doi:10.3969/j.issn.2097-3853.2026.02.003

贝叶斯系统地理学:语系起源与扩散研究新视野

俞健,邓晓华

(福建理工大学 人文学院,福建 福州 350118)

摘要:贝叶斯系统地理学以贝叶斯统计为推理工具、系统地理学为理论基础,依托语言系统发育树与地理位置数据重建语系起源及扩散路径,弥补了语言古生物学、语言多样性原则等传统方法的局限。2010 年以来,该方法被广泛应用于语系研究,成功重建了南美洲 Arawak 语系、印欧语系、日本阿伊努语、非洲班图语系、澳大利亚 Pama-Nyungan 语系及亚洲汉藏语系等的起源与扩散历程,为若干长期争议问题提供了新的定量证据和解释框架。相关研究明确了其核心操作步骤,包括语言数据收集与同源编码、同源进化模型选择、先验信息设置、系统发育树推算及语系扩散模型构建,同时通过评估不同语言扩散模式下主流模型的适配性,划分了模型适用性等级与对应条件。为应对东亚语系复杂演化场景,该领域应推动贝叶斯系统地理学与语言速度场方法互补融合,构建“地理定位+时间校准”双重验证体系,整合语言、基因、考古数据,以进一步深化东亚诸语系跨学科研究。

关键词:贝叶斯系统地理学;语系;起源;扩散;重建

中图分类号:H0-0;O212.8　　**文献标志码**:A　　**文章编号**:2097-3853(2026)02-0118-11

Bayesian phylogeography: a new perspective on origins and diffusion of language families

YU Jian, DENG Xiaohua

(School of Humanities, Fujian University of Technology, Fuzhou 350118, China)

Abstract: Bayesian phylogeography, using Bayesian statistics as a reasoning tool and phylogeography as its theoretical foundation, reconstructs the origin and diffusion of language families using linguistic phylogenetic trees and geographical data, overcoming limitations of traditional methods like linguistic palaeontology and the linguistic diversity principle. Since 2010, it has been widely applied in the study of language families. It has successfully reconstructed the origin and diffusion process of Arawak language family in South America, Indo-European language family, Ainu language family in Japan, Bantu language family in Africa, Pama-Nyungan language family in Australia and Sino-Tibetan language family in Asia, which provides new quantitative evidence and interpretation framework for several long-term controversial issues. Relevant researches have clarified its core operation steps, including linguistic data collection and cognate coding, cognate evolution model selection, prior information setting, phylogenetic tree inference, and diffusion model construction. Adaptability evaluations of mainstream models across different diffusion patterns have clarified their applicability levels and conditions. To address the complex evolution of East Asian language families, the method should be integrated with the language velocity field approach to form a “geographical positioning + time calibration” dual verification system, integrating linguistic, genetic, and archaeological data for deeper interdisciplinary research of East Asian languages.

Keywords: Bayesian phylogeography; language family; origin; diffusion; reconstruction

收稿日期:2026-01-11

项目基金:国家社会科学基金项目(22CYY049);广西高校人文社科重点研究基地重大项目(2025JDZD012)

第一作者简介:俞健(1993—),男,浙江仙居人,讲师,博士,研究方向:语言人类学。

通信作者:邓晓华(1957—),男,福建连城人,教授,博士,研究方向:语言人类学。

语系起源和扩散的研究是理解人类及其文化历史的重要途径。语言古生物学与语言多样性原则是历史比较语言学定位语系起源的核心传统方法。语言古生物学通过重建原始语言,考察其词汇中特定地理区域的动植物术语及有考古对应物的文化项目来锁定起源地[1];语言多样性原则则将语系起源地与语言多样性最高的地区关联,认为语言分歧最显著的区域即为可能起源地[2]。然而,这两种方法存在明显局限。语言古生物学难以获取可靠的特定地理区域原始词汇,且古今气候地理环境变迁导致现代语言地理区域重建的原始词汇无法反映原始语言分布范围;语言多样性的形成不仅与人群迁徙分化、语言接触相关[3],还受人口密度、资源多样性、环境承载能力等因素影响[4-5],同时可能因空间饱和、稳定共存导致起源地多样化率放缓,或因后期迁徙造成多样性丧失,故无法简单以分歧数量界定语言多样性与起源地的直接关联。鉴于传统方法的缺陷,贝叶斯系统地理学提供了显著改进:它以重建的语言系统发育树为基础,结合现存语言点的地理位置,还原过去祖先语言的地理分布。2010 年以来,国外历史比较语言学界运用该方法在语系起源与扩散研究中成效卓著,解决了诸多长期争议问题。但该方法尚未引起国内语言学界的足够重视,目前仅少数学者借助其对汉藏语系汉语族、藏缅语族的起源地[6]及壮侗语族的扩散模式[7]开展研究,尚无其他学者将其应用于其他语系的相关问题研究。为此,本文在介绍贝叶斯系统地理学理论背景与相关模型迭代历史的基础上,梳理近十年国内外学界利用该方法开展的语系起源与扩散重建工作,总结其核心操作步骤,并对该方法的性能与适用条件进行评估,以期推动国内相关领域研究的发展。

一、贝叶斯系统地理学简介

贝叶斯系统地理学以贝叶斯统计为推理工具,以系统地理学为理论基础,旨在推断系统发育树的空间扩展,即所有祖先节点的空间位置以及生物或语言改变其位置的扩散速率[8],其中树主要由序列数据决定(如 DNA 序列、语言同源数据或形态字符),空间位置构成了地理扩散的路径,扩散速率将扩散路径缩放到地理和时间背景当中。

(一)理论背景

贝叶斯统计是一种统计学理论,源自 18 世纪数学家兼神学家 Thomas Bayes,他首次用贝叶斯推理对统计数据分析的重要问题进行数学处理。数学家 Pierre Simon 开创并普及了贝叶斯概率统计,进一步发展了其概率解释。20 世纪,因哲学与实践层面的考量,该方法未获多数统计学家认可。但随着强大计算机及马尔可夫链蒙特卡罗算法(Markov chain Monte Carlo,MCMC)等新算法的出现,贝叶斯方法在 21 世纪的统计学中应用愈发广泛。[9]其核心是借助贝叶斯定理,结合先验知识,如先前实验结果、考古事件、文献记录等[10],在获取新数据后计算并更新概率,将概率视为对事件的信任程度,且能直接把量化置信度的概率分布分配给参数或参数集。

系统地理学(phylogeography)源于群体遗传学与系统学的交叉研究[11],其核心是在群体遗传学基础上,探究谱系(尤其种内水平谱系)从古至今地理分布的过程与成因[12]。随着地理变异遗传标记研究范围的扩大,该学科目标进一步明确为理解地理或时空背景下的微观进化与物种形成[13],并能通过相关视角推断人口扩张、迁移、地域差异等历史事件[14]。该学科发展初期存在明显局限,彼时更侧重“系统”(phylo)成分,对空间“地理”(geographic)成分的处理较为简单。这一现象的关键原因是,21 世纪初地理信息系统(geographic information system,GIS)尚未被演化生物学家广泛开发应用[15]。21 世纪后,GIS 在进化生物学中的重要性日益凸显,学界开始探索地理信息与种群系统发育的深度融合路径。[13]传统研究中,仅能将基于变异数据得到的系统发育树、网络等进化关系,与环境变化、景观结构等外部信息进行定性比较,进而推断历史情景。而二者融合后,GIS 可从外部信息中构建结构化历史假设,同时从基因或基因性状的地理变异中精准推断历史,地理背景也由此成为构建真实生物进化模型的关键要素。

(二)统计模型的迭代

早期系统地理学常采用系统发育树的嵌套分支(nested clade)分析线粒体数据[16],但该方法未基于统计模型,无法量化历史估计的统计可信度[17],且生态信息等数据难以直接纳入分析。随着景观遗传学、溯祖理论的发展,以及生态位与古

气候模型工具的开发,系统地理学实现了精细地理尺度的研究突破,能够探究地理、气候、生态等因素对基因流动的影响及种群地理结构的形成[18],进而明确基因流动的可能地理路径与迁移潜在障碍[19]。不过,这些新方法虽提升了统计严谨性,却仅适用于可生成简单先验假设的系统,难以应对复杂系统地理历史场景。为此,Lemmon A R 和 Lemmon E M 提出连续景观下的系统地理学历史似然框架,既可检验明确先验假设,也能在无假设时估计系统地理历史及其统计可信度。[20]该框架借助系统发育树与树枝叶子节点的个体地理坐标,通过空间显式迁移随机行走模型(random walk,RW),描述谱系树内部节点个体的地理坐标及平均每代扩散距离,最终推算出抽样个体祖先的地理位置。

经典系统地理学方法存在显著局限,其忽略进化过程与时空的相互作用,通常先重建不含空间信息的系统发育树,再基于系统地理条件进行推断,难以洞察特定时间背景下的空间动态。[21-22]而要明确系统发育树起源地及不同地点间的演化关系,需完整重建进化史上的扩散模式与过程,进化概率模型为此提供了突破路径。该方法借助明确的特征进化模型,可呈现整个系统发育过程中的完整特征历史,便于得出统计推论。[23]在此基础上,Lemey 开发了基于贝叶斯框架的地理发育系统。[24]其通过对特定时间尺度的系统发育进行采样,在贝叶斯软件中实现特征映射(character mapping),进而重建物种时空传播模式,同时适配系统发育的不确定性。该模型采用连续时间马尔可夫链模型(continuous time Markov chain,CTMC)处理离散特征(discrete characters)进化,基于最大似然重建计算系统发育叶子节点(leaf nodes)观察到的条件概率,涵盖各祖先特征的分支长度差异。但该离散特征进化模型存在不足,未明确模拟连续空间中的扩散过程,且对共同祖先地理位置的推断局限于采样点集合,而样本多为连续分布,并不适配离散抽样方案。对于连续地理坐标,布朗扩散模型被认为与马尔可夫链转换模型相近[25],但严格的布朗扩散模型假设系统发育在空间运动中保持均匀性。为解决这一问题,Lemey 等放宽分子钟模型的速率恒定假设[26],基于贝叶斯框架构建多变量布朗扩散模型,将其与同源序列演化标准模型相拟合,把连续空间的系统地理学嵌入完整概率模型,提出贝叶斯松弛随机行走模型(relaxed random walk,RRW)以推断连续系统地理历史。该模型中系统发育树各分支的扩散率标量均从潜在离散化率分布中独立绘制。[27]此外,De Maio 等针对离散特征模型推断迁移速率和根位置时的极不可靠性及对有偏抽样的敏感性问题[28],在 BEAST2 中引入贝叶斯结构聚结近似框架(Bayesian structured coalescent approximation)。该方法融合了结构化溯祖方法的准确性[29]与处理大规模种群数据的高计算效率,核心思路是有效集成所有可能的迁移历史,聚焦探索关键参数,减少总体计算量。不过,连续空间的系统地理学仍有局限,如系统发育树叶子节点的地理位置常为大致范围,但上述模型均以地理质心作为位置点估计值。若要提升真实性,需将叶子节点位置在不同地理区域内整合,并纳入地理景观异质性数据。

值得注意的是,Lemey、De Maio 等提出的贝叶斯地理系统发育模型,均依赖平面扩散过程,未考虑地球球形性质,导致类似墨卡托投影的扭曲失真。对此,Bouckaert 提出基于球面扩散的贝叶斯系统地理分析框架[30],给出可有效计算的球面扩散近似方法,可在 MCMC 算法框架中使用并在 BEAST2 中实现。该框架支持从区域采样叶子节点,能为树分支地理位置设置先验信息,实现发育树与地理信息联合推理,且结合非匀速随机行走模型,可区分陆地与水域、森林与沙漠等不同景观的扩散速率,更贴合真实扩散过程。

(三)数据可视化

将系统发育与系统地理学模型可视化到地图背景,是系统地理学的关键环节。[31]常见方式为基础地图上的简单制图,以饼图、缩放点或等值线呈现特征数据,Hoffmann 等就提供了遗传特征与分化指数的制图示例[32],此类图表可通过 GIS 创建。系统发育树和网络的可视化常依托地理背景,借助 cartographer、phylogeographer 等软件导入含地理坐标的谱系树并显示,通常需扭曲树枝长度以适配地理空间。另一种思路是扭曲地图适配系统发育树,依据生物体基因流动可视化地理扩散,通过颜色编码树与地图的链接窗口,实现复杂地理层次的交互式可视化,选择树的层级可生成对应树枝地理分布图[33]。此外,多数 GIS 具备三维空间数据可视化工具,可实现空间二维与时间

二维的时空数据可视化。例如 Kidd 和 Ritchie 就将三种墨西哥淡水鱼的系统发育树,与中新世、上新世及现代三个时间层次的地理海拔、河流和湖泊数据进行了可视化拟合。[13] 随着贝叶斯系统地理学方法的发展,系统地理学可视化被提出了更高的要求。为此,Bielejec 等开发了 SPREAD 跨平台应用程序[34],用于分析和可视化物种时空扩散的贝叶斯系统地理学重建。SPREAD 提供了离散树(discrete tree)、离散贝叶斯因子(discrete Bayes factors)、连续树(continuous tree)和时间切片器(time slicer)四个模块来分析和可视化系统地理扩散的不同方面。

整体而言,贝叶斯系统地理学模型迭代围绕三个方面展开:第一,突破空间处理局限,从离散到连续、从平面到球面;第二,强化统计可信度,从无量化到概率推断;第三,适配复杂场景,从单一速率到异质性速率,从简单数据到大规模数据。早期嵌套分支分析无统计支撑,仅适用于简单离散数据;后续模型逐步拓展至连续空间,引入概率推断提升结论可靠性;球面扩散框架则攻克平面投影扭曲难题,同时适配速率异质性与大规模数据研究。各模型针对性弥补前序局限,为语系起源与扩散研究提供分层适配的量化工具。

二、贝叶斯系统地理学重建语系的起源和扩散及其操作步骤

贝叶斯系统地理学最初是为了重建病毒的空间扩散[26,35],,其应用于语言与生物演化研究存在本质区别。生物演化以遗传物质垂直传递为核心,扩散与种群迁徙、地理隔离等自然过程强绑定,适配连续匀速的自然扩散逻辑。而语言作为社会文化系统,演化兼具垂直谱系分化与水平接触融合属性,深受语言替代、文化认同等因素影响,并非单纯“人口迁移函数”。生物扩散的连续性与布朗扩散等模型假设契合,但语言扩散常呈“非连续性”“跳跃式替代”特征,违背了模型对均匀连续空间运动的预设。此外,生物“特征丢失”多为自然选择结果,而语言同源词消失、语法趋同可能源于语言接触而非谱系分化,这要求将语言学问题意识纳入模型适配与解读,避免直接套用生物或病毒地理学框架而忽略语言的社会文化本质。尽管存在语言数据与贝叶斯系统地理学的非兼容性,该方法仍被广泛应用于探索语系的扩散历史,不同程度上重建了南美洲的 Arawak 语系[36]、印欧语系[37]、日本的阿伊努语[38]、非洲的班图语系[39]、澳大利亚的 Pama-Nyungan 语系[40]和亚洲的汉藏语系[6-7]等语系起源和扩散的历史。在介绍贝叶斯系统地理学重建这些语系起源和扩散的基础上,本文对其重建的基本步骤进行总结,具体如下。

(一)语系起源与扩散的贝叶斯系统地理学重建

1.Arawak 语系

Arawak 是一个分布在南美洲低地、地理上分散的语系。Walker 和 Ribeiro[36] 通过编码 60 种 Arawak 语言和方言的基本词汇的同源词来生成系统发育树,然后使用一个贝叶斯松弛随机行走模型(RRW),以现存语言的经纬度位置信息作为推断扩散过程的最终结果,推断出 Arawak 语系有大西洋海岸和亚马逊西部两个潜在起源地。然后其使用 Bayes Traits 平台[41] 中的伽马分布的可逆跳跃超先验(reversible-jump hyper prior,RJHP)重建一个在系统发育上的离散扩散过程,离散的 RJHP 模型在亚马逊西部起源上给出了最高的平均后验概率。亚马逊西部由于受到 RRW 和 RJHP 两个地理系统扩散模型的支持而被判定为 Arawak 语系最可能的起源地。基于对 Arawak 语系贝叶斯系统地理学的推断,Walker 和 Ribeiro 提出了 Arawak 语系地理时空扩张最可能的历史过程:Arawak 语系在亚马逊西部起源后沿着马代拉河(Madeira)或普鲁斯河(Purus)顺流到亚马逊河,继而扩散到东北部的帕利库尔(Palikur)和马拉万(Marawan)河口附近,然后向南沿陆路迁移至巴西南部和中部,向北走水路迁移至加勒比(Caribbean)及其周边地区,最后从亚马逊中部河流区域出发沿里约内格罗河(Rio Negro)扩散至亚马逊西北部。

2.印欧语系

印欧语系的起源有庞蒂克草原起源[42] 和安纳托利亚起源[43] 两种假说。为了在两种假说之间进行测试,Bouckaert 等[37] 改编并扩展了 Lemey[26] 等人提出的从分子序列数据研究病毒暴发起源的贝叶斯系统地理学推理框架,将 103 种古代和现代印欧语言的基本同源词汇和地理范围匹配为数据集,将语系的系统发育过程建模为同源词在时间维度上的获得与丢失过程,将系统发育

推断方法与松弛随机行走模型(RRW)相融合,把语言位置视为一个连续的向量(经度和纬度),沿着树的分支随时间演变,进而推断系统发育树内部节点和根节点上祖先语言的地理位置。值得注意的是,为了模拟语系时空扩张的真实感,Bouckaert 等扩展了 RRW 模型。[37] 首先,使用语言的地理范围而不是具体的经纬度信息以避免语言位置分配的不确定性。其次,考虑到地理异质性,在根节点和内部节点位置上给定空间先验分布,如纳入有关欧亚大陆形状的先验信息以及对通过水域位置分配零概率等。在 RRW 模型下,印欧语谱系树的根位置的估计后验分布位于现在土耳其安纳托利亚地区,并且推断出内部节点上印欧语系各祖先语支的地理范围。

3.阿伊努语

阿伊努语是一种曾经在日本北部繁盛的土著群体所使用的濒危语言。传统观点认为现代阿伊努人是来自距今 1 万年前的东南亚[44],但考古学和遗传学证据的最新证据表明,阿伊努人是西伯利亚狩猎采集者鄂霍次克人在大约 900—1 600 年前迁移到北海道北部后重大基因遗传和文化贡献的结果[45-46]。Lee 和 Toshikazu[38] 从语言学角度根据保存的 19 种阿伊努语,使用 SPLITSTREE4 生成了阿伊努人系统发育树,然后在 BEAST 平台中使用连续随机行走模型(continuous random walk,CRW)对阿伊努语时空演变进行贝叶斯系统地理学方法重建,结果与鄂霍次克人扩张并塑造阿伊努人及其文化的假设吻合,它们是大约 1 300年前从北海道北部传播的共同祖先的后代。

4.班图语系

历史上的人类迁徙是沿着熟悉的栖息地还是相对独立于这些栖息地,这一问题对判定人类迁徙路径至关重要。对班图语起源与扩散路径的争论,为这一问题提供了参考。Grollemund 等[39] 利用贝叶斯系统发育方法生成了 409 种班图语同源词汇集的系统发育树,并通过考古数据进行日期校准。然后他们在 BayesTraits 软件中使用布朗运动模型,从每种当代班图语的地理位置数据出发,推断后验样本中树的内部节点可能的祖先地理位置,最后使用这些重建节点地理信息来推断班图语人群的起源和传播路径。班图语人群的起源地最终定位在喀麦隆西北部的大草原,首先向东南方向移动,然后沿着刚果雨林南部边界向东穿过。研究发现,班图语人群的扩张并不是随机行走,而是沿着新兴的草原走廊迁徙,当人群确实从大草原迁移到雨林时,迁移速度便会减慢。Grollemund 等的研究反驳了 Currie 等[47] 同样使用贝叶斯系统地理学推断班图语穿过刚果雨林的扩张路径,说明不熟悉的栖息地会极大地改变人类迁徙的路线和速度。

5.Pama-Nyungan 语系

澳大利亚大陆有 28 个语系,其中 27 个限制于北部地区,而 Pama-Nyungan 语系覆盖了 90%的澳洲大陆,是全球最大的狩猎采集者的语系。该语系起源何处以及如何占领澳洲大陆大部分地区,语言学领域为此提出了快速替代假说[48]、全新世早期强化假说[49]、南极后冷逆转假说[50] 和初始殖民假说[51]。Bouckaert 等[40] 将 306 种 Pama-Nyungan 语言的基本词汇数据与贝叶斯系统地理方法相结合,模拟了该语系在澳洲的扩展,并对 4 种起源假说进行测试,发现 Pama-Nyungan 语系起源于全新世中期卡奔塔利亚湾平原地区(Gulf Plains of Carpentaria)并快速扩散替换其他语系的假说得到强有力支持。Bouckaert 等[40] 还引入系统地理学的创始人扩散模型来更自然地捕捉 Pama-Nyungan 语系扩散的过程。该模型模拟语言分化时一个支系迁移殖民新区域、另一个支系留守,更贴合真实族群迁徙。

6.汉藏语系

汉语族和藏缅语族的共同祖先起源有中国北方起源[52]、中国四川起源[53]、印度东部起源[54] 三种假说的争议。张梦翰团队[6] 使用贝叶斯系统发育方法建立了 109 种汉藏语的系统发育树并估计根的时间深度,同时基于 Bayes Traits 平台,利用系统地理学方法对汉藏语系起源地的进行概率密度估计,最终得出汉藏语系起源于新石器时代晚期中国北方。此外,壮侗语族的起源和扩散存在贵州内陆起源假说[55] 和华南沿海起源假说[56] 两个相互争议的假说。为了推断壮侗语人群的起源与扩散路线,张梦翰团队[7] 将 100 种壮侗语言的地理分布分为贵州内陆、云南内陆、沿海地区(广西—广东)、海南岛和大陆东南亚五个不同的区域,基于 Bayes Traits 平台,采用贝叶斯系统发育比较方法进行了离散系统地理学推断,发现沿海地区是壮侗语最可能的起源地。张梦翰团队进一步利用贝叶斯可逆跳跃马尔可夫链蒙特卡

罗方法(Bayesian Reversible-jump Markov Chain Monte Carlo)推断壮侗语的传播路径,最佳模型强烈支持这样一种情景:以两广交界作为扩散中心,壮侗语族跨越琼州海峡传播到海南岛,向西北扩散到云南省和贵州省的内陆地区,再向西南扩展到大陆东南亚,还有一些从沿海地区扩散到大陆东南亚。

综上,不同语系的扩散模式在驱动逻辑、空间路径与景观适配性上呈现显著差异。部分语系以地理廊道为核心依托,如班图语系和 Arawak 语系;部分由人群定向迁徙主导,如阿伊努语;还有的以快速替换为核心特征,如 Pama-Nyungan 语系。扩散方向上,既有单一起源地的渐进式定向延伸,如印欧语系,也有核心区域的跨域多路径辐射,如壮侗语族。这些差异是地理环境、人群活动与语言互动的综合体现,既展现了语系演化的多样性,也凸显了贝叶斯系统地理学根据具体场景灵活适配模型的优势。

(二)贝叶斯系统地理学重建语系起源与扩散的操作步骤

1.语言数据收集与同源编码

语言的复杂性在系统发育研究中通常被分解为一组变量,变量可以是同源形式的语法特征、语音特征或词汇特征,但一般选用同源词汇特征。因此数据收集就是同源词词汇数据库的建构,故尽可能搜集目标语系各个语支所有语言的基本词汇表,用存在(1)或不存在(0)编译同源集合的二进制矩阵。

2.模型选择

估计树的可能性需依托同源进化模型,描述同源词在树节点上的存在或不存在概率,作为祖先节点的状态和计算分支的长度的函数。目前主流的同源进化模型主要有三种。[26]一是二进制连续时间马尔可夫链模型(CTMC),其假设数据由两种状态的时间可逆马尔可夫过程生成,“0”和“1”的分布遵循稳态分布,二者可相互转换。反映在同源词进化上,即每种语言能以固定比率获得或丢失同源词集,比率基于当前观察数据确定。二是协变模型(Covarion),包含两层状态:表层为“0”或“1”,隐藏层含 slow-0、fast-0、slow-1、fast-1 四种状态。转换可能发生在快速状态和慢速状态之间,也可以发生在“0”和“1”状态之间。该模型允许特征稳态时以慢速“背景”速率演化,外部事件(如语言接触)触发下转为快速爆发式变化,更贴合同源词长期稳定、偶发突变的真实演化状态。三是随机多洛模型(Pseudo-Dollo),假设特征以速率 λ 仅获得一次,以速率 μ 永久丢失,其演化具有定向性,既不满足时间可逆性,也不处于平衡状态。

3.先验信息设置

在 BEAUti 程序中选定进化模型并导入含特征矩阵的#nexus 文件后,需完成先验信息设置,核心包含四个方面。(1)位点模型(site model):语言在不同谱系、数据子集的变化率存在显著差异[57],可通过修改 Gamma 类别计数,实现树的各位点与分支间变化速率的异质性控制。(2)时钟模型(clock model):核心功能是将观察到的变化数量转换为时间维度,进而精准估计系统发育上亚群的年龄。(3)树先验(tree prior):常用 Yule 模型与 Birth-death 模型,前者为纯出生过程,假设数据集包含该语系所有语言,含灭绝语言则会导致多样化率与节点年龄估计偏倚;后者应用更广,兼具出生率(控制多样化事件)与死亡率(控制语言灭绝)参数,支持基于不完整语言抽样估计树的谱系结构与时间深度。(4)时间校准(time calibrations):需可靠校准点估算所有节点年龄,可通过设置语言子群最近共祖时间先验或采用古代语言年龄实现,校准点信息源自历史文献、铭文、人群迁徙事件等,若子群属于独立地理单元,还可参考考古记录。

4.系统发育树推算

在完成模型选择和先验信息设置后,需要在 BEAUti 程序中选择 MCMC 选项中的迭代次数,然后将不同模型与先验参数的组合导入 BEAST 程序进行迭代运算。BEAST 使用 MCMC 算法来探索树和参数的空间并返回其后验分布的样本。从初始随机树和一组参数开始,马尔可夫链依次迭代地对树和参数提出小的更改。在初始“磨合”阶段,MCMC 算法需要时间来找到后验概率最高的区域。此后,每一步都是根据后验分布从树和参数空间中抽取的样本。由于这些是高度相关的,BEAST 将它们记录到#log 文件和#trees 文件中,这些文件即为后验样本。最后,需在所有模型组合中选择一个与数据集的拟合度最高的模型组合。贝叶斯模型选择基于边际似然(marginal likelihood),可以通过 Tracer 分析窗口的模型比较

选项中的 Harmonic Mean Estimator 来实现。[58] 基于上述步骤即可得到目标语系的贝叶斯系统发育树和时间深度。

5.语系扩散模型

语系扩散模型基于某一语系的贝叶斯系统发育树和语言采样点的地理信息。首先收集目标语系每种语言的近似质心的经纬度数据或者地理范围数据,然后将语言扩散建模为沿着树分支的连续空间中随机扩散,最好是通过球体上的扩散来捕捉。这种方法将系统发育树叶子节点上的采样语言的同源数据和地理位置数据结合起来,共同推断祖先关系和祖先节点的位置。通过这种方式,同源词和地理学可共同告知树的后验分布和推论祖先节点位置在树拓扑中的不确定性。但是随机行走并不意味着每种语言都是随机移动而不受社会、政治或生态因素的影响,为了识别和模拟与景观特征相关的迁移速率变化,可以加入景观感知迁移模型。

三、贝叶斯系统地理学重建语系起源和扩散的适用性评估

虽然贝叶斯系统地理学方法已成功应用于诸多语系起源地重建,但是该方法对不同语言空间扩散场景的适配性差异显著,部分重建结论与传统假说冲突,其适用性边界亟待明确。核心在于如何区分语言扩散模式,进而划分各类模型适用性等级和使用条件。

(一)重建模型与语言扩散模式

Neureiter 等[8] 通过模拟扩张(expansion)与迁移(migration)两种语言空间扩散场景(均含不同方向趋势),针对三类主流贝叶斯系统地理学模型(松弛随机游走模型,relaxed random walk;恒定定向随机游走模型,constant directional random walk;松弛定向随机游走模型,relaxed directional random walk)的性能展开评估。在扩张模式中,语言以网格区域增长(grid-based region-growing)的方式逐步占据邻近区域,扩散方向受山脉、海洋、沙漠等地理约束限制,原居地始终有语言群体留存(如班图语系随农耕技术的扩散)。此模式下,三类模型均展现出高适用性,即便存在显著的方向趋势,重建结果的偏差仍能稳定在较低水平。值得注意的是,RRW 模型表现尤为稳健,模型适用性不受树规模、语言区域重叠等因素的显著影响。在迁移模式中,语言群体完全脱离原居地,以定向随机游走(directional random walk)的方式转移至新区域,原居地语言因迁徙或语言替代而消失(如南岛语族从华南大陆向台湾岛的迁移)。此模式下,三类模型的适用性均大幅下降:缺乏历史数据(如古代语言位置记录)时,重建偏差随扩散方向趋势呈线性增长,模型对不确定性的估计易出现极端偏差。为辅助真实语言研究中的模型选择,Neureiter 等[8] 提出三类描述性统计量:分支重叠度(clade overlap)衡量不同系统树分支语言的空间重叠程度,多样性—空间依赖性(diversity-space dependence)反映语言分化与空间扩散的关联强度,树不平衡度(tree imbalance)量化系统树拓扑的不对称性。三者结合可明确目标语系的扩散模式,若统计量符合扩张模式特征,优先选用松弛随机游走模型;若倾向迁移模式,则需谨慎使用贝叶斯系统地理学方法,或通过补充全时段历史数据、结合考古与遗传证据提升结果可信度。

(二)重建模型的适用性等级和适用条件

Wichmann 和 Rama[59] 通过合成数据(模拟含20种语言的语系、100项词表及语言群体空间迁移)系统测试6种语言发源地检测方法及3种基线方法,并给出适用性等级和适用条件。第一层级为最优适用模型,包含两类核心模型,适用于数据条件较完备或需快速生成可靠假说的场景。一是 Bayes Traits 固定速率模型,其以语言系统发育树、各语言地理坐标及同源词数据为输入,通过将经纬度转换为三维笛卡儿坐标,适配地理运动的布朗运动假设,能输出具有明确可信度的起源地范围,尤其适合对推断精度要求高的研究。二是最小距离基线(minimal distance baseline),无需任何语言相关数据,仅通过计算每种语言到其他所有语言的平均地理距离,将平均距离最小的语言位置认定为起源地,操作简便且数据依赖度低,可作为起源地假说的快速生成工具。第二层级为中等适用性模型,涵盖五类模型,适用于数据条件有限或可接受一定推断偏差的场景:一是 RevBayes 固定速率随机游走模型(revbayes fixed rate random walk)与 RevBayes 可变速率随机游走模型(revbayes variable rate random walk),二者均以系统发育树和地理坐标为基础,核心差异在于对分支演化速率的假设需同源词数据支撑,适用于需灵活调整速率假设的研究;二是 Bayes Traits 可变

速率模型，通过松弛布朗运动模型允许分支长度伸缩，同样依赖同源词与系统树数据，适合语言群体迁移速率存在差异的场景；三是 BEAST2 松弛随机游走模型（BEAST2 relaxed random walk），采用联合推断框架同步构建系统树与推断起源地，需词表与地理坐标数据，适用于系统发育树尚未明确的研究；四是多样性法（diversity method），无需同源词与系统树，仅以语言词表为基础，通过计算基于归一化莱文斯坦距离（levenshtein distance normalized）的语言距离与地理距离比值，确定多样性指数最高的语言位置为起源地，适合无同源词数据的场景。第三层级为低适用性模型，包含两类基线工具，仅适用于缺乏关键数据、需最低限度参考依据的场景：一是随机基线（random baseline），通过随机选择某一语言的位置作为起源地，结果与实际数据无关联性，仅可作为性能对比的参照；二是质心基线（centroid baseline），以现存所有语言分布范围形成的多边形质心作为起源地，仅依赖地理分布的宏观特征，结果可靠性较低，难以支撑严谨的起源地推断。

四、东亚诸语系起源与扩散研究展望

贝叶斯系统地理学的核心优势在于强大的统计推理能力与跨领域数据整合潜力，能依托语言系统发育树与地理坐标，精准量化语系起源地与扩散路径。目前，生态、古气候、景观遗传等多维度数据已逐步纳入其分析框架，为解析语言演化的复杂时空动态提供了更全面支撑。然而，该方法存在明显的结构性局限与方法论风险：一是同源编码强依赖研究者对语音、语义演变的经验积累，缺乏统一标准，细微判定差异会引发数据偏差，影响推断精度；二是谱系树模型与语言演化现实不兼容，在语言演化高度网络化、存在强烈接触扩散的语区，易产生系统性偏误；三是地理先验设置对结果约束性强，空间范围界定、地理障碍权重分配等依赖现有考古或族群认知，其不确定性可能引导模型向预设方向收敛，形成循环论证，削弱结论客观性。这些局限在东亚语系研究中尤为凸显，东亚大陆上汉、藏缅、南亚、苗瑶、侗台等语言群体演化场景复杂，历史上频繁的族群接触导致语言横向传递普遍存在[60]，东亚岛屿上的南岛语系起源于华南沿海却完全退出该区域[61-62]，传统贝叶斯系统地理学依赖的谱系树模型难以适配此类复杂场景，亟须结合前沿方法突破现有局限。

语言、基因与考古的跨学科深度融合是未来东亚语系起源与扩散研究的核心方向。在过往研究中，考古数据与基因数据多被视为旁证，或仅用于贝叶斯系统地理学模型的年代校准，未能与语言特征数据形成有机整合，导致跨学科研究的解释力未能充分释放，难以全面破解东亚地区复杂的语言演化与人群迁徙互动关系。Yang 等[63]提出的语言速度场（Language velocity field）方法，为这一困境提供了创新解决方案。该方法无需依赖传统贝叶斯系统地理学所必需的系统发育树，而是通过主成分分析对语言特征进行降维，结合动态模型捕捉语言垂直分化（如谱系树反映的亲缘关系）与水平接触（如词汇借用、语法趋同）的双重信号。在汉藏语系这类高频接触场景中，语言速度场的重建结果展现出显著优势，其推断的汉藏语系扩散路径与古代 DNA 揭示的人群迁徙路线、考古农耕遗址的空间分布高度契合，大幅提升了跨学科证据的关联性与一致性。针对部分语系起源地的长期争议（如南岛语系的“台湾起源说”与“华南沿海起源说”），可构建“语言速度场地理定位 + 贝叶斯时间校准”的双重验证体系：利用语言速度场的辐射型扩散中心识别功能，确定语系可能的地理起源范围；结合贝叶斯系统地理学的时间校准模块，为起源地赋予精确的时间维度，形成“地理—时间”双轴验证。对于已灭绝语言（如华南沿海古代百越语言）导致的起源地信号缺失问题，语言速度场同样提供了新的解决思路：通过分析灭绝语言周边现存语言（如百越语言相关的壮侗语族和南岛语族语言）的扩散轨迹，利用语言速度场的动态模型反推灭绝语言的可能分布范围；结合该区域考古遗址的空间分布、出土器物与基因遗存，补充起源地推断的关键线索，填补因语言灭绝造成的研究空白。

未来研究需以统一跨学科统计框架为核心，推动贝叶斯系统地理学与语言速度场方法深度互补。具体而言，将语言、基因、考古三类数据纳入同一统计推断框架，实现联合建模。语言数据作为核心输入，提供谱系亲缘关系与地理分布基础信号，通过同源特征降维捕捉垂直分化与水平接触双重动态；基因数据约束地理扩散路径参数，以人群遗传迁徙轨迹界定语言扩散潜在廊道，同时

通过遗传分化时间校准语言谱系树关键节点年龄;考古数据优化时间校准精度,以文化层年代锚定语言分化区间,借景观适应性指标调整扩散速率异质性参数。这一逻辑可有机融合三类数据,突破传统谱系树模型局限,精准适配东亚语言接触频繁、人群互动复杂的演化场景,填补灭绝语言研究空白,推动多学科深度交叉融合,为东亚诸语系起源与扩散研究提供更具说服力的量化解释框架。

参考文献:

[1] DIFFLOTH G. The contribution of linguistic palaeontology to the homeland of Austroasiatic[C]//SAGART L,BLENCH R,SANCHEZ-MAZAS A. (Eds). The Peopling of East Asia:Putting Together the Archaeology,Linguistics and Genetics. London: Routledge Curzon, 2005:77-80.

[2] SAPIR E. Time perspective in aboriginal American culture:a study in method[C]//Geological Survey Memoir 90: No. 13, Anthropological Series. Ottawa: Government Printing Bureau,1916.

[3] COUPE A R. Linguistic diversity and language contact in Nagaland[M]. Through the Ages:The Naga Condition,2021.

[4] GREENHILL S J. Demographic correlates of language diversity[C]//BOWERN C, & EVANS B (Eds.). The Routledge Handbook of Historical Linguistics. Abingdon,UK:Routledge, 2014:555-578.

[5] PACHECO COELHO M T,PEREIRA E B,HAYNIE H J,et al. Drivers of geographical patterns of North American language diversity[J]. Proceedings of the Royal Society B,2019,286(1899):20190242.

[6] ZHANG M H,YAN S,PAN W Y,et al. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic [J]. Nature,2019,569(7754):112-115.

[7] TAO Y X,WEI Y C,GE J Q,et al. Phylogenetic evidence reveals early Kra-Dai divergence and dispersal in the late Holocene[J]. Nature Communications,2023,14:6924.

[8] NEUREITER N,RANACHER P,VAN GIJN R,et al. Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? [J]. Royal Society Open Science,2021,8(1):201079.

[9] FIENBERG S E. When did Bayesian inference become "Bayesian"? [J]. Bayesian Analysis,2006,1(1):1-40.

[10] MCELREATH R. Statistical rethinking:a Bayesian course with examples in R and Stan[M]. Boca Raton: Chapman and Hall/CRC, 2018:18-41.

[11] AVISE J C,ARNOLD J,BALL R M,et al. Intraspecific phylogeography:the mitochondrial DNA bridge between population genetics and systematics[J]. Annual Review of Ecology and Systematics,1987,18 (1987):489-522.

[12] AVISE J C. Phylogeography:the history and formation of species[M]. Cambridge:Harvard University Press, 2000.

[13] KIDD D M,RITCHIE M G. Phylogeographic information systems:putting the geography into phylogeography[J]. Journal of Biogeography,2006,33(11):1851-1865.

[14] AVISE J C. The history and purview of phylogeography:a personal reflection[J]. Molecular Ecology,1998,7(4):371-379.

[15] SWENSON N G,HOWARD D J. Do suture zones exist? [J]. Evolution,2004,58(11):2391-2397.

[16] TEMPLETON A R. Nested clade analyses of phylogeographic data:testing hypotheses about gene flow and population history[J]. Molecular Ecology,1998,7(4):381-397.

[17] PANCHAL M,BEAUMONT M A. The automation and evaluation of nested clade phylogeographic analysis[J]. Evolution, 2007,61(6):1466-1480.

[18] CORANDER J,SIRÉN J,ARJAS E. Bayesian spatial modeling of genetic population structure[J]. Computational Statistics, 2008,23(1):111-129.

[19] STORFER A,MURPHY M A,EVANS J S,et al. Putting the "landscape" in landscape genetics[J]. Heredity,2007,98 (3):128-142.

[20] LEMMON A R,LEMMON E M. A likelihood framework for estimating phylogeographic history on a continuous landscape [J]. Systematic Biology,2008,57(4):544-561.

[21] TEMPLETON A R. Statistical phylogeography:methods of evaluating and minimizing inference errors[J]. Molecular Ecology,2004,13(4):789-809.

[22] HOLMES E C. The phylogeography of human viruses[J]. Molecular Ecology,2004,13(4):745-756.

[23] PAGEL M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies[J]. Systematic Biology,1999,48(3):612-622.

[24] LEMEY P,RAMBAUT A,DRUMMOND A J,et al. Bayesian phylogeography finds its roots[J]. PLoS Computational Biology,2009,5(9):e1000520.

[25] SCHLUTER D,PRICE T,MOOERS A Ø,et al. Likelihood of ancestor states in adaptive radiation[J]. Evolution,1997,51(6):1699-1711.

[26] LEMEY P,RAMBAUT A,WELCH J J,et al. Phylogeography takes a relaxed random walk in continuous space and time[J]. Molecular Biology and Evolution,2010,27(8):1877-1885.

[27] DRUMMOND A J,HO S Y W,PHILLIPS M J,et al. Relaxed phylogenetics and dating with confidence[J]. PLoS Biology,2006,4(5):e88.

[28] DE MAIO N,WU C H,O'REILLY K M,et al. New routes to phylogeography:a Bayesian structured coalescent approximation[J]. PLoS Genetics,2015,11(8):e1005421.

[29] VAUGHAN T G,KÜHNERT D,POPINGA A,et al. Efficient Bayesian inference under the structured coalescent[J]. Bioinformatics,2014,30(16):2272-2279.

[30] BOUCKAERT R. Phylogeography by diffusion on a sphere:whole world phylogeography[J]. PeerJ,2016,4:e2406.

[31] PARKS D H,PORTER M,CHURCHER S,et al. GenGIS:a geospatial information system for genomic data[J]. Genome Research,2009,19(10):1896-1904.

[32] HOFFMANN M H,GLAß A S,TOMIUK J,et al. Analysis of molecular data of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) with Geographical Information Systems (GIS)[J]. Molecular Ecology,2003,12(4):1007-1019.

[33] WHITE D,SIFNEOS J C. Regression tree cartography[J]. Journal of Computational and Graphical Statistics,2002,11(3):600-614.

[34] BIELEJEC F,RAMBAUT A,SUCHARD M A,et al. SPREAD:spatial phylogenetic reconstruction of evolutionary dynamics[J]. Bioinformatics,2011,27(20):2910-2912.

[35] MAGEE D,SCOTCH M. The effects of random taxa sampling schemes in Bayesian virus phylogeography[J]. Infection,Genetics and Evolution,2018,64:225-230.

[36] WALKER R S,RIBEIRO L A. Bayesian phylogeography of the Arawak expansion in lowland South America[J]. Proceedings Biological Sciences,2011,278(1718):2562-2567.

[37] BOUCKAERT R,LEMEY P,DUNN M,et al. Mapping the origins and expansion of the Indo-European language family[J]. Science,2012,337(6097):957-960.

[38] LEE S,HASEGAWA T. Evolution of the Ainu language in space and time[J]. PLoS One,2013,8(4):e62243.

[39] GROLLEMUND R,BRANFORD S,BOSTOEN K,et al. Bantu expansion shows that habitat alters the route and pace of human dispersals[J]. Proceedings of the National Academy of Sciences of the United States of America,2015,112(43):13296-13301.

[40] BOUCKAERT R R,BOWERN C,ATKINSON Q D. The origin and expansion of Pama-Nyungan languages across Australia[J]. Nature Ecology & Evolution,2018,2(4):741-749.

[41] PAGEL M,MEADE A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo[J]. The American Naturalist,2006,167(6):808-825.

[42] ANTHONY D W. The horse,the wheel,and language:how Bronze-Age riders from the Eurasian steppes shaped the modern world[M]. Princeton:Princeton University Press, 2008.

[43] RENFREW C. Archaeology and language:the puzzle of Indo-European origins[M]. London:Cambridge University Press, 1987.

[44] TURNER C G. Dentochronological separation estimates for Pacific rim populations[J]. Science,1986,232(4754):1140-1142.

[45] MASUDA R,AMANO T,ONO H. Ancient DNA analysis of brown bear (ursus arctos) remains from the archeological site of rebun island,Hokkaido,Japan[J]. Zoological Science,2001,18(5):741-751.

[46] SATO T,AMANO T,ONO H,et al. Mitochondrial DNA haplogrouping of the Okhotsk people based on analysis of ancient DNA:an intermediate of gene flow from the continental Sakhalin people to the Ainu[J]. Anthropological Science,2009,117

(3):171-180.
[47] CURRIE T E,MEADE A,GUILLON M,et al. Cultural phylogeography of the Bantu Languages of sub-Saharan Africa[J]. Proceedings of the Royal Society B:Biological Sciences,2013,280(1762):20130695.
[48] MCCONVELL P. Backtracking to babel:the chronology of pama-nyungan expansion in Australia[J]. Archaeology in Oceania,1996,31(3):125-144.
[49] WILLIAMS A N,ULM S,TURNEY C S M,et al. Holocene demographic changes and the emergence of complex societies in prehistoric Australia[J]. PLoS One,2015,10(6):e0128661.
[50] CLENDON M. Reassessing Australia's linguistic prehistory[J]. Current Anthropology,2006,47(1):39-61.
[51] DIXON R M W. The Australian Linguistic Area[C]//AIKENVALD A Y, DIXON R M W. (Eds.). Areal diffusion and genetic inheritance: problems in comparative linguistics. Oxford:Oxford University Press,2001:64-104.
[52] MATISOFF J. Sino-Tibetan linguistics:present state and future prospects[J]. Annual Review of Anthropology,1991,20:469-504.
[53] VAN DRIEM G. Tibeto-Burman vs. Indo-Chinese:Implications for population geneticists, archaeologists and prehistorians [C]//SAGART L,BLENCH R,SANCHEZ-MAZAS A. (Eds.). The Peopling of East Asia:Putting Together the Archaeology, Linguistics and Genetics. London:Routledge Curzon,2005:81-106.
[54] BLENCH R,POST M W. Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages[C]// HILL N,OWEN-SMITH T. (Eds.). Trans-Himalayan Linguistics:Historical and Descriptive Linguistics of the Himalayan Area. Berlin:De Gruyter Mouton,2013:71-104.
[55] 李壬癸. 侗傣语族的祖居地、扩散及其史前文化[C]//上海师范大学语言研究所. 东方语言学:第十九辑. 上海:上海教育出版社, 2020:1-13.
[56] 龚群虎. 汉泰关系词的时间层次[M]. 上海:复旦大学出版社,2002.
[57] GREENHILL S J,WU C H,HUA X,et al. Evolutionary dynamics of language systems[J]. Proceedings of the National Academy of Sciences,2017,114(42):E8822-E8829.
[58] SUCHARD M A,WEISS R E,SINSHEIMER J S. Bayesian selection of continuous-time Markov chain evolutionary models [J]. Molecular Biology and Evolution,2001,18(6):1001-1013.
[59] WICHMANN S,RAMA T. Testing methods of linguistic homeland detection using synthetic data[J]. Philosophical Transactions of the Royal Society of London Series B,Biological Sciences,2021,376(1824):20200202.
[60] 邓晓华,高天俊. 演化语言学的理论、方法与实践[J]. 山西大学学报(哲学社会科学版),2014,37(2):72-75.
[61] 邓晓华. 从语言推论壮侗语族与南岛语系的史前文化关系:谨以此文悼念恩师严学窘教授[J]. 语言研究,1992,12(1):110-122.
[62] 邓晓华. 南方汉语中的古南岛语成分[J]. 民族语文,1994(3):36-40.
[63] YANG S Z,SUN X R,JIN L,et al. Inferring language dispersal patterns with velocity field estimation[J]. Nature Communications,2024,15(1):190.

(责任编辑:王圆圆)