的试验单位要避免在时间和空间上各自集聚在一起。

这种散置事实上限制了随机化。试验设计者可能为了最大程度的散置,使整个试验方案呈现有规则的布局(系统设计)。这种做法被 R. Fisher 激烈反对、而被 W. Gosset(学生氏 *t*-检验的创立者)公开支持。两人就此展开了长达 13a 的持续争论,直至 Gosset 去世。Hurlbert 显然支持充分的空间散置,而且讨论了这种做法的统计学依据:完全随机的目的是为了正确估计犯第一类错误的概率 α,充分空间散置虽然使正确估计 α 的值成为不可能,但是这种做法导致犯第一类错误的概率小于纯粹随机化试验的 α 值,从而保证了各个具体试验的正确性[2]。

如果没有充分地“过滤”掉时空变异效应,就把试验结果解释为处理效应,即使没有进行统计检验也可以视为错误。这一点与伪重复不同。以下是这种错误的案例:

[案例 5][23]这篇文章的研究目标在于分析春季浮游植物群落结构与动态。每次取样取溪水的两个样

品,每 2 周取 1 次,一共取了 6 个月(2005 年 1 ~6 月)。对于水体的取样,虽然取样点是相对固定的,但是由于水是流动的,在同一位置不同时间的取样类似于陆地的非定点取样,空间变异的效应与群落结构的动态混合起来发生作用。因此虽然试验者看到了在取样时间序列上存在群落结构的差异,但是无法让人信服这是由群落动态所导致,而不是取样变异所致。如果试验者设置了重复的采样位置,并且各采样位置存在相似的消长规律,那就有更充分的把握说明这种消长变化确实反映了群落动态。但是这个案例不能归为伪重复,因为并没有进行统计检验,但是这个案例中的问题可以通过设置重复得以解决。

[案例 6][24]这篇文章的一个重要内容是试图证明群落内物种间、植物功能型(PFG)之间存在补偿作用,从而为多样性促进稳定性提供支持性的材料。研究方法是:在内蒙古草原两个固定样地连续 24a 在生长季内分期取样,每期取 5 个样方;再把历年生长盛季的一期取样汇总,每个样地得到 120 个样方数据。用这 120 个样方的数据分析得出优势种与亚优势种之间、优势种与非优势种之间存在负相关关系,重要 PFG 之间也存在负相关关系;而且两个样地情况相似。据此,文章认为两个群落存在物种和 PFG 之间的补偿作用。这部分推理过程的错误之一就是,混淆了取样的空间格局效应与补偿作用的效应。对于破坏性的取样方法,任何两个样方都几乎不会来自同一个点,因此这 120 个样方几乎注定来自 120 个不同的采样点。而不同采样点之间本来就存在着物种和 PFG 之间的负相关[25],这是空间格局的背景效应,因而本文结果不能排它地解释为时间上的补偿效应。必须控制空间效应才能正确判断补偿效应是否存在,可以采用偏相关分析(而且要把空间因素作为被控制的变量)。或者通过设置永久性样方并且采用非破坏性取样的试验方法去掉空间效应后再分析。

这类问题的根本是:干扰因素没有得到控制。上面的两个案例比较隐蔽,但是有的时候,这种错误却很明显;例如[26],为了研究镉如何影响宝山堇菜的光合作用,以未处理的紫花地丁作为对照。

4 讨论与建议

对两个主要期刊各一期的调查发现,有上述 3 个问题的文章占《生态学报》和《植物生态学报》该期试验性文章的 17.9% ~42.9% 和 14.3% ~42.9%(表 1)①。还可以看出,简单伪重复仍然是中国生态学试验中普遍存在的问题,占本次调查的试验性文章的 8.6% ~22.9%(两杂志合并后结果)。在 Hurlbert[2] 所指出的几种伪重复情形中,有的是非常隐蔽因而难以识破的,但是简单伪重复并非如此,是稍加注意就可以避免的。造

表 1 三类错误以及描述不充分情况的统计

Table 1 Statistics of three types of error and insufficient description in the reviewed articles

期刊 Journal	试验文章总数(测度性试验) Total no. reviewed (Mensurative)	描述不充分 Insufficiently Described	简单伪重复(疑似) Error-I (Suspected)	反复当重复(疑似) Error-II (Suspected)	第三类问题(疑似) Error – III (Suspected)
生态学报 Acta Ecologica Sinica	28(16)	10	3(3)	2(1)	0(3)
植物生态学报 Journal of Plant Ecology	7(6)	4	0(2)	0(0)	1(0)

(1)试验总数等于该期杂志试验性报道总和,本次调查中每篇文章对应一个试验;括号内数字是指试验性报道中测度性试验数目;(2)最后三列中括号中的数字是指疑似该类错误的文章数;(3)疑似简单伪重复是指:那些没有设置真正的重复,在数据分析过程中也并没有使用统计检验,但是在图表或者文字表述以某种方式暗示处理效应达到统计显著水平;或者是符合伪重复的特征但由于描述不充分而不能肯定的;疑似第 2 和疑似第 3 类错误的情形是:根据文章对试验的描述很可能犯了这两类错误之一,但是由于描述不充分而不能肯定 (1) Total number of papers reviewed is the same as the number of experiments, i.e. one experiment per paper. Numbers in parenthesis in second column are the numbers of mensurativ experiment; (2) The numbers parenthesized in the last three columns are the numbers of experiments which were suspected in the corresponding error; (3) Suspected simple pseudoreplication means either the case in which treatments were not replicated and inferential statistics was not used but in the tables or figures of the paper the author suggested there were significant treatment effects, or the case in which experiment design was possible a simple pseudoreplication one but could not be confirmed because of insufficient description in the paper. The latter case is the same as that of the suspected error II and suspected error III.

① 疑似错误全部没有犯错构成区间的下限,疑似错误全部为真实错误构成区间的上限。

成本文案例中简单伪重复的原因大抵可归结为试验者对伪重复认识不清或重视不够。这一点与迄今只有一篇生态学文献正确引用 Hurlbert 文章[2]的事实相互印证。

生态学试验中往往涉及过程观测,对同一对象进行反复测量的情况比较常见;如果没有注意到“反复”与“重复”的区别就有可能错误地使用统计方法,因为这种数据产生方式违反了许多统计检验的“独立性”的前提假设。虽然本次调查中只发现一个案例,但是这种错误在生态学研究中并不罕见。这种错误往往比本文中的其它两种更容易被发现,一般也可以通过专门数据分析方法得到纠正。因此,产生此类错误的主要原因是主观上未注意。

第三类错误是一种试验设计(或者数据解释)的逻辑错误,因为各试验单位没有得到公平的对待;与伪重复不同,它不需要进行统计检验就构成错误。如果进行统计检验,这种错误体现为随机误差估计有偏倚。混淆时空间变异与处理效应的可能情形非常多,有时非常隐蔽。但是试验者随时问自己这样两个问题有助于避免此类错误:(1)当前的试验设计(或者数据分析方法)是否能够把处理效应从可能的时空变异格局中区分出来呢?(2)是不是即使没有处理效应、纯粹的时空变异也会得到当前的结果?

数据产生过程与数据解释过程是一个有机的整体。试验者要总体控制,瞻前顾后,才能保证整个过程的逻辑结构正确。一方面在数据产生过程中要注意避免违反拟采用的统计方法的前提假设(顾后);另一方面在数据解释过程中要根据数据产生的方式来选择合适的分析方法(瞻前),以避免得到虚假的结论。

凡是被收入本文案例的文章,都是对试验的描述比较规范、因而可以把试验设计准确还原的;在我们这次调查中也发现一些试验报道性文章对试验设计、实施和数据处理过程描述得不够充分,因而难以还原试验和分析过程。这当然会降低读者对其结果的信任程度,也不利于发表数据的再利用。

没有正确设置重复的试验报道,如果试验描述足够充分,也许有被再利用的价值[8]。前提是,作者不只是报道了平均值,而且——以某种形式——报道了变异程度(可以是方差、标准差或者均值的标准误)以及取样量,否则被 Meta-分析利用的价值也大打折扣。当然,正确的试验报道本来就应该至少包括这 3 项信息[27],一个完整的科学报道应该允许读者自己去判断假设是否得到检验[28]。

此文章在我国被引频次较高的《生态学报》上发表,目的就是引起国内广大研究者在进行试验设计或数据分析时,注意试验的逻辑结构、避免伪重复,使研究方法和评价手段大大提高,准确再现研究的科学内涵,推进学科发展。

References:

[1] Montgomery D C. Design and Analysis of Experiments (6th ed.). New York : Wiley & Sons Inc, 2005. 19－21.

[2] Hurlbert S H. Pseudoreplication and the design of ecologyical field experiments. Ecological Monographs, 1984, 54(2): 187－211.

[3] Hurlbert S H. On misinterpretation of pseudoreplication and related matters: a reply to Oksanen. OIKOS, 2004, 104(3): 591－597.

[4] Hurlbert S H. Dragging statistical malpractice into the sunshine-a citation-classic commentary on pseudoreplication and the design of ecological field experiments by Hurlbert, S. H. Current Contents/Art and Humanities, 1993, (7 MAR 29): 18－18

[5] Cottenie K, Luc De M. Comment to Oksanen (2001): Reconciling Oksanen (2001) and Hurlbert (1984). OIKOS, 2003, 100(2): 394－396.

[6] Hargrove W W and John P. Pseudoreplication: a *sin qua non* for regional ecology. Landscape Ecology, 1992, 6(4):251－258.

[7] Carpenter S R. Large-scale perturbations: Opportunities for innovation. Ecology, 1990, 71(6): 2038－2043.

[8] Oksanen L. Logic of experiments in ecology: is pseudoreplication a pseudoissue? OIKOS, 2001, 94(1): 27－38

[9] Heffner R A, Butler M J IV and Reilly C K. Pseudoreplication revisited. Ecology, 1996, 77(8): 2558－2562.

[10] Kroodsma D E, Byers B E, Goodale E, Johnson S, Liu W C. Pseudoreplication in playback experiments, revisited a decade later. Animal Behaviour, 2001, 61: 1029－1033.

[11] Kozlov M V. Pseudoreplication in Russian ecological research. Bulletin of the Ecological Society of America, 2003, 84: 45－47.

[12] Zhang S R, Pan D Y, Reto J. Strasser. A review of progress in studies of plant ecophysiology: controlled experiments and instrumention. Journal of Plant Ecology, 2007, 31(5): 982－987.

[13] Li J Z. Some problems on quantitative study in geography. Geographical Research,1988, 7(2):1－5.

[14] Zhang F, Wu Y Z, Zhang G P, Ru W M. Analysis on some mistakes in the application of statistics in ecological researches. Journal of Plant

Ecology, 2006, 30(2):361 －364.

[15] Krebs C. Ecological Methodology (2nd ed.) Menlo Park: Addison Welsey Publishers Inc, 1999, 231.

[16] Li Y H, Wang X, Kong D Z, Ye Q S. Effects of long-term CO_2 enrichment on photosynthesis and plant growth in *Anthurium andraeanum* L. seedlings. Acta Ecologica Sinica,2007, 27(5):1852 －1857。

[17] Chen Z, Wang X K, Duan X N, Feng Z Z, Wu Q B. Ozone effects on wheat root and soil microbial biomass and diversity. Acta Ecologica Sinica, 2007, 27(5):1803 －1808。

[18] Luo Z S, Xiang C H, Mu C L. The litterfall of major forests in Guansi river watershed in Mianyang City, Sichuan Province. Acta Ecologica Sinica, 2007, 27(5):1772 －1781.

[19] Chang J G, Liu S R, Shi Z M, Chen B Y, Zhu X L. Soil respiration and its components partitioning in the typical forest ecosystems at the transitional area from the northern subtropics to warm temperature, China. Acta Ecologica Sinica, 2007, 27(5):1791 －1802.

[20] Wang X G, Zhu B, Wang Y Q, Zheng X H. Soil respiration and its sensitivity to temperature under different land use conditions. Acta Ecologica Sinica, 2007, 27(5): 1960 －1968.

[21] Cao Y, Wang G X. Effects of soil water content on germination and seedlings growth of sweet flag. Acta Ecologica Sinica, 2007, 7(5):1748 －1757.

[22] Cox D R. Planning of experiments. New York: Wiley & Sons Inc, 1958. 85 －90.

[23] Li Q H, Hu R, Han B P. Spring dynamics of the phytoplankton community of an oligotrophic reservoir in the southern subtropics of China. Journal of Plant Ecology, 2007, 31(2)313 －319.

[24] Bai Y F, Han X G, Wu J G, Chen Z Z and Li L H. Ecosystem stability and compensatory effects in the Inner Mongolia grassland. Nature,2004, 431: 181 －184.

[25] Wang S P, Niu H S, Cui X Y, Jiang S, Li Y H, Xiao X M, Wang J Z, Wang G J, Huang D H, Qi Q H and Yang Z G. Ecosystem stability in Inner Mongolia. Nature, 2005, 435: E5 －E6.

[26] Deng P Y, Liu W, Han B P. Photosynthesis of *Viola baoshanensis* under Cd stress. Acta Ecologica Sinica, 2007, 27(5):1858 －1862.

[27] Ellison A M. Exploratory data analysis and graphic display. In: Scheiner S. M. & Gurevitch J. eds. Design and analysis of ecological experiments (2nd ed.). Oxford: Oxford University Press, 2001. 37 －62.

[28] Scheiner S M. Theories, hypothesis, and statistics. In: Scheiner S. M. & Gurevitch J. ed. Design and analysis of ecological experiments (2nd ed.). Oxford: Oxford University Press, 2001. 3 －13.

[29] Weiss, Neil A. Introductory statistics (6th ed.). Beijing: Higher Education Press, 2004,791.

参考文献:

[12] 张守仁, 樊大勇, Reto J. Strasser. 植物生理生态学研究中的控制实验和测定仪器新进展. 植物生态学报, 2007, 31(5): 982 ~987.

[13] 李钜章. 地理学定量研究中的若干问题. 地理研究, 1988, 7(2): 1 ~5.

[14] 张峰, 武玉珍, 张桂萍, 茹文明. 植物生态学报, 2006, 30(2):361 ~364.

[16] 李永华, 王献, 孔德政, 叶庆生. 长期 CO_2加富对苗期红掌(*Anthurium andraeanum.*) 植株生长和光合作用的影响. 生态学报, 2007, 27(5): 1852 ~1857.

[17] 陈展, 王效科, 段晓男, 冯兆忠, 吴庆标. 臭氧浓度升高对盆栽小麦根系和土壤微生物功能的影响, 生态学报, 2007, 27(5): 1803 ~1808.

[18] 骆宗诗, 向成华, 慕长龙. 绵阳官司河流域主要森林类型凋落物含量及动态变化. 生态学报, 2007, 27(5):1772 ~1781.

[19] 常建国, 刘世荣, 史作民, 陈宝玉, 朱学凌. 北亚热带 9 南暖温带过渡区典型森林生态系统土壤呼吸及其组分分离. 生态学报, 2007, 27(5):1791 ~1802.

[20] 王小国, 朱波, 王艳强, 郑循华. 不同土地利用方式下土壤呼吸及其温度敏感性. 生态学报, 2007, 27(5) : 1960 ~1968.

[21] 曹昀, 王国祥. 土壤水分含量对菖蒲萌发及幼苗生长发育的影响. 生态学报, 2007, 27(5):1748 ~1757.

[23] 李秋华,胡韧,韩博平. 南亚热带贫营养水库春季富有植物群落结构与动态. 植物生态学报, 2007, 31(2)313 ~319.

[26] 邓培雁, 刘威, 韩博平. 宝山堇菜(*Viola baoshanensis*)镉胁迫下的光合作用. 生态学报, 2007, 27(5):1858 ~1862.

[29] 韦斯. 编计学导论, 第 6 版影印版. 北京: 高等教育出版社, 2004. 791.