文章编号:1671-4229(2024)05-0069-07

数据缺失情形下的空间平均值及其估计

黄晓韵
(广州大学 经济与统计学院, 广东 广州 510006)

摘 要: 自空间平均被创立以来,无论是在科研还是在实践中都得到了广泛的应用,并取得了一定的成果。空间平均中权重系数的估计问题一直是众多学者研究分析的课题。在缺失数据的情况下,空间平均是关于两个随机变量 $R = \sum_i \beta_i s_i r_i$ 以及 $S = \sum_i \beta_i s_i$ 的比率。文章假设在权重 β_i 已知的情况下,利用"Delta 方法"推导了用于估计真实空间平均的估计量 r 的平方偏差、方差和均方误差的近似公式,分析了估计量 r 的偏差以及方差来源,并给出最小化偏差的空间平均估计权重,最后将所得结果运用于全球氨氧化古菌相对丰度数据。

关键词: 空间平均; 最优平均; 缺失数据
中图分类号: O212 **文献标志码**: A

Spatial average and its estimation in the case of missing data

HUANG Xiao-yun
(School of Economic & Statistics, Guangzhou University, Guangzhou 510006, China)

Abstract: Spatial averages have been widely used in scientific research and practical applications since they were proposed. How to estimate the weight coefficient in the spatial average has been a problem studied by many scholars. In the case of missing data, the spatial average is the ratio of the two random variables $R = \sum_i \beta_i s_i r_i$ as well as $S = \sum_i \beta_i s_i$. In this paper, we use the "Delta method" to derive approximate formulas for the squared bias, variance, and mean squared error of the estimator r used to estimate the true spatial average, assuming that the weight β_i is known. The bias of the estimator r and the source of variance are analyzed, and the spatial average estimated weight to minimize the bias is given. Finally, the obtained results are applied to the global relative abundance data of ammoniating archaea.

Key words: spatial average; optimal average; missing data

目前,空间平均在气候科学等领域中得到了广泛的应用,例如空间平均值已被用于计算全球或者区域的平均温度或平均降水[1-2]。空间平均值作为描述空间分布的指标之一,是通过赋予空间区域上的点观测不同的权重计算出来的[3]。在对空间平均值进行计算时,如何确定权重就显得非常重要。在计算过程中,通常会根据调查研究目的来选择不同的空间平均值计算权重[1,4]。一般而言,权重的选择不仅对空间平均值点估计的计算起着重要作用,对了解偏差、方差等也有着不可忽视的重要作用[5-7]。

在空间统计学中,如何求出最优空间平均值

收稿日期: 2023-11-22; **修回日期**: 2024-01-03
作者简介: 黄晓韵(1998—),女,硕士研究生. E-mail:2112164130@e.gzhu.edu.cn
引文格式: 黄晓韵. 数据缺失情形下的空间平均值及其估计[J]. 广州大学学报(自然科学版),2024,23(5):69-75.

是一个很有意义的问题。在寻找最优空间平均值过程中,其关键问题是确定空间平均的计算权重,并根据所得到的权重对观测值进行平均,使得空间平均值的某些统计量达到最小。1997年,Kagan[1]广泛地分析了如何利用离散观测值来估计空间或时间平均值的问题以及最小化MSE的最优空间平均问题,Kagan考虑并选择权重β_i使得线性加权平均$\sum_i \beta_i r_i$有最小的均方误差。

在实际应用中,受到设备故障、数据输入错误等因素的影响,原始数据中可能会存在缺失值,此时,空间平均值的计算通常只能受限于非缺失数据集。数据的缺失不仅增加了空间平均值的计算难度,也增加了其统计分析的复杂性。缺失数据插补作为解决数据缺失的一种方法,能够在一定程度上解决数据缺失所带来的数据信息丢失、估计值偏差大等问题。也就是说,倘若数据中出现缺失值,可以利用辅助信息为缺失值寻找一个替代值来补全数据。而Kagan[1]认为,当一组数据从一个时间段变化到另一个时间段时,若对每个时间段的数据分别计算空间平均最优权重系数,便可仅仅通过非缺失数据求得每个时间段的最优空间平均值。后续研究中,有学者讨论了该方法替代最优插值法的可行性[8-10]。

由于不同时期非缺失数据观察值通常是不同的,所以先前关于缺失数据情况下计算空间平均值最优平均问题的讨论主要集中于如何对每个时期的观测提出不同的最优平均方法,进而计算不同时间段的最优空间平均权重[10]。2018年,针对空间域上非缺失观察值会随着时间的变化而变化以及存在缺失数据的情况,Seshadri[11]找到了一组可用于不同时期的固定权重β_i,并且利用Delta方法推导出空间平均值的偏差和方差的近似估计,同时检验了所得到估计的准确性。在参考文献[11]中,作者主要考虑的是数据的缺失独立于所有观测变量的情况,这是一个相对严格的假设。在实际运用中,造成数据缺失的原因有很多,例如在医疗诊断中,病人由于收入的限制而选择不进行某项检查及病人检查数据的获取需要时间等。也就是说,数据的缺失往往会受到其他观测变量的影响。为了对缺失数据下的空间平均计算进行更加系统的理论分析,本文在假设数据条件缺失概率已知的情况下,利用Delta方法推导出空间平均值的偏差和方差的近似估计,并在最小化空间平均偏差的情况下,得到一组能应用于所有时期的固定空间平均估计权重β_i,结果发现,如果观测值r_i的期望越接近真实空间平均值,那么在该位置的观测将被赋予更高的权重。

本文第1节在假设权重β_i已知的情况下,推导了描述空间平均估计量r的平方偏差、方差的近似公式;第2节讨论了如何为最小化空间平均的偏差确定权重β_i;第3节在观测值以条件概率π_i缺失的前提下,将第2节所得结论应用于全球氨氧化古菌相对丰度数据中,以求出其空间平均相对丰度所对应的权重。

1 空间平均统计量

考虑某个空间区域上的点观测$r_i(1\leqslant i\leqslant n)$,对每个观测值$r_i$都赋予固定的权重$\beta_i$且$\sum_i^n \beta_i = 1$。对于数据中存在缺失的情况,可以利用随机变量$s_i$表示数据是否存在缺失,若$s_i=0$,那么说明位于位置$i$处的点观测$r_i$缺失;若$s_i=1$,则位置$i$处的点观测$r_i$不缺失,并对数据是否缺失做进一步的假设,从而观测值r_i缺失与否不会影响另一点的观测值是否缺失,即s_i和s_j相互独立。假设数据缺失机制是随机的,令$Pr(s_i=1)=\alpha$,给定随机变量x_i,则条件缺失概率为

$$\pi(x_i)=E(s_i|x_i)=Pr(s_i=1|x_i), \tag{1}$$

空间平均统计量为

$$r\equiv\frac{R}{S}=\frac{\sum_i^n \beta_i\pi_i r_i}{\sum_i^n \beta_i\pi_i}, \tag{2}$$

其中,$\pi_i=\pi(x_i)$,该统计量用于估计真实的空间平均值υ。在公式(2)中,空间平均统计量r的分子以及分母都是随机变量。通常来说,υ和r都是关于时间的函数。

1.1 通过“Delta方法”近似均方误差

首先,将空间统计量表示成关于变量R、S的函数形式$r=f(R,S)=\frac{R}{S}$。假设函数$f(R,S)$是光滑的且有限矩存在,那么,$f(R,S)$在(ER,ES)处的

二阶泰勒展开式如下:

$$f(R,S)\approx f(ER,ES)+(R-ER)\frac{\partial f}{\partial R}+(S-ES)\frac{\partial f}{\partial S}+\frac{1}{2}(R-ER)^2\frac{\partial^2 f}{\partial R^2}+\frac{1}{2}(S-ES)^2\frac{\partial^2 f}{\partial S^2}+(R-ER)(S-ES)\frac{\partial^2 f}{\partial R\partial S}。$$

(ER,ES)处的$f(R,S)$关于 R 和 S 的偏导数分别为$\frac{\partial f}{\partial R}=\frac{1}{ES},\frac{\partial f}{\partial S}=-\frac{ER}{(ES)^2},\frac{\partial^2 f}{\partial R^2}=0,\frac{\partial^2 f}{\partial S^2}=\frac{2ER}{(ES)^3},\frac{\partial^2 f}{\partial R\partial S}=-\frac{1}{(ES)^2}$,两边同时取期望得

$$Ef(R,S)\approx\frac{ER}{ES}+\frac{ER}{(ES)^3}E(S-ES)^2-\frac{1}{(ES)^2}E(R-ER)(S-ES)。\quad(3)$$

真实的空间平均值用 υ 表示,那么估计 υ 的均方误差(Mean Squared Error, MSE)为

$$\mathrm{MSE}=E(f(R,S)-\upsilon)^2=E(f(R,S)-Ef(R,S))^2+E(Ef(R,S)-\upsilon)^2。$$

令方差为 $V_r=E((f(R,S)-Ef(R,S))^2$ 以及偏差的平方为 $B_r=E(Ef(R,S)-\upsilon)^2$,那么估计真实空间平均值的 MSE 可以表示为

$$\mathrm{MSE}=V_r+B_r。$$

1.1.1　不含缺失数据的情况

因为 $\sum_i^n\beta_i=1$ 且 $\pi_i=1$,所以 $S=\sum_i\beta_i=1$,此时,与 S 有关的一些项不再存在,包括 S,$E(S-ES)^2$ 以及 $E(R-ER)(S-ES)$,那么空间平均被简化为

$$R=\sum_i^n\beta_i r_i,$$

对应的方差以及平方差分别被简化为如下形式:

$$V_r=ER^2-(ER)^2,$$

$$B_r=E(ER-v)^2。$$

1.1.2　含有缺失数据的情况

回到一般的情况,即个别的观测值以 π_i 的条件概率缺失。当 $1-a\ll1$ 时,有 $Ef(R,S)\cong\frac{ER}{ES}$,因为 $Ef(R,S)$ 中的其他项相对于$\frac{ER}{ES}$而言比较小,所以 $V_r=E(f(R,S)-\frac{ER}{ES})^2$。为了避免在 V_r 中出现 R 和 S 的高阶矩,故利用 $f(R,S)$ 的一阶近似,得方差

$$V_r=E\left\{(R-ER)\frac{\partial f}{\partial R}+(S-ES)\frac{\partial f}{\partial S}\right\}^2。\quad(4)$$

将前面列出的偏导数带入式(4)中,得到

$$V_r=\frac{1}{(ES)^2}E(R-ER)^2+\frac{(ER)^2}{(ES)^4}E(S-ES)^2-2\frac{ER}{(ES)^3}E(R-ER)(S-ES),$$

或者可以表示成如下形式:

$$V_r=\frac{\sigma_R^2}{\mu_S^2}+\frac{\mu_R^2}{\mu_S^4}\sigma_S^2-2\frac{\mu_R}{\mu_S^3}\sigma_{RS}^2,\quad(5)$$

其中,μ 和 σ^2 表示下标变量的均值和标准差(或协方差)。同样的,偏差可以表示成如下形式:

$$B_r=E(Ef(R,S)-v)^2=E\left(\left(\frac{\mu_R}{\mu_S}-v\right)+\left(\frac{\mu_R}{\mu_S^3}\sigma_S^2-\frac{\sigma_{RS}^2}{\mu_S^2}\right)\right)^2。\quad(6)$$

1.2　偏差和方差的渐近表示

1.2.1　不含缺失数据的情况

对于数据完整的情况,有 $\pi_i=1$ 且 $S=\sum_i^n\beta_i=1$,那么空间平均估计量简化为 $R=\sum_i\beta_i r_i$,其对应的方差为

$$V_r=\sum_i\beta_i^2\sigma_{r_i}^2+2\sum_{i<j}\beta_i\beta_j Cov(r_i,r_j),$$

其中,$\sigma_{r_i}^2=Er_i^2-(Er_i)^2$,$Cov(r_i,r_j)=Er_ir_j-Er_iEr_j$。若用 $\boldsymbol{S}_r$ 表示观测值的协方差矩阵以及 $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_n)^{\mathrm{T}}$ 表示权重向量,那么方差可以表示如下:

$$V_r=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}_r\boldsymbol{\beta}。$$

空间平均估计量的偏差可以用时间序列的平均进行表示:

$$B_r=\frac{1}{N}\sum_t(\sum_i\beta_i(Er_i(t)-v(t)))^2,$$

令 $\boldsymbol{D}_1=\frac{1}{N}\sum_t\boldsymbol{d}_1(\boldsymbol{t})\boldsymbol{d}_1(\boldsymbol{t})^{\mathrm{T}}$,$\boldsymbol{d}_1(\boldsymbol{t})=(Er_1(t)-v(t),Er_2(t)-v(t),\cdots,Er_n(t)-v(t))^{\mathrm{T}}$,其中,$\boldsymbol{D}_1$ 为 $n\times n$ 的矩阵,N 是时间 t 的总周期数。那么,偏差可以表示为

$$B_r=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_1\boldsymbol{\beta}。$$

1.2.2　含有缺失数据的情况

在使用式(5)和式(6)时,需要计算 R、S 的一

阶矩和二阶矩,下面逐一给出其具体形式。对于一般情况,个别的观测值以 π_i 的条件概率缺失。此时,R 的方差为

$$\sigma_R^2 \equiv E(R-ER)^2 = ER^2 - (ER)^2。 \quad (7)$$

公式(7)的第一项可以展开成

$$ER^2 = E\sum_i \beta_i^2\pi_i^2 r_i^2 + 2E\sum_{i<j}\beta_i\beta_j\pi_i\pi_j r_i r_j。$$

由于 π_i 和 r_i 相互独立,π_i 和 π_j 相互独立,那么 $E\pi_i^2 = E(\pi_i-\alpha)^2+\alpha^2 = \sigma_\pi^2+\alpha^2$,$E\pi_i = \alpha$,有

$$ER^2 = \sigma_\pi^2\sum_i \beta_i^2 Er_i^2 + \alpha^2\sum_i \beta_i^2 Er_i^2 + 2\alpha^2\sum_{i<j}\beta_i\beta_j Er_i r_j。$$

将公式(7)的第二项展开并代入条件缺失概率的期望

$$(ER)^2 = \alpha^2\sum_I \beta_i^2 (Er_i)^2 + 2\alpha^2\sum_{i<j}\beta_i\beta_j Er_i Er_j。$$

σ_R^2 可推导成如下形式:

$$\sigma_R^2 = \sigma_\pi^2\sum_i \beta_i^2 Er_i^2 + \alpha^2\sum_i \beta_i^2\sigma_{r_i}^2 + 2\alpha^2\sum_{i<j}\beta_i\beta_j Cov(r_i,r_j), \quad (8)$$

其中,$\sigma_{r_i}^2 = Er_i^2 - (Er_i)^2$,$Cov(r_i,r_j) = Er_i r_j - Er_i Er_j$。

下面计算 S 的方差。由于

$$ES^2 = \sigma_\pi^2\sum_i \beta_i^2 + \alpha^2\sum_i \beta_i^2 + 2\alpha^2\sum_{i<j}\beta_i\beta_j,$$

$$(ES)^2 = \alpha^2\sum_I \beta_i^2 + 2\alpha^2\sum_{i<j}\beta_i\beta_j,$$

故 σ_S^2 可推导成如下形式:

$$\sigma_S^2 = \sigma_\pi^2\sum_i \beta_i^2。 \quad (9)$$

注意到,σ_S^2 与观测值的条件缺失概率方差有关,条件缺失概率的方差越大,σ_S^2 就越大。同理,R 和 S 之间的协方差是

$$\sigma_{RS}^2 = \sigma_\pi^2\sum_i \beta_i^2 Er_i。 \quad (10)$$

为了更好地说明空间平均值的偏差以及方差来源,接下来将利用观测值 r_i 与真值 v_i 的关系对 σ_R^2、σ_S^2 做进一步的推导。观测值 r_i 与该点的真值 v_i 的关系为

$$r_i = v_i + \varepsilon_i,$$

其中,ε_i 是随机误差项。假设,$E\varepsilon_i = 0$ 且 ε_i 与 v_i 相互独立,那么

$$Er_i^2 = Ev_i^2 + E\varepsilon_i^2 + 2Ev_i\varepsilon_i = Ev_i^2 + \sigma_\varepsilon^2,$$

其中,σ_ε^2 是 ε_i 的方差,σ_ε^2 对于不同的位置而言是常数,因为测量不同位置的数据时,所用的仪器都相同,所以认为 σ_ε^2 是恒定的。假设 v_i 和 ε_i 相互独立且随机误差项相互独立,即 $E\varepsilon_i\varepsilon_j = E\varepsilon_i E\varepsilon_j = 0$。因此,$r_i$ 的期望和协方差可以分别表示为 $Er_i = Ev_i$ 以及 $Cov(r_i,r_j) = Cov(v_i,v_j)$。将 r_i 的期望和协方差带入 R 的方差 σ_R^2 中,得到

$$\sigma_R^2 = \sigma_\pi^2\sum_i \beta_i^2 Ev_i^2 + \alpha^2\sum_i \beta_i^2\sigma_{v_i}^2 + 2\alpha^2\sum_{i<j}\beta_i\beta_j Cov(v_i,v_j) + \sigma_\pi^2\sigma_\varepsilon^2\sum_i \beta_i^2 + \alpha^2\sigma_\varepsilon^2\sum_i \beta_i^2。 \quad (11)$$

σ_R^2 的第一项与 Ev_i^2 有关,第二项以及第三项分别是由 v_i 的方差以及协方差所带来的,而最后两项则是由随机误差的方差引起的。同样的,R 和 S 的协方差可以表示成如下形式:

$$\sigma_{RS}^2 = \sigma_\pi^2\sum_i \beta_i^2 Ev_i, \quad (12)$$

σ_{RS}^2 与其采样位点真实值的期望值有关。易得,R 的期望以及 S 的期望分别为 $\mu_R = \alpha\sum_i\beta_i Ev_i$ 和 $\mu_S = \alpha$,将 σ_S^2 和 σ_{RS}^2 代入 B_r,得到

$$B_r = E\left(\left(\sum_i \beta_i Ev_i - v\right) + \left(\frac{\sigma_\pi^2}{\alpha^2}\right)\left(\mu_v\sum_i \beta_i^2 - \sum_i \beta_i^2 Ev_i\right)\right)^2, \quad (13)$$

其中,$\mu_v \equiv \sum_i\beta_i Ev_i$。从公式(13)中可以看出,计算空间平均值的偏差来源是有限样本的抽样误差以及由缺失数据所带来的误差,并且缺失数据所带来的误差与$\frac{1}{\alpha^2}$成比例关系。

至于方差,其推导式如下:

$$V_r = \frac{\sigma_\pi^2}{\alpha^2}\sum_i \beta_i^2 Ev_i^2 + \sum_i \beta_i^2\sigma_{v_i}^2 + 2\sum_{i<j}\beta_i\beta_j Cov(v_i,v_j) + \sigma_\varepsilon^2\sum_i \beta_i^2 + \frac{\sigma_\varepsilon^2\sigma_\pi^2}{\alpha^2}\sum_i \beta_i^2 + \frac{\mu_v^2\sigma_\pi^2}{\alpha^2}\sum_i \beta_i^2 - 2\frac{\mu_v\sigma_\pi^2}{\alpha^2}\sum_i \beta_i^2 Ev_i。 \quad (14)$$

注意到,方差公式的前五项是由随机变量 R 引起的,第六项是由 S 的方差引起的,最后一项是由 R 和 S 的协方差引起的,这说明空间平均值的方差主要是由 R 和 S 的方差以及它们的协方差带来

的。进一步的,公式(14)表示成如下形式:

$$V_r = \sum_i \beta_i^2\sigma_{v_i}^2 + 2\sum_{i<j}\beta_i\beta_j Cov(v_i,v_j) + \sigma_\varepsilon^2\sum_i\beta_i^2 + \frac{\sigma_\pi^2}{\alpha^2}\sum_i\beta_i^2(\sigma_{v_i}^2+\sigma_\varepsilon^2) + \frac{\sigma_\pi^2}{\alpha^2}\sum^i\beta_i^2(Ev_i-\mu_v)^2,$$

由于 $\sigma_{v_i}^2+\sigma_\varepsilon^2=\sigma_{r_i}^2$,故

$$V_r = \sum_{i,j}\beta_i\beta_j Cov(r_i,r_j) + \frac{\sigma_\pi^2}{\alpha^2}\sum_i\beta_i^2\sigma_{r_i}^2 + \frac{\sigma_\pi^2}{\alpha^2}\sum_i\beta_i^2(Er_i-\mu_r)^2, \quad (15)$$

其中,$\mu_r = \sum_i\beta_i Er_i$。令 $\boldsymbol{S}_r$ 为观测值 r_i 的协方差矩阵,$\boldsymbol{S}_r^d$、$\boldsymbol{F}_r$ 是第 i 个对角元素分别为 $\sigma_{r_i}^2$、$(Er_i-\mu_r)^2$ 的对角矩阵,那么方差可以表示成如下形式:

$$V_r = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\beta}$$

其中,$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_n)^{\mathrm{T}}$,$\boldsymbol{C}=\boldsymbol{S}_r+\frac{\sigma_\pi^2}{\alpha^2}(\boldsymbol{S}_r^d+\boldsymbol{F}_r)$。

令数据缺失机制为完全随机缺失情况下,空间平均值方差为 V_r',那么

$$V_r' = \sum_{i,j}\beta_i'\beta_j' Cov(r_i',r_j') + \frac{\alpha(1-\alpha)}{\alpha^2}\sum_i\beta_i'^2\sigma_{r_i'}^2 + \frac{\alpha(1-\alpha)}{\alpha^2}\sum_i\beta'^2 i(Er_i'-\mu_{r'})^2。$$

与 Seshadri 所讨论的方差 V_r' 相比,在数据缺失机制为随机缺失的情况下,如果 $\sigma_\pi^2<\alpha(1-\alpha)$,那么空间平均的方差 V_r 会更小。

本文在这一节中给出了个别观测值以条件概率 π_i 缺失的情况下,估计真实空间平均估计量 r 的平方偏差、方差和均方误差的近似公式,并简单分析了估计量 r 的偏差以及方差来源。根据最优平均理论,此时需要通过最小化估计量的偏差、方差以及均方误差,算出空间平均最优权重。之前已有研究[1-2]讨论了不存在数据缺失的情况下,如何通过最小化空间平均估计量的偏差、方差来求得最优估计。也有相关研究[11]考虑了观测值以 α 的概率缺失的情况,并通过最小化偏差、方差或 MES 给出了空间平均值所对应的最优权重。目前,对于观测值以条件概率 π_i 缺失的情形,空间平均最优权重的估计问题讨论得还比较少。接下来,笔者将通过最小化估计量的偏差计算空间平均的估计权重。

2　最优平均

推导出空间平均估计量的偏差以及方差的近似公式后,接下来将考虑如何确定,最小化偏差的权重 β_i,也就是最优平均问题。本文的不同之处在于,将问题扩展到每个观测值以 π_i 的条件概率缺失。

以下将从偏差的两个贡献来考虑如何求得最小化偏差的权重 β_i。首先,当数据不存在缺失时,偏差的贡献来源于有限样本的抽样,此时,偏差为 $B_r=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_1\boldsymbol{\beta}$。因为受到条件 $\sum_i\beta_i=1$ 的约束,权重 β_i 必须满足 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta}=1$ 以及 $\boldsymbol{\beta}\geqslant 0$ 的条件,其中,$\boldsymbol{u}=(1,1,\cdots,1)^{\mathrm{T}}$。为了最小化偏差,引入函数

$$g_{b,1}(\lambda,\boldsymbol{\beta})=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_1\boldsymbol{\beta}+2\lambda(1-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta})+2\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\beta},$$

该函数需要在稳定点处达到最小。为了使不等式约束条件 $\boldsymbol{\beta}\geqslant 0$ 成立,求得的解还需要满足互补松弛条件 $\rho_i\beta_i=0$。对向量 $\boldsymbol{\beta}$ 求导,得

$$\boldsymbol{D}_1\boldsymbol{\beta}=\lambda\boldsymbol{u}-\boldsymbol{\rho}。$$

当数据不存在缺失时,观测值 r_i 的期望越接近真实空间平均值,则该点在计算空间平均值时的权重就越大。

回到缺失数据的情况,考虑偏差的第一个贡献很小的极限情况,即 μ_r 为近似真实空间平均值的期望 Ev。在实际应用中,将利用观测值对真实空间平均值做近似。在此情况下,偏差可以表示为

$$B_r = \left(\frac{\sigma_\pi^2}{\alpha^2}\right)^2\left(\sum_i\beta_i^2(Ev-Er_i)\right)^2。$$

令 $\boldsymbol{D}_2$ 是 $n\times n$ 阶的对角矩阵,其中,$D_{2,ii}=E(r_i-v)$。为了最小化偏差,引入函数

$$g_{b,2}(\lambda,\boldsymbol{\beta})=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_2\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_2\boldsymbol{\beta}+4\lambda(1-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta})。$$

对 β 求导并令其等于 0,得

$$\boldsymbol{D}_2\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}_2\boldsymbol{\beta}=\lambda\boldsymbol{u}$$

当 $n=3$ 时,

$$\begin{pmatrix} D_{2,11}^2\beta_1^3+D_{2,11}D_{2,22}\beta_1\beta_2^2+D_{2,11}D_{2,33}\beta_1\beta_3^2 \\ D_{2,11}D_{2,22}\beta_1^2\beta_2+D_{2,22}^2\beta_2^3+D_{2,22}D_{2,33}\beta_2\beta_3^2 \\ D_{2,11}D_{2,33}\beta_1^2\beta_3+D_{2,22}D_{2,33}\beta_2^2\beta_3+D_{2,22}^2\beta_3^3 \end{pmatrix}=\begin{pmatrix}\lambda\\ \lambda\\ \lambda\end{pmatrix},$$

其中,$D_{2,ii}$表示 $n\times n$ 矩阵主对角线上的第 i 个元素。将每一个线性方程减去前一个线性方程,得到以下方程:

$(D_{2,11}\beta_1^2 + D_{2,22}\beta_2^2 + D_{2,33}\beta_3^2)(D_{2,11}\beta_1 - D_{2,22}\beta_2) = 0$,

$(D_{2,11}\beta_1^2 + D_{2,22}\beta_2^2 + D_{2,33}\beta_3^2)(D_{2,22}\beta_2 - D_{2,33}\beta_3) = 0$。

考虑 $D_{2,11} < 0$ 以及 $D_{2,22}$、$D_{2,33} > 0$ 的情况,令 $D_{2,22}\beta_2 - D_{2,33}\beta_3 = 0$ 或者

$$\frac{\beta_2}{\beta_3} = \frac{D_{2,33}}{D_{2,22}},$$

另外,因为 $D_{2,11}\beta_1^2 + D_{2,22}\beta_2^2 + D_{2,33}\beta_3^2 = 0$,所以有

$$\frac{\beta_1}{\beta_2} = \left(-\frac{D_{2,22}}{D_{2,11}}\left(1 + \frac{D_{2,22}}{D_{2,33}}\right)\right)^{\frac{1}{2}},$$

根据条件 $\sum_i \beta_i = 1$,便可求出 β_1,β_2 和 β_3。

对于一般的情况,

$$\begin{pmatrix} D_{2,11}\beta_1 \sum_{k=1}^{n} D_{2,kk}\beta_k^2 \\ \cdots \\ \cdots \\ D_{2,nn}\beta_n \sum_{k=1}^{n} D_{2,kk}\beta_k^2 \end{pmatrix} = \begin{pmatrix} \lambda \\ \cdots \\ \cdots \\ \lambda \end{pmatrix}。$$

同样的,将每一个线性方程减去前一个线性方程,得到以下方程:

$$\begin{pmatrix} (D_{2,11}\beta_1 - D_{2,22}\beta_2) \sum_{k=1}^{n} D_{2,kk}\beta_k^2 \\ \cdots \\ (D_{2,n-1,n-1}\beta_{n-1} - D_{2,nn}\beta_n) \sum_{k=1}^{n} D_{2,kk}\beta_k^2 \end{pmatrix} = \begin{pmatrix} 0 \\ \cdots \\ 0 \end{pmatrix}。$$

在计算上述 $n-1$ 个方程时,首先根据 $D_{2,ii} < 0$ 以及 $D_{2,ii} > 0$ 将 $n-2$ 个方程分成两组,再由 $D_{2,j-1,j-1}\beta_{j-1} - D_{2,jj}\beta_j = 0, j = 2,3,\cdots,n$ 得出 $n-2$ 个 β_{j-1} 与 β_j 的比例关系,最后利用 $\sum_{k=1}^{n} D_{2,kk}\beta_k^2 = 0$ 以及 $\sum_i \beta_i = 1$ 便可解出 β_i,其中,$i = 1,2,\cdots,n$。注意到,位置 i 处观测值 r_i 的期望越接近真实空间平均值,意味着该观测值的权重 β_i 越大。

总的来说,无论是在非数据缺失情形下还是数据缺失情形下,在位置 i 处的观测值 r_i 的期望越接近真实空间平均值,那么在计算空间平均时,该位置的观测值就有着越大的空间平均估计权重。

3 实 证

本文选取全球氨氧化古菌相对丰度空间平均计算问题作为空间平均值参数估计的应用案例进行研究。为确保样本的代表性,本文选用2009—2012年太平洋、大西洋、印度洋、地中海、红海等区域的氨氧化古菌相对丰度来计算全球平均值,并将其视为真实的相对丰度空间平均值,数据来源于海洋基因图谱数据库 https://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/。

一般来说,氨氧化古菌难以在富营养化的表层海水和中层海水中生存,环境复杂的深海为氨氧化古菌提供了广阔的生存空间,所以推测可能存在部分具有特定结构的氨氧化古菌能够适应从陆地向深海扩展的环境变化[12]。因此,受到氨氧化古菌对不同海水深度的适应性影响,在采集不同海水深度的样品时,可能会出现多个氨氧化古菌物种相对丰度缺失的问题。作为影响氨氧化古菌相对丰度数据中是否缺失的重要因素,本文将选用海水深度来预测氨氧化古菌相对丰度的缺失概率。当相对丰度数据被观测时,$s_i = 1$;当相对丰度数据缺失时,$s_i = 0$。为了估计相对丰度的缺失率,本文基于海水深度 x_i 对相对丰度的缺失概率建立如下 Logistic 回归模型:

$$\pi(x_i) = \frac{e^{\theta_0 + \theta_1 x_i}}{1 + e^{\theta_0 + \theta_1 x_i}}, \tag{16}$$

其中,θ_0 和 θ_1 是未知的回归参数。根据公式(16)以及本文提出的方法,计算空间平均值的估计参数,结果如表1及表2。

表1 数据缺失情况下的 Logistic 模型计算结果

Table 1 Results of logistic model calculation with missing data

参数	估计值	标准差	Z 值	P 值
θ_0	-0.903 1	0.048 0	-18.80	<2e-16
θ_1	0.004 2	0.000 2	21.48	<2e-16

表2 空间平均的估计权重

Table 2 The estimated weights of the spatial average precipitation

区域	估计权重
北冰洋	0.152 1
北太平洋	0.095 9
北大西洋	0.080 6
地中海	0.043 7
红海	0.096 0
印度洋	0.161 3
南太平洋	0.095 9
南大西洋	0.151 4
南极洋	0.123 0

由表 1 可知,参数 θ_1 的估计值为 0.004 2,对应的 P 值远远小于显著性水平 0.05,说明深度对不同氨氧化古菌物种相对丰度的缺失有显著的正效应。从表 2 可以看到,氨氧化古菌相对丰度空间平均的估计权重中,印度洋所对应的估计权重最大,结合第 2 节得到的理论结果,说明印度洋观测值的期望最接近真实的空间平均值;而地中海的估计权重最小,说明地中海观测值的期望与真实空间平均值的差异最大。

参考文献:

[1] Kagan R L. Averaging of meteorological fields[M]. Dordrecht: Kluwer Academic Publishers, 1997.

[2] Vinnikov K Y, Robock A, Qiu S, et al. Optimal design of surface networks for observation of soil moisture[J]. Journal of Geophysical Research Atmospheres, 1999, 104(D16):19743-19749.

[3] 孟斌,王劲峰,张文忠,等. 基于空间分析方法的中国区域差异研究[J]. 地理科学,2005,25(4):393-400.

[4] Benedetti R, Palma D. Optimal sampling designs for dependent spatial units[J]. Environmetrics, 1995, 6(2):101-114.

[5] Casella G, Berger R L. Statistical inference[M]. Boston: Duxbury/Thomson Learning, 2001.

[6] Cressie N A C. Statistics for spatial data[M]. Hoboken:John Wiley & Sons, 2015.

[7] Morrissey M L, Maliekal J A, Greene J S, et al. The uncertainty of simple spatial averages using rain gauge networks[J]. Water Resources Research, 1999, 31(8):2011-2017.

[8] Krige D G. Two-dimensional weighted moving average trend surfaces for ore-evaluation[J]. Journal of the South African Institute of Mining and Metallurgy, 1966, 66:13-38.

[9] Stein M L. Minimum norm quadratic estimation of spatial variograms[J]. Journal of the American Statistical Association, 1987, 82(399): 765-772.

[10] Wiener N. Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications[M]. Cambridge:MIT Press, 1964.

[11] Seshadri A K. Statistics of spatial averages and optimal averaging in the presence of missing data[J]. Spatial Statistics, 2018,25:1-21.

[12] Liu L T, Xiao X, Zhang Y. Environmental adaptation basis and ecological function of deep-sea ammonia-oxidation archaea [J]. Acta Microbiologica Sinica, 2021, 61(9): 2643-2662.

【责任编辑: 卓祯雨】