·基础理论与方法·

监测数据统计分析模型在生态学研究中的应用

赵哲[1]　王海涛[1]　姜宝法[1,2]

[1]山东大学公共卫生学院流行病学系，济南 250012；[2]山东大学气候变化与健康研究中心，济南 250012

通信作者：姜宝法，Email：bjiang@sdu.edu.cn

【摘要】 近年来，环境监测、疾病监测等各种监测网络不断健全，监测系统成为开展生态学研究的重要数据来源。监测数据类型包括了横断面数据、时间序列数据和面板数据，涉及暴露、结局和混杂3个维度。针对该数据的信息属性和结构特点，相关统计学方法逐渐发展完善，出现了一些新的方法、模型。基于数据的时空属性，本文对监测数据在生态学研究中常用模型的原理、适用条件及优劣进行了综述。

【关键词】 监测数据；生态学研究；统计模型；环境流行病

基金项目：国家科技基础资源调查专项（2017FY101202）

DOI：10.3760/cma.j.issn.0254-6450.2019.08.026

Applications of statistical models on surveillance data in ecological study

Zhao Zhe[1], Wang Haitao[1], Jiang Baofa[1,2]

[1]Department of Epidemiology, School of Public Health, Shandong University, Jinan 250012, China; [2]Climate Change and Health Center, Shandong University, Jinan 250012, China

Corresponding author: Jiang Baofa, Email: bjiang@sdu.edu.cn

【**Abstract**】 In recent years, with the improvement of various surveillance network, surveillance system has become an important data source for ecological study. Different data types, including cross-sectional data, time series data and panel data, containing abundant information involving exposure, outcome and confoundings. Gradually, some new statistical methods have been developed or improved for the special structural characteristics of surveillance data. In this paper, we summarized the principles of these models, preconditions, as well as their advantages and limitations.

【**Key words**】 Surveillance data; Ecological study; Statistical models; Environmental epidemiology

Fund program: Special Foundation of Basic Science and Technology Resources Survey of Ministry of Science and Technology (2017FY101202)

DOI:10.3760/cma.j.issn.0254-6450.2019.08.026

监测数据是利用相关测量手段对反映研究对象内外状态及其影响因素的各项指标进行连续地、动态地、定量观察形成的数据集，数据来源于疾病、环境等监测系统，包括疾病的结局信息和致病因素的暴露信息等，无需额外投入，是描述、评判人群内外环境状态的重要资料来源。随着监测系统的健全和信息与计算科学的发展，利用监测数据建立统计分析模型开展生态学研究成为热点。

生态学研究是从人群及其所生活的自然、社会生态系统出发，以其影响群体健康状态的内外环境因素为切入点，对监测数据进行挖掘，广泛应用在环境流行病学研究领域。它利用“自然实验”的方式，科学评估全人群健康影响因素的平均暴露水平对健康指标的影响，符合流行病学群体的特征，结论更具有公共卫生学意义。根据时空特性，所收集数据主要可分为3类：横断面数据（cross-sectional data）、时间序列数据（time series data）和面板数据（panel data）[1-3]，涉及暴露、结局和混杂3个属性维度。如何利用好监测数据，选择合适的模型，充分挖掘信息，合理提取有效特征，得到科学、准确的结论，是需要解决的难题。对此，基于数据的时空属性，本文对监测数据在生态学研究中常用模型的原理、适用条件及优劣进行了综述。

一、时间属性数据的分析

生态学研究中，在某一地理或行政区域对外环境状况和特定人群内环境健康状态进行连续地动态

监测，形成同一个群体在不同时点的数据集，称为时间序列数据，包括某段时间内暴露因素、结局效应各自序列的变动情况及二者间的短期变化关系。由于同一变量内部在时间尺度上存在自相关，不同变量间在时序性上存在前后的变动关系，所以既可以用来预测，又可以研究暴露对结局的效应。时间序列是随机过程（stochastic process）的特例[4]，序列中包含的有效信息与无效信息的比值称为信噪比（signal-to-noise ratios）。在该数据集中，不仅存在着如性别、民族等不随时间改变的固定特征，也存在着随时间变化的随机特征，如空气质量指数、传染病的日发病数等。根据随机过程的因素分解法，数据的有效信息可分解为季节性变化、周期性波动和长期趋势[5]。在环境流行病学的分析中，通常借助特定的模型在较小的信噪比下完成病因的探讨[6]。因此，如何控制季节性和长期趋势等时间因素带来的强混杂，并且最大可能地保留因暴露变量变化而引起的短期波动，识别较小的效应值[6-7]，是选择和建立时间序列回归模型所面临的重要挑战。根据自变量和因变量的资料性质及纳入回归的形式，常用的模型可以分为广义线性模型（generalized linear models，GLM）、广义相加模型（generalized additive models，GAM）、分布滞后非线性模型（distributed lag non-linear model，DLNM）及处理零膨胀数据（zero-inflated data）的模型等。

1. GLM：统计学中许多经典理论模型是假定自变量和因变量间存在线性关系而建立的，其中因变量为连续变量，可分为系统成分和误差成分，模型需要满足严苛的适用条件，才能实现最佳线性无偏估计（best linear unbiased estimator）。而流行病学研究中自变量和因变量间存在大量非线性关系，同时因变量的资料性质可以为计数资料，也可以是计量资料，从而带来一般线性模型应用上的局限性，GLM拓展了这一应用条件的适用范围。GLM由3个部分组成：随机部分、系统部分、连接函数[8]。通过假定因变量服从某种特定分布，建立单调、可导的非线性连接函数，对因变量进行变换，从而将自变量线性组合与因变量的概率分布进行连接，使得变量间的关系在GLM的架构下变得清晰[8]。GLM将部分非线性问题转化为线性关系进行处理，将经典线性模型中因变量的正态假设放宽为具有散布参数的指数型分布，极大的扩充了模型的应用场景，同时实现了对参数的无偏估计。GLM的基本形式为：$g(Y)=g(\mu_i)+\varepsilon_i=\eta+\varepsilon_i=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_ix_i+\varepsilon_i$。其中，$Y$为服从某一概率分布的因变量，$E(Y)\equiv\mu_i$，$x_i$为自变量，$\beta_i$为偏回归系数，$\eta$为系统部分，即自变量的线性组合，$\varepsilon_i$为残差，$g(\cdot)$为对应特定指数分布族的连接函数，见表1。

表1 常用分布与连接函数

分布	连接函数	模型名称
正态分布	单位连接	多元线性回归
二项分布	Logit连接	logistic回归
负二项分布	倒数连接	负二项回归
泊松分布	Log连接	泊松回归

在监测数据中，由于反映健康结局监测数据常以计数资料的形式出现，如在特定时间、特定区域某急性疾病的发病数、慢性疾病的入院数等，因此泊松回归模型更为常用。将对结局事件有不同贡献的变量纳入回归方程，以评估其效应。在拟合泊松回归模型时，如存在异常值、变量取值的自相关及遗漏变量等影响因素[9-11]，则出现方差大于均数的过离散现象（over-dispersion），从而违背了泊松分布期望等于方差的前提假定，继续使用泊松回归模型会增大犯统计学Ⅰ类错误的概率。判断是否存在过离散可以使用O检验，基于泊松回归的残差检验，得分检验和拉格朗日乘子检验（Lagrange multiplier test，LM）等方法[9,12-13]。确定存在过离散后，可用类泊松回归（quasi-Poisson regression）或负二项回归等进行拟合，以改进泊松分布中关于等离散（条件均值等于条件方差）的约束条件[10,14]，但有研究表明类泊松回归其参数的标准误差往往会大于泊松回归模型的标准误差，负二项分布为非嵌套的模型，不容易检验拟合优度的效果，以广义泊松分布理论基础的广义泊松回归解决了上述不足[15-17]。

时间序列的GLM用参数方法拟合随时间变化的自变量，通过加入如正弦函数等周期性函数控制长期趋势和季节性，合理的反映自变量对因变量的影响，可以借助互相关分析或Spearman相关分析确定最佳滞后。但由于参数方法的确定性和时间趋势的不确定性，使得其在估计自变量的系数时不够灵活[18]，需要事先选择特定的指数分布族，且可用分布类型有限，如无法进行合理变换，模型拟合结果会存在较大偏差[8]。

2. 处理零膨胀数据的模型：处理表征人群健康结局事件的因变量时，由于其多属于计数资料，在季节性、政策法规等其他外部自然、社会因素的影响下，一段时期观测中可能会出现较多预期结局为零的事件，即“零膨胀”现象，将此类单位时间或空间观

察事件组成的计数资料数据集中含有较多的零的数据称为零膨胀数据。对是否存在零膨胀现象，可以通过Vuong统计量对其进行判断[19-20]。当处理零膨胀数据时，仍采用泊松回归或负二项回归模型进行分析，会造成模型参数估计的结果与实际值偏差较大，低估事件的概率，此时常采用零膨胀模型（zero-inflated models，ZIM）或Hurdle模型等进行拟合，以期获得无偏结果。

ZIM混合概率分布模型，将数据生成过程看成混合分布，包括零计数过程（服从二项分布，用于因素的定性分析）和泊松/负二项/广义泊松分布计数过程（服从泊松/负二项/广义泊松分布，用于因素的定量分析）2部分，所形成的模型分别称为零膨胀泊松回归（zero inflated Poisson regression，ZIP）、零膨胀负二项回归（zero inflated negative binormal regression，ZINB）和零膨胀广义泊松回归（zero inflated generalized Poisson regression，ZIGP）。有研究对处理零膨胀问题的方法与机器学习中的方法进行了对比，机器学习方法的效果优于ZIP、ZINB等经典的计数模型[20]。

Hurdle模型又称两部分模型（two-part model）[21]，将整个事件的发生看成2个过程，第1个过程为二项分布，为是否发生零事件的判定过程，发生零事件记为0，不发生零事件记为1；当不发生零事件，认为可以跨过“栅栏”（hurdle），进入计数过程，该过程服从泊松分布或负二项分布，当服从泊松分布时，上述2部分模型称为泊松Hurdle模型（Poisson hurdle model，PH），服从负二项分布时称负二项分布Hurdle模型（negative binormal hurdle model，NBH）。

ZIM与Hurdle模型是处理零膨胀数据的两类重要方法，ZIM从数据分布的混合性出发，用于解决由于错分带来的结构零的问题，结构零的产生源于并不存在导致结局发生的因素[22]。Hurdle模型从数据生成的整体性出发，将数据生成过程看成两个步骤，解决抽样在抽样过程中产生的抽样零[23-24]，抽样零为暴露于相关因素但仍未出现阳性结局事件。

3. GAM：当自变量与因变量的散点图呈现非线性的关系时，常采用非线性、非参数的形式拟合。GAM是在GLM和加性模型的基础上发展而来的非参数模型，同样由随机部分，加性部分和连接函数组成[25]，唯一不同的是GAM可以通过非参数平滑函数拟合自变量和因变量间的非单调、非线性关系，对非参数化的数据进行探索分析[26]，以相加的形式纳入模型，作为模型的解释成分，常借助样条函数（spline functions）来表达，样条函数包括惩罚样条（penalized splines）、自然立方样条（natural cubic splines）等。样条函数是数值分析中进行插值运算的常用函数，名称最初来源于工程师制图时将弹性木条用固定于设定好的样点上而形成的平滑曲线，后被形象地借用，其分段光滑，区间内可导，具有较好的数值稳定性和收敛性的特点，常用的有二次样条和三次样条[26-27]。GAM基本表达式为：$g(Y)=g(\mu_i)+\varepsilon_i=\eta+\varepsilon_i=\beta_0+f(x_1)+f(x_2)+\cdots+f(x_i)+\varepsilon_i$。$g(\cdot)$代表连接函数，可以是上述指数族分布中的任一种分布；x_i为自变量，$f(x_i)$为各种平滑函数，η为系统部分，ε_i为残差。GAM通过非参数函数对生态学研究中时间序列数据普遍存在的长期趋势、季节性、“星期几效应”以及与研究目的有关的其他协变量进行控制，从而可以探讨非线性、非单调的关系，使数据处理更加灵活多样。在GAM构建中，如何确定节点的个数即样条函数的自由度以选择最优的平滑函数是建立模型和寻找规律的关键，也是模型拟合的难点。部分研究采用数据驱动的建模策略，以广义交叉验证（generalized cross-validation，GCV）等方法确定自由度[28-29]，可用R软件中“mgcv”包实现。模型的求解采用迭代加权最小二乘法（iteratively reweighted least squares）和回切算法（back-fitting algorithm）组成的局部得分过程（local-scoring procedure）实现[25,30-31]，与最小二乘法（ordinary least square，OLS）得到解析解不同，该参数估计只能求得数值解。

GAM弥补了GLM中自变量与因变量间可用函数关系有限的不足，用非参数方法来拟合有时间变化趋势的预测变量，表征自变量对因变量的贡献，具有较高的灵活性，尤其在探索性研究中可以降低建模难度，节约大量的时间[18]，被广泛应用于生态学研究中。但利用GAM开展的生态学研究，仅只能表示某一特定时间的暴露-反应关系，无法精确地联合评估由于滞后效应存在而导致的暴露-滞后-反应关系，同时，平滑函数中自由度、节点等参数的如何设定以滤过信息中的噪声尚无统一的标准，原则上自由度既要足够大，以保证效应值的提取，又要足够小，以消除周期性，需要慎重权衡取舍。

4. DLNM：“时间序列设计的目标是评估随暴露而改变的健康结局序列的短期变换”[32]，暴露和结局之间的关系可能在暴露的频率、强度或持续时间改变后立即出现，但更多情况下存在时间上的滞后。因此，如何恰当的反映暴露变量的滞后结构是评估

滞后效应的核心问题[7,29]。DLNM既能表达非线性的暴露反应关系又能反映二者间的滞后关系，使环境暴露与健康结局的关系评估更为准确[33-34]。DLNM的核心模型为GLM、GAM或广义估计方程[35]（generalized estimator equations，GEE）等，利用非参数函数控制时间序列资料的长期趋势、季节趋势和其他与时间长期变异有关的混杂因素的基础上，建立暴露与结局及滞后效应的两个基函数，然后通过交叉基（cross-basis）过程建立模型。交叉基是上述两个基函数的张量积（tensor product），可以同时拟合暴露-反应的非线性关系及暴露因素的滞后效应[33-36]，是DLNM的建模中的重要一步。常用的基函数为样条函数，主要包括自然立方样条、B样条等。自然立方样条是参数样条，用于以GLM为基础的建模过程；B样条为非参数样条，用于以GAM为基础的建模过程，可借助R软件中的“dlnm”包实现[33,35,37]。DLNM可以通过残差正态性检验评价模型拟合，3-D图（坐标轴为所研究变量、滞后及相对危险）反映滞后效应，也可以绘图得到累积效应[33,37]。

时间序列的生态学研究广泛应用在评价某一特定区域空气污染、极端天气和气象因素等短期环境暴露与健康结局间的关系，在评估效应，确定剂量反应关系及预测疾病发展趋势等方面发挥了积极作用[7]。但它将单个研究区域看成一个整体，忽视了整体内部变量因空间属性不同带来的异质性。由于是对一个小区域整体进行追踪研究，空间尺度较小，忽视了嵌套在更大区域整体内所带来的同质性，同时，因为研究目的的关系，也忽视了小区域整体对大区域整体异质性的贡献。

二、空间属性数据的分析

生态学研究的时间序列分析解决了属性变量在特定区域跨时域的动态变化问题，但不能解决因特征属性的地域分异所带来的健康效应的不同。空间统计学以空间相依性和异质性为前提条件，区别于独立性、随机性的传统前提假设，借助地理信息系统（geographic information system，GIS）研究区域化变量空间分布及空间关系，是经典统计学的有效补充[38-40]。它在横截面数据集原有属性信息的基础上增加了空间位置特征，通过空间权重矩阵的设定反映属性变量在不同地理位置的分布变化情况。

空间数据分析的理论依据来自于地理学第一定律（Tobler's First Law of Geography）：地理单位间的变化可以相互影响，任何事物都是与其他事物相关的，只不过空间上相近的事物关联更紧密[41]，空间自相关的概念据此产生。类似于时间序列中的某一变量在临近时间上的自相关，空间自相关用来表征一个区域单位上的某种属性与邻近区域单位上同一属性的相似程度，为寻找病因假设提供线索[32]。与时间序列分析前后观测值在时序的一维性不同，空间数据的变化是多方向的，建模时是否在传统横截面模型上加入空间特征，需要借助空间自相关统计量以判断属性变量是否存在空间关系[39-40,42]。相比于传统的横截面模型，空间数据建模复杂，需要根据不同区域间的距离函数进行空间权重矩阵的设定，以反映变量的空间效应。同时，由于监测数据来源于监测站点，受空间特征的异同和成本等原因限制，布点有限，在数据上表现为离散的点，进行空间分析前，常利用空间插值法连点成面，获得全域的属性数据[43]。

生态学研究的空间属性数据分析包括探索性分析，检验、验证性分析（统计模型分析）[44]，以建立、检验和运用模型为核心，旨在发现疾病、健康事件及其影响因素的空间聚集与分散，控制分布在空间上的混杂因素，找到原因，进行合理的预测和干预。

1. 探索性分析：探索性分析可以挖掘属性特征在空间上的同质性和异质性，常用空间聚类的方法。聚类分析属于机器学习领域中无监督学习的范畴[45]，空间聚类是聚类分析在空间领域的应用，是根据空间坐标数据进行聚类分析，按照样本间的亲疏关系，将数据集中的对象分成由相似对象组成的类，使同类对象的属性值有较高的相似度，不同类的对象间差异较大。分为基于划分的聚类，基于层次的聚类，基于网格的聚类基于密度的聚类，基于模型的聚类等，是研究疾病及其影响因素地理分区及分类的重要方法，以层次聚类法和k-均值聚类法等最为常用[46-48]。通过聚类分析可以发现健康结局与因素的聚集性，有利于病因假设的提出。

2. 检验、验证性分析：空间回归是将空间自相关的概念与一般线性模型融合，根据地理学第一定律，通过空间位置建立资料间的统计关系[39,49-50]。它从地理的角度研究疾病发病（或患病、死亡等）空间分布与自变量（环境因素如空气、水、土壤等，和社会经济学因素）间的关系，将疾病空间位置关系和空间属性数据结合起来，对疾病的影响因素进行更全面的探索，可以分为空间回归全局模型和空间回归局部模型两大类。空间回归全局模型包括空间滞后模型（spatial lag model，SLM）、空间误差模型（spatial error model，SEM）和空间杜宾模型（spatial Durbin

model, SDM）；空间回归局部模型包括地理加权回归模型（geographical weighted regression）等[51-55]。

全局模型针对于空间自相关问题，其回归系数β不因空间位置的改变而发生变化；SLM适用于相邻区域内因变量存在空间自相关的情形；SEM适用于相邻区域自变量存在空间自相关的情况；SDM适用于自变量和因变量均不满足独立假设的情况。局部模型针对于空间异质性问题，采用局部加权最小二乘法（locally weighted least squares）进行估计，同时考虑了空间相关性和异质性。

空间回归可充分地利用数据，将空间地理信息数据与属性数据结合，进一步量化各因素在空间区域的效应大小，对可能导致疾病发生的因素进行更全面、深入地探索。但在不同尺度的数据合并时，应注意尺度效应对空间信息带来的偏倚，避免产生“生态学谬误”。

空间属性数据的分析在评价不同区域健康问题及影响因素关系方面发挥了重要作用，能够体现区域间的同质性和异质性，广泛应用在生态学研究中[56]，能从空间视角发现疾病的分布及流行规律，通过对空间中混杂的控制，发现因素和健康事件在空间分布上的关系，完成有效信息的识别，为发现病因线索、寻找病因假说提供依据，是对传统回归分析的有效补充。但由于数据类型属于横截面数据，仅能表现出各属性信息跨区域同一时点的状态特征，无法观察到各属性在时间轴上的动态变化，在因果推断上存在较大局限性。

三、时间和空间属性数据的分析

随着监测系统的健全与完善，所收集的具有时空属性的数据日趋增长。同时，健康事件和其影响因素有着时空变化规律，单纯从时间或空间角度探讨会忽视异质性，损失信息，如何从其中挖掘信息，建立跨时间、跨空间的模型，充分利用更大的信息量、更多的变异，系统地评价在不同时空条件下疾病分布及流行情况，合理的控制混杂因素带来的噪声，发现传统方法不易发现的暴露和结局的变化规律，得到更为可靠的结果，已逐渐成为生态学研究的重要领域。在时空属性的生态学研究中，对信息的利用常采用以数据为驱动（data-driven）的技术路线[57]，面板数据较为常见，另外贝叶斯的方法也越来越得到关注。

1. 普通面板数据：面板数据是针对监测区域的不同监测点，在一段时间内跟踪观察某些暴露或结局特征，形成既有横断面的个体维度，又有纵向时间维度的数据集。将时间所跨维度与个体所含数量进行比较，分为“短面板”（short panel）和“长面板”（long panel）；当因变量包含自变量的滞后值，称为动态面板（dynamic panel），相反为静态面板（static panel）；如果因变量为分类变量或者计数变量等，称为非线性面板[58]，包括面板Logit回归、面板泊松回归和面板负二项回归等[1]。面板数据不仅能同时反映变量在截面和时间二维空间上的变化规律，它还可以通过控制空间区域的异质性、内生性以及变量间的共线性从而提高参数估计的有效性。

针对面板数据的分析，常假定个体回归方程有相同的斜率而可以有不同的截距，称为个体效应模型（individual specific effects model）。个体效应包括固定效应（fixed effects）和随机效应（random effects）。面板数据的模型为：$y_{it}=\beta_i x_{it}+\delta_i z_i+\mu_i+\varepsilon_{it}$。其中，$i$代表区域，$t$为时间，$x_{it}$为随时间和个体变化的特征，$z_i$为固定不变的特征，$\mu_i+\varepsilon_{it}$为随机扰动。$\mu_i$为空间的异质性截距，$\varepsilon_{it}$为时间的随机项。当$\mu_i$与某个自变量有关时称固定效应模型，与所有自变量均无关时为随机效应模型，数据经平稳性和协整性检验后选择合适的模型[59]。固定效应和随机效应的选择可以通过豪斯曼检验（Hausman test）判断[59]，传统的豪斯曼检验不适用于异方差的情况，当数据存在异方差时，可通过辅助回归结合聚类稳健的标准误来进行假设检验予以解决。

普通面板模型将不同的行政区域看成独立的个体，虽然没有设定空间权重矩阵，但充分利用了表征空间非结构效应（spatial unstructured effect）[39, 58]的截距μ_i，能够反映变量在横断面维度和时间维度的变化规律，有助于解决不随时间改变的个体异质性（heterogeneity）所引起的遗漏变量的问题，提供更多观察单位的动态信息，有利于提高估计精度，但在模型的效应设定和数据的收集中如选取不当会引起结果出现较大的偏差[1-3, 59]。

2. 空间面板数据：空间面板是从普通面板的基础上发展起来的，它是包含了一些地理单位的时间序列观测值的数据[60]。空间面板模型考虑了时间和空间的异质性，可以结合人群及健康影响因素的时空特征进行综合分析，发现内在规律，使结果更为准确、可靠。常用的空间面板模型包括SLM和SEM。与普通面板数据不同，空间面板模型通过设定空间权重矩阵反映了空间相依性；与横截面的SLM和SEM不同，空间面板在原有模型的基础上增加了时间维度，能更好的反映随时间的变化趋势。判断空

间相关性可用LM，采用LM-Lag和LM-Error统计量进行判断[61]。

3. 其他时空方法：层次贝叶斯模型以贝叶斯定理为依托，建立后验分布与先验分布、似然函数间的关系。当先验概率难以确定时，可以将其看成是随机变量α的函数，从而构造一个α的先验概率，称为超先验。“由先验和超先验构成一个新的先验，即多层先验[43]”，表现为相互存在异质性的特定空间区域与时点，从而形成层次贝叶斯模型。模型常用马尔科夫链蒙特卡洛（Markov chain Monte Carlo，MCMC）的方法求数值解，同时关注了时间、空间上的异质性与相关性，减少了信息损失。

综合时间和空间属性的生态学研究从全局角度出发，研究人群中健康事件及影响因素的动态变化规律，充分考虑了数据间的自相关和异方差的问题。相比于仅从时间角度或只从空间角度探究疾病及影响因素的关系，利用信息全面，能够从大尺度的宏观角度揭示在一定的时空范围内、微观尺度上不易发现的事物发生、发展规律[39]。

利用监测数据开展生态学研究也存在一些问题。作为反映研究对象特征的数据，其收集过程的规范化、系统化是得到正确结论的前提。监测数据质量的提升将会极大的促进生态学研究的发展。目前的信息收集过程中仍存在信息孤岛、过程繁杂、标准不统一等问题，从而导致在链接、建库等数据处理过程中容易出现错误；同时，目前缺乏统一的监测数据清洗标准，无法保证同类研究结果的一致性。在利用生态学研究进行数据分析解释中时，由于以群体水平为观察和分析的单位，容易产生生态学谬误，造成研究结果与真实情况不符，需要在研究中应予以充分考虑。

本文综述了监测数据在生态学研究中常用模型及其基本原理，对不同时空属性的模型功能、用途、优劣进行了简要介绍，目的是在群体水平上对暴露的效应值大小及剂量-反应模式进行科学评估，建立相应的疾病预警系统，得到更为准确的研究结论，为病因假说的提出、疾病预防和控制策略措施的制定提供依据。在后续的研究应用中，需要根据数据结构特点、模型适用条件及优缺点，选择合适的模型进行研究，从而为病因探索提供更有力的支持。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] 陈强. 高级计量经济学及Stata应用[M]. 2版. 北京：高等教育出版社，2014.
Chen Q. Advanced econometrics and Stata application[M]. 2nd ed. Beijing: Higher Education Press, 2014.

[2] 杰弗里·M·伍德里奇. 计量经济学导论：现代观点[M]. 张成思，李红，张步昙，译. 5版. 北京：中国人民大学出版社，2015.
Wooldridge JM. Introductory econometrics: a modern approach[M]. Zhang CS, Li H, Zhang BT, trans. 5th ed. Beijing: China Renmin University Press, 2015.

[3] 达摩达尔·N·古扎拉蒂. 计量经济学基础[M]. 费剑平，孙春霞，译. 4版. 北京：中国人民大学出版社，2005.
Gujarati DN. Basic econometrics[M]. Fei JP, Sun CX, trans. 4th ed. Beijing: China Renmin University Press, 2005.

[4] 赵志，周倩，张晋昕. 时间序列分析方法及其进展[J]. 中国卫生统计，2015，32(6)：1087-1090.
Zhao Z, Zhou Q, Zhang JX. Methods on times series analysis and its progress[J]. Chin J Health Stat, 2015, 32(6): 1087-1090.

[5] 张丽，相晓妹，宋辉，等. 西安市空气污染物与出生缺陷的时序变化及相关性的生态学研究[J]. 西安交通大学学报：医学版，2017，38(3)：353-358，401. DOI：10.7652/jdyxb201703007.
Zhang L, Xiang XM, Song H, et al. Ecological research on time series changes in air pollutants and birth defects in Xi'an and their correlation[J]. J Xi'an Jiaotong Univ: Med Sci, 2017, 38(3): 353-358, 401. DOI: 10.7652/jdyxb201703007.

[6] Dominici F, Wang C, Crainiceanu C, et al. Model selection and health effect estimation in environmental epidemiology[J]. Epidemiology, 2008, 19(4): 558-560. DOI: 10.1097/EDE.0b013e31817307dc.

[7] Bhaskaran K, Gasparrini A, Hajat S, et al. Time series regression studies in environmental epidemiology[J]. Int J Epidemiol, 2013, 42(4): 1187-1195. DOI: 10.1093/ije/dyt092.

[8] 乔治·H·邓特曼. 广义线性模型[M]. 林毓玲，译. 上海：上海人民出版社，2011.
Dunteman GH. Introduction to generalized linear models[M]. Lin YL, trans. Shanghai: Shanghai People's Publishing House, 2011.

[9] 乔舰，范淑芬. 广义线性模型中过离散成因的理论证明及检验[J]. 统计与决策，2016(16)：68-70. DOI：10.13546/j.cnki.tjyjc.2016.16.017.
Qiao J, Fan SF. Theoretical proof and test of the cause of over-dispersion in the generalized linear models[J]. Statistics & Decision, 2016(16): 68-70. DOI: 10.13546/j.cnki.tjyjc.2016.16.017.

[10] 曾平，赵晋芳，刘桂芬. Poisson回归中过度离散的检验方法[J]. 中国卫生统计，2011，28(2)：211-212. DOI：10.3969/j.issn.1002-3674.2011.02.036.
Zeng P, Zhao JF, Liu GF. Test for overdispersion on Poisson regression[J]. Chin J Health Stat, 2011, 28(2): 211-212. DOI: 10.3969/j.issn.1002-3674.2011.02.036.

[11] Chebon S, Faes C, Cools F, et al. Models for zero-inflated, correlated count data with extra heterogeneity: when is it too complex?[J]. Stat Med, 2017, 36(2): 345-361. DOI: 10.1002/sim.7142.

[12] Carrivick PJW, Lee AH, Yau KKW. Zero-inflated Poisson modeling to evaluate occupational safety interventions[J]. Saf Sci, 2003, 41(1): 53-63. DOI: 10.1016/S0925-7535(01)00057-1.

[13] 丁丞. 不同统计模型在公共卫生研究中的应用[D]. 杭州：浙江大学，2015.
Ding C. Different statistical models applied to the researches in public health[D]. Hangzhou: Zhejiang University, 2015.

[14] Payne EH, Gebregziabher M, Hardin JW, et al. An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data [J]. Commun Stat Simul Comput, 2018, 47(6): 1722-1738. DOI: 10.1080/03610918.2017.1323223.

[15] Ismail N, Jemain AA. Handling overdispersion with negative binomial and generalized Poisson regression models [C]. Presented at Casualty Actuarial Society Forum. Baltimore: Casualty Actuarial Society Forum, 2007: 103-158.

[16] 徐昕. 广义泊松回归模型的推广及其在医疗保险中应用[J]. 数理统计与管理, 2017, 36(2): 215-225. DOI: 10.13860/j.cnki.sltj.20160621-003.
Xu X. Generalization of generalized Poisson regression model and its medicare application [J]. J Appl Stat Manag, 2017, 36(2): 215-225. DOI: 10.13860/j.cnki.sltj.20160621-003.

[17] 戴林送, 林金官. 广义泊松回归模型的统计诊断[J]. 统计与决策, 2013(21): 29-33. DOI: 10.13546/j.cnki.tjyjc.2013.21.002.
Dai LS, Lin JG. Statistical diagnoses of generalized Poisson regression model [J]. Statistics & Decision, 2013(21): 29-33. DOI: 10.13546/j.cnki.tjyjc.2013.21.002.

[18] 王彤, 贾彬, 王琳娜. 广义可加模型稳健估计在空气污染对健康影响评价中的应用[J]. 中国卫生统计, 2007, 24(3): 245-247, 270. DOI: 10.3969/j.issn.1002-3674.2007.03.007.
Wang T, Jia B, Wang LN. Robust estimation in generalized additive models and its application for health effects of air pollution [J]. Chin J Health Stat, 2007, 24(3): 245-247, 270. DOI: 10.3969/j.issn.1002-3674.2007.03.007.

[19] 唐欣然, 黄耀华, 王杨, 等. Vuong检验在临床研究中的应用及SAS实现[J]. 中华疾病控制杂志, 2013, 17(7): 613-616.
Tang XR, Huang YH, Wang Y, et al. Clinical application and SAS implementation of Vuong test [J]. Chin J Dis Control Prev, 2013, 17(7): 613-616.

[20] 刘振球, 严琼, 左佳鹭, 等. 零膨胀计数数据回归模型的选择与比较及R语言的实现[J]. 中国卫生统计, 2018, 35(2): 310-312.
Liu ZQ, Yan Q, Zuo JL, et al. Methods on selection and comparison of regression models on zero-inflated count data and its applications with R software [J]. Chin J Health Stat, 2018, 35(2): 310-312.

[21] Zorn CJ. Evaluating zero-inflated and hurdle Poisson specifications [C]//Presented at Midwest political science association. San Diego: Midwest Political Science Association, 1996.

[22] He H, Tang W, Wang WJ, et al. Structural zeroes and zero-inflated models [J]. Shanghai Arch Psychiatry, 2014, 26(4): 236-242.

[23] 赵丽华, 刘桂芬, 原静, 等. Hurdle模型及其在居民就诊影响因素中的应用[J]. 中国卫生统计, 2010, 27(2): 149-151. DOI: 10.3969/j.issn.1002-3674.2010.02.012.
Zhao LH, Liu GF, Yuan J, et al. Hurdle model and its application in hospitalizing factors of residents [J]. Chin J Health Stat, 2010, 27(2): 149-151. DOI: 10.3969/j.issn.1002-3674.2010.02.012.

[24] 原静, 刘桂芬, 薛玉强. 零膨胀计数资料模型选择与比较[J]. 中国卫生统计, 2011, 28(4): 354-356, 360. DOI: 10.3969/j.issn.1002-3674.2011.04.001.
Yuan J, Liu GF, Xue YQ. The selection and comparison of zero-inflated Count data model [J]. Chin J Health Stat, 2011, 28(4): 354-356, 360. DOI: 10.3969/j.issn.1002-3674.2011.04.001.

[25] Hastie T, Tibshirani R. Generalized additive models [J]. Stat Sci, 1986, 1(3): 297-310. DOI: 10.1080/00401706.1992.10484913.

[26] 李源培. 东洞庭湖区水位和气候因素对日本血吸虫中间宿主钉螺分布的影响及其孳生地探测[D]. 上海: 复旦大学, 2011.
Li YP. Impact of water level and climatic factors on the distribution of Schistosoma japonicum intermediate host *Oncomelania hupensis* and the identification of snail habitats in eastern Dongting Lake Areas [D]. Shanghai: Fudan University, 2011.

[27] 李庆扬, 王能超, 易大义. 数值分析[M]. 5版. 北京: 清华大学出版社, 2008.
Li QY, Wang NC, Yi DY. Numerical analysis [M]. 5th ed. Beijing: Tsinghua University Press, 2008.

[28] Liu Y, Sun JJ, Gou YN, et al. A multicity analysis of the short-term effects of air pollution on the chronic obstructive pulmonary disease hospital admissions in Shandong, China [J]. Int J Environ Res Public Health, 2018, 15(4): 774. DOI: 10.3390/ijerph15040774.

[29] Li RZ, Lin HL, Liang YM, et al. The short-term association between meteorological factors and mumps in Jining, China [J]. Sci Total Environ, 2016, 568: 1069-1075. DOI: 10.1016/j.scitotenv.2016.06.158.

[30] 贾彬. 广义可加模型及其在医学中的应用[D]. 太原: 山西医科大学, 2005.
Jia B. Generalized additive models and its application in medical fields [D]. Taiyuan: Shanxi Medical University, 2005.

[31] Dominici F, McDermott A, Zeger SL, et al. On the use of generalized additive models in time-series studies of air pollution and health [J]. Am J Epidemiol, 2002, 156(3): 193-203. DOI: 10.1093/aje/kwf062.

[32] Baker DB, Nieuwenhuijsen MJ. Environmental epidemiology: study methods and application [M]. Oxford: Oxford University Press, 2008.

[33] Gasparrini A. Distributed lag linear and non-linear models in R: the package dlnm [J]. J Stat Softw, 2011, 43(8): 1-20. DOI: 10.18637/jss.v043.i08.

[34] Liu ZD, Li J, Zhang Y, et al. Distributed lag effects and vulnerable groups of floods on bacillary dysentery in Huaihua, China [J]. Sci Rep, 2016, 6: 29456. DOI: 10.1038/srep29456.

[35] 杨军, 欧春泉, 丁研, 等. 分布滞后非线性模型[J]. 中国卫生统计, 2012, 29(5): 772-773, 777.
Yang J, Ou CQ, Ding Y, et al. Distributed lag non-linear models [J]. Chin J Health Stat, 2012, 29(5): 772-773, 777.

[36] Liu ZD, Ding GY, Zhang Y, et al. Analysis of risk and burden of dysentery associated with floods from 2004 to 2010 in Nanning, China [J]. Am J Trop Med Hyg, 2015, 93(5): 925-930. DOI: 10.4269/ajtmh.14-0825.

[37] Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models [J]. Stat Med, 2010, 29(21): 2224-2234. DOI: 10.1002/sim.3940.

[38] 王劲峰, 徐成东. 地理探测器: 原理与展望[J]. 地理学报, 2017, 72(1): 116-134. DOI: 10.11821/dlxb201701010.
Wang JF, Xu CD. Geodetector: principle and prospective [J]. Acta Geogr Sin, 2017, 72(1): 116-134. DOI: 10.11821/dlxb201701010.

[39] 周晓农. 空间流行病学[M]. 北京: 科学出版社, 2009.
Zhou XN. Spatial epidemiology [M]. Beijing: Science Press, 2009.

[40] 王远飞，何洪林. 空间数据分析方法[M]. 北京：科学出版社，2007.
Wang FY, He HL. Methods for spatial data analysis [M]. Beijing: Science Press, 2007.

[41] Westlund H. A brief history of time, space, and growth: Waldo Tobler's first law of geography revisited[J]. Ann Reg Sci, 2013, 51(3): 917-924. DOI: 10.1007/s00168-013-0571-3.

[42] 姜庆五，赵飞. 空间自相关分析方法在流行病学中的应用[J]. 中华流行病学杂志，2011，32(6)：539-546. DOI：10.3760/cma.j.issn.0254-6450.2011.06.002.
Jiang QW, Zhao F. Application of spatial autocorrelation method in epidemiology[J]. Chin J Epidemiol, 2011, 32(6): 539-546. DOI: 10.3760/cma.j.issn.0254-6450.2011.06.002.

[43] 彭迪迪. 空间流行病学及分层贝叶斯模型的应用[D]. 上海：华东师范大学，2015.
Peng DD. Spatial epidemiology and application of hierarchical Bayesian model[D]. Shanghai: East China Normal University, 2015.

[44] Ward MP, Carpenter TE. Analysis of time-space clustering in veterinary epidemiology [J]. Prev Vet Med, 2000, 43(4): 225-237. DOI: 10.1016/s0167-5877(99)00111-7.

[45] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.
Zhou ZH. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[46] 席景科，谭海樵. 空间聚类分析及评价方法[J]. 计算机工程与设计，2009，30(7)：1712-1715. DOI：10.16208/j.issn1000-7024.2009.07.025.
Xi JK, Tan HQ. Spatial clustering analysis and its evaluation[J]. Comput Eng Des, 2009, 30(7): 1712-1715. DOI: 10.16208/j.issn1000-7024.2009.07.025.

[47] 段明秀. 层次聚类算法的研究及应用[D]. 长沙：中南大学，2009. DOI：10.7666/d.y1536795.
Duan MX. The study of hierarchical clustering algorithm and its applications [D]. Changsha: Central South University, 2009. DOI: 10.7666/d.y1536795.

[48] 曾祥嫚，王立英，伍卫平，等. 我国非青藏高原流行区囊型棘球蚴病聚类分析[J]. 中国血吸虫病防治杂志，2014，26(2)：180-183. DOI：10.16250/j.32.1374.2014.02.019.
Zeng XM, Wang LY, Wu WP, et al. Cluster analysis of cystic echinococcosis in non Tibetan Plateau regions [J]. Chin J Schistoso Control, 2014, 26(2): 180-183. DOI: 10.16250/j.32.1374.2014.02.019.

[49] 伍劲屹，周艺彪，姜庆五. 空间回归模型在公共卫生中的应用[J]. 中华流行病学杂志，2013，34(11)：1151-1153. DOI：10.3760/cma.j.issn.0254-6450.2013.011.024.
Wu JY, Zhou YB, Jiang QW. Application of spatial regression model in public health[J]. Chin J Epidemiol, 2013, 34(11): 1151-1153. DOI: 10.3760/cma.j.issn.0254-6450.2013.011.024.

[50] 刘明. 空间回归模型设定方法探讨[J]. 统计与决策，2018(1)：5-9. DOI：10.13546/j.cnki.tjyjc.2018.01.001.
Liu M. Discussion on the specification of spatial regression model [J]. Stat Decis, 2018(1): 5-9. DOI: 10.13546/j.cnki.tjyjc.2018.01.001.

[51] 姜磊. 空间回归模型选择的反思[J]. 统计与信息论坛，2016，31(10)：10-16. DOI：10.3969/j.issn.1007-3116.2016.10.002.
Jiang L. The choice of spatial econometric models reconsidered in empirical studies[J]. Stat Inf Forum, 2016, 31(10): 10-16. DOI: 10.3969/j.issn.1007-3116.2016.10.002.

[52] 黄秋兰，唐咸艳，周红霞，等. 四种空间回归模型在疾病空间数据影响因素筛选中的比较研究[J]. 中国卫生统计，2013，30(3)：334-338.
Huang QL, Tang XY, Zhou HX, et al. Comparison of four spatial regression models for screening disease factors[J]. Chin J Health Stat, 2013, 30(3): 334-338.

[53] Lopez D, Gunasekaran M, Murugan BS, et al. Spatial big data analytics of influenza epidemic in Vellore, India [C]// Proceedings of 2014 IEEE international conference on big data. Washington: IEEE, 2014: 19-24. DOI: 10.1109/BigData.2014.7004422.

[54] Mahara G, Wang C, Yang K, et al. The association between environmental factors and scarlet fever incidence in Beijing region: using GIS and spatial regression models[J]. Int J Environ Res Public Health, 2016, 13(11): 1083. DOI: 10.3390/ijerph13111083.

[55] 肖雄，杨长虹，谭柯，等. 地理加权回归模型在传染病空间分析中的应用[J]. 中国卫生统计，2013，30(6)：833-836，841.
Xiao X, Yang CH, Tan K, et al. A study on the application of geographically weighted regression model in spatial analysis of infectious disease[J]. Chin J Health Stat, 2013, 30(6): 833-836, 841.

[56] 饶华祥，徐莉立，蔡芝锋，等. 空间截面回归模型在肺结核病社会影响因素生态学分析中的应用[J]. 中国卫生统计，2018，35(5)：646-649，654.
Rao HX, Xu LL, Cai ZF, et al. The application of space cross-section regression model in the ecological analysis of tuberculosis related social factors[J]. Chin J Health Stat, 2018, 35(5): 646-649, 654.

[57] 沈体雁，冯等田，孙铁山. 空间计量经济学[M]. 北京：北京大学出版社，2010.
Shen TY, Feng DT, Sun TS. Spatial econometrics[M]. Beijing: Peking University Press, 2010.

[58] 孙烨，方立群，曹务春. 山东、安徽、江苏省2006—2013年秋冬型恙虫病流行特征及影响因素研究[J]. 中华流行病学杂志，2016，37(8)：1112-1116. DOI：10.3760/cma.j.issn.0254-6450.2016.08.012.
Sun Y, Fang LQ, Cao WC. Study on the epidemiological characteristics and influencing factors of scrub typhus in the autumn-winter natural foci, from 2006 to 2013 [J]. Chin J Epidemiol, 2016, 37(8): 1112-1116. DOI: 10.3760/cma.j.issn.0254-6450.2016.08.012.

[59] 冯国双，于石成，胡跃华. 面板数据模型在手足口病与气温关系研究中的应用[J]. 中国预防医学杂志，2013，14(12)：910-913. DOI：10.16506/j.1009-6639.2013.12.008.
Feng GS, Yu SC, Hu YH. Application of panel data model in the study of the relationship between reported hand-foot-mouth morbidity and temperature[J]. Chin Prev Med, 2013, 14(12): 910-913. DOI: 10.16506/j.1009-6639.2013.12.008.

[60] Elhorst JP. Spatial econometrics: from cross-sectional data to spatial panels [M]. Berlin Heidelberg: Springer, 2014. DOI: 10.1007/978-3-642-40340-8.

[61] Mátyás L, Sevestre P. The econometrics of panel data: fundamentals and recent developments in theory and practice [M]. Berlin Heidelberg: Springer-Verlag, 2008. DOI: 10.1007/978-3-540-75892-1.

（收稿日期：2018-12-06）
（本文编辑：李银鸽）