

# 贝叶斯时空模型在布鲁氏菌病疫情数据分析中的应用

郑杨 冯子健 李晓松

**【导读】** 以2000—2007年内蒙古地区布鲁氏菌病(布病)疫情数据为例,运用空间统计学和传染病流行病学的相关理论,应用贝叶斯理论框架建立时空模型,分析布病在时间和空间上呈现的格局及其演变,以及与之相关联的协变量及其变化,并与传统分析方法进行比较。结果显示,拟合协变量的贝叶斯时空模型相对较佳(离差信息准则值最小,为2388.000)。2000—2007年内蒙古自治区101个旗县的布病疫情呈现较强的空间相关性,时空格局存在较明显的共变现象,每年空间相关性不尽相同,空间相关系数后验中位数位于0.968~0.973之间,总体上随时间变化略呈下降趋势。地区类型和牛羊存栏数量与内蒙古布病流行可能有关,且牛羊存栏数量对布病的影响随年份而变化。与传统描述流行病学分析方法比较,贝叶斯方法对布病发病相对危险度的估计更加稳定。

**【关键词】** 布鲁氏菌病; 贝叶斯统计; 时空建模

**Application of Bayesian spatio-temporal modeling in describing the brucellosis infections**  
ZHENG Yang<sup>1</sup>, FENG Zi-jian<sup>2</sup>, LI Xiao-song<sup>3</sup>. 1 Department of Health Statistics, Shanghai Municipal Center for Disease Control and Prevention, Shanghai 200336, China; 2 Office for Disease Control and Emergency Response, Chinese Center for Disease Control and Prevention; 3 Department of Health Statistics, West China School of Public Health Sichuan University  
Corresponding author: LI Xiao-song, Email: lixiaosong1101@126.com

This work was supported by a grant from the Health Research Program of China (No. 200802133).

**【Introduction】** Based on the number of brucellosis cases reported from the national infectious diseases reporting system in Inner Mongolia from 2000 to 2007, a model was developed. Theories of spatial statistics were used, together with knowledge on infectious disease epidemiology and the frame of Bayesian statistics, before the Bayesian spatio-temporal models were respectively set. The effects of space, time, space-time and the relative covariates were also considered. These models were applied to analyze the brucellosis distribution and time trend in Inner Mongolia during 2000–2007. The results of Bayesian spatio-temporal models was expressed by mapping of the disease and compared to the conventional statistical methods. Results showed that the Bayesian models, under consideration of space-time effect and the relative covariates (deviance information criterion, DIC=2388.000), seemed to be the best way to serve the purpose. The county-level spatial correlation of brucellosis epidemics was positive and quite strong in Inner Mongolia. However, the spatial correlation varied with time and the coefficients ranged from 0.968 to 0.973, having a weakening trend during 2000–2007. Types of region and number of stock (cattle and sheep) might be related to the brucellosis epidemics, and the effect on the number of cattle and sheep changed by year. Compared to conventional statistical methods, Bayesian spatio-temporal modeling could precisely estimate the incidence relative risk and was an important tool to analyze the epidemic distribution patterns of infectious diseases and to estimate the incidence relative risk.

**【Key words】** Brucellosis; Bayesian statistics; Spatio-temporal modeling

信息技术的快速发展,使疾病及其相关数据的收集手段发生了突飞猛进的变化,海量的疾病数据

呈现出复杂的时间和空间关系,大多数时空模型是从空间模型扩展而来。随着计算机技术的迅速发展和马尔科夫链蒙特卡罗方法(Markov Chain Monte Carlo, MCMC)的引入,贝叶斯统计这一经典统计方法得到迅速发展,并在气象、环境、生物和医学等领域得到广泛应用,从贝叶斯空间模型扩展而来的贝叶斯时空模型也成为当前疾病时空数据分析和研究

DOI: 10.3760/cma.j.issn.0254-6450.2011.01.016

基金项目: 国家卫生行业科研专项经费(200802133)

作者单位: 200336 上海市疾病预防控制中心卫生统计科(郑杨); 中国疾病预防控制中心疾病控制与应急响应办公室(冯子健); 四川大学华西公共卫生学院卫生统计学教研室(李晓松)

通信作者: 李晓松, Email: lixiaosong1101@126.com

的热点之一。但我国该模型在疾病领域尤其是对传染病时空数据分析仍处于起步阶段。

在我国,法定报告传染病布鲁氏菌病(布病)近年来疫情反弹严重,在内蒙古自治区尤其为甚,成为重要的公共卫生问题<sup>[1]</sup>。本研究基于我国传染病疫情报告系统中内蒙古地区布病及其相关数据的收集,建立贝叶斯时空模型分析 2000—2007 年内蒙古自治区布病的时空分布格局及其演变,并与传统方法比较,探讨贝叶斯时空模型在传染病数据时空分析中的应用价值。

### 基本原理

贝叶斯时空模型是在贝叶斯统计思想的框架下,为分析时空数据资料中蕴含的时间和空间信息而建立的数学模型。贝叶斯统计是不同于经典统计(频率学派)的另一统计学派,基于总体信息、样本信息和先验(prior)信息进行统计推断<sup>[2]</sup>。与传统的建模分析相比较,贝叶斯时空模型认为模型中的所有未知参数为随机变量,需要利用先验分布进行描述,发病或死亡危险度在空间和时间上的变异用随机效应表示。

针对不同类型的疾病时空数据,建立的贝叶斯时空模型有所不同。目前,大多数疫情数据收集形式多为区域数据(areal data),空间和时间的相关性在疾病区域数据的分析中有着重要的作用,也是建模分析的目的之一。

设区域*i*某个时间点*t*的某疾病发病(死亡)人数为 $y_{it}$ ,当发病(死亡)率很低时,通常可认为 $y_{it}$ 服从 Poisson 分布,则有  $E(y_{it}) = \lambda_{it} = e_{it} \theta_{it}$ ;  $e_{it}$  表示第*i*个区域第*t*个时间点的疾病期望发病(死亡)人数; $\theta_{it}$ 表示第*i*个区域第*t*个时间点疾病实际发病(死亡)人数与期望发病(死亡)人数之比,即疾病发病(死亡)的相对危险度(RR),也是研究者关心的未知参数<sup>[3]</sup>。用 $\theta_{it}$ 的 log 连接函数建立贝叶斯时空模型。

1. 时间效应和空间效应相互独立的贝叶斯时空模型:

$$\log(\theta_{it}) = \mu_0 + u_i + v_t \quad (\text{模型 1})$$

式中 $\mu_0$ 为截距, $u_i$ 和 $v_t$ 分别代表空间效应和时间效应,分别主要解释空间和时间的关联性,用相关系数(-1~1)表示。假定空间效应 $u_i$ 服从条件自回归(conditional autoregressive process, CAR)过程中的 proper CAR 分布形式<sup>[4]</sup>,即每一个区域的空间效应受相邻区域空间效应的影响,则每个区域 $u_i$ 的条件分布为正态分布,表示为

$$u_i | u_{j, j \neq i} \sim N(\gamma \sum_{j=1}^{m_i} w_{ij} u_j, \sigma_u^2 / m_i)$$

式中 $u_j$ 表示第*i*个区域第*j*( $j=1, 2, \dots, m_i$ )个相邻区域的空间效应, $m_i$ 为第*i*个区域相邻区域的个数。 $\gamma$ 为空间相关系数。 $w_{ij}$ 为空间权重,如果区域相邻,则 $w_{ij}=1/m_i$ ,否则 $w_{ij}=0$ 。 $\sigma_u^2$ 为空间效应的方差。假定时间效应 $v_t$ 服从一阶自回归[first-order autoregressive process, AR(1)]过程<sup>[5]</sup>,第*t*个时间的效应 $v_t$ 与其前一个时间的效应 $v_{t-1}$ 有关,即

$$v_t = \rho v_{t-1} + \varepsilon_t$$

式中 $\varepsilon_t$ 为残差,服从均数为 0、方差为 $\sigma_\varepsilon^2$ 的正态分布,即 $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ 。

2. 时空交互效应的贝叶斯时空模型(即空间效应随时间变化而变化<sup>[6]</sup>):

$$\log(\theta_{it}) = \mu_0 + u_{it} \quad (\text{模型 2})$$

式中 $u_{it}$ 为时空随机效应,即空间相关性随时间变化,如同 Model 1 中空间效应的先验分布假定,各个区域每个时间点的空间效应 $u_{it}$ 服从正态分布。

模型中时空相关的方差先验分布通常考虑为伽玛分布或均匀分布<sup>[7]</sup>,一般应根据模型和研究数据特点比较选择合适的方差先验<sup>[8]</sup>。根据变异大小和模型复杂程度,假设 $\tau$ ( $\tau=1/\sigma^2$ ,表示精度)服从伽玛分布时,采用 $\tau \sim G(0.001, 0.001)$ 或 $\tau \sim G(0.1, 0.1)$ 分布;假设标准差 $\sigma$ 服从均匀分布时,采用 $\sigma \sim U(0, 10)$ 或 $\sigma \sim U(0, 100)$ 分布。

3. 贝叶斯时空模型(为更准确反映时间和空间效应,进一步纳入可能的影响因素):

$$\log(\theta_{it}) = \alpha_0 + \beta X_{it} + C \quad (\text{模型 3})$$

式中 $\alpha_0$ 为截距, $\beta$ 表示相关因素的回归系数, $X_{it}$ 为第*i*个区域第*t*个时间的相关因素, $C$ 表示拟合的时空效应,可为时间和空间的独立效应,也可为时空交互效应。

贝叶斯时空模型的运算依靠 MCMC 算法,目前可通过多种软件如 R、S-plus、Matlab、MLwin 等实现,其中交互式软件 WinBUGS 使 MCMC 算法更易实现,较为常用。模型拟合优度常采用离差信息准则(deviance information criterion, DIC)<sup>[9]</sup>,对同一数据而言, DIC 值越小表示模型拟合越好。

### 实例分析

利用 2000—2007 年内蒙古自治区布病疫情报告数据,并结合人口、地理信息、布病相关因素等数据,建立贝叶斯时空模型分析内蒙古布病发病时间分布及其演变。根据研究目的和资料可获得性,本

研究中选取可能与布病发病相关的 2 个主要因素:地区类型(牧区、半牧半农区、农区和城市)和平均牛羊存栏数量(万头),以各旗县为基本空间单元、年为基本时间单位进行分析。选择 2000—2007 年内蒙古全区平均发病率作为标准,采用间接法对每年各旗县布病发病例数进行不同性别年龄标准化,计算标准化发病比。标准化发病比即 *RR* 的极大似然估计值(LLR)<sup>[10]</sup>。

获取的内蒙古布病报告发病数据为集合形式,布病在各旗县每年的发病率相对较低,且人与人直接传染鲜有实证,因此假定各旗县每年的布病发病例数服从 Poisson 分布,基于“区域数据”建立贝叶斯时空模型。根据模型 1、2 分别拟合时间、空间的独立效应和时空交互效应。进一步在拟合优度较好的贝叶斯时空模型中纳入可能影响因素即平均牛羊存栏数量和地区类型,并在模型 3 基础上考虑平均牛羊存栏数量随时间变化效应,建立贝叶斯时空模型:

$$\log(\theta_{it}) = \alpha_0 + \beta_1 X_{it} + \beta_2 X_{2it} + C \quad (\text{模型 4})$$

式中  $\beta_1$  表示第 *t* 年平均牛羊存栏数量的回归系数,  $X_{it}$  为第 *i* 县第 *t* 年的平均牛羊存栏数量;  $\beta_2$  为地区类型的回归系数,  $X_{2it}$  为第 *i* 县第 *t* 年的地区类型。

本研究运用 WinBUGS 软件拟合上述贝叶斯时空模型,采用 DIC 值反映模型拟合优度,并结合参数估计值的精度对方差的不同先验分布进行比较<sup>[8]</sup>。利用拟合最佳的模型结果分析 2000—2007 年内蒙古布病的时空格局及其演变,并与经典统计结果进行比较。

结果:4 种方差先验假定下 2 个贝叶斯时空模型 DIC 值如表 1 所示,各参数估计值及其 95%CI 比较结果如图 1、2 所示。考虑方差服从均匀分布  $\sigma \sim U(0, 100)$  的 DIC 值均最低,各参数估计值 95%CI 均相对较窄,精度更好。因此,选用  $\sigma \sim U(0, 100)$  的方差先验假定,模型拟合最佳。

由图 1 可知,假设时间和空间效应独立,2000—2007 年内蒙古 101 个旗县布病发病的时间相关系数较小,95%CI 包含 0 且较宽,时间变化趋势不太明

显。比较表 1 中两种模型 DIC 值,由于时空交互效应的纳入, DIC 值明显下降,模型的拟合优度得到较大提高。因此,内蒙古布病发病的时空效应拟合为时空交互形式更符合数据特点。

在时空交互效应的时空模型基础上纳入相关因素, DIC 值下降为 2388.000,模型拟合相对最佳(表 2)。2000—2007 年内蒙古以旗县为单位的布病发病空间相关系数后验中位数为 0.968 ~ 0.973,存在较强的空间正相关性,提示相邻旗县具有相似的布病疫情;空间相关系数总体上随时间变化有减小趋势,提示每年布病发病的空间格局发生变化,疫情空间聚集性减小,略呈蔓延趋势。以旗县为单位的布病发病与平均牛羊存栏数量和地区类型总体均呈现正相关性。但年份不同,牛羊存栏数量对布病发病的影响差异较大。地区类型以农区和城市作为对照,牧区对布病发病的影响大于半牧半农区。

采用相对最佳贝叶斯时空模型对 2000—2007 年内蒙古各旗县布病发病的 *RR* 估计值和标准化发病比分别绘制地图(图 3,文中仅列举 2001、2004 和 2007 年的结果),图中颜色越深,表示发病的 *RR* 值越大。

从图 3 发现:①随时间推移,布病疫情范围不断扩大,呈蔓延趋势;发病空间格局随时间变化由相对集中于少数旗县向“多极化”发展,由牧区向半农半牧区以及农区和城市扩散。②贝叶斯估计值地图由于借助了邻近区域的发病信息,地图形态更为平滑。例如位于最西部的阿拉善盟布病标准化发病比为 0 显然不合理,贝叶斯估计值则对此进行了修正。

进一步将贝叶斯估计结果与标准化发病比进行比较发现,贝叶斯法估计值 95%CI 明显比似然法估计值 95%CI 范围窄,精度更高,在发病 *RR* 值较小的旗县,此种差异更为突出。本研究仅列举 2007 年疫情最严重的苏尼特左旗发病 *RR* 值估计情况,如表 3

表 1 单纯时间空间效应和时空交互的贝叶斯时空模型结果比较

模型	不同方差先验的 DIC 值			
	Prior 1	Prior 2	Prior 3	Prior 4
	$\tau \sim G(0.001, 0.001)$	$\tau \sim G(0.1, 0.1)$	$\sigma \sim U(0, 10)$	$\sigma \sim U(0, 100)$
时间、空间效应独立	13 301.500	13 301.400	13 301.700	13 299.500
时空交互	2 782.540	2 777.220	2 784.350	2 775.700

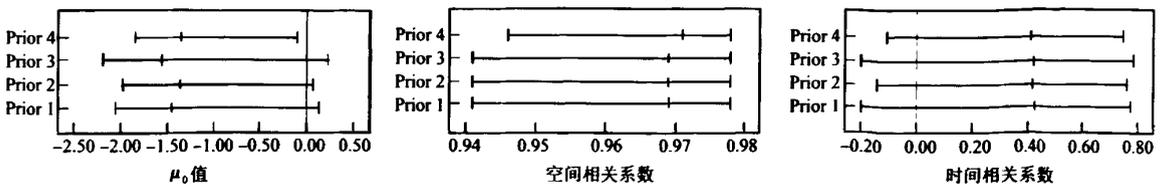


图 1 模型 1 中  $\mu_0$  值、空间和时间相关系数不同先验估计结果

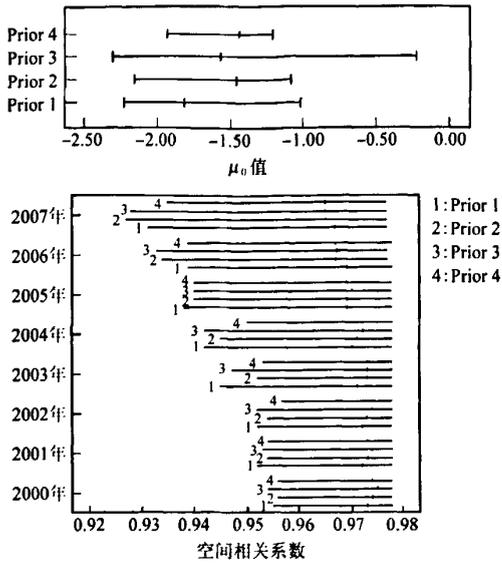


图2 模型2中 $\mu_0$ 值和空间相关系数不同先验估计结果  
 中标化死亡比(SMR)为似然法估计值,  $\theta$  为运用最佳贝叶斯时间模型估计 RR 的后验均数。

讨 论

本研究建立贝叶斯时空模型分析内蒙古自治区布病疫情数据,从定量角度研究其发病在空间、时间以及时空交互维度上的相关性程度,这些相关性信息蕴含在疫情时空数据中,是布病发病规律和分布格局的综合反映。模型结果表明 2000—2007 年布病发病空间相关性较强,即相邻旗县疫情的相似性较强,布病在内蒙古自治区的空间聚集性程度较高。相邻旗县发病的相似性随时间变化略有减弱,

表2 最佳贝叶斯时空模型参数估计结果

指 标	估计值(95%CI)
截距	-2.7120(-3.236, -2.142)
牛羊存栏数量	
2000年	-0.0130(-0.0160, -0.0070)
2001年	0.0021(0.0018, 0.0025)
2002年	0.0051(0.0050, 0.0054)
2003年	0.0062(0.0054, 0.0064)
2004年	0.0050(0.0049, 0.0052)
2005年	0.0047(0.0042, 0.0049)
2006年	0.0035(0.0031, 0.0036)
2007年	0.0028(0.0024, 0.0031)
地区类型	
农村和城市	-
半牧区	0.904(0.869, 0.920)
牧区	1.766(1.603, 1.828)
空间相关系数	
2000年	0.973(0.949, 0.978)
2001年	0.969(0.929, 0.978)
2002年	0.972(0.951, 0.978)
2003年	0.971(0.948, 0.978)
2004年	0.970(0.947, 0.978)
2005年	0.970(0.943, 0.978)
2006年	0.969(0.943, 0.978)
2007年	0.968(0.938, 0.978)

表3 2000—2007 年内蒙古苏尼特左旗布病发病 RR 值两种估计结果的比较

年份	SMR(95%CI)	$\theta$ (95%CI)
2000	0.00(0.00, 0.61)	0.00(0.00, 0.04)
2001	0.00(0.00, 0.60)	0.00(0.00, 0.10)
2002	2.10(1.15, 3.52)	1.94(1.16, 3.23)
2003	0.90(0.33, 1.96)	0.84(0.35, 1.72)
2004	22.01(18.60, 25.86)	21.92(18.67, 25.70)
2005	41.23(36.48, 46.43)	40.99(36.48, 45.99)
2006	40.96(36.24, 46.15)	40.92(36.33, 46.13)
2007	77.66(69.90, 84.77)	77.60(71.30, 84.59)

考虑与全区疫情整体上升迅速,聚集范围扩大有关,但聚集性整体上仍然维持在较高水平。通过观察布病发病危险度地图,老疫区疫情活跃,新疫区不断出现,疫情呈蔓延趋势,此分布特点与传统方法对内蒙古自治区布病疫情的描述相同<sup>[11]</sup>。但是,鉴于全区

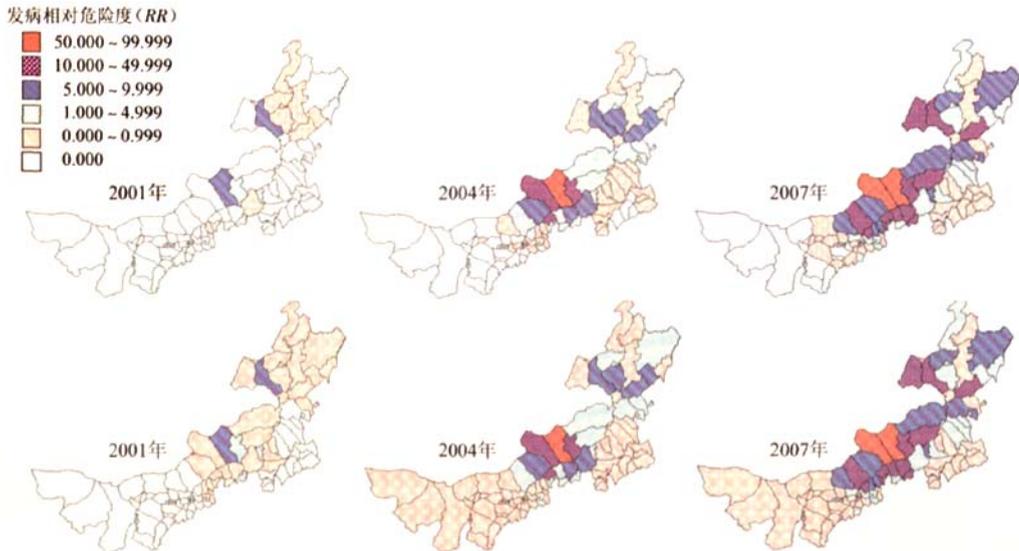


图3 内蒙古自治区不同年份布病发病的似然法(上)和贝叶斯法(下) RR 估计值地区分布

布病空间相似程度较高,疫情严重的区域保持相对聚集分布,疫情蔓延方向由严重区域到周边旗县呈放射状特点,值得相关部门关注。

内蒙古自治区布病疫情存在较强空间相关性,提示相邻区域可能存在共同的危险因素。而运用贝叶斯时空模型调整布病本身时空相关性效应后,可能会更加真实反映其他相关因素的效应。通过模型协变量的分析,布病流行与地区类型和牛羊存栏数量相关联,符合布病发病特点<sup>[12]</sup>。地区类型对布病发病影响较大,布病疫情空间相关性较强,可能与内蒙古地区布病发病分布高度集中有关。但是,牛羊存栏数量对布病发病的影响较小,且此种影响随时间变化趋势与布病发病和牛羊数量的变化均不大相符,提示可能存在其他重要的相关因素,值得进一步研究。本研究中2000年牛羊存栏数量对布病的影响作用为负值可能与当年大多数旗县发病数为0有关。

与经典统计方法相比较,建立贝叶斯时空模型分析内蒙古自治区布病时空格局及其演变时,优势主要体现在:①实现时空两个维度的同时分析,即布病发病的时间趋势和空间分布以及时空共变现象的分析及与时空关联的相关因素分析。②可修正经典标准化法所估计的RR值。它将数据蕴含的时空信息和参数的不确定性(先验分布)整合于同一模型中,借助邻近旗县和时点的布病发病信息对发病危险度进行估计,解决了空间结构带来的估计偏差,修正了标准化发病比在发病数较低时所暴露的弱点,使估计值更加稳定和可靠,从疾病地图上表现为发病危险度在全区范围内更加平滑<sup>[13]</sup>。

建立贝叶斯时空模型应注意:①关于相关参数先验分布的假定问题。贝叶斯时空模型中对参数先验分布范围的确定十分重要,关系到模型的可识别性问题<sup>[8]</sup>。正确合适的先验分布,能使所有的变量在后验分布中得到识别,但识别的程度取决于确定的变异范围。参数范围的确定必须要有尽可能多的变量能够被识别,包含可能的变异。即使是无信息先验,也应该根据模型的特点来选择合适的分布范围。因此,在本研究中,对时间、空间效应和时空交互效应的方差先验不能确定时,对不同的先验分布进行比较分析,选择最适合模型特点和数据的先验分布。②关于空间和时间效应的拟合问题。目前对于拟合何种形式的时空效应进行建模尚无定论。许多学者在贝叶斯空间模型的基础上拟合多种形式的时空效应以及时空交互效应,提出了多种形式的贝叶斯时空模型。本研究选择应用最广泛的两种形

式,即时间、空间效应相互独立和时空交互形式,其中时空交互形式更符合本研究数据特点。此外,贝叶斯时空模型的运算时间与模型拟合的复杂程度有关,本研究采用较高配置的计算机运行时间、空间独立效应的贝叶斯模型,每迭代1000次约需6 min;而运行时间、空间独立效应的贝叶斯模型,每迭代1000次约需40 min,收敛则需7 h。因此,拟合贝叶斯时空模型应考虑运行效率。

本研究受资料可获得性限制,目前最小分析单位只到旗县一级,时间单位为年,可能在某种程度上掩盖发病时空格局的真实特点。如可能,进一步的研究可选取不同的研究尺度,根据数据呈现的时空特点,考虑拟合多种形式的时间、空间以及时空交互效应。此外,本文对发病相关因素的探讨仅局限在地区类型和平均牛羊存栏数量,进一步深入探讨可考虑纳入其他相关因素。根据贝叶斯时空模型的优势,可用于挖掘蕴含在大量而复杂时空数据中的传染病流行特征和规律,但在不同传染病中的建模应用还有待研究。

#### 参 考 文 献

- [1] Cui BY. Epidemic surveillance and control of brucellosis in China. *Dis Surveil*, 2007, 22(10):649-651. (in Chinese)  
崔步云. 中国布鲁氏菌病疫情监测与控制. *疾病监测*, 2007, 22(10):649-651.
- [2] Mao SS. Bayesian statistics. Beijing: Chinese Statistics Publishing House, 1999: 1-34. (in Chinese)  
茆诗松. 贝叶斯统计. 北京: 中国统计出版社, 1999, 1-34.
- [3] Lawson AB. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Boca Raton: CRC Press, 2008: 257-262.
- [4] Thomas A, Best N, Lunn D, et al. GeoBUGS user manual, version 1.2. 2004. Available from: <http://www.mrc-bsu.cam.ac.uk/bugs>.
- [5] Box GEP, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control. 3<sup>rd</sup> ed. San Francisco, CA: Holden-Day, 1994.
- [6] Staubach C, Schmid V, Knorr-Held L, et al. A Bayesian model for spatial wildlife disease prevalence data. *Prev Vet Med* 2002, 56(1):75-87.
- [7] Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 2006, 1:515-533.
- [8] Elliott P, Wakefield JC, Best NG, et al. *Spatial epidemiology: methods and applications*. London: Oxford University Press, 2004:267-285.
- [9] Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit (with discussion). *J Roy Statist Soc B*, 2002, 64:583-639.
- [10] Armstrong BG. Comparing standardized mortality ratios. *Ann Epidemiol*, 1995, 5(1):60.
- [11] Guo WD, Chi HY. Epidemiological analysis of human brucellosis in Inner Mongolia Autonomous Region from 2002-2006. *Chin Trop Med*, 2008, 8(4):604-606. (in Chinese)  
郭卫东, 池海谊. 内蒙古自治区2002-2006年人间布鲁氏菌病流行特征. *中国热带医学*, 2008, 8(4):604-606.
- [12] Chen WJ, Cui BY, Zhang QH. Analysis of epidemic characteristics on brucellosis in Inner Mongolia. *Chin J Control Endem Dis*, 2008, 23(1):56-58. (in Chinese)  
陈文婧, 崔步云, 张庆华. 内蒙古自治区布鲁氏菌病流行50年特征分析. *中国地方病防治杂志*, 2008, 23(1):56-58.
- [13] Wang JF. *Spatial analysis*. Beijing: Science Publishing House, 2006:47-52. (in Chinese)  
王幼峰. 空间分析. 北京: 科学出版社, 2006:47-52.

(收稿日期:2010-08-06)

(本文编辑:张林东)