地理大数据挖掘的本质

裴 韬[1,2],刘亚溪[1,2],郭思慧[1,2],舒 华[1,2],杜云艳[1,2],马 廷[1,2],周成虎[1,2]
(1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101;
2. 中国科学院大学,北京 100049)

摘要: 针对地理大数据的内在本质以及地理大数据挖掘对于地理学研究的意义,本文解释了地理大数据的含义,并在大数据"5V"特征的基础上提出了粒度、广度、密度、偏度和精度等"5度"的特征,揭示了地理大数据的本质特点。在此基础上,从地理大数据的表达方式、地理大数据挖掘的目标、地理模式的叠加与尺度性、地理大数据挖掘与地理学的关系等4个方面阐述了地理大数据挖掘的本质与作用,并从挖掘目标的角度对地理大数据挖掘方法进行分类。未来地理大数据挖掘的研究将面临地理大数据的聚合、挖掘结果的有效性评价以及发现有价值的知识而非常识等几方面的挑战。

关键词: 空间模式;空间关系;空间分布;流空间;时空异质性;知识发现
DOI: 10.11821/dlxb201903014

1 引言

早在30年前,计算机领域的研究者就已经预见到海量数据将会给计算机科学及其他学科的发展带来的挑战与机遇,提出了"数据挖掘"一词。1995年,李德仁院士率先倡导从GIS数据库中发现知识[1]。之后,Harvey等提出"地理数据挖掘与知识发现"(Geographic Data Mining and Knowledge Discovery)[2],标志着地理学与数据挖掘技术的实质性交叉,地理数据挖掘作为发现地理学规律的重要手段,已被地理学者所承认。然而,之后的10多年里,地理数据挖掘虽在方法研究中取得了显著的进展,但对地理学领域新知识的揭示仍未取得令人信服的成就。随着大数据时代到来,一系列重量级的研究相继涌现:基于手机数据的人类行为预测[3]、利用搜索引擎对流感的预测[4]以及深度学习算法对于人类思维能力的挑战[5-6]等。这些发现不仅颠覆了传统的认识,更为重要的是,它们证明了大数据对于科学发现的潜在推动力。

毫无例外,大数据对于地理学也形成了巨大的冲击,迫使地理学者思考一系列问题:地理大数据挖掘的本质是什么?其与地理学之间的关系如何?在地理学发展中究竟能够起到何种作用?为了回答以上问题,本文拟从以下几个方面进行论述:首先,将阐述地理大数据的内涵与外延是什么;其次,系统分析地理大数据的特点;第三,从地理大数据挖掘的核心问题入手揭示其本质;第四,根据挖掘目标对地理大数据挖掘方法进行分类;最后,对地理大数据挖掘的发展和面临的挑战进行展望。

收稿日期: 2018-10-08; **修订日期:** 2019-02-15
基金项目: 国家自然科学基金项目(41525004, 41421001) [**Foundation:** National Natural Science Foundation of China, No.41525004, No.41421001]
作者简介: 裴韬(1972-), 男, 研究员, 博士生导师, 主要从事地理大数据挖掘研究。E-mail: peit@lreis.ac.cn

2 地理大数据的内涵及外延

大数据虽已成为当前学界的热词，但关于大数据内涵以及外延的界定一直未有定论。实际上，给大数据以确切定义其意义并非在于明确地圈定哪些数据属于大数据，而是在于指导如何进行大数据分析以及如何在应用中避免大数据的局限性。Mayer-Schonberger等曾经在《Big Data: A Revolution That Will Transform How We Live, Work, and Think》中给出了大数据的价值（Value）定义[7]，Marr总结出大数据的“5V”特征[8]，即：Volume（大量）、Velocity（更新快）、Variety（多样）、Value（价值）、Veracity（真实性）。大数据的产生主要源于传感器、网络和计算技术的突破，因而体现出数据量大、更新快以及种类多（前3个“V”）的特征；而另一方面，大数据的获取多为传感器用户的自发性上传（如微博和微信数据的获取）或非目的性记录（如手机信令、公交刷卡记录等），如以数据产生的主体为研究对象，则此大数据当属非目的性的观测数据，故通常含有大量噪声，最终导致价值密度低、真实性差（后2个“V”）的特征。其实，“5V”的刻画也仅仅是大数据的表象，并非大数据真正的定义。

本文中，大数据的本质被认为是针对研究对象的样本“超”覆盖，当然，此处并非指完全没有遗漏的样本覆盖，而是指超出目的性采样（也可称为“小数据”）范畴的、趋向于全集的信息获取（只有在极端情况下，“超“覆盖才可能是全集样本）。大数据的本质所导致的这种信息覆盖，突破了目的性和局部性的传统采样的局限，必然带来思维方式和认识上的变革。由此可以推论，地理大数据就是针对地理对象的“超”覆盖样本集，此处的“超”覆盖涉及时间、空间与属性维度。同样地，地理大数据也具备“5V”特征，但地理大数据同时还具有自己独特的性质，这将在后面的章节进一步论述。地理大数据的内涵至少表明，其辨识度集中体现在以下两点：① 地理大数据与其他大数据之间的差别在于是否具有时空属性；② 地理大数据与小数据的区别在于样本的覆盖度。

地理大数据内涵的确定是基于获取信息的模式，而其外延的划分则依赖于信息采集的手段。根据所使用的传感器类型以及数据所记录对象的不同，可将地理大数据分为对地观测大数据和人类行为大数据两类。其中，对地观测大数据记录地表要素的特征，获取信息的传感器类型主要包括航天、航空以及地表监测传感器等，以主动的获取方式为主，对应的数据包括：卫星遥感、无人机影像以及各类监测台站（网）的数据等。人类行为大数据记录人类移动、社交、消费等各种行为的信息，信息获取的传感器种类繁多，包括：手机终端、智能卡、社交媒体应用、导航系统等，以被动的获取方式居多，可视为人类活动的足迹（footprint），产生的数据包括：手机信令数据、出租车轨迹数据、物联网数据以及社交媒体数据等。两类大数据直接关注的主要对象分别为“地”和“人”。人类发展与地理环境之间的关系一直是地理学的核心论题，而地理大数据的爆发，使得对地观测与人类行为大数据的全面结合成为可能，从而为地理学中人地关系的研究提供了新资源、新动力和新视角。两类数据关注的角度各异，数据结构、粒度和表达方式又不尽相同，继而为地理大数据的分析与处理提出了新命题。

3 地理大数据的特征

前已述及大数据区别于小数据的主要特征，但作为具有特定内涵与外延的地理大数据，是否包含一般大数据共性之外的特征对于地理大数据的分析处理至关重要。为此，本文将从地理大数据产生机理的角度着重讨论其内在的特征。一方面，相对于小数据，

地理大数据样本的“超”覆盖主要体现在3个方面：粒度更细、密度更高、范围更大；另一方面，地理大数据，尤其是人类行为大数据的获取大多属于非目的性，从而导致有偏性和不确定性。因此，地理大数据可以总结为时空粒度、时空广度、时空密度、时空偏度和时空精度等“5度”的特征。

3.1 时空粒度

如果将地理信息承载单元的大小称为粒度，那么地理大数据的出现，则让地理信息的承载粒度由大变小。由于不同类型大数据的获取方式不同，因此粒度对于不同数据的含义也不一样。在对地观测大数据中，粒度是指数据所代表的（地表）范围大小，粒度的变化体现在由对地观测大数据反演得到的地物单元不断地细化。例如，城市影像分辨率的提升使得由其反演得到的地物单元从粗粒度的地块细化到具体的建筑。而在人类行为大数据中，粒度是指记录和统计单元的大小[9]，粒度的变细表现为用以记录和统计的单元的缩小。以人口统计为例，中国实施的人口普查方案中，普查小区为人口统计的最小粒度。普查小区在城市中多为街道的尺度，而在农村中则为乡镇的级别。普查小区的大小范围从几平方千米到几十平方千米，某些区域甚至更大。而手机数据的应用，为人口的精细化估计提供了可能。图1即为利用北京市手机用户数据进行精细人口估计的结果[10]。图1中人口信息的基本单元为基站小区（可近似为以手机基站位置划分的泰森多边形）。在城市人口的密集区，基站小区的尺度约为200 m左右。同样，利用浮动车轨迹数据针对城市道路拥堵状况的评估可以精细到任意时刻和任意路段[11-13]；融合微信请求数据、出租车定位数据、兴趣点（Point of Interest, POI）数据以及Quickbird高分影像可以将城市功能区的识别粒度细化至建筑物[14]；利用住户智能水电表信息可以对年龄、工作状态和收入的估计细化到家庭[15]。地理大数据粒度的精细化可以使我们从微观的角度观察地理现象，为研究其细部特征和机理提供了新的可能性。

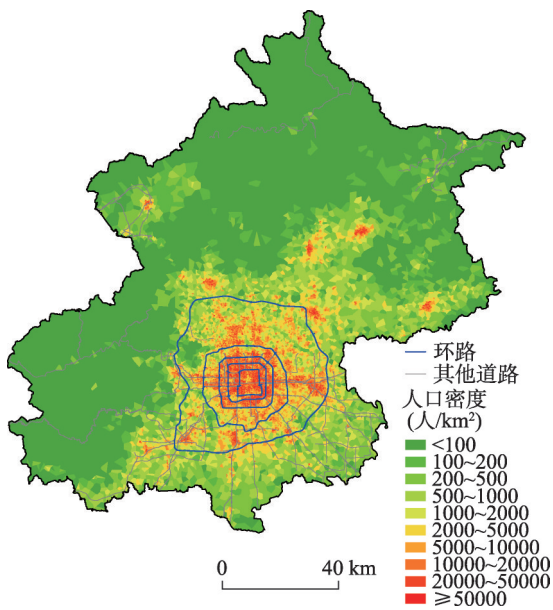

图1 基于手机数据的城市精细人口估计[10]

Fig. 1 Fine-grained population estimation using mobile phone data

3.2 时空广度

传统的地理小数据因受到信息获取手段和成本的限制，往往只能集中于局部的区域，或者需要在研究粒度与范围之间进行权衡，即在选择较大范围的同时不得不采用较粗的粒度。而在大数据时代，部分IT公司借助互联网的优势，可获取较大范围，甚至全国直至全球范围内的数据及其衍生的产品，同时又保持较小的时空粒度，从而使其研究范围在“豁然开朗”的同时又保持着“高清晰度”。对地观测大数据中全球性的数据产品已涉及多个研究领域，如全球夜光遥感数据产品[16]，国产30 m分辨率的全球土地利用数据[17]，全球长时间序列叶面积指数产品[18]等。而在人类行为大数据中，数据覆盖范围之广也是前所未有：百度发布的全国（不含港澳台地区）春运人口迁徙图（http://qianxi.baidu.com），滴滴发布的全国出租车（不含港澳台地区）运营状态图（https://www.

didiglobal.com/）、Facebook发布的全球用户网络（http://fbmap.bitaesthetics.com/）等。地理大数据提供了观察大尺度下地理现象和规律的可能性，为研究全球变化、宏观社会行为提供了宝贵的素材。

3.3 时空密度

由于成本的原因，传统的地理学研究对于地理现象的观测除了受限于范围的局部性，样本的密度也相对稀疏。因此，在有限样本基础上进行地理现象的刻画通常需要借助空间估计和推断的方法，如克立格插值[19-20]、地理加权回归[21-22]、环境因子模型[23-24]等。由空间统计方法获得的分布特征，虽然可通过空间相关性在一定程度上弥补样本稀疏的缺憾，但估计的结果毕竟无法取代属性的真实分布。与此相反，地理大数据的基本特征之一就是面向地理对象的高密度样本。在对地观测大数据中，数据的密度是随着粒度的变细而不断增加的。随着传感器分辨率的提升以及无人机等技术的广泛应用，影像像素分辨率不断提高，使得像素密度相应增加，混合像元信息不断裂解细化，导致像元所代表的信息更加精细；随着全球对地观测台网的逐步升级，对地监测的台站数目也不断增加，其中，气象台站从20世纪60年代的8000多个[25]增加到现在的超过100000个[26]，平均密度已达每1490 km^2就有1个观测站；对海洋观测的Argo浮标从2000年开始布设，数目已增加到2018年7月的3762个[27]。相比于人类行为大数据，以问卷调查得到的传统“小数据”虽然粒度也小，但密度很低，而以手机通话和信令为代表的大数据，用户已覆盖了城市的大部分人口，与此类似的还有腾讯的QQ及微信用户。随着智能卡和互联网应用的普及，人类行为大数据中样本的密度也越来越高。地理大数据样本密度的提升使得对地理现象的观测更加细致与逼真。

3.4 时空偏度

虽然地理大数据在粒度、广度以及密度等方面较传统小数据具有明显的优势，但同时也普遍存在着缺陷，而使其饱受诟病。需特别说明的是，人类行为大数据普遍存在有偏的现象，集中体现为数据载体在时间、空间和属性等几个方面的有偏性。以微博数据为例，很多研究使用其进行城市功能和人群行为的研究。而实际上，微博的使用者，在年龄属性上主要集中在18~30岁的年龄段；在性别上，女性用户的比例更大[28]；而在空间上，沿海地区较中西部使用率更高；不仅如此，微博所含的内容更加偏重于娱乐、教育、财经等方面的热点事件[29]。针对地理大数据有偏性对统计结果的影响，Zhao等应用手机数据进行了研究[30]，结果显示：由部分抽样的手机数据得到的移动距离、回旋半径、移动熵的数值与全样本之间存在显著的差异。由此可见，将有偏的大数据的规律推断为全体性质存在风险。偏度的普遍性存在导致其所得到的规律往往表现出一定程度的“偏见”，故在使用地理大数据时需要谨慎甄别。

3.5 时空精度

地理大数据另一个不容忽视的缺陷是其精度较差。精度问题在空间数据中普遍存在，而地理大数据的精度问题尤为突出，有时甚至会影响到计算结果的可信度。对地观测数据中的精度问题已经为众多研究所揭示[31-33]，在此不再赘述。对于人类行为大数据，由于其在获取过程中的被动性（例如，用于估计城市精细人口的手机信令数据并非为估计人口而设计收集）和自发性（例如，用于度量城市心情的微博数据由用户自发上传），数据中往往充斥着各种类型的误差，这种误差同样会存在于空间、时间以及属性中。以手机信令数据为例，由于城市建筑物的遮挡以及基站容量的限制，手机在通话时并不一定与其最近的基站发生通信，此时若将用户位置归于最近基站的小区内，则会产生空间误差。同样，在社交媒体数据中，用户所上传的事件位置、时间和文本内容，往往并不

能代表事件发生的真实状况。因此，与目的性采样的小数据不同，地理大数据中的误差除了技术原因之外，很多源于数据产生主体的不可控因素，有时甚至是一些主观故意造成的时空位置信息的改变[34]。地理大数据中误差的存在，往往会引发认识的偏差，甚至导致谬误的发生，谷歌公司对于流感预测的成功与失败就是例证[4, 35]。

地理大数据所具有的冲击力源于其粒度细、广度宽和密度大，这些都是传统小数据所不具备的；然而，地理大数据的偏度重和精度差同样也是小数据所力求避免的，传统的采样理论和误差理论就是针对偏度和精度而产生的模型体系，可以有效地限制偏度和控制精度。由此可见，地理大数据与小数据之间各有优劣，在现有条件下，一方不能完全取代另外一方，二者的结合可扬长避短，而在地理大数据的应用中，应注重其局限性，避免错误的产生与滥用。

4 地理大数据挖掘的核心问题

数据的价值在于隐匿其中的规律[36-38]，而数据挖掘的主要目的就是发现其中的知识。对于地理大数据所蕴藏的地理特征，数据挖掘方法如何应对？本文认为以下4个方面需要重点关注。① 对地观测大数据的获取是以对客体的观测为主要方式，故数据易于结构化，而人类行为大数据以主体记录为主，由记录产生的数据结构复杂、特征多变、类型多样，因此，如何进行表达成为地理大数据挖掘的前提。② 地理大数据繁冗复杂，需要确定挖掘的目标及其本质，唯此，地理大数据挖掘方有可能发展成为地理信息科学的分支乃至独立的学科。③ 由于地理大数据所具有的粒度、广度和密度等特征，地理现象从微观到宏观诸多尺度特征贯穿于地理大数据中，这是传统小数据所无法比拟的，因此，需要阐述清楚在挖掘过程中如何处理地理大数据内含的尺度性。④ 面对当前地理学研究的重要素材——地理大数据，有必要弄清地理大数据挖掘与地理学之间的关系，尤其是地理大数据挖掘在地理学的发展中能够起到何种作用。

4.1 地理大数据的表达:位空间和流空间

地理大数据中对地观测大数据所聚焦的对象是地表要素，而人类行为大数据的主体是人，二者间的作用可以视为主体与环境之间存在的关系。对地表要素观测的数据以位置为核心，属性的变化都以位置为支点，在研究中可视为影响人类行为与活动的环境要素。这种以时空位置为核心的数据类型可用位空间为框架进行表达，而位空间是指以位置为基本表达单元，以欧氏距离作为基本测度的空间[39]。在其中，地理要素表达的基础是位置，地理现象以瞬时状态为表现形式[40]。位空间是传统地图表达的框架，位置及其相对关系就是地理现象时空格局的内涵，而地理大数据挖掘的任务之一就是在位空间中揭示这种格局。

人类行为大数据是人类各种活动的反映。在与人相关的活动和关系中，流可以作为基本要素（流可定义为包含起点和终点的点对），这其中包括：人流、物流、信息流、资金流、关系流等。流可以视为两个结点（位置）之间的流动或交互[41-42]。在人类行为大数据中，流的起点与终点之间，距离不再是唯一衡量其关系的测度，而是与时间、成本、吸引力等多种测度并存[43]。人类行为大数据中对于人的关注不仅限于位置上的变化，而是以各种出行、社会关系等为核心。距离效应的减弱以及时空模式关注点的改变使得位空间已无法满足人类行为大数据更深层次的表达与分析，而需要借助流空间的概念。在流空间中，流作为基本单元，流和流之间的交错形成网络。流空间中表达的是交互关系，而流空间中数据挖掘的目的是提取位置之间的交互模式。目前传统的地图方式尚难

以有效表达流的模式，而全息地图和虚拟现实技术将有可能成为其新的载体。流空间与位空间的测度不一样，性质也不一样，其分析模型也存在本质的区别，而针对其时空特征提取方法的研究则是地理大数据挖掘重要的发展方向。

4.2 地理大数据挖掘的内容:模式与关系

本文将地理数据挖掘的目标定义为寻找地理对象之间、地理对象与环境之间存在的规则和异常。据此，地理大数据挖掘的内容也分为两个部分：① 地理时空模式的挖掘，其本质是发现地理对象的分布规则与时空分布；② 地理时空关系的挖掘，其本质是发现地理对象与不同环境因子之间的关系。由于地理大数据的特点，挖掘内容较之“小数据”也有所改变。

4.2.1 地理时空模式 地理学中目前公认的定理是空间相关性与空间异质性定理[44-45]。两个定理表述的意义看似相向，但实际是从两个侧面共同描述了地理现象：相近者相似，但彼此相异。在位空间中，地理学第一定律表现为属性相似度与距离的关系，而异质性则表现为空间上的非平稳性。在流空间中，空间相关性表现为空间网络结构的存在，即具有相近起点和终点的流构成了位置之间的联系，且联系的强度与距离等变量相关；而异质性则表现为不同单元之间流的差异性。地理大数据时空模式挖掘的本质是揭示地理对象因时空相关与异质性而形成的“异—同”规则及由此产生的时空分布。所谓“异”，是指地理对象之间的差别，而“同”则是指不同对象的共性。以地震数据的模式挖掘为例，一方面，需要确定提取丛集地震的“异—同”规则，从而将其与背景地震区分开来，并判别它们各自的统计分布类型（如泊松分布或威布尔分布等）；另一方面，在找出“异—同”规则的基础上，还要确定丛集地震和背景地震的空间分布范围和特征。前者属于“异—同”规则的推断，“同”类地震属于相同的统计分布，相“异”的地震分属不同的统计分布；后者属于时空分布的提取，而实际上，丛集地震和背景地震的时空分布可视为时空相关和异质性定律综合、直观的反映。针对时空模式，传统地理数据挖掘的主要任务包括：时空异质性的判别、地理时空异常模式的提取、空间分布模式的识别、地理时空演化趋势提取等。地理大数据所带来的改变集中体现在模式的类型及尺度两个方面：对于模式的类型，除了传统的栅格、要素、场的异质性与分布之外，地理大数据挖掘将更加关注序列、流与网络的结构与异质性等复杂模式；对于模式的尺度，由于具有的粒度、广度与密度的特征，地理大数据的挖掘将会产生更宏观、更综合、更精细的模式。

4.2.2 地理时空关系 地理对象与环境因子之间通常表现为相关或关联关系。相关关系通常用以刻画地理对象属性与环境因子之间的定量关系，例如：铅污染的程度与高速公路的远近[46]；而关联通常描述地理对象同时出现或存在的某种依赖关系，例如：盗窃与入室抢劫案件之间的关系[47]。地理时空关系中通常蕴藏着两方面的因素，以铅污染与高速公路之间的关系为例，一方面是变量之间的作用机制，即高速公路上汽车的尾气排放导致周围土壤中铅含量增加；而另一方面是这种土壤铅含量的变化与污染源远近之间的关系，即距离高速公路越近，铅的含量越高。针对时空关系的挖掘，地理大数据所带来的改变主要体现在关系的类型以及关系的转换上。一方面，变量之间关系的类型更加多样和复杂，非线性、不确定性及多元的时空关系成为大数据挖掘的重点之一[48]；另一方面，除了同类型空间下的时空关系挖掘，不同类型空间（如：社交空间、现实空间、情感空间）之间信息的反演与延伸成为大数据挖掘的主要特点之一，由此而导致的关系的转换也成为大数据思维的核心体现，例如：通过遥感数据反演经济状况[49]、利用搜索热词预测流感趋势[4]、应用手机数据反演城市土地利用[50]等。

需要说明的是，由于地理大数据的密度大、粒度小，相较于传统小数据，数据间具有很强的时空相关性，因此，从中容易“发现”各种时空关系，而这些关系往往涉及非常复杂的成因，是否具有因果关系需要仔细甄别。以盗窃与入室抢劫案的相伴发生为例，二者共现的实际原因可能是某些区域自然与社会环境较差而导致各种类型案件的高发，而并非存在明显的因果关系[47]。

4.3 地理模式的内在结构:尺度与叠加

前已述及地理大数据挖掘的目的是提取时空模式与时空关系。众多研究表明，地理格局、分布与过程的发生等都是尺度依存的。换句话说，地理模式都是在一定尺度下出现的，而地理大数据挖掘也离不开尺度。具体地，模式的挖掘就是要找出内部相“同”、对外相“异”的若干分布，而异质性与均匀性这种看似矛盾的性质可随尺度的变化而发生转换。这种转换正是地理格局尺度特征的重要原因。图2就揭示了点过程中异质性与均匀性在不同观测尺度下的转换，即在大尺度下呈异质性的点数据（图2a），其某个局部在小尺度下是均匀的分布（图2b）。由此，我们认为大尺度的复杂模式可以视为若干局部均匀模式的叠加。而对于地理大数据中的时空关系，也是尺度依存的。具体表现为，地理要素的尺度特征决定了相关关系的尺度性，大尺度的要素特征决定了总体趋势，而较小尺度的要素特征决定了局部的相关性，不同尺度上规律的叠加最终形成了总体上复杂的关系。例如，地形和气候要素的大尺度分布确定了中国人口东密西疏的宏观格局，而中尺度的要素（如局部的区位、地形、交通等）特征决定了人口的局部分布。多因子的多尺度叠加最终导致地理现象的复杂空间分布。

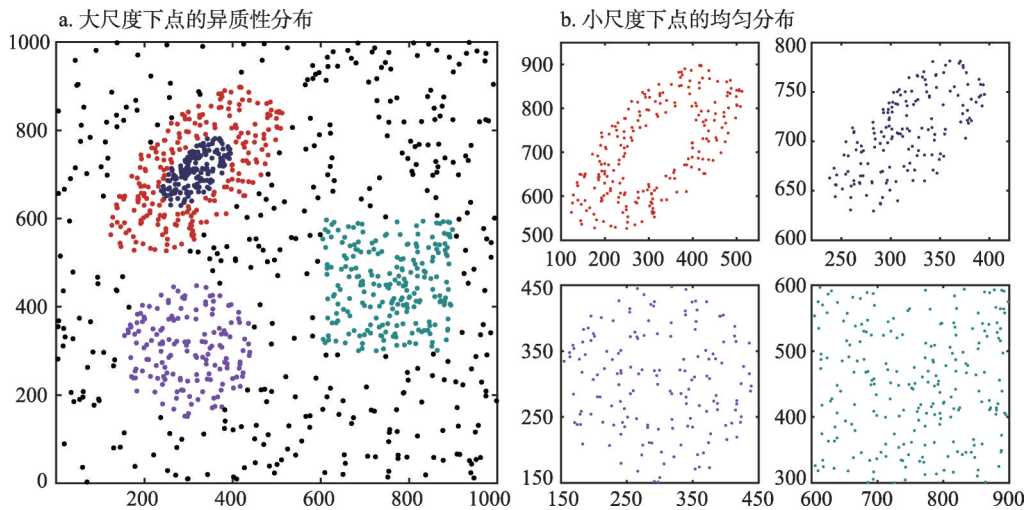

图2 地理点过程中均匀性与异质性随尺度的变化

Fig. 2 Transformation between homogeneity and heterogeneity of geographical point processes at difference scales (a. heterogeneity at large scale; b. homogeneity at small scale)

数据挖掘的尺度性体现在不同观察尺度下所挖掘出的模式不同，为此，裴韬等[51-52]提出点过程的混合模式分解理论。其主要思想为，地理点集现象可以视为不同时空尺度、不同性质地理过程叠加的结果。一方面，其时空特征可视为不同尺度下模式的叠加，而另一方面，地理现象产生的原因是不同时空尺度下因子相互叠加作用的结果。例如，一个地区的地震分布可以视为背景地震、断裂带地震以及地震序列的叠加，同时，地震的形成也是不同尺度构造运动叠加的结果。由于地理大数据所具有的特点，在一套地理大

数据中通常存在着从细—粗的多尺度特征。既然地理的模式与机制是综合叠加的结果，那么反过来，地理大数据挖掘就可以看成是对模式和关系进行分解的过程。

4.4 地理大数据挖掘的知识—地理模式背后的人地关系

地理数据挖掘一般都会伴随着未知模式的发现及其原因的探索两个阶段。针对对地观测数据的挖掘，所提取的模式为地表要素的格局，而针对人类行为数据的挖掘，提取的是人的行为模式。而模式的背后究竟是何种机制在起作用？地理大数据，尤其是人类行为大数据的出现，构成了从人地关系中揭示地理模式之机制的完备条件。地表要素的模式，表面是地的特征，其后则是人类行为的结果，例如，土地利用的分布与变化表面上属于地表要素的特征，而实际上则是人类活动的印记。手机通话数据看似反映人类行为的特征，在空间上展现出不同模式，但换角度观之却是城市功能区差异的体现[50]。地理大数据背后的模式，其机理都可以归结为人地关系，地的模式中蕴藏着人的因素，而人的行为模式受到地的制约。人和地的关系类似于中国古代易经中“阴”和“阳”的关系，对立统一，互有你我。通过人可知地、通过地可知人，而从地理模式中解析出的人地关系则是地理大数据挖掘的内涵。

地理学的发展经历了从经验范式（第一范式）—实证范式（第二范式）—系统仿真范式（第三范式）—数据驱动范式（第四范式）的演进[48]，现阶段地理学发展更加依赖于地理大数据及其分析方法。由此不难看出，地理大数据挖掘已成为地理学规律发现的重要工具。然而需要说明的是，地理大数据挖掘虽可产生地理模式及其与环境的相关性等“规律”，但“规律”中模式的真实性以及相关性中是否存在因果关系仍然需要通过观察、实验以及模拟等手段进行验证。

5 地理大数据的挖掘方法

地理大数据的挖掘方法非常多，其分类方案也存在多种标准，目前主要有以下几种：按照是否依赖于先验知识，可将其分为模型驱动和数据驱动两类挖掘方法[53-54]；依据挖掘任务，又可将其分为：数据总结、聚类、分类、关联规则、序列模式、依赖关系、异常以及趋势的挖掘方法[55-56]；根据挖掘对象可分为：关系数据、对象数据、图像数据、文本数据、多媒体数据、网络数据的挖掘方法等[57-58]；根据挖掘模型的特征可分为：机器学习方法、统计方法、神经网络方法和数据库方法等[59]。上述方案虽然都对数据挖掘方法进行了划分，但仍存在一定缺憾。针对任务、对象和特征的分类，均属于不完备的方案，一旦有新的挖掘任务、挖掘对象和挖掘模型，原有的分类方案就必须增补。不仅如此，划分方案中存在重叠，具体为：以挖掘对象进行分类的对象数据、图像数据与多媒体数据，以挖掘任务进行分类的异常与聚类模式的挖掘模型、趋势与序列模式的挖掘方法，以模型特征分类的机器学习与统计学方法等。

本文根据地理大数据挖掘的目标，将挖掘方法分为两类：时空分类的挖掘方法和时空关系的挖掘方法。前者用于区分地理对象的异同，旨在提取时空模式，而后者用于寻找时空变量的相关性，旨在挖掘地理对象与环境之间的时空关系。分类挖掘方法包括：空间聚类[60-61]、空间分类[62]、空间决策树[63]、点过程分解[51]等。时空关系的挖掘方法包括：关联规则挖掘[64-65]、主成分分析[66-67]、回归分析方法[68-69]等。此外，还有一些方法既可用于时空分类，也可用于时空关系挖掘，如神经网络[70-71]、支持向量机[72-73]，随机森林[74-75]等，可视具体算法模型而定。除了上述挖掘方法之外，还有部分方法是优化模型，用于

参数的估计，并辅助于数据挖掘方法，如深度学习策略[76]、EM算法[77]、MCMC算法[78]等。由于地理现象的复杂性，人工智能方法已经广泛应用于地理学的研究中，而人工智能与地理大数据的结合，将为地理大数据挖掘的发展提供新的动力源。

由于地理大数据的“5V”和“5度”特征，传统的数据挖掘方法在应用时需要顾及地理大数据的特点。首先，数据量对数据挖掘方法提出了严峻挑战，大数据必然带来更大的计算量，现有算法如何进行并行化和分布式计算是首当其冲的问题。其次，对于复杂场景的应用，大数据的出现往往会产生两类效应，一方面会简化模型，例如，对于复杂交通环境下时间最短路径的计算，传统研究会采用复杂的预测模型，而浮动车GPS大数据的出现，则使得时间最短路径的计算转化为简单的查询[79-80]；另一方面则会催生出更加复杂的模型，例如，浮动车GPS大数据的出现又会带来“如何进行出租车共享以节省资源”[81]等问题；第三，由于地理大数据的偏度以及不等精度的特点，对于地理大数据计算所得结果的真伪需要更为谨慎的评价，这样才能保证方法应用的效果。

6 结论

大数据的时代已经来临，地理学的发展也沐浴其中。作为一种特殊的大数据，地理大数据的挖掘所面临的挑战主要来自以下三个方面。第一，多源地理大数据的聚合问题成为深入解析时空模式和时空关系的瓶颈。地理大数据的种类众多，其时空粒度、模态性质和数据结构不尽相同，如何从时空和属性维度进行不同大数据的“垂直”融合和“横向”贯通成为未来深度挖掘地理大数据的关键。第二，地理大数据的有偏性和不等精度给地理大数据的知识发现带来诸多不确定性，如何有效评价和应用其挖掘结果是地理大数据研究不可回避的问题。第三，产生“有价值而非常识”的知识是地理大数据挖掘所面临的艰巨任务。大数据挖掘的目的就是知识的提取，目前的研究虽然在统计物理、人工智能等领域已有若干颠覆传统认识的成果，但在地理学的领域，数据挖掘所发挥的作用仍然缺乏说服力，正如腾讯位置大数据所展示的中国人口分布的不均匀模式一样（https://heat.qq.com/index.php），虽然震撼，但其显示的人口分界线却早已被几十年前诞生的“胡焕庸线”所揭示。

面向未来的挑战，地理大数据挖掘今后的发展脉络也不难梳理。首先，地理大数据挖掘应从更大尺度、更细粒度、更全维度出发解决地理学的基础问题，全球变化及其影响、人类行为特征、地表要素的社会特征、城市动力学演化等都是未来地理大数据关注的热点。其次，地理数据分析模型的发展将会更加顾及地理大数据的“5V”和“5度”特征，一方面，以人工智能为代表的数据挖掘方法，通过大数据样本的训练，在面对复杂地理问题时，其进展令人期待；另一方面，只有效率更高、更加稳健的算法方能适应地理大数据挖掘。最后，遥感研究从传统的对地表要素的观测将会延伸至对社会的观测，产生更多的科学和商业应用；而针对人类行为大数据的探索，则将感知社会拓展到对地表要素特征的反演，从而应用于与城市相关的研究中；最终两类大数据的结合，将成为揭示地理学中人地关系的重要突破口。

参考文献(References)

[1] Li Deren, Cheng Tao. Knowledge discovery from GIS databases. Acta Geodaetica et Cartographica Sinica, 1995, 24(1): 37-44. [李德仁, 程涛. 从GIS数据库中发现知识. 测绘学报, 1995, 24(1): 37-44.]

[2] Harvey J M, Han J W. Geographic Data Mining and Knowledge Discovery. London: CRC Press, 2009.

[3] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility. Science, 2010, 327(5968): 1018-1021.
[4] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data. Nature, 2009, 457(7232): 1012-1015.
[5] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484-489.
[6] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550 (7676): 354-359.
[7] Mayer-Schonberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. London: John Murray, 2013.
[8] Marr B. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. Chichester, UK: John Wiley & Sons, 2015.
[9] Liu Yu. Revisiting several basic geographical concepts: A social sensing perspective. Acta Geographica Sinica, 2016, 71 (4): 564-575. [刘瑜. 社会感知视角下的若干人文地理学基本问题再思考, 地理学报, 2016, 71(4): 564-575.]
[10] Liu Z, Ma T, Du Y, et al. Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. Transactions in GIS, 2018, 22(2): 494-513.
[11] Zheng Y, Liu Y, Yuan J, et al. Urban computing with taxicabs//Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, September 17-21, 2011: 89-98.
[12] Castro P S, Zhang D, Li S. Urban traffic modelling and prediction using large scale taxi GPS traces//Proceeding of Pervasive'12 Proceedings of the 10th International Conference on Pervasive Computing, Newcastle, UK, June 18-22, 2012: 57-72.
[13] Kong X, Xu Z, Shen G, et al. Urban traffic congestion estimation and prediction based on floating car trajectory data. Future Generation Computer Systems: The International Journal of EScience, 2016, 61: 97-107.
[14] Niu N, Liu X, Jin H, et al. Integrating multi- source big data to infer building functions. International Journal of Geographical Information Science, 2017, 31(9): 1871-1890.
[15] Newing A, Anderson B, Bahaj A B, et al. The role of digital trace data in supporting the collection of population statistics-the case for smart metered electricity consumption data. Population, Space and Place, 2016, 22(8): 849-863.
[16] NASA. New night lights maps open up possible real-time applications. https://www.nasa.gov/feature/goddard/2017/new-night-lights-maps-open-up-possible-real-time-applications, 2017.
[17] Chen J, Ban Y, Li S. China: Open access to Earth land-cover map. Nature, 2015, 514(7523): 434.
[18] Liu Yang, Liu Ronggao. Retrieval of global long-term leaf area index from LTDR AVHRR and MODIS observations. Journal of Geo-Information Science, 2015, 17(11): 1304-1312. [刘洋, 刘荣高. 基于LTDR AVHRR和MODIS观测的全球长时间序列叶面积指数遥感反演. 地球信息科学学报, 2015, 17(11): 1304-1312.]
[19] Oliver M A, Webster R. Kriging: A method of interpolation for geographical information systems. International Journal of Geographical Information System, 1990, 4(3): 313-332.
[20] Stein M L. Interpolation of Spatial Data: Some Theory for Kriging. New York: Springer Science & Business Media, 2012.
[21] Brunsdon C, Fotheringham A S, Charlton M E. Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis, 1996, 28(4): 281-298.
[22] Brunsdon C, Fotheringham S, Charlton M. Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician), 1998, 47(3): 431-443.
[23] Zhu A, Yang L, Li B, et al. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma, 2010, 155(3/4): 164-174.
[24] Zhu A, Qi F, Moore A, et al. Prediction of soil properties using fuzzy membership values. Geoderma, 2010, 158(3/4): 199-206.
[25] Zhang Wenjian. WMO integrated global observing system (WIGOS). Meteorological Monthly, 2010, 36(3): 1-8. [张文建. 世界气象组织综合观测系统(WIGOS). 气象, 2010, 36(3): 1-8.]
[26] NOAA/National Centers for Environmental Information. Global Historical Climate Network Daily - Description. https://www.ncdc.noaa.gov/ghcn-daily-description, 2018.
[27] Qian Chengcheng, Cheng Ge. Big data science for ocean: Present and future. Bulletin of Chinese Academy of Sciences, 2018, 33(8): 884-891. [钱程程, 陈戈. 海洋大数据科学发展现状与展望. 中国科学院院刊, 2018, 33(8): 884-891.]

[28] Yuan Y, Wei G, Lu Y. Evaluating gender representativeness of location-based social media: A case study of Weibo. Annals of GIS, 2018, 24(3): 163-176.

[29] Data Center of Sina Micro-blog. 2017 User Development Report of Sina Micro-blog. http://data.weibo.com/report/reportDetail?id=404, 2017. [新浪微博数据中心. 2017微博用户发展报告. http://data.weibo.com/report/reportDetail?id=404, 2017.]

[30] Zhao Z, Shaw S L, Xu Y, et al. Understanding the bias of call detail records in human mobility research. International Journal of Geographical Information Science, 2016, 30(9): 1738-1762.

[31] Congalton R G. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 1991, 37(1): 35-46.

[32] Zandbergen P A. Positional accuracy of spatial data: Non-normal distributions and a critique of the national standard for spatial data accuracy. Transactions in GIS, 2008, 12(1): 103-130.

[33] Cheng R, Emrich T, Kriegel H P, et al. Managing uncertainty in spatial and spatio-temporal data//Proceedings of the IEEE 30th International Conference on Data Engineering (ICDE), Chicago, IL, USA, Mar 31-Apr 04, 2014: 1302-1305.

[34] Zhao B, Zhang S. Rethinking spatial data quality: Pokémon go as a case study of location spoofing. The Professional Geographer, 2018. Doi: 10.1080/00330124.2018.1479973.

[35] Lazer D, Kennedy R, King G, et al. The parable of Google flu: Traps in big data analysis. Science, 2014, 343(6176): 1203-1205.

[36] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, 39(11): 27-34.

[37] Benz U C, Hofmann P, Willhauck G, et al. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS Journal of Photogrammetry and Remote Sensing, 2004, 58(3/4): 239-258.

[38] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. Analytics, 2011.

[39] Sun Zhongwei, Lu Zi. A geographical perspective to the elementary nature of space of flows. Geography and Geo-Information Science, 2005, 21(1): 109-112. [孙中伟, 路紫. 流空间基本性质的地理学透视. 地理与地理信息科学, 2005, 21(1): 109-112.]

[40] Han Zhigang, Kong Yunfeng, Qin Yaochen. Research on geographic representation: A review. Progress in Geography, 2011, 30(2): 141-148. [韩志刚, 孔云峰, 秦耀辰. 地理表达研究进展. 地理科学进展, 2011, 30(2): 141-148.]

[41] Castells M. Grassrooting the space of flows. Urban Geography, 1999, 20(4): 294-302.

[42] Goodchild M F, Yuan M, Cova T J. Towards a general theory of geographic representation in GIS. International Journal of Geographical Information Science, 2007, 21(3): 239-260.

[43] Batty M. The New Science of Cities. Cambridge, MA: Mit Press, 2013.

[44] Tobler W R. A computer movie simulating urban growth in the Detroit region. Economic Geography, 1970, 46(suppl.1): 234-240.

[45] Goodchild M F. The validity and usefulness of laws in geographic information science and geography. Annals of the Association of American Geographers, 2004, 94(2): 300-303.

[46] Du Zhenyu, Xing Shangjun, Song Yumin, et al. Lead pollution along expressways and its attenuation by green belts in Shandong province. Journal of Soil and Water Conservation, 2007, 21(5): 175-179. [杜振宇, 邢尚军, 宋玉民, 等. 山东省高速公路两侧土壤的铅污染及绿化带的防护作用. 水土保持学报, 2007, 21(5): 175-179.]

[47] Chen Long, Stuart Neil, Mackaness A Williams. Cluster and hot spot analysis in Lincoln, Nebraska, USA. Geomatics and Spatial information Technology, 2015, 38(3): 189-192. [陈龙, Stuart Neil, Mackaness A Williams. 美国内布拉斯加州林肯市犯罪行为的聚类及热点分布分析. 测绘与空间地理信息, 2015, 38(3): 189-192.]

[48] Cheng Changxiu, Shi Peijun, Song Changqing, et al. Geographic big-data: A new opportunity for geography complexity study. Acta Geographica Sinica, 2018, 73(8): 1397-1406. [程昌秀, 史培军, 宋长青, 等. 地理大数据为地理复杂性研究提供新机遇. 地理学报, 2018, 73(8): 1397-1406.]

[49] Keola S, Andersson M, Hall O. Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth. World Development, 2015, 66: 322-334.

[50] Pei T, Sobolevsky S, Ratti C, et al. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science, 2014, 28(9): 1988-2007.

[51] Pei T, Gao J, Ma T, et al. Multi-scale decomposition of point process data. GeoInformatica, 2012, 16(4): 625-652.

[52] Pei Tao, Li Ting, Zhou Chenghu. Spatiotemporal point process: A new data model, analysis methodology and viewpoint

for geoscientific problem. Journal of Geo-Information Science, 2013, 15(6): 793-800. [裴韬, 李婷, 周成虎. 时空点过程: 一种新的地学数据模型、分析方法和观察视角. 地球信息科学学报, 2013, 15(6): 793-800.]

[53] Niemeijer D. Developing indicators for environmental policy: data-driven and theory-driven approaches examined by example. Environmental Science & Policy, 2002, 5(2): 91-103.

[54] Miller H J, Goodchild M F. Data-driven geography. GeoJournal, 2015, 80(4): 449-461.

[55] Li Deren, Wang Shuliang, Shi Wenzhong, et al. On spatial data mining and knowledge discovery. Geomatics and information science of Wuhan university, 2001, 26(6): 491-499. [李德仁, 王树良, 史文中, 等. 论空间数据挖掘和知识发现. 武汉大学学报(信息科学版), 2001, 26(6): 491-499.]

[56] Li Deren, Wang Shuliang, Li Deyi, et al. Theories and technologies of spatial data mining and knowledge discovery. Geomatics and Information Science of Wuhan University, 2002, 27(3): 221-233. [李德仁, 王树良, 李德毅, 等. 论空间数据挖掘和知识发现的理论与方法. 武汉大学学报(信息科学版), 2002, 27(3): 221-233.]

[57] Chen M, Han J, Yu P. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and data Engineering, 1996, 8(6): 866-883.

[58] Džeroski S. Relational Data Mining. Boston, MA: Springer, 2009: 887-911.

[59] Wang Haiqi, Wang Jingfeng, Research on Progress of Spatial Data Mining. Geography and Geo-Information Science, 2005(4): 6-10. [王海起, 王劲峰. 空间数据挖掘技术研究进展. 地理与地理信息科学, 2005(4): 6-10.]

[60] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise//Proceeding KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, Aug 02-04, 1996: 226-231.

[61] Han J, Lee J G, Kamber M. An overview of clustering methods in geographic data analysis. Geographic data mining and knowledge discovery, 2009, 2: 149-170.

[62] Koperski K, Han J, Stefanovic N. An efficient two-step method for classification of spatial data//Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98), Vancouver, BC, Canada, July 11-15, 1998: 45-54.

[63] Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment, 1997, 61(3): 399-409.

[64] Huang Y, Shekhar S, Xiong H. Discovering colocation patterns from spatial data sets: A general approach. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1472-1485.

[65] Koperski K, Han J. Discovery of spatial association rules in geographic information databases//Proceedings of the 4th International Symposium on Large Spatial Databases (SSD 95), Portland, ME, USA, Aug 06-09, 1995: 47-66.

[66] Byrne G F, Crapper P F, Mayo K K. Monitoring land-cover change by principal component analysis of multitemporal Landsat data. Remote Sensing of Environment, 1980, 10(3): 175-184.

[67] Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. Nature, 2008, 456(7218): 98-101.

[68] Beale C M, Lennon J J, Yearsley J M, et al. Regression analysis of spatial data. Ecology Letters, 2010, 13(2): 246-264.

[69] McMillen D P. Geographically weighted regression: The analysis of spatially varying relationships. American Journal of Agricultural Economics, 2004, 86(2): 554-556.

[70] Atkinson P M, Tatnall A R L. Neural networks in remote sensing: Introduction. International Journal of Remote Sensing, 1997, 18(4): 699-709.

[71] Li X, Yeh A G O. Neural-network-based cellular automata for simulating multiple land use changes using GIS. International Journal of Geographical Information Science, 2002, 16(4): 323-343.

[72] Pal M, Mather P M. Support vector machines for classification in remote sensing. International Journal of Remote Sensing, 2005, 26(5): 1007-1011.

[73] Brereton R G, Lloyd G R. Support vector machines for classification and regression. Analyst, 2010, 135(2): 230-267.

[74] Gislason P O, Benediktsson J A, Sveinsson J R. Random forests for land cover classification. Pattern Recognition Letters, 2006, 27(4): 294-300.

[75] Mutanga O, Adam E, Cho M A. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. International Journal of Applied Earth Observation and Geoinformation, 2012, 18: 399-406.

[76] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.

[77] Moon T K. The expectation-maximization algorithm. IEEE Signal Processing Magazine, 1996, 13(6): 47-60.

[78] Andrieu C, De Freitas N, Doucet A, et al. An introduction to MCMC for machine learning. Machine Learning, 2003, 50

(1/2): 5-43.

[79] Yuan J, Zheng Y, Xie X, et al. T-drive: Enhancing driving directions with taxi drivers' intelligence. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(1): 220-232.

[80] Dai J, Yang B, Guo C, et al. Personalized route recommendation using big trajectory data. Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE), Seoul, South Korea, Apr 13-17, 2015: 543-554.

[81] Vazifeh M M, Santi P, Resta G, et al. Addressing the minimum fleet problem in on-demand urban mobility. Nature, 2018, 557(7706): 534-538.

Principle of big geodata mining

PEI Tao[1, 2], LIU Yaxi[1, 2], GUO Sihui[1, 2], SHU Hua[1, 2], DU Yunyan[1, 2], MA Ting[1, 2], ZHOU Chenghu[1, 2]

(1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper reveals the principle of geographic big data mining and its significance to geographic research. In this paper, big geodata are first categorized into two domains: earth observation big data and human behavior big data. Then, another five attributes except for "5V", including granularity, scope, density, skewness and precision, are summarized regarding big geodata. Based on this, the essence and effect of big geodata mining are uncovered by the following four aspects. First, as the burst of human behavior big data, flow space, where the OD flow is the basic unit instead of the point in traditional space, will become a new presentation form for big geodata. Second, the target of big geodata mining is defined as revealing the spatial pattern and the spatial relationship. Third, spatio-temporal distributions of big geodata can be seen as the overlay of multiple geographic patterns and the patterns may be changed with scale. Fourth, big geodata mining can be viewed as a tool for discovering geographic patterns while the revealed patterns are finally attributed to the outcome of human-land relationship. Big geodata mining methods are categorized into two types in light of mining target, i.e. classification mining and relationship mining. The future research will be facing the following challenges, namely, the aggregation and connection of big geodata, the effective evaluation of mining result and mining "true and useful" knowledge.

Keywords: spatial pattern; spatial relationship; spatial distribution; flow space; spatio-temporal heterogeneity; knowledge discovery