南京郵電大學
Nanjing University of Posts and Telecommunications

Inferring language dispersal patterns with velocity field estimation
基于速度场估计推断语言扩散模式

nature communications，11 December 2023

报告人： 王巍强

2024.5.13

Research Background

研究背景 01

Background

在过去的一万年里，随着农业技术的加强，人口之间发生了大量的学术传播和文化传播。与此同时，世界各地的语族和语群也出现了起源和扩散。鉴于人类是语言的载体，而语言又是文化的载体，人类遗传学的技术进步使我们能够追踪不同语言人口的复杂人口动态。另一方面，语言进化的历史可以为文化创新的起源和传播提供惊人的见解。

然而，经常用于推断语言传播模式的系统地理学方法存在局限性，主要是因为系统发生树不能完全解释语言之间的水平接触所导致的语言进化，如借用和区域扩散。在此，本文引入了不依赖于系统发生树的语言速度场估计，来推断语言的传播轨迹和中心。

Research methods

研究方法 02

语言数据

Linguistic data: binary-coded linguistic trait that 1 denotes the presence of this trait while 0 denotes the absence

Trait 1	Trait 2	Trait 3
0	1	1
1	1	0
0	0	1

Geographic data: two-dimensional coordinate with longitude and latitude for each language sample

Lon	Lat	Trait 1	Trait 2	Trait 3
30	50	0	1	1
24	65	1	1	0
10	15	0	0	1

Data conversion: perform kNN algorithm to convert the binary state of each linguistic trait into frequency state based on geographic data

Trait 1	Trait 2	Trait 3
0.3	0.6	0.2
0.4	0.1	0.5
0.1	0.2	0.3

本文的语言数据集来源于四个语言家族和群体的公共词汇数据集。它们包含遵循特定单词列表(如Swadesh 100或200单词列表)的几个词汇。这些词已经被以前的语言学专家很好地编码为不同的词汇同源词。每个词汇都包含几个同源词，这些同源词表现出相同的意义和系统的声音对应。为了计算，每个同源词都被进一步重新编码为一个新的二进制编码语言特征，其中1表示该同源词在语言中存在，而0表示不存在。因此，本文的语言数据集包含103个印欧语言样本的5995个语言特征，109个汉藏语言样本的949个语言特征，420个班图语言样本的3859个语言特征，以及60个阿拉瓦克语言样本的693个语言特征。此外，还为每种语言样本分配了经纬度地理坐标。

缺失值的估算&编码特征的转化

Linguistic data: binary-coded linguistic trait that 1 denotes the presence of this trait while 0 denotes the absence

Trait 1	Trait 2	Trait 3
0	1	1
1	1	0
0	0	1

Geographic data: two-dimensional coordinate with longitude and latitude for each language sample

Lon	Lat	Trait 1	Trait 2	Trait 3
30	50	0	1	1
24	65	1	1	0
10	15	0	0	1

Data conversion: perform kNN algorithm to convert the binary state of each linguistic trait into frequency state based on geographic data

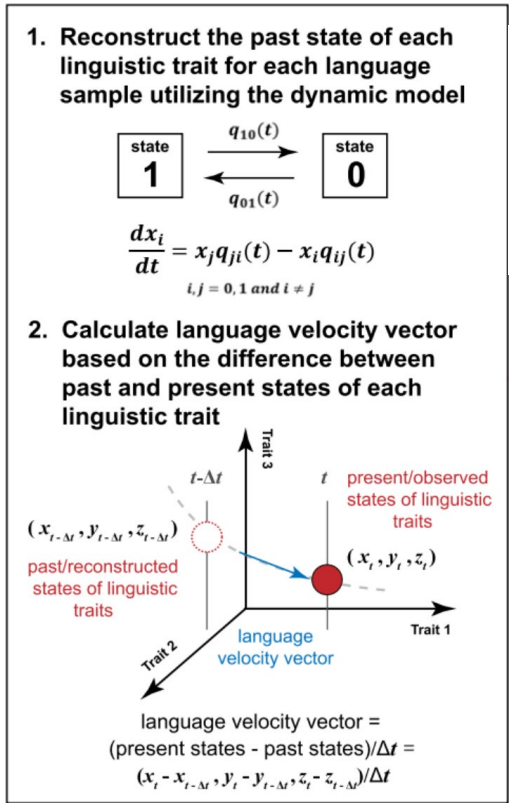
Trait 1	Trait 2	Trait 3
0.3	0.6	0.2
0.4	0.1	0.5
0.1	0.2	0.3

缺失值的估算：本文首先去除缺失值大于75%的语言特征。然后，使用模式值归因方法对剩余语言特征的缺失值进行归因。为了评估模值估算方法的效率，本文采用余弦相似度度量来衡量采用和不采用模值估算的速度场之间的相似度。同样，本文还评估了三种估算方法估计的速度场的一致性:频率值估算、零值估算和模式值估算。并且利用Procrustes分析来检验这三种方法所估算的语言特征的PC值之间的一致性。所有的评估都表明，缺失值的估算不会影响速度场的估计。

编码特征的转化：本文使用k近邻(k-NN)算法将每个二进制编码的语言特征转换为范围为0到1的频率特征。首先，本文选择了地理上最接近给定语言样本(包括其本身)的k个语言样本。其次，本文计算了这k个语言样本中每个语言特征表现为状态1和状态0的频率。由于在这k个语言样本中，每个语言特征的不同状态频率之和等于1，因此只要给定状态1的频率，就可以确定每个语言特征的状态0的频率。因此，将语言特征的二进制状态转换为状态1或状态0的频率是没有区别的。在本研究中，对于每个语言样本，我们将每个语言特征的二值转换为其在k个最近的语言样本中显示状态1的频率。

语言特征演化的动态模型

假设：本文提出了关于语言特征演化的三个模型假设。首先，每一种语言特征在进化过程中都可以以异质速率在不同状态之间进行多次转换。第二，一种语言中每一种语言特征的变化都可能受到邻近语言的影响。特别是，这种影响可能来自邻近语言拥有相同语言特征的不同国家之间的竞争。第三，每个特征状态在某一特定领域具有特定的社会语言学声望。根据这些假设，本文提出了一个简单的动态模型。



$$\begin{cases} \frac{dx_0^i}{dt} = x_1^i q_{10}(x_0^i, s_0^i) - x_0^i q_{01}(x_1^i, s_1^i) \\ \frac{dx_1^i}{dt} = x_0^i q_{01}(x_1^i, s_1^i) - x_1^i q_{10}(x_0^i, s_0^i) \end{cases} \quad (1)$$

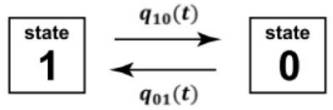
威望参数的估计

为了估计动态模型中的威望参数，本文设计了一个参数估计原理，该原理来源Felsenstein提出的遗传学DNA替代模型。这种DNA取代模型建立在泊松过程的基础上，假设在DNA进化过程中，每个碱基都可以以不均匀的速率多次转变为其他碱基(例如，A转变为T或C转变为G) 。这类似于本文的模型假设，即每个语言特征在进化过程中可以以异质速率在增益和损失之间进行多次转换。因此，本文也使用泊松过程来模拟每种语言特征的增益和损失。威望参数可以使用等式2来估计。

$$\begin{cases} s_1^i = e^{-\lambda} + (1 - e^{-\lambda})\,\pi_1^i \\ s_0^i = e^{-\lambda} + (1 - e^{-\lambda})\,\pi_0^i \end{cases} \tag{2}$$

重建语言特征的过去状态频率&建立速度场

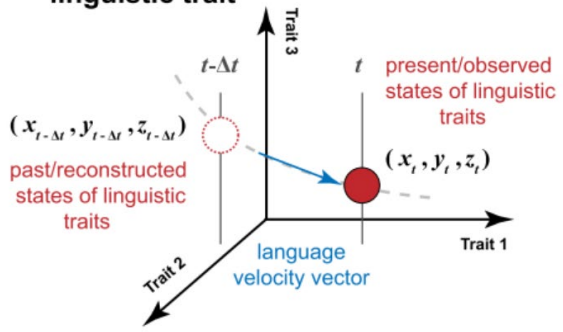
1. **Reconstruct the past state of each linguistic trait for each language sample utilizing the dynamic model**



$$\frac{dx_i}{dt} = x_j q_{ji}(t) - x_i q_{ij}(t)$$

$$i, j = 0, 1 \text{ and } i \neq j$$

2. **Calculate language velocity vector based on the difference between past and present states of each linguistic trait**



language velocity vector =
(present states - past states)/Δt =
$(x_t - x_{t-\Delta t}, y_t - y_{t-\Delta t}, z_t - z_{t-\Delta t})/\Delta t$

$$x_1^i(-m) = \left[1 + \left(\frac{1}{x_1^i(0)} - 1\right) e^{(s_1^i - s_0^i)m}\right]^{-1} \quad (3)$$

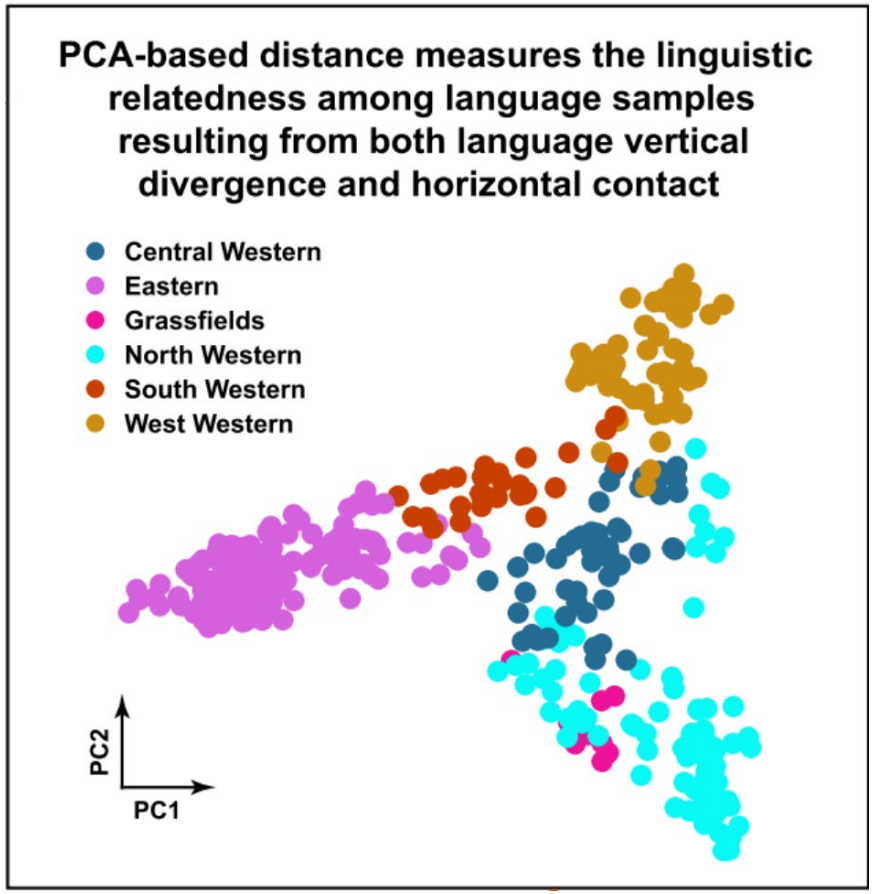
使用等式3，本文重建了每个语言特征的过去状态频率。

$$\mathbf{V}_l = \frac{1}{m}[\mathbf{X}_l(0) - \mathbf{X}_l(-m)] \quad (4)$$

本文建立了一个高维速度场来量化语言特征的历时进化轨迹。这个速度场是由一组速度矢量组成的。语言l的速度矢量(Vl)近似为其语言特征的过去和现在状态频率之差除以重构时间。对于每个语族或语群，n个语言样本的速度向量可以组成一个高维速度场，表示为矩阵V。

$$\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n]^T \quad (5)$$

● 速度场的PCA投影



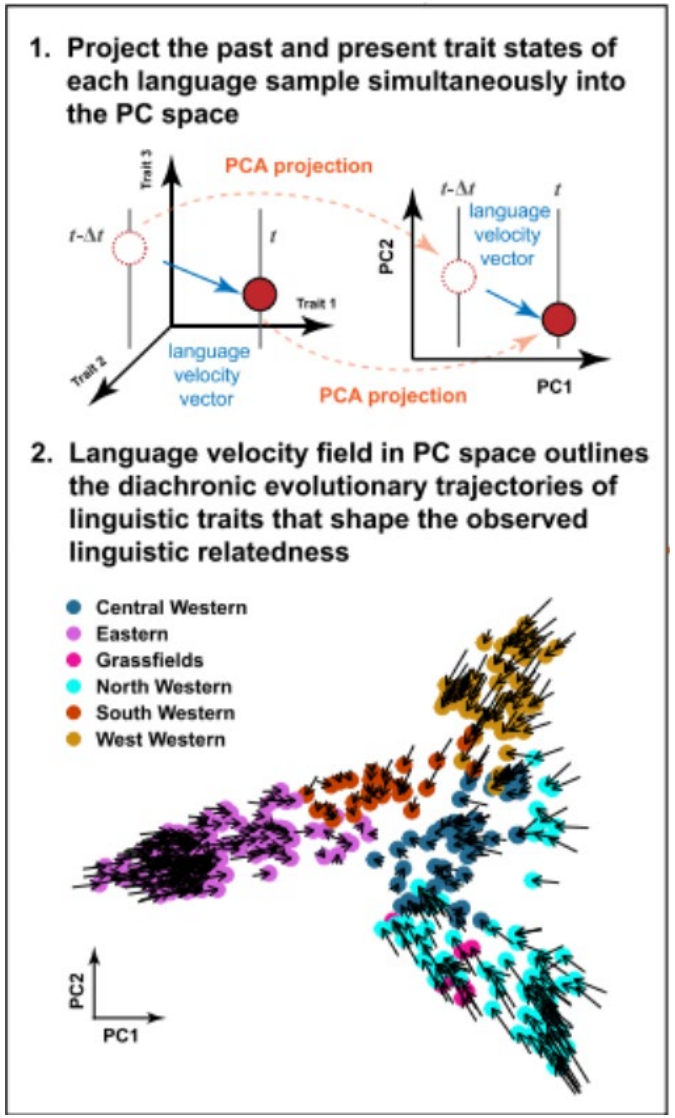
$$[\mathbf{PC1}, \mathbf{PC2}] = \mathbf{PC} = \begin{bmatrix} \mathbf{PC}_1^T \\ \mathbf{PC}_2^T \\ \vdots \\ \mathbf{PC}_n^T \end{bmatrix} = \mathbf{DA}_2 \qquad (6)$$

将高维速度场V投射到PC空间，以描绘语言特征的历时进化轨迹，这些特征塑造了观察到的语言相关性。首先，本文对二值编码的语言数据进行PCA，利用等式6将二值编码的语言特征重新排列为两个最优新特征(即PC1和PC2)。

$$\mathbf{V}^{PC} = \begin{bmatrix} (\mathbf{V}_1^{PC})^T \\ (\mathbf{V}_2^{PC})^T \\ \vdots \\ (\mathbf{V}_n^{PC})^T \end{bmatrix} = \mathbf{VA}_2 = \begin{bmatrix} \mathbf{V}_1^T\mathbf{A}_2 \\ \mathbf{V}_2^T\mathbf{A}_2 \\ \vdots \\ \mathbf{V}_n^T\mathbf{A}_2 \end{bmatrix} = \frac{1}{m}\left(\begin{bmatrix} \mathbf{X}_1^T(0) \\ \mathbf{X}_2^T(0) \\ \vdots \\ \mathbf{X}_n^T(0) \end{bmatrix} \mathbf{A}_2 - \begin{bmatrix} \mathbf{X}_1^T(-m) \\ \mathbf{X}_2^T(-m) \\ \vdots \\ \mathbf{X}_n^T(-m) \end{bmatrix} \mathbf{A}_2 \right) \qquad (7)$$

其次，利用等式7将高维速度场V投影到二维PC空间。这种投影可以看作是将每个语言样本中语言特征的现在和过去的状态频率同时映射到PC空间，然后用它们的差除以重建时间。

● 速度场的PCA投影



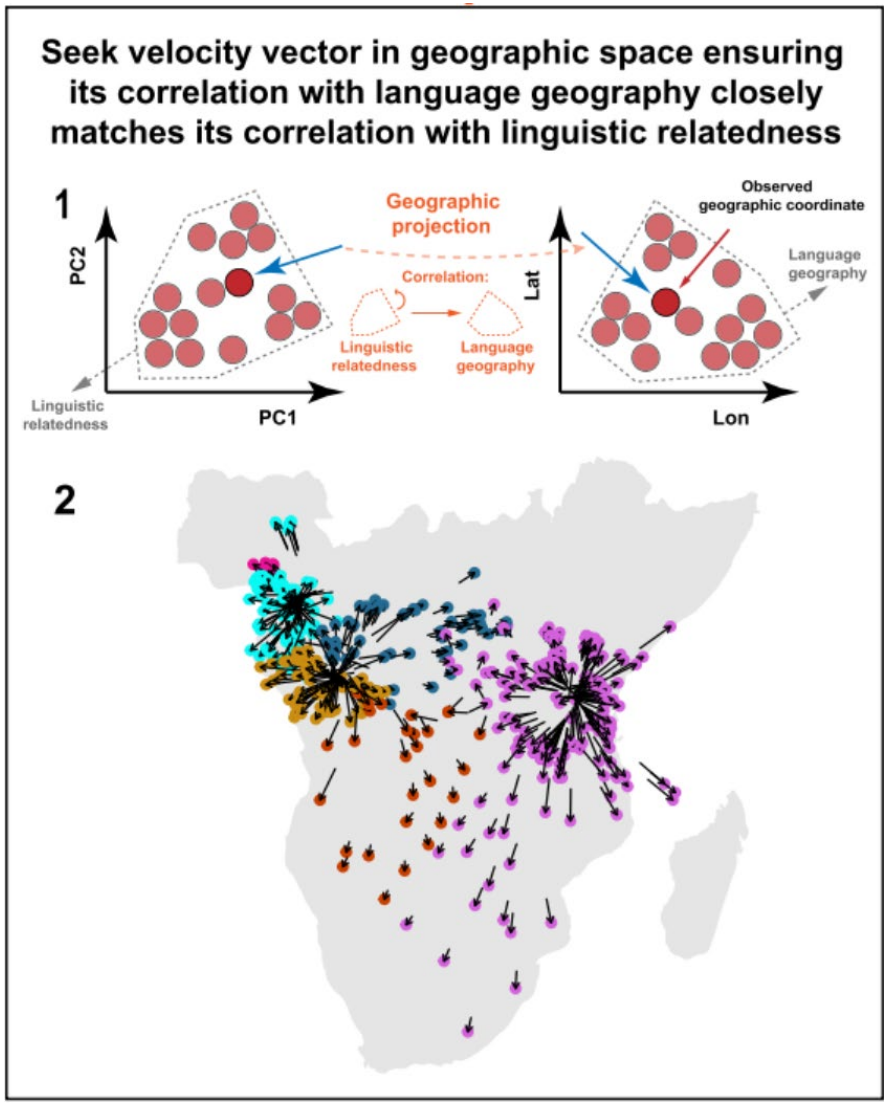
$$[\mathbf{PC1}, \mathbf{PC2}] = \mathbf{PC} = \begin{bmatrix} \mathbf{PC}_1^T \\ \mathbf{PC}_2^T \\ \vdots \\ \mathbf{PC}_n^T \end{bmatrix} = \mathbf{DA}_2 \qquad (6)$$

将高维速度场V投射到PC空间，以描绘语言特征的历时进化轨迹，这些特征塑造了观察到的语言相关性。首先，本文对二值编码的语言数据进行PCA，利用等式6将二值编码的语言特征重新排列为两个最优新特征(即PC1和PC2)。

$$\mathbf{V}^{PC} = \begin{bmatrix} (\mathbf{V}_1^{PC})^T \\ (\mathbf{V}_2^{PC})^T \\ \vdots \\ (\mathbf{V}_n^{PC})^T \end{bmatrix} = \mathbf{VA}_2 = \begin{bmatrix} \mathbf{V}_1^T\mathbf{A}_2 \\ \mathbf{V}_2^T\mathbf{A}_2 \\ \vdots \\ \mathbf{V}_n^T\mathbf{A}_2 \end{bmatrix} = \frac{1}{m}\left(\begin{bmatrix} \mathbf{X}_1^T(0) \\ \mathbf{X}_2^T(0) \\ \vdots \\ \mathbf{X}_n^T(0) \end{bmatrix} \mathbf{A}_2 - \begin{bmatrix} \mathbf{X}_1^T(-m) \\ \mathbf{X}_2^T(-m) \\ \vdots \\ \mathbf{X}_n^T(-m) \end{bmatrix} \mathbf{A}_2 \right) \qquad (7)$$

其次，利用等式7将高维速度场V投影到二维PC空间。这种投影可以看作是将每个语言样本中语言特征的现在和过去的状态频率同时映射到PC空间，然后用它们的差除以重建时间。

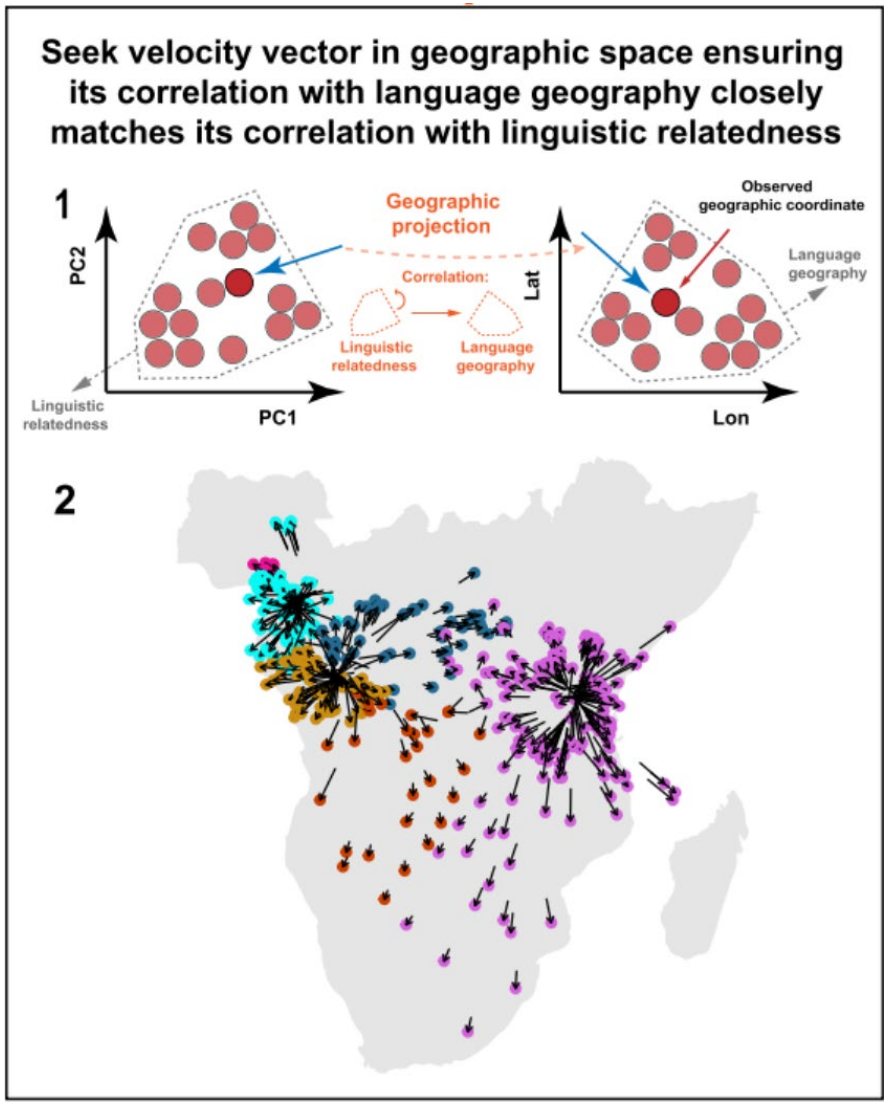
PC空间中速度场的地理投影



根据观察到的语言亲缘性与语言地理之间的相关性，采用La Manno等人提出的核投影(kernel projection)方法，将速度场从PC空间投影到地理空间。核投影在地理空间中寻找每个速度向量，确保其与PC空间中的语言分布的相关性与地理空间中的相关性密切一致。通过核投影，可以根据La Manno等人提出的等式8计算语言样本在地理空间内的速度向量。同时，地理空间中的速度场反映了语言样本是从何处扩散到当前地理位置。

$$\mathbf{v}_l^{Geo} = \sum_{j=1}^{s} \left(P_{lj} - \frac{1}{s} \right) \frac{\mathbf{C}_j - \mathbf{C}_l}{||\mathbf{C}_j - \mathbf{C}_l||} \tag{8}$$

● 速度场的空间和网格平滑



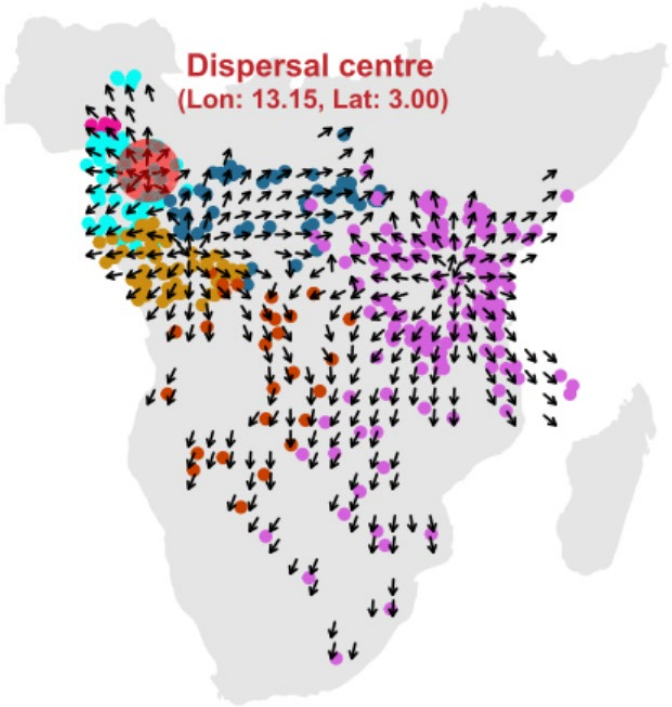
本文采用空间平滑和网格平滑方法来更好地可视化地理空间内的速度场。这些平滑方法有助于更好地显示速度场，同时保留速度场中反映的原始语言分布模式。对于空间平滑，首先使用等式9将每个速度向量的长度单独缩放为与PC空间中的速度向量的长度相同。其次，使用等式10通过加权其他语言样本的速度向量的长度来进一步调整其长度。

$$\mathbf{V}_l^{Geo-scale} = \frac{\mathbf{V}_l^{Geo}}{||\mathbf{V}_l^{Geo}||}||\mathbf{V}_l^{PC}|| \tag{9}$$

$$\mathbf{V}_l^{Geo-scale-smooth} = \frac{\mathbf{V}_l^{Geo-scale}}{||\mathbf{V}_l^{Geo-scale}||}\sum_{j=1}^{n} K_\sigma(\mathbf{C}_l,\mathbf{C}_j)||\mathbf{V}_j^{Geo-scale}|| \tag{10}$$

● 分散中心推断

1. **Grid smoothing:** visualize a vector field showing language velocity vectors evaluated on regular grids
2. **Dispersal centre inference:** infer the geographic coordinate of language dispersal centre based on grid-smoothed language velocity field



为了推断语言传播中心，本文设计了一个基于地理空间内网格平滑速度场的简单策略。鉴于地理空间内的速度矢量描绘了语言传播方向，本文假设传播中心周围的速度矢量应呈现向外辐射模式。

根据这一假设，测量了每个网格点周围网格平滑速度矢量向外辐射模式的程度。这种模式的程度是通过计算这些速度矢量在每个维度上的方差(平均方差)的平均值来衡量的，如等式12，等式13所示。显示出最高平均方差的网格点，即邻近速度矢量向外辐射模式最强的网格点，被视为语言扩散中心。

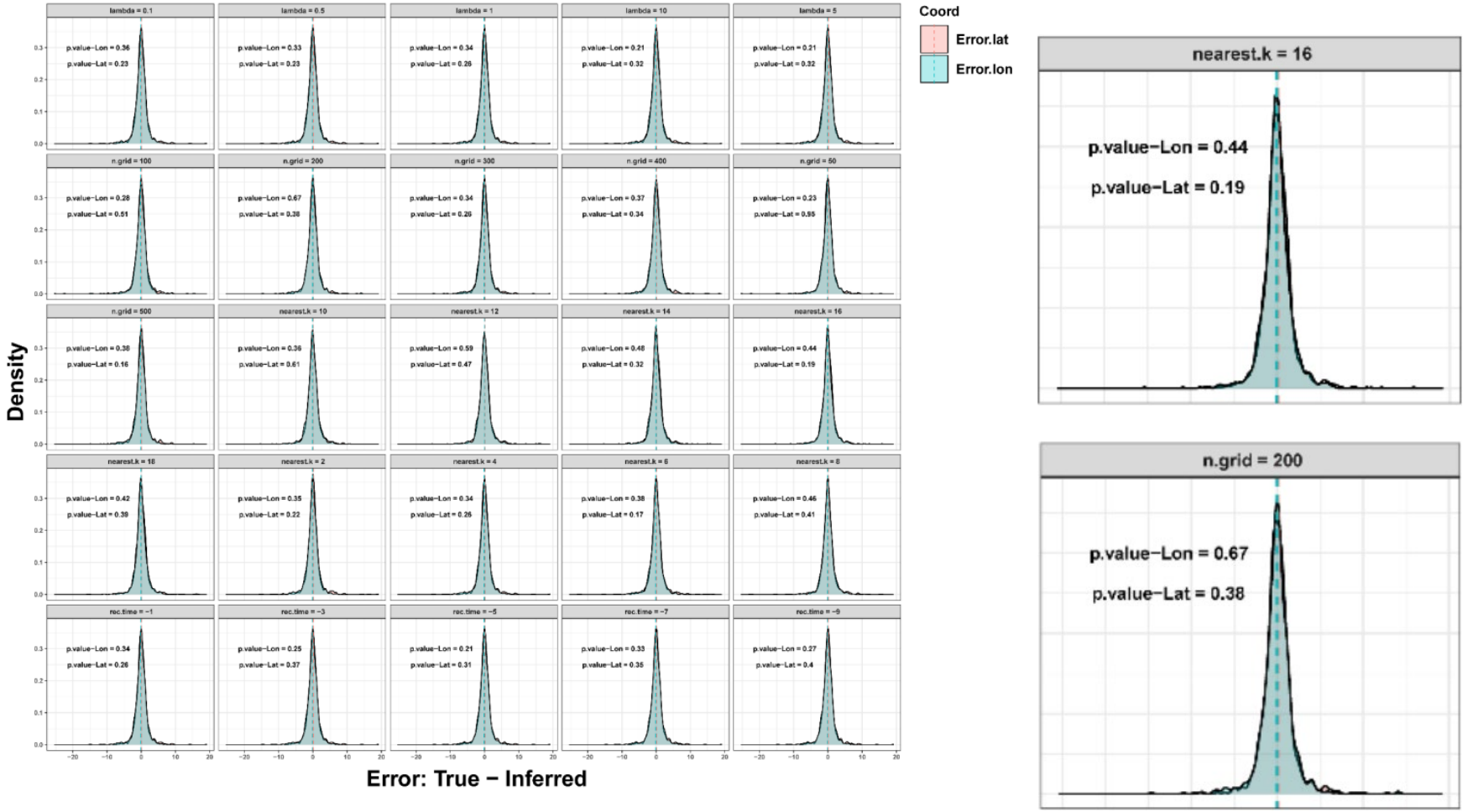
$$\mathbf{V}_g^{Grid-scale} = \frac{\mathbf{V}_g^{Grid}}{||\mathbf{V}_g^{Grid}||} \tag{12}$$

$$\sigma_g^2 = \frac{1}{2(s-1)} tr\left[\left(\mathbf{V}^{Grid-scale}\right)^T \left(\mathbf{E}_s - \frac{\mathbf{1}\mathbf{1}^T}{s} \right) \left(\mathbf{V}^{Grid-scale}\right) \right] \tag{13}$$

Experience

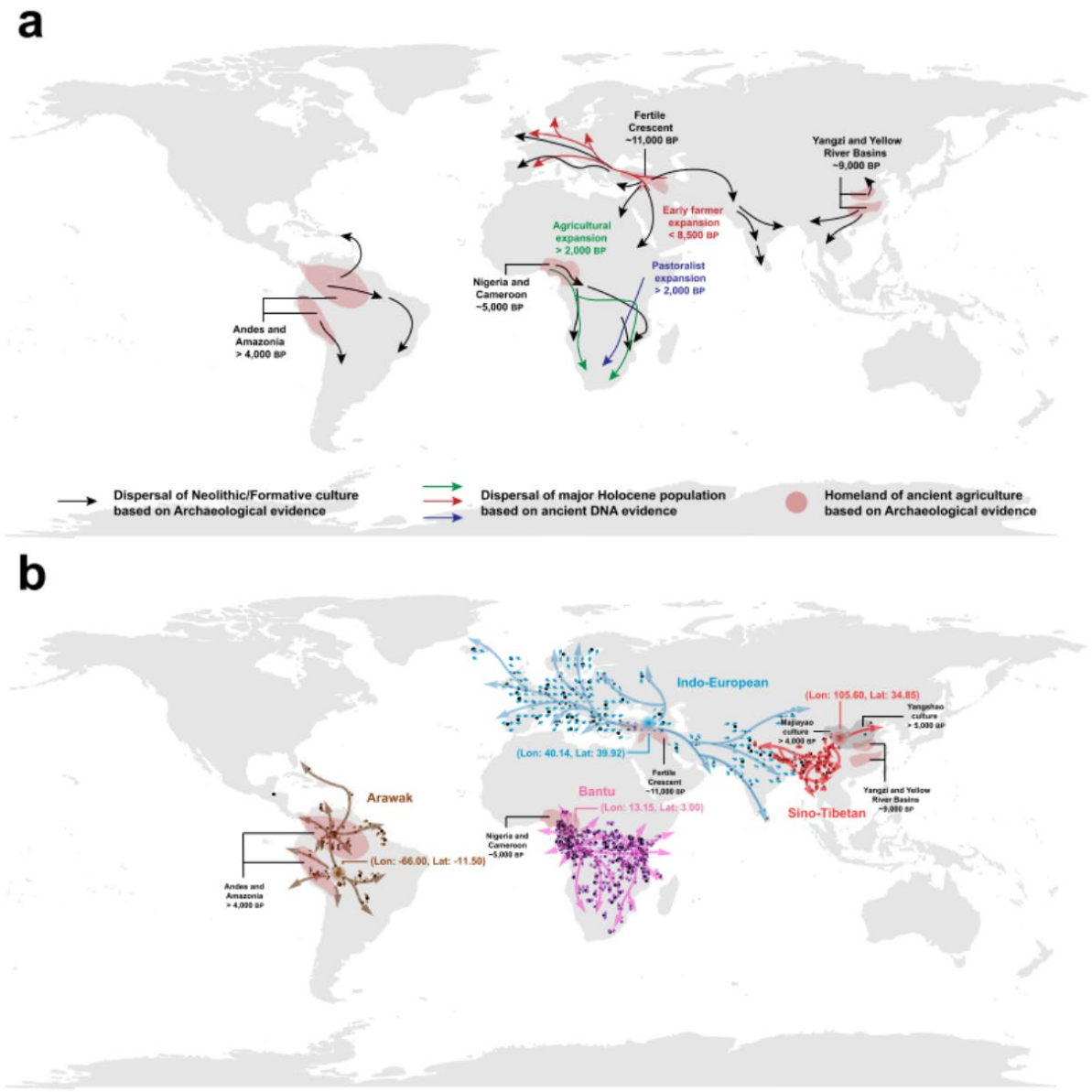
实验 03

LVF的仿真验证

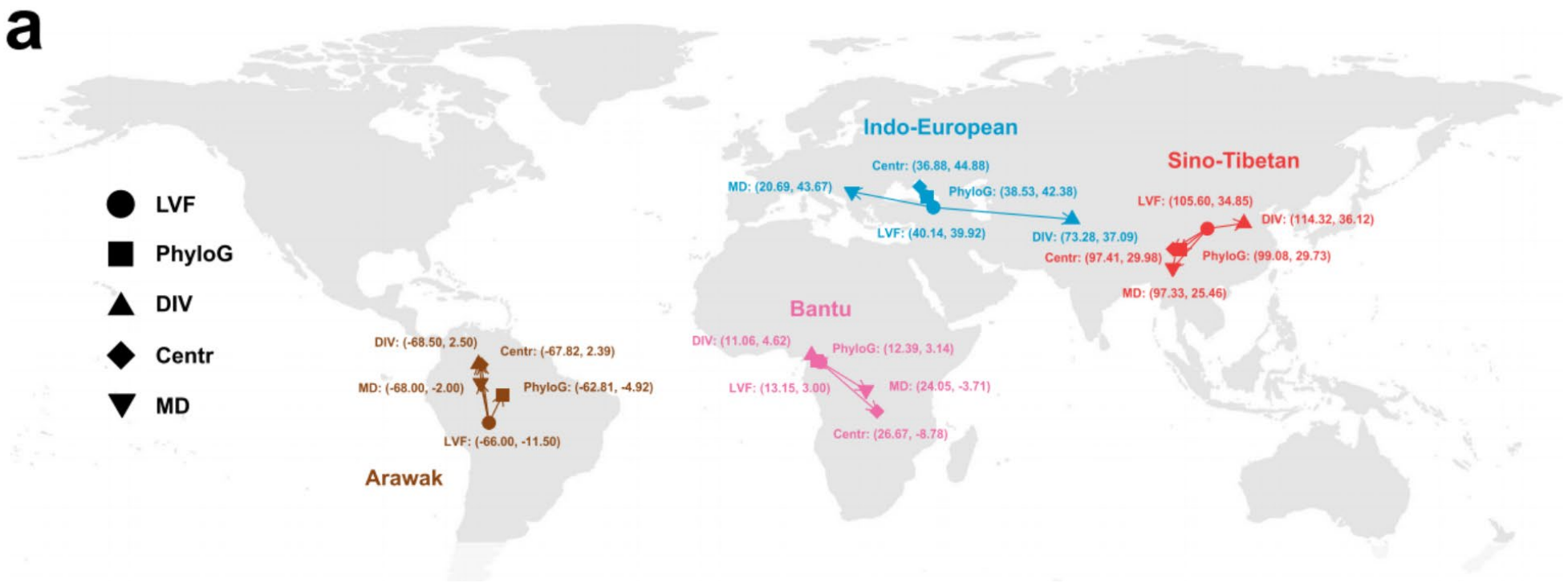


为了验证LVF的有效性和鲁棒性，我们将其应用于1000个模拟语言数据集。为了评估LVF的有效性，我们采用双侧Wilcoxon秩和检验来检验特定参数设置(k = 10， λ = 1, m = 1)下给定和推断的分散中心坐标之间的差异。k表示k近邻，λ表示泊松过程的突变率，m表示重建时间。结果表明，在k = 10， λ = 1, m = 1的参数设置下，LVF估计的扩散中心与给定的扩散中心没有显著差异。这表明在这些参数设置下LVF的有效性很高。此外，我们还进行了双侧Wilcoxon秩和检验，以检验在不同参数设置下给定和推断的分散中心坐标之间的差异。具体来说，当将LVF应用于模拟数据集时，我们改变了k (k = 2,4,6，…，18)，λ (λ = 0.1, 0.5, 1,5,10)和m (m = 1,3,5,7,9)的值。结果表明，在不同参数设置下，语言传播中心的推断坐标与给定坐标没有显著差异。这表明LVF在不同的参数设置下仍然有效。因此，我们设置k = 10， λ = 1和m = 1作为LVF的默认参数值。

语言速度场估计的实证应用



LVF与其他空间重建方法的比较

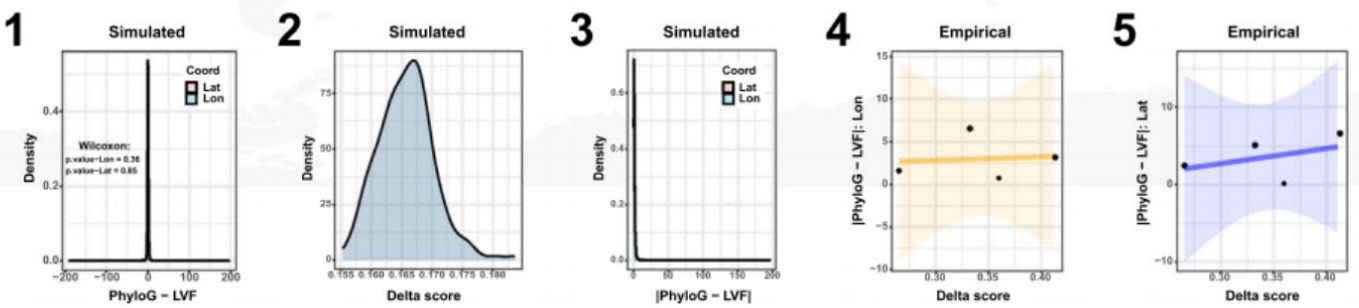


本文将LVF与其他三种无系统地理方法进行了比较。它们是多样性(DIV)、质心(Centr)和最小距离(MD)方法44,55。这些方法建立在与LVF和系统地理学方法完全不同的理论基础之上。

具体地说，多样性方法假定传播中心应该位于包含最大语言多样性的区域。语言多样性是指某一区域内语言特征之间的差异程度，其值越高意味着差异越大。质心方法假定当前语言地理位置的延伸所形成的多边形的中心应为分散中心。最小距离方法假定，与其他语言的平均地理距离最小的语言的位置应该是分散中心。我们将这三种基本方法应用于四个实证案例。结果表明，LVF推断的分散中心与这三种方法推断的分散中心有显著差异。这突出了LVF和这些无系统发育方法之间的根本区别。

LVF与其他空间重建方法的比较

b



6

		Simulation		Indo-European		Sino-Tibetan		Bantu		Arawak	
Tree-likeness	Delta score	0.1657		0.2656		0.3324		0.3598		0.4129	
	p-value	-		< 0.001		< 0.001		< 0.001		< 0.001	
		Lon	Lat	Lon	Lat	Lon	Lat	Lon	Lat	Lon	Lat
|PhyloG - LVF|	Difference	1.55	0.94	1.61	2.46	6.52	5.12	0.76	0.14	3.19	6.58
	p-value	-	-	0.158	0.069	0.020	0.012	0.443	0.878	0.058	0.007
		R^2	*p*-val	R^2	*p*-val	R^2	*p*-val	R^2	*p*-val	R^2	*p*-val
Linguistic relatedness explanatory power	PCA-based distance	0.90	0.001	0.37	0.001	0.44	0.001	0.65	0.001	0.53	0.001
	Phylogenetic tree	0.93	0.001	0.39	0.001	0.05	0.160	0.38	0.001	0.09	0.057

总结

该研究引入了不依赖系统发生树的语言速度场估计法来推断语言的扩散轨迹和中心，其有效性和稳健性通过模拟和实证研究得到了验证。与系统地理学方法类似，语言速度场也包括两个主要方面，首先是建立一个速度场，以描述语言特征的非同步进化轨迹，从而形成所观察到的语言相关性。这个速度场的功能类似于系统发生树，但它还能捕捉到横向联系的归因。然后是根据语言相关性与语言地理之间的相关性，将这一速度场投射到地理空间中。这类似于系统发生树的地理投影，有助于勾勒出语言在地理空间中的传播轨迹。之后进行速度场的空间和网格平滑处理，最后通过计算向外辐射程度来推断语言扩散中心和扩散轨迹。

THANKS