

地理学论文的数据要求

姚鲁烽, 何书金, 赵 歆

(《地理学报》编辑部, 北京 100101)

在地理学论文中, 数据作为论据的主要组成部分, 对于论题的分析、论点的阐述、论理的发挥起着重要的支撑作用。在地理学报的来稿中, 数据的选择、使用和说明是审稿中发现问题最多的类型之一。因此, 我们将《地理学报》1998-2007 年审稿意见中涉及数据的问题进行分类整理, 供初写论文的青年学者参考。

地理学论文的数据要求包括: 数据的类型组合、来源说明、时间尺度和空间范围等几个方面。

1 数据的类型组合

1.1 地理学数据的类型

地理学研究中使用的数据依据其获取途径大致可分为观测数据、实验数据和统计数据三种类型。

1.1.1 观测数据 观测数据是指通过观测仪器获取的数据, 包括: 台站观测数据、定点观测数据、遥感观测数据等。

(1) 台站观测数据是指在气象台和水文站等台站常年观测得到的数据, 如降水、气温、径流量、含沙量等数据。

(2) 定点观测数据是指在野外科研实验站对特定地表要素进行定点观测取得的数据, 如对冰川、冻土、沙丘、泥石流等的移动和变化进行观测记录的数据。

(3) 遥感观测数据是指利用卫星和飞机搭载仪器对地观测获得的数据。如对卫片、航片等影像处理所得资料数据。

1.1.2 实验数据 实验数据是指利用实验仪器设备分析样品或模拟环境动力得到的数据。主要包括: 样品分析数据和模拟实验数据。

(1) 样品分析数据是指将野外采集的土样、水样、冰芯等样品经过物理、化学分析所得到的沉积粒径、元素含量、年代测定等分析数据。

(2) 模拟实验数据是指在实验室内通过模拟不同自然环境动力得到数据, 主要包括土壤侵蚀模拟、风洞模拟实验等得到数据。

1.1.3 统计数据 统计数据是指通过全面统计或随机抽样调查获得的数据。主要包括统计年鉴数据、抽样调查数据、测量统计数据。

(1) 统计年鉴数据是指各级政府统计部门在统计年鉴中公布的经济统计数据, 如产值、产量、人口、收入、消费等数据。

(2) 抽样调查数据是指通过社会抽样调查统计获得的数据, 如流动人口构成、购买意愿及消费行为等数据。

(3) 测量统计数据是指野外定点测量的数据, 如草丛高度、树木胸径、地层厚度、砾石倾向等数据。

1.2 数据的组合使用

地理学的很多研究需要多种类型的数据相互支持、共同印证。

(1) 在自然地理研究中,分析长期历史气候环境时,最好用地层分析、考古资料、历史文献等多种数据;在研究地层、冰芯等地表沉积中所含化学元素时,选取的类型要有针对性。应该有多种类型元素的百分比含量变化对比,以确定各种元素的相关关系和对环境变化反映的差别。

(2) 在人文地理研究中,社会调查数据应与统计数据相结合;在人口构成、人员流向等人文地理抽样调查中,抽样调查的人员组成类型要有广泛性和代表性。

(3) 在地理信息中,遥感观测数据要与地面实测数据进行对比和校正。在分析植被、土地利用等要素的区域分布规律时,要有该地区80%以上的植被或土地利用类型参与对比。

1.3 数据的组合匹配

研究地表环境变化,有时需要不同时期不同类型的数据对比。有些学科在选取不同时间的数据时,如果采用了不同的数据源,就需要说明不同类型数据的组合匹配情况。

1.3.1 不同遥感数据的匹配 在遥感研究中,如果同时采用TM和MSS两种资料,要注意两种遥感数据的可比性;采用长时间序列的AVHRR NDVI遥感数据,应对数据质量进行分析,特别是后期NOAA卫星及AVHRR传感器的改变是否会造成前后两个阶段NDVI值系统性的差值,应对数据做标准化处理,减小数据源的误差对计算结果的影响。

1.3.2 不同测绘数据的匹配 采用不同时期不同比例尺的地形图研究地貌演化时,测绘数据精度差异会对地形破碎程度研究产生影响,要说明如何套绘不同比例尺的地形图。

1.3.3 遥感数据与测绘数据的匹配 采用不同时代地形图与遥感影像数据研究景观变化时,地形图的数字化提取与从遥感影像解译获得的景观数据上存在一定的差异。要分析这种景观斑块变化是反映了两种类型的数据源的差异还是景观实际变化。

1.3.4 地理信息数据的匹配 对于从已有图件数字化得来的数据,如年日照时数图、土地资源图、年平均降水量分布图等,应说明这些图件数据的比例尺、生产时间,不同类型的数据在空间尺度上是否存在差异。如果存在差异,运用何种办法实现匹配,这是GIS工作中最重要的环节。利用各类环境资源的多年平均值数据时,应有各类数据的时段说明。

2 数据的来源说明

对数据来源的说明是保证数据可靠性的重要步骤,包括:引用处处的标注、采集环境的阐述、测定方法的说明和生成过程的解释。

2.1 引用数据的出处标注

在论文中引用他人的研究数据时,必须在论文中引用数据的句子后明确标注参考文献出处,以此说明数据的原始获取过程,同时表示尊重他人的研究成果。数据来源的标注要求包括:

2.1.1 数据的原始出处标注 标注数据出处应该直接引用该数据的最初文献出处,而不是间接引用后人的再使用文献,这样有利于读者了解数据的原始获取过程。

2.1.2 数据的正式发表出处 标注数据出处要尽量引用已正式出版的期刊论文或专著。在引用博士、硕士学位论文数据时,要检索一下该论文和数据是否已正式发表。

2.1.3 共享数据的使用说明 在引用共享数据时应该署名共享数据的制作单位、人员和共享途径。

2.1.4 数据提供者的署名方式 对提供数据的合作者,可以依据数据的多少,在作者名中署名或在文中提出致谢。

2.1.5 避免使用未发表数据 为使审稿人能及时了解数据的可靠性,一般不宜引用待刊

论文的数据。否则审稿专家难以判别数据的可靠性。

2.1.6 不用标注的数据类型 使用多年不变的公用数据可以不列数据出处,例如:流域面积、河流长度、山峰高度等。

2.2 野外数据的采样环境

由于自然环境的多样性,不同区域的自然要素分布规律对比及其影响因素是复杂的。只有说明数据的野外采集条件、方法,才能使数据具有说服力和可比性。

2.2.1 样点布局的设置说明

(1) 样点布局的文字说明。应说明采样线路的布设原则和目的,以及样点的区域代表性或类型代表性。采样点布设是地统计的基础,如果采样点布设明显不均匀,应当关注和讨论因采样点布设可能对空间统计结果带来的误差。

(2) 样点布局的地图说明。最好在环境类型分区的底图上标出样点位置,这样可以更直观地反映样点的空间分布状况。例如,在研究不同土壤类型的分区时,不仅要说明土壤样点数量,还要在土壤类型的底图上显示说明每一个土壤类型上分布多少个样点和土壤剖面,以表明采样的代表性。

(3) 样点布局的表格说明。利用表格说明不同样点的环境差异,可以清楚说明选择样点的目的性。例如,对山地样点的环境背景说明,可以用表格按不同海拔高度的序列进行列举,这样可以更有条理地说明采样布局原则和样点环境背景。

2.2.2 取样地点的环境说明

(1) 数据水平空间位置的说明。为分析地表要素的环境影响,对数据的平面空间位置要给予必要的说明。例如:要说明样点与河流、湖泊或海洋的距离;说明气象站距离城区的距离等。

(2) 数据垂直空间位置的说明。使用气象资料要说明气象站海拔高度;采用地下水资料要说明采样点的地下水位深度。在建立计算地下水存储量的数学模型时,要考虑不同沉积物的厚度等因素。

(3) 数据属性空间位置的说明。很多地理研究不仅要说明数据的空间位置,还要说明数据的环境类型属性。例如,土壤样品的采集不仅要说明采样的地下深度顺序,还要说明土壤发生层的A、B、C层序列,才能得到土壤的发生学特征;对冰芯数据不仅要说明海拔高度,还应该简要说明冰川成冰带分布状况及冰芯钻取地点所在成冰带、冰芯钻取地点的主要成冰过程;在确定冰芯的积累年数后,要说明冰芯钻取地点的年降水量或年积累量;在建立地表温度的遥感监测模型时,数据的选取不能按行政区县来选取,而要按照地表类型来选取,如按照工业区、居住区、农田区等来分区建立监测模型。这样得到的模型才能在其他地区进行使用、比较和验证。

(4) 采样环境的时间说明。对日变化量大的数据,有必要说明具体的取样时间。例如,在对比不同地区土壤含水量、蒸散量时,需要考虑土壤取样的时间。因为降雨前后、土地灌溉前后取样、多日干旱后取样的结果是完全不一样的;在分析土壤养分变化时要说明采集的季节是否相同,因为土壤速效养分与采样时间的关系很大。

(5) 模拟数据的环境。在验证模型时,要检查实测范围内是否有各种要素类型存在。如果只取小范围内的单一类型数据,数学模型就会缺少可靠性和广泛应用性。因此必须保证校准模型的实测数据范围内有不同要素类型存在,使得数学模型中各种系数可靠,并能得以在广大地域上应用。此外,还要检查数据的地域选取范围是否有针对性。

2.2.3 样点的环境变化说明

(1) 数据位置的移动。对于城市气象站的长期观测数据分析,要说明气象站是否有变动。因为随着我国近50年的城市快速发展,有些气象站的位置有变化。

(2) 环境背景的改变。要说明样地是代表自然环境还是受人影响的。例如:林地

可能是由原生的,也可能是由其他用地转化而成;荒地可能是自然形成的,也可以是农地等转变而成的。

(3) 环境变化的过程。第四纪环境变化过程对剖面数据的影响要注意。例如,影响钻孔位置取在湖泊边缘,会出现低水位时的沉积缺失,使第四纪环境变化分辨率和记录的完整性受到影响。

2.3 实验数据的测定方法

依据实验数据作为论据的文章,需要在论文中说明获取数据的测试仪器和测试方法。

2.3.1 实验目的的说明 对不同元素的分析、测试,要简要说明各种元素的环境指示意义和原理,以此表明实验研究的学术意义,并列出具体的参考文献。

2.3.2 实验仪器的说明 应简要说明实验仪器的产地、型号、精度以及实验单位场地等。为保证数据的有效性和精确性,应该采用最新的实验仪器和可靠试剂。

2.3.3 测定方法的说明 要有简要的完整实验过程说明,最好能绘出技术路线图。如果数据的测试方法是前人做过和发表过的,应列出相关的参考文献。

2.3.4 实验环境的说明 实验环境对于实验数据会产生影响,应该加以说明。使其他研究者可以在工作中检验和对比该项数据。例如,用 CH_4 、 N_2O 气相色谱仪测定 CH_4 、 N_2O 的排放量,测定时的温度等条件应该标明以供参考。

2.4 数据生成的修订过程

当研究中有的数据是通过其他相关数据转换生成时,需要说明转换过程中的数据修订方法。例如,气象站温度观测结果受海拔高度的影响,利用温度资料与太阳总辐射的拟合结果来确定辐射温度修正系数也会受海拔的影响,需要进行系数修订说明。再如,在进行海洋初级生产力的遥感研究中,采用修正的VGPM模式反演估算初级生产力,模式中日均积分初级生产力的计算由5个重要参数所控制。对于边缘海域而言,其中叶绿素-a浓度和真光层深度是2个最重要的控制参数。这2个参数的变化很大程度上控制了初级生产力的变化状况。对于真光层深度是如何获取、叶绿素-a浓度如何修订,以及他们各自对初级生产力的影响与贡献应交代清楚。

3 数据的时间要求

数据的时间特征要求包括:数据即时程度和时段长度的情况;数据时间的完整性、同步性、代表性特征;数据的时序间隔和时序插补要求。

3.1 数据的即时程度

数据的即时程度是指稿件中获取的最新数据时间。由于近年来气候环境的明显变化和我国社会经济的快速发展,使得地理学研究中地理环境的最新状况的分析变得格外重要。论文中采用最新数据,会提高论文的研究时效和对读者的吸引力。

3.1.1 每年都有统计发布的数据 如气象、水文、国民经济统计资料等数据在论文中最好含有3年前的最新数据。

3.1.2 每隔5-10年的全国普查数据 如人口、工业、矿产、土壤等全国性普查资料,在论文中应包含最近一次的统计数据。

3.1.3 遥感数据应该包含所得的最新数据。

3.2 数据的时段长度

为探讨各类自然要素的时间演变规律,许多论文需要分析有关资料数据在一定时间尺度的变化过程。各学科所需要数据分析的时间尺度不尽相同。

3.2.1 历史记载数据的时间尺度 研究历史自然灾害、历史地方病区分布、历史人口变化等方面的论文需要至少近600年的资料,因为明清以来600年是我国地方志记载相对

较全的时期。清末以来,西方测绘技术传入我国,中央和地方政权开始有了各类相关仪器测量数据的记载。在东部沿海地区有100年来的定量气候、地形观测记载数据。

3.2.2 观测与统计数据的时间尺度 研究气象、水文、海洋、部门产业结构变化等需要近50年的历史,这是由于1950年后我国在全国范围内建立了气象站、水文站和各类经济统计部门。而利用遥感数据研究地表变化的时段长度最好在20年以上。

3.2.3 实验数据的时间尺度 在利用地层钻孔剖面、海洋珊瑚层面、洞穴碳酸钙纹理、寒区冰芯沉积等获取的实验分析数据中,要检查实验数据是否包括了样品生成的全部时代。分析数据一般应占样品生成时段长度的80%以上。剖面要有测年数据。

3.2.4 周期分析的时间尺度 在研究气候等要素波动周期时,长周期的波动研究需要利用长时段尺度的数据。一般数据时段长度应为最大波动周期的3倍。例如,研究气象要素20年变化周期需要有至少60年的气象数据。因此,数据的时间尺度越大,自然要素演变过程越完整,研究者所能分析的要素变化周期就越长。

3.2.5 模拟预测的时间尺度 在气候、水文等自然要素预测性的论文中,实测数据时段应是预测时段的5倍以上,推测的可靠性才能较高。例如,推断某地区未来5年经济发展趋势,至少需要利用过去25年的经济资料。

3.3 数据的时间完整性

3.3.1 日变化数据的完整性 要有24小时的全日连续记录数据,以显示研究要素一日内的完整变化情况。一般不能只有12小时的半日数据或18小时数据。

3.3.2 年内变化数据的完整性 要有12个月的全年连续记录数据,以显示研究要素年内的完整变化情况。不能只有6个月的数据或10个月的大半年数据。对季节变化的研究,即使没有全年的数据也要有1、4、7、10月的四季典型月份数据。

3.3.3 年际变化数据的完整性 一般要有研究要素在较长时段中最高值、最低值年份的数据,以及平均值、平均变化值的计算说明。

3.4 数据的时间同步性

时间是环境变化研究的基准点之一,所以论文中数据的时间序列要同步。

3.4.1 进行研究要素相关分析时,相关分析数据的时段应一致 例如:进行区域降水-径流相关分析时,气象和水文观测数据的时段应一致。

3.4.2 进行分区分级研究时,要使用相同年份的数据 例如,在进行经济区划研究时,各类经济统计数据的时间段应一致。在进行土壤分区时,20年前土壤普查数据与2年前采集的新土壤数据不能混合使用。因为20年中土壤的许多特性都发生了显著变化,混合使用就无法说明是土壤特性是20年前还是目前的状况。

3.4.3 利用地面实测数据验证遥感数据精度时,地面数据的实测日期与遥感影像日期要一致 这样的数据才能匹配,否则就难以起到验证的作用。

3.5 数据的时间代表性

3.5.1 遥感数据 利用卫星影像估算森林覆盖率要说明是用何月的或者还是多年的平均值,1月份和7月份卫星影像估算的森林覆盖率、水域面积等的结果会存在很大的差距。再如,植被的NDVI需要说明是指平均NDVI、累积NDVI、还是最大NDVI。

3.5.2 水文数据 在对比不同河段水体中泥沙或元素含量时,洪水期、枯水期的含量也大不一样。因此,取样时机的一致性和可比性是至关重要的,必须在文中说明。

3.5.3 气象数据 对气象数据不能用秋末冬初之类的表述,要用具体月份的说明。

3.6 数据的时序间隔

数据的时间密度选取要根据不同的研究要素变化特征来考虑。

3.6.1 气象、水文等观测数据变化数据要逐月的变化数据。

3.6.2 古气候研究的实验数据需要有年度数据 如冰芯、纹泥、树木年轮、珊瑚影像密

度、碳酸钙沉积层理等研究需要有年度变化数据。

3.6.3 地表要素的时间变化间隔,一般取决于人为因素影响大小和变化快慢。

(1) 沙地、湖泊、河道、海岸等人为干扰相对不大的地貌变化可取10年以上的间隔数值;

(2) 土地面积、植被类型、冰川进退等受人为干扰的可选用5年间隔的数据;

(3) 比较10年以内的环境变化要用各年数据。不宜仅用序列首尾2个年份数值来说明整个时段的变化状况。

3.7 数据的时序插补

在使用长时段数据序列使,有时会遇到其中有些年份的数据因各种原因而缺失的情况,需要通过数据插补来满足时间序列的完整。

3.7.1 数据时间序列插补的方法和原则 建立研究要素变化的序列时,应该给出影响要素的情况,因为很多要素变化的主导因子是人类活动,而人类活动在时空上皆有强烈的差异。如果插补没有确定一定的原则和方法,插补就会有大的随意性。

3.7.2 不同季节的数据插补 在利用空间化技术后对暖冬、冷夏分析。要考虑暖冬、冷夏等极端气候事件的影响,因为插值本身容易导致极端值被降低。

3.7.3 不同时段的数据插补 用不同时段、不同站点数据做插值,要说明采用什么方法插值、如何计算全国平均状况、采用多大的格点为基本单元计算,由于不同时段的站点数不同,格点的空间分辨率是否一致、这样利用格点数据计算与利用站点数据计算的差别和优劣如何。

4 数据的空间要求

数据使用、分析和说明的空间要求主要包括:数据的空间范围、空间密度、空间插补、空间代表、空间类型等方面内容。

4.1 数据的空间范围

数据的空间范围关系到数据的代表性是否全面。地理环境的区域差异使得同一自然要素在不同地域有不同的演变形式。如海岸带变化要检查是否考虑了平原海岸、基岩海岸、侵蚀海岸、堆积海岸、沉降海岸、抬升海岸等各种类型海岸的变化数据。

另外,许多自然要素的空间变化,只有在相对较大的范围才能体现出来。例如,气候、土壤、植被等要素的空间分布特征,需要放在相对广泛的空间范围来考察。

对获取辐射温度修正系数时应用的气象站资料的拟合结果要做详细交代,是仅用某一个站点的温度还是用多个气象站平均后的温度。

在进行2种以上要素的相关分析中,要考虑数据的空间范围。数据的空间范围过大或过小都会影响要素间相关分析的科学性和适用性。

4.2 数据的空间密度

4.2.1 数据空间密度的区域代表性 例如,由于我国地域辽阔,降水、温度等自然要素的变化受经度、纬度、海拔高度的影响,区域差异很大。仅用全国平均降水量、温度、日照时数来分析要素间的联系是不够的,必须用不同区域的数据进行分别对比;在分析河水化学离子时,对干流、支流、不同季节、不同河型、不同地表覆盖类型区都应采样对比。

4.2.2 数据空间密度的区域差异性 在利用遥感数据研究地形起伏度和植被指数时,分析窗口的大小是关键环节。 $8\text{ km} \times 8\text{ km}$ 数据不适用于破碎的深切割山区。应该用更高空间分辨率的数据。

4.3 数据的空间插补

进行不同要素空间相关分析时,有时需要对空间密度数据不足区域进行插补。要注意不同地区、不同地形、不同方法的数据插补。任何插值结果的可靠性直接受插值站点的数量及其空间分布的影响。要说明插补的数据占多大比例;用什么方法插补;插值方法有否物理意义;其精度误差是否经过验证;是否影响分析结果等。

4.3.1 不同地区的数据插补 中国西部气象站点稀疏,做多年平均值内插尚可,如果具体到每年做内插则误差较大。

4.3.2 不同地形的数据插补 地形对气象要素的空间分布有很大影响,气象要素在空间插值时要考虑地形的影响。对于黄土高原等复杂地形内的气象站密度应足够,在黄土高原周边地区选择一定数量的站点资料,对边缘地区的插值精度是有重要影响的。多个气象站的平均值可能会有气象站高差不同的影响,应该每个站先做距平,然后平均在一起,这样可以消除高差的部分影响。

4.3.6 不同方法的数据插补 运用站点资料计算结果绘制等值线时,不同的插值方法绘制的等值线形式是不同的,应交代所选择的插值方法。

4.4 数据的空间代表性

(1) 水平空间的代表性 在分析植物生长量、地表径流、农业产量等要素与气温、降雨的关系时,如果取全国的气温、降雨的均值与全国植物生长量均值进行比较,就会出现相关结果不准确的问题。因为,我国地域辽阔,各地气温与降雨的年际变化与年内变化都是不同步的。大范围的均值会掩盖地域的自然差异。如果取华北、东北、华东等气温降雨相对一致的区域均值分别与某一要素进行对比,相关分析结果才有可信性和科学性。

(2) 垂直空间的代表性 在研究山地、峡谷、盆地等高程差异较大地区时,要注意说明数据的垂直空间代表性;在研究地层、土层的片面差异时,也要说明层位的代表性。例如,土壤数据分析中要说明图表计算的结果代表的是整个土壤剖面还是某一层位。

4.5 数据的空间类型

地理学研究的数据按照空间分布状态,可分为点状数据、线状数据和面状数据等三种类型。在进行不同数据的对比分析时,要考虑到数据的空间分布类型。

4.5.1 点状数据与面状数据的差异 在利用遥感数据与观测数据进行气象资料对比时,用均方根偏差为指标得出 NCEP/NCAR 低估若干数值的变化幅度要考虑自动气象站观测资料为点上观测资料, NCEP/NCAR 再分析资料为格网数据,代表一定区域的平均值,两者空间尺度不一致,经过区域平均处理后均方差变小是必然的,未必完全是低估的原因。

4.5.2 线状数据与面状数据的差异 在生态需水量中考虑根据排污量计算的环境需水,在水资源管理的实践上是不合理的,因为排污总量需要根据水的自净能力、环境容量来限定,相反不能根据排污量提供纳污水量。就后者而言,把沿河道线状分布的河道内生态需水量插值到空间面上是没有意义的。

4.5.3 点状数据与线状数据的差异 对河流水质的研究,要解释干流与支流、干流上下游之间的差别,分析水中物质元素的来源,应量化流域雨水、土壤和母岩风化等对水体元素的贡献率。