DOI:10.13203/j.whugis20200488 文章编号:1671-8860(2020)12-1919-11

异质稀疏分布时空数据插值、重构与预测方法探讨

程诗奋[1,2] 彭 澎[1,2] 张恒才[1,2] 陆 锋[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:时空数据挖掘是地理信息科学的核心研究命题。大数据时代,地理时空数据的爆炸性增长对时空知识发现提出了迫切的需求,促进了时空数据挖掘技术不断发展。然而,时空大数据普遍存在的异质性与稀疏分布特征制约了时空数据挖掘算法的实现,显著影响了自然和社会复杂系统刻画与分析能力。鉴于此,围绕异质稀疏分布时空数据表达与应用过程中面临的系列瓶颈问题开展研究,探讨了缺失时空数据插值、稀疏时空数据重构、时空状态预测等时空数据挖掘重点任务的研究现状和存在问题,凝练了关键的科学问题,并提出了相应的解决方案,以期丰富时空数据挖掘领域的方法体系,提升时空数据建模的质量与应用价值。

关键词:时空自相关;时空异质性;时空插值;时空预测;多任务多视图学习

中图分类号:P208 **文献标志码**:A

随着传感器网络、移动定位技术的不断普及和发展,数据采集与计算单元的外延不断扩展,地球科学经历了一场从数据贫乏到数据丰富的重大革命[1-2]。随着时空维度的不断增长和刻画粒度的不断细化,海量的泛化时空数据已成现实[3]。这些泛化时空数据蕴含着丰富的信息,对时空知识发现提出了迫切的需求,促进了时空数据挖掘技术的不断普及和发展[4-9]。

时空数据挖掘旨在从海量的时空数据中发现之前未知,但潜在有用的知识、结构、关系或模式[10-11]。给定一个时空数据集,首先需要进行数据的预处理工作,包括去除噪声、缺失数据插值等。在此基础上,选用合适的时空数据挖掘算法分析时空数据并输出时空模式。输出的时空模式根据研究问题的不同可分为时空聚类、时空预测、时空异常检测、频繁模式挖掘等[12]。

时空数据的缺失和稀疏分布是普遍存在的现象。在目前的大数据时代,由于数据分析需求不断深化和时空粒度不断细化,时空数据缺失和稀疏分布问题更加突出。而时空数据插值与重构精度及易用性对后续的时空数据挖掘过程具有重要影响。时空数据插值是偶发性缺失数据的推断过程。现有的时空插值方法虽然考虑了时空异质性[13],但未考虑时空数据的缺失模式、插值样本的高效选择、时间和空间的非线性交互关系,插值精度有待提升。时空数据的稀疏重构是系统性的数据加密或重采样过程。研究者提出了多种方法来解决数据稀疏性问题[14-15]。统计和机器学习方法通过考虑时空依赖性提高稀疏重构精度,但模型求解复杂,通常难以部署。轻量级的模型易于构建,但无法捕获地理空间数据的时空依赖性,重构精度有限。

时空数据挖掘算法受到时空自相关性和时空异质性的统计约束[16]。现有的统计和机器学习方法通常未全面描述这些时空特征,导致难以获取细粒度的时间非平稳性变化特征和复杂地理过程的周期性和趋势性。此外,时空异质性导致的局部模型结构无法描述预测任务之间的全局时空相关性,使预测模型丧失了全局预测能力。预测模型的参数优化同样也存在问题,极大地限制了时空数据建模能力。

可以看到,在时空数据挖掘过程的几个关键环节,现有方法均存在一些不足。因此,有必要探索新的建模方法,以提升现有时空数据模型的

收稿日期:2020-11-06

项目资助:国家自然科学基金(41631177,41771436);中国博士后科学基金(2019M660774,2020T130644)。

第一作者:程诗奋,博士,博士后,研究方向为时空数据挖掘。chengsf@lreis.ac.cn

通讯作者:陆锋,博士,研究员,博士生导师。luf@lreis.ac.cn

学习能力、预测精度以及应用价值。鉴于此,本文首先回顾和总结了缺失时空数据插值、稀疏时空数据重构、时空数据预测等主题的研究现状和存在问题,然后探讨了所要解决的关键科学问题,提出了对应的解决方案,并展望了未来的研究方向。

1 时空数据插值研究现状

1.1 缺失时空数据插值

缺失数据插值是时空数据分析的关键步骤。业界提出了很多方法来处理时空数据缺失问题[17-19]。空间维度插值方法主要考虑数据之间的空间相关性来内插缺失数据,包括反向距离权重算法(inverse distance weight,IDW)[20-21]、克里金模型[22]、BSHADE(biased sentinel hospitals-based area disease estimation)点估计模型(point estimation model of BSHADE, PBSHADE)[23]等,一定程度上考虑了数据分布的空间自相关和空间异质性。时序数据插值方法则利用时间序列预测方法来内插缺失数据,如指数平滑模型(simple exponential smoothing, SES)[24]、自回归整合移动平均[25]。考虑到单维度插值方法只考虑了空间或者时间维度信息,难以达到满意的插值效果,近年来一些学者将单维度插值方法扩展到时空维度,产生了一系列的时空缺失数据插值方法,包括时空反距离加权模型[26]、时空克里金模型(spatiotemporal-Kriging, ST-Kriging)[27]、顾及时空异质性的插值模型(spatiotemporal heterogeneous covariance method, ST-HC)[13]。然而,当前的时空插值算法未全面考虑时空数据的缺失模式、时空异质性、时空非线性关系,对插值精度产生了一定的影响。

1.2 稀疏时空数据重构

地理空间数据呈现爆炸性增长,更加精细的时空分析粒度需求也在同步增长,时空数据稀疏问题仍然存在,甚至变得更加紧迫。稀疏分布时空数据重构问题也可以理解为系统性稀疏分布时空数据插值问题。目前的时空数据稀疏问题解决方案可以粗略分为机器学习方法和统计方法。基于机器学习的方法如矩阵分解[28]、张量分解[29]、半监督学习[30]、基于插值的算法等,通常需要构建求解的目标函数,采用梯度下降等数值计算方法迭代训练模型以达到最优的重构精度。复杂的空间和时空统计插值方法如高精度曲面模型[31-32]、克里金模型、BSHADE点估计模型、顾及时空异质性的插值模型等,通常需要逐点求解偏微分方程来计算插值样本最优权重。这些方法在重构精度上取得了一定的效果,但是由于模型求解的复杂性,通常部署难度较大。经典的IDW算法和SES算法等轻量级模型虽然简单易用,但单一的轻量级模型无法同时捕获时空依赖性和非线性时空关系,因此难以满足稀疏时空数据重构精度的要求。

1.3 时空序列预测

时空序列预测的基本目标是学习从输入特征(也称为自变量)到输出变量(也称为因变量)的映射[33],可以分为基于统计的参数模型和基于机器学习的非参数模型[34]。时空序列预测模型的常用统计方法大多是从经典的空间统计方法扩展而来,通过进一步考虑时间维度信息,构成时空预测模型,包括时空自回归移动平均[35-36]、时空地理加权回归模型[37-38]等。通过模型的不断改进,现有的统计模型初步实现了时空自相关和时空异质性的表达。然而,现有模型中时空关系通常是以线性方式表达,并且由于时空异质性的存在,模型将每个地理单元当作单独的预测任务,忽略了地理单元之间的全局时空相关性。基于机器学习的非参数模型近年来在时空预测建模领域得到广泛应用,包括时空K近邻模型(spatiotemporal K near neighbor, STKNN)[39-40]、深度学习模型[41-42]等。然而,机器学习模型的样本独立同分布假设,有悖于时空数据的时空异质性。例如,STKNN模型忽略了空间异质性的存在,通常采用全局固定的模型结构,具体表现为在整个研究区域每个空间对象都具有固定的空间邻居、时间窗口、时空权重和时空参数,从而难以描述不同空间对象差异性的变化模式。针对整个时间范围建模的方式忽略了时空过程的非平稳变化。现有的大多数深度学习模型,如时空残差神经网络[43]、图卷积神经网络[44]等,仅考虑时空数据的自相关特性和时空非线性交互,对空间异质性和时间非平稳性刻画不足。

2 时空数据挖掘核心问题

针对上述研究现状,本文凝练了现有时空数据挖掘算法在几个关键环节需要解决的关键科学问题。

1)时空异质性和缺失模式对插值模型的作用和机理分析

时空数据缺失是普遍存在的现象,由于各种

原因可能存在多种缺失模式,不同的缺失模式对插值精度具有不同的影响。另外,空间异质性和时间非平稳性从空间和时间层面对插值过程进行约束,而时间和空间层面又存在复杂的非线性时空关系。因此,如何在插值过程中消除缺失模式对插值精度的影响,分解和耦合时空异质性,时空数据的精确插值,是亟待解决的关键问题。

2)实现模型的稀疏重构精度和易用性之间的均衡

目前存在多种统计和机器学习方法来解决时空数据稀疏问题。现有稀疏重构模型求解过程的复杂性使模型的易用性难以保证,时空数据的海量涌现使得该问题更为突出。轻量级模型简单易用,但无法精确地表达时空依赖关系,使稀疏重构精度难以满足要求。如何改进轻量级稀疏重构模型,实现重构精度和易用性之间的均衡,满足实际应用需求,是当前的关键问题。

3)时空预测任务中空间异质性和时间非平稳性的统一表达

空间异质性和时间非平稳性是时空数据的本质特性。现有机器学习方法基于样本独立同分布假设,采用全局静态的模型结构,在整个时空范围采用统一的时间区间剖分方式,难以充分表达复杂地理过程的内在机理和作用机制。如何在现有预测模型中实现空间异质性和时间非平稳性的统一表达,是拓展当前机器学习模型的适用性,提升时空数据建模准度与应用能力的基础问题。

4)时空预测任务中全局时空相关性和时空异质性之间的矛盾关系

考虑到时空异质性的存在,现有的统计和机器学习方法针对每个同质的子区域或者不同的空间对象来构建局部的预测模型,以表达地理要素的非平稳变化。局部模型结构使得各个地理要素之间相互独立,无法表征地理要素之间的全局时空相关性。然而,区域与区域之间存在空间溢出效应,对象与对象之间存在空间关联,忽略这种全局时空相关性将难以充分表达复杂的地理过程。如何在现有预测模型中实现全局时空相关性和时空异质性的深度耦合和协同,形成一体化的时空建模框架,是亟待解决的关键问题。

3　时空数据重构解决方案

针对上述异质稀疏分布地理空间数据在表达与应用中需要解决的关键问题,本文提出以下解决方案。

3.1　时空缺失数据的渐进式插值方法

时空数据可能存在多种缺失模式[45]。不考虑时空数据缺失模式的时空插值方法精度有限[46],有些情况下甚至无法完全修复缺失数据。例如BSHADE点估计模型在数据连续缺失的情况下,会导致缺失数据矩阵在求解过程中出现奇异值甚至无法求解的问题,从而产生较大的插值误差。因此,可以考虑采用渐进式插值思想,首先对缺失数据集作粗粒度插值,以避免连续数据缺失对后续细粒度插值算法执行的影响,然后在此基础上考虑空间异质性、时间异质性以及时空非线性关系,从不同层面约束插值过程。可采用集成学习方法,将插值过程分解为空间维度插值、时间维度插值以及时空整合,整合多个视角的不同知识,更加全面和精确地描述复杂的插值过程。时空缺失数据的渐进式插值方法(two-step method for spatiotemporal missing data reconstruction,ST-2SMR)的整体架构如图1所示[47]。

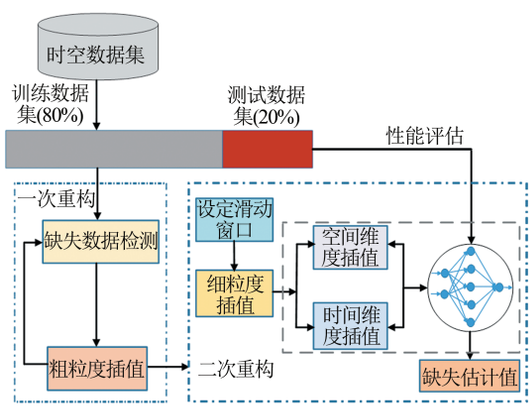

图 1　时空缺失数据的渐进式插值方法[47]

Fig.1　A Two-step Method for Missing Spatiotemporal Data Reconstruction[47]

首先,分别在时间和空间维度上构造异质协方差函数,通过最大化目标函数得到缺失数据在时间维度和空间维度的插值结果。然后,通过未缺失数据的时空插值结果构造样本训练神经网络模型,挖掘时间和空间维度的非线性关系。在得到神经网络模型后,输入缺失数据的时空插值结果,即可得到整合后的最终插值结果。采用北京市 2014-05-01—2015-04-30 的缺失空气质量数据集作为实验数据对该方法进行了验证,这些数据从北京市 36 个空气质量监测站点按小时间隔收集,包含 8 579 条记录。如图 2 所示,ST-2SMR 方法全面考虑时空数据缺失模式、样本选择和时空关系,获得了比现有方法更高的插值精度,并

且可以保证完全修复缺失数据。图2中，MAE（mean absolute error）表示平均绝对误差，MRE（mean relative error）表示平均相对误差，RC（ratio of construction）表示重构率。

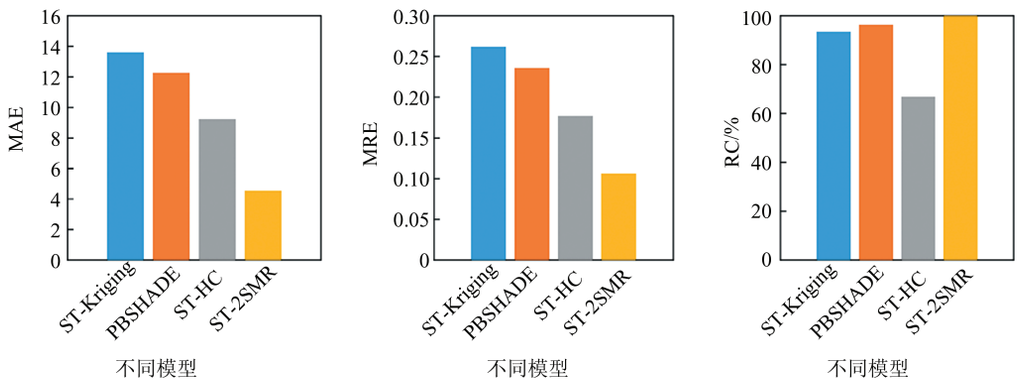

图2 城市空气质量数据集上的插值性能

Fig.2 Interpolation Performance of the ST-2SMR Using Urban Air Quality Dataset

3.2 轻量级稀疏时空数据重构方法

针对现有时空数据稀疏重构方法无法保证模型重构精度和易用性之间均衡的问题，可通过集成多个轻量级模型来保证模型的易用性，合理量化时空依赖性来提高模型的稀疏重构精度。轻量级的集成时空重构方法（lightweight ensemble spatiotemporal interpolation，ST-ISE）的整体解决思路如图3所示[48]。

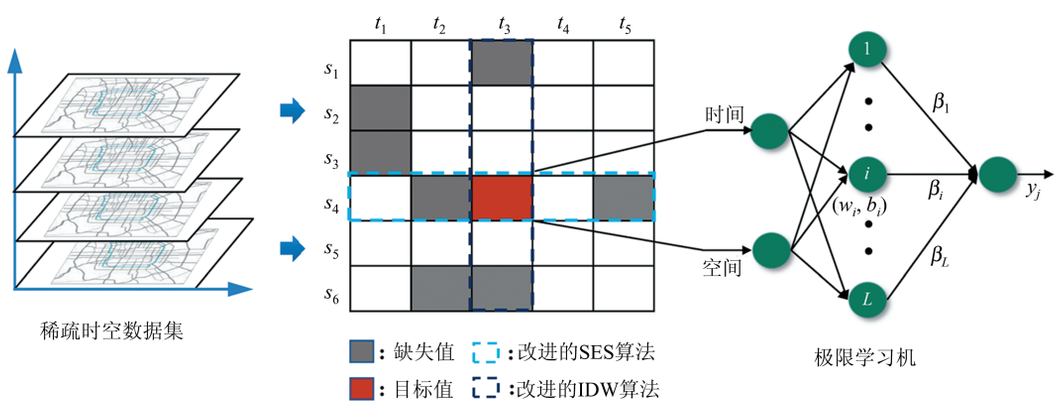

图3 轻量级稀疏时空数据重构方法

Fig.3 A Lightweight Ensemble Reconstruction Model for Sparse Spatiotemporal Data

在轻量级模型方面，经典的IDW和SES应用广泛。然而，传统SES算法仅使用缺失数据所在时间间隔之前的样本数据进行插值计算，忽略了之后的样本数据，且无法自动选择插值的时间窗口，会导致过多的不相关数据参与计算，从而降低插值精度。因此，在时间维度上，可在传统简单指数平滑算法中引入动态滑动窗口，捕获地理过程的时空演化特征，从而提高传统简单指数平滑算法表达时间依赖性的能力。传统的IDW算法通常采用空间对象之间的欧氏距离来刻画空间相关性。这一方法可以很好地描述研究区域的物理属性，但忽略了空间对象关联的时空模式变化。另外，当空间对象缺乏精确的空间坐标时，很难用欧氏距离准确描述空间相关性。因此，在空间维度上，可考虑空间对象的时空模式，采用相关距离来取代传统反向距离权重算法中的欧氏距离，从而提高传统反向距离权重算法表达空间依赖性的能力。为了加快模型的训练速度以及保证模型的轻量化，可引入极限学习机拟合时空非线性关系来整合时空重构结果。采用北京2018-04-15 5 min采样频率的浮动车速度数据集对ST-ISE模型进行了评估。如图4所示，与现有模型相比，该方法的MAE降低了10.93%～52.48%，表现出更高的重构精度，证明了在稀疏时空数据重构过程中精确表达时空依赖关系的重要性。

3.3 顾及时空异质性的动态预测模型

时空预测旨在通过有效表达时空数据内在的时空自相关性和时空异质性，描述时空变量之间的关系，推测未知地理现象或过程[16]。从建模

角度看,时空异质性导致样本的分布在整个研究区域随时间和空间动态变化,因此需要设计时空动态模型,使模型的参数随时间空间自适应变化[49]。

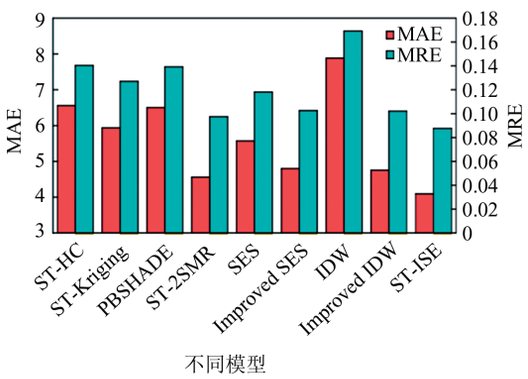

图 4　不同模型在浮动车速度数据集上的重构精度

Fig.4　Reconstruction Accuracy of Different Models for the Floating Vehicle Speed Dataset

考虑到地理现象时间非平稳性以及周期性的特点,在特定的时间区间内,时空模式的变化趋于平稳,具有统计同质性。在不同的时间区间,时空模式表现出明显的差异性[50-51]。因此,如何自动识别地理空间的时空模式,合理区分时空模式在不同时间区间的演化特征,是精确表达时间非平稳性的关键[49]。针对时空模式识别,可利用传统的时间序列聚类算法,把具有相似行为的时空对象划分到同一组。在获取每个时空模式的时间序列后,可进一步利用聚类算法实现时间区间的剖分。然而,传统的聚类算法,如 k-means 聚类和层次聚类,在聚类过程中忽略了时间序列的连续性特点,使得先前的信息(即来自先前连续时间间隔的测量结果)无法用于建模并生成地理变量的预测值。因此,可采用 Warped k-means 算法,通过强加顺序约束来改进传统 k-means 算法,针对每种时空模式,自动剖分时间区间,用于比较不同时间区间时空模式的时间非平稳变化。在自动识别地理空间的时空模式以及获取每种时空模式的分区策略后,考虑空间异质性的存在,可针对不同模式下不同时间区间中的不同空间对象构建自适应的模型[52],从而构建顾及时空异质性的动态预测模型。

考虑到交通数据具有明显的时空异质性,因此以本文构建的动态的 STKNN 模型(dynamic STKNN, D-STKNN)实现短时交通的高效预测来展开阐述,整体架构如图 5 所示[49]。通过刻画每个路段的历史交通状况特征,采用近邻传播聚类算法来识别路段相似的交通模式。针对每种交通模式,利用 Warped k-means 算法自动划分时间区间,通过比较不同路段在不同交通模式的时间区间下交通状况的差异性,来刻画时间非平稳性特征。在此基础上,为每个路段构建自适应 STKNN 过程(Adaptive-STKNN),来适应不同路段交通模式的变化,包括自适应的空间邻居、时间窗口、时空权重、时空参数[53],进一步刻画空间异质性。如图 6 所示,利用北京浮动车速度数据集对该框架进行了评估,数据采集频率为 5 min,时间周期从 2012-03-01—2012-04-30。可以看出,D-STKNN 模型的预测性能优于现有的模型,证明了在短时交通预测建模中同步考虑空间异质性和时间非平稳性的重要性。图 6 中,MAPE(mean absolute percentage error)表示平均绝对百分比误差,RMSE(root mean square error)表示均方根误差。

3.4　基于多任务多视图的时空建模框架

时空数据变化通常具有周期性和趋势性特点。仅考虑时空邻近性难以全面反映时空交互过程。更重要的是,时空异质性导致的局部建模方式存在较大问题。针对单个地理单元建模的方式将每个地理单元当作单独的预测任务,忽略了预测任务之间的全局时空相关性。多个子模型的存在使得构建的预测模型丧失了全局预测能力[54]。因此,可先利用多任务多视图特征学习范式,在保证时空异质性的前提下,构建统一的时空预测模型,实现对地理空间中地理现象未知属性的同步预测[55]。从多视图的视角出发,针对每个空间对象构建时空邻近视图、周期视图和趋势视图,全面刻画不同时段的历史时空状态对当前时空状态的影响,实现时空异质性的统一表达[56]。由于空间异质性的约束,使得不同空间对象具有不同空间邻居和时间窗口,从而产生不同维度的时空状态矩阵。若直接利用时空状态信息,无法适应多任务多视图特征学习的范式。因此,可先利用多核学习来挖掘地理空间相似的时空模式,得到每个视图的预测结果。然后,将每个视图的预测结果进行高层的语义映射,作为时空多任务多视图学习模型的输入特征。同时将每个空间对象的预测当作一个任务,可使每个任务具有一致的特征维度;再采用多任务多视图特征学习机制,在保证时空异质性的前提下,建立单一空间对象的多视图模型联立训练过程,学习任务之间的全局相关性和视图之间的一致性[57],同时限制所有空间对象选择一组共享特征,即可

解决时空异质性和全局时空相关性的协同和耦合问题。

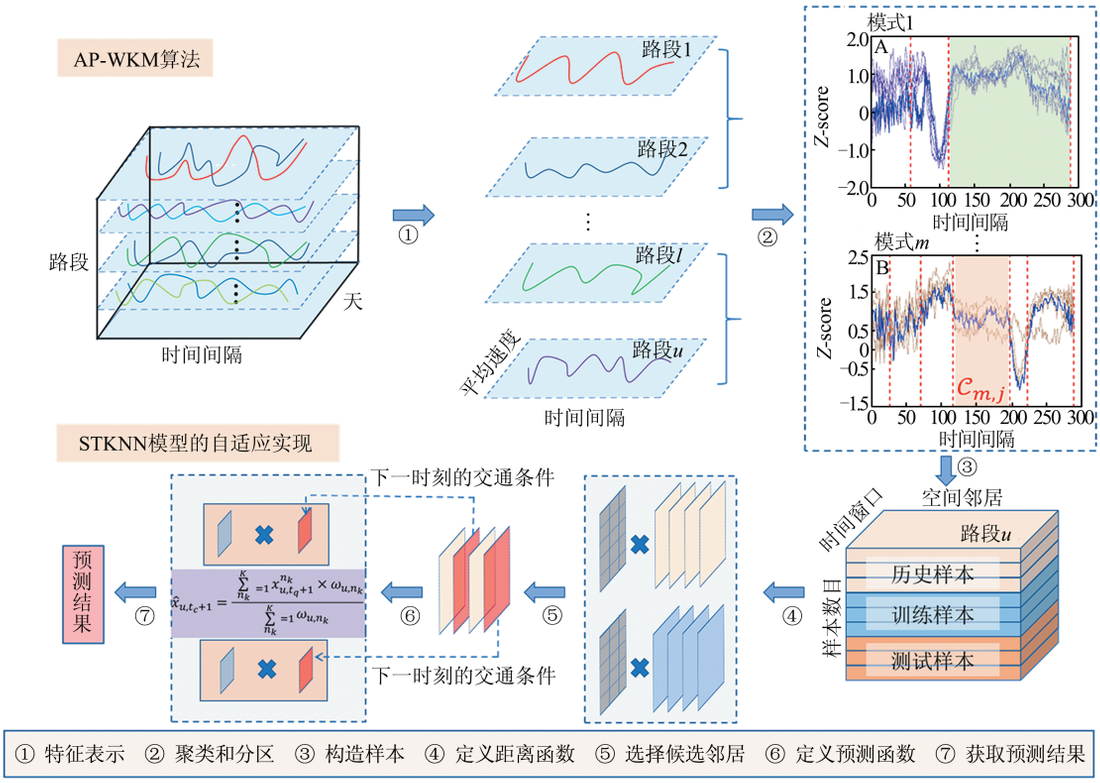

图5 D-STKNN模型整体架构图

Fig.5 Schematic Diagram of the D-STKNN Model

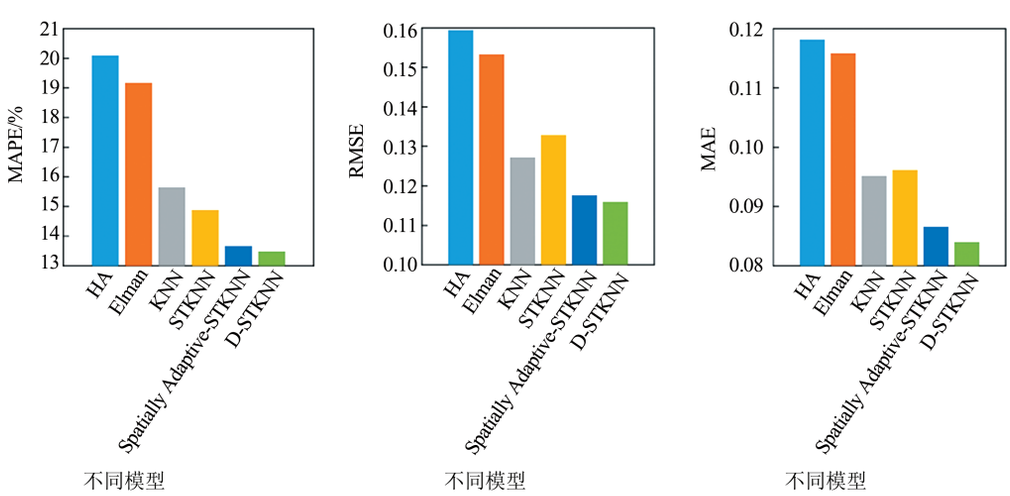

图6 不同模型在浮动车速度数据集上的预测精度

Fig.6 Forecast Accuracies of Six Models for the Floating Vehicle Speed Dataset

本文以短时交通预测为例,介绍一种基于多任务多视图特征学习的预测模型(spatiotemporal multi-task and multi-view feature learning model, stRegMTMV)[55],整体架构见图7。在道路网络中路段的交通状况具有时空约束性。在空间维度,路段的交通状态常受其周围路段的影响,如某路段的拥堵经过一段时间会传播到周围路段,导致区域拥堵。且这种影响具有空间异质性,即不同的路段其受影响的周围路段数目不一致[50-51]。在时间维度,由于交通模式的存在,路段的交通状态通常和历史时刻的交通状态存在关联,具有时间的邻近性、周期性、趋势性[56]。因此,在stRegMTMV模型中,可将每个路段的历史交通状况时间序列进行重组,通过时间和空间维度信息的融合,形成时空邻近矩阵、时空周期矩阵、时空趋势矩阵来对路段的时空依赖关系进行全面刻画。

在此基础上,采用Adaptive-STKNN作为核函数分别对3个视图建模,分别得到邻近视图、周

期视图和趋势视图的预测结果,将每个视图的预测结果作为时空多任务多视图学习模型的输入特征,从而实现整个道路网络交通状况的同步预测。如图 8 所示,利用§3.3 的浮动车速度数据集对该框架进行了评估。结果表明,基于该框架的短时交通预测精度优于现有的 9 种时空预测模型,凸显了实现全局时空相关性和时空异质性统一表达的重要性。

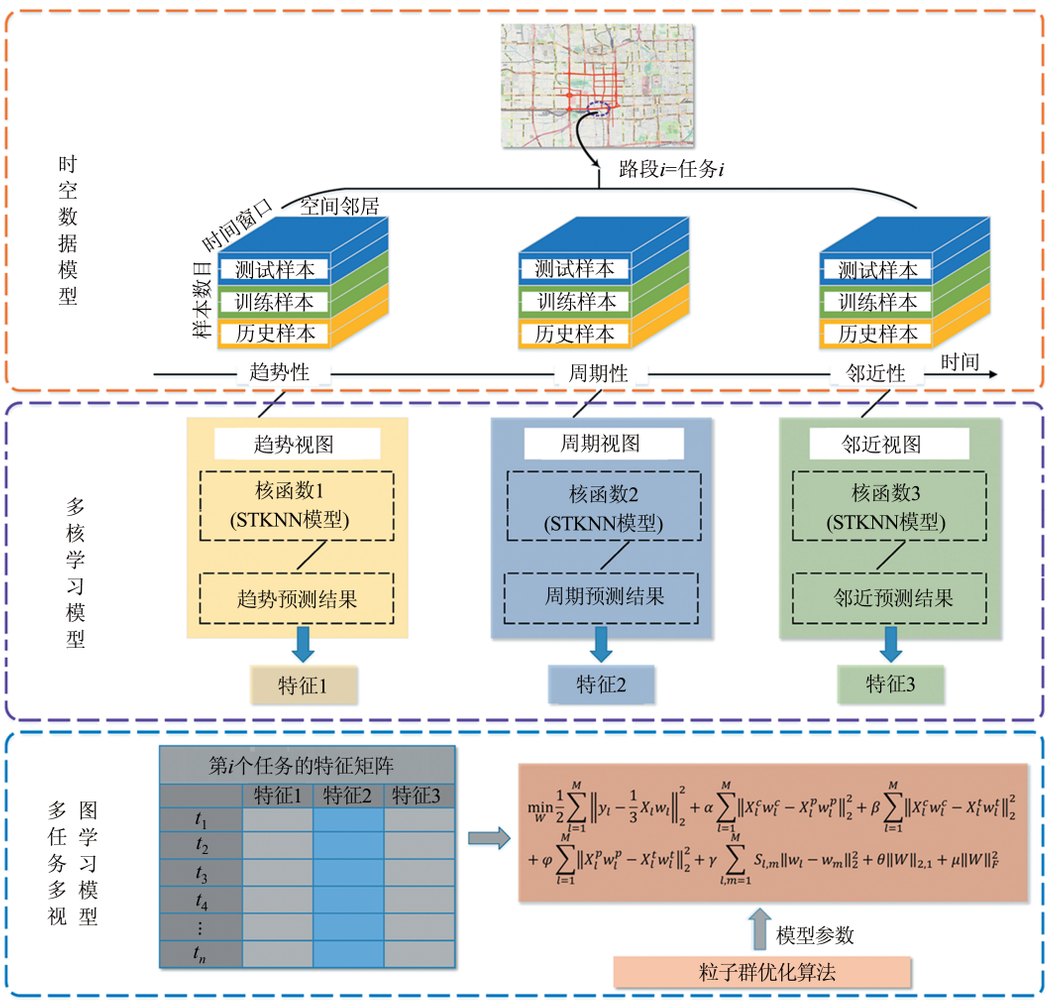

图 7　stRegMTMV 模型的整体架构

Fig.7　Schematic Diagram of the stRegMTMV Model

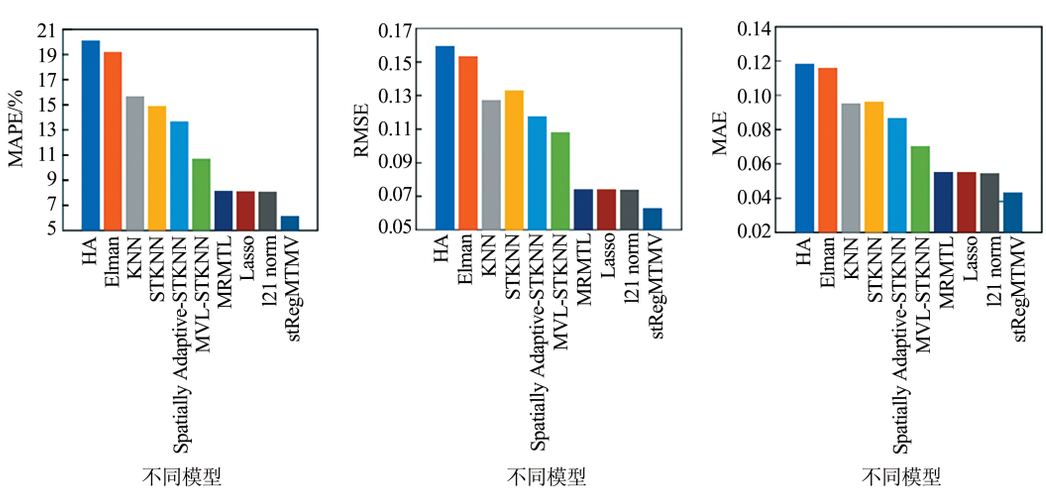

图 8　9 种不同模型在浮动车速度数据集上的预测精度

Fig.8　Forecast Accuracies of Nine Models for the Floating Vehicle Speed Dataset

4 总结与展望

地理时空数据的不断积累与分析尺度的不断细化之间存在永恒矛盾。时空数据缺失与稀疏分布依然是当前地理空间大数据挖掘面临的普遍问题。此外,时空数据具有空间异质性和时间非平稳性的本质特征,给传统的统计和机器学习方法带来了巨大挑战。本文系统回顾了时空数据挖掘过程的几个关键环节的研究现状和存在问题,凝练了4个关键的科学问题,并提出了相应的解决方案。

针对时空数据重构与预测,仍然有很多问题值得进一步深入探讨。

1)空间异质性包括空间局域异质性和空间分层异质性[58]。本文给出了空间局域异质性的建模思路,例如在构建时空缺失数据的渐进式插值方法时,利用期望比来刻画不同位置属性值的差异;在构建时空异质性的动态预测模型时,检验了不同位置的空间邻居、时间窗口、时空参数的异质性。如何实现空间分层异质性和现有的机器学习模型的有机融合,实现不同层之间的交互需要进一步研究。此外,空间异质性还可表现为空间各向异性,如何在时空预测模型中表达各向异性的空间依赖性,也是值得进一步研究的问题。

2)时空数据具有时空尺度依赖特性[59]。时空尺度是地理学的基本概念,地理现象的分布、格局和模式在不同的观测尺度会呈现不同的特征,而不同时空尺度的对象之间又存在依赖性[12]。因此,在时空数据挖掘过程中,尺度效应是不可忽视的关键因素。如何在现有时空统计和机器学习方法中实现地理空间尺度的深度耦合和协同,是理解复杂地理过程的关键环节[60]。

3)时空数据挖掘包含庞大的分支体系。本文仅对时空缺失数据插值、时空稀疏数据重构、时空预测、多任务多视图建模等问题进行了探讨。事实上,时空聚类、时空异常分析、时空关联分析等问题同样受到时空自相关和时空异质性的统计约束[8]。因此,如何将本文提出的建模思路应用于其他时空数据挖掘任务中,拓展其应用价值,是后续需要进一步研究的工作。

参 考 文 献

[1] Zhou Chenghu, Zhu Xinyan, Wang Meng, et al. Panoramic Location Based Map[J]. *Progress in Geography*, 2011, 30(11): 1 331-1 335(周成虎,朱欣焰,王蒙,等. 全息位置地图研究[J]. 地理科学进展, 2011, 30(11): 1 331-1 335)

[2] Li Deren. Towards Geo-Spatial Information Science in Big Data Era[J]. *Acta Geodaetica et Cartographica Sinica*, 2016, 45(4): 379-384(李德仁. 展望大数据时代的地球空间信息学[J]. 测绘学报, 2016, 45(4): 379-384)

[3] Li Deren. The Intelligent Processing and Service of Spatiotemporal Big Data[J]. *Journal of Geo-Information Science*, 2019, 21(12): 1 825-1 831(李德仁. 论时空大数据的智能处理与服务[J]. 地球信息科学学报, 2019, 21(12): 1 825-1 831)

[4] Karpatne A, Ebert-Uphoff I, Ravela S, et al. Machine Learning for the Geosciences: Challenges and Opportunities[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 31(8): 1 544-1 554

[5] Atluri G, Karpatne A, Kumar V. Spatio-Temporal Data Mining: A Survey of Problems and Methods[J]. *ACM Computing Surveys (CSUR)*, 2018, 51(4): 1-41

[6] Lu Feng, Zhang Hengcai. Big Data and Generalized GIS[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 645-654(陆锋,张恒才. 大数据与广义GIS[J]. 武汉大学学报·信息科学版, 2014, 39(6): 645-654)

[7] Li Qingquan, Li Deren. Big Data GIS[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 641-644, 666(李清泉,李德仁. 大数据GIS[J]. 武汉大学学报·信息科学版, 2014, 39(6): 641-644, 666)

[8] Shekhar S, Jiang Z, Ali R Y, et al. Spatiotemporal Data Mining: A Computational Perspective[J]. *ISPRS International Journal of Geo-Information*, 2015, 4(4): 2 306-2 338

[9] Wang Jinfeng, Ge Yong, Li Lianfa, et al. Spatiotemporal Data Analysis in Geography[J]. *Acta Geographica Sinica*, 2014, 69(9): 1 326-1 345(王劲峰,葛咏,李连发,等. 地理学时空数据分析方法[J]. 地理学报, 2014, 69(9): 1 326-1 345)

[10] Cheng T, Haworth J, Anbaroglu B, et al. Handbook of Regional Science[M]. Berlin, Heidelberg: Springer, 2014: 1 173-1 193

[11] Li Deren, Yao Yuan, Shao Zhenfeng. Big Data in Smart City[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 631-640(李德仁,姚远,邵振峰. 智慧城市中的大数据[J]. 武汉大学学报·信息科学版, 2014, 39(6): 631-640)

[12] Pei Tao, Liu Yaxi, Guo Sihui, et al. Principle of Big Geodata Mining[J]. *Acta Geographica Sinica*, 2019, 74(3): 586-598(裴韬,刘亚溪,郭思慧,等.

地理大数据挖掘的本质[J]. 地理学报, 2019, 74(3): 586-598)

[13] Deng M, Fan Z, Liu Q, et al. A Hybrid Method for Interpolating Missing Data in Heterogeneous Spatio-Temporal Datasets[J]. *ISPRS International Journal of Geo-Information*, 2016, 5(2): 13

[14] Chen X, Wei Z, Li Z, et al. Ensemble Correlation-based Low-rank Matrix Completion with Applications to Traffic Data Imputation[J]. *Knowledge-Based Systems*, 2017, 132: 249-262

[15] Asif MT, Mitrovic N, Dauwels J, et al. Matrix and Tensor Based Methods for Missing Data Estimation in Large Traffic Networks[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2016, 17(7): 1 816-1 825

[16] Deng Min, Cai Jiannan, Yang Wentao, et al. Spatio-temporal Analysis Methods for Multi-modal Geographic Big Data[J]. *Journal of Geo-Information Science*, 2020,22(1):41-56(邓敏,蔡建南,杨文涛,等. 多模态地理大数据时空分析方法[J]. 地球信息科学学报, 2020, 22(1): 41-56)

[17] Tak S, Woo S, Yeo H. Data-Driven Imputation Method for Traffic Data in Sectional Units of Road Links[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2016, 17(6): 1 762-1 771

[18] Tonini F, Dillon W W, Money E S, et al. Spatio-Temporal Reconstruction of Missing Forest Microclimate Measurements[J]. *Agricultural and Forest Meteorology*, 2016, 218: 1-10

[19] Durán-Rosal A, Hervás-Martínez C, Tallón-Ballesteros A, et al. Massive Missing Data Reconstruction in Ocean Buoys with Evolutionary Product Unit Neural Networks[J]. *Ocean Engineering*, 2016, 117: 292-301

[20] Fan Zide, Li Jialin, Deng Min. An Adaptive Inverse Distance Weighting Spatial Interpolation Method with the Consideration of Multiple Factors[J]. *Geomatics and Information Science of Wuhan University*, 2016, 41(6): 842-847(樊子德,李佳霖,邓敏. 顾及多因素影响的自适应反距离加权插值方法[J]. 武汉大学学报·信息科学版, 2016, 41(6): 842-847)

[21] Yan Jinbiao, Duan Xiaoqi, Zheng Wenwu, et al. An Adaptive IDW Algorithm Involving Spatial Heterogeneity[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(1): 97-104(颜金彪,段晓旗,郑文武,等. 顾及空间异质性的自适应IDW插值算法[J]. 武汉大学学报·信息科学版, 2020, 45(1): 97-104)

[22] Pesquer L, Cortés A, Pons X. Parallel Ordinary Kriging Interpolation Incorporating Automatic Variogram Fitting[J]. *Computers & Geosciences*, 2011, 37(4): 464-473

[23] Xu C D, Wang J F, Hu M G, et al. Interpolation of Missing Temperature Data at Meteorological Stations Using P-BSHADE[J]. *Journal of Climate*, 2013, 26(19): 7 452-7 463

[24] Gardner E S. Exponential Smoothing: The State of the Art-Part Ⅱ[J]. *International Journal of Forecasting*, 2006, 22(4): 637-666

[25] Yozgatligil C, Aslan S, Iyigun C, et al. Comparison of Missing Value Imputation Methods in Time Series: The Case of Turkish Meteorological Data[J]. *Theoretical and Applied Climatology*, 2013, 112(1): 143-167

[26] Li L, Losser T, Yorke C, et al. Fast Inverse Distance Weighting-Based Spatiotemporal Interpolation: A Web-Based Application of Interpolating Daily Fine Particulate Matter PM2.5 in the Contiguous US Using Parallel Programming and K-D Tree[J]. *International Journal of Environmental Research and Public Health*, 2014, 11(9): 9 101-9 141

[27] Bhattacharjee S, Mitra P, Ghosh S K. Spatial Interpolation to Predict Missing Attributes in GIS Using Semantic Kriging[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, 52(8): 4 771-4 780

[28] Shang J, Zheng Y, Tong W, et al. Inferring Gas Consumption and Pollution Emission of Vehicles Throughout a City[C]. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2014

[29] Wang Y, Zheng Y, Xue Y. Travel Time Estimation of a Path Using Sparse Trajectories[C]. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2014

[30] Zheng Y, Liu F, Hsieh H. U-Air[C]. The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2013

[31] Yue T, Du Z, Song D, et al. A New Method of Surface Modeling and Its Application to DEM Construction[J]. *Geomorphology*,2007,91(1): 161-172

[32] Yue T, Wang S. Adjustment Computation of HASM: A High-Accuracy and High-Speed Method[J]. *International Journal of Geographical Information Science*, 2010, 24(11): 1 725-1 743

[33] Li Y, Shahabi C. A Brief Overview of Machine Learning Methods for Short-term Traffic Forecasting

and Future Directions[J]. *SIGSPATIAL Special*, 2018, 10(1): 3-9

[34] Deng M, Yang W, Liu Q, et al. Heterogeneous Space-Time Artificial Neural Networks for Space-Time Series Prediction[J]. *Transactions in GIS*, 2018, 22(1): 183-201

[35] Kamarianakis Y, Prastacos P. Space-Time Modeling of Traffic Flow[J]. *Computers and Geosciences*, 2005, 31(2): 119-133

[36] Cheng T, Wang J, Haworth J, et al. A Dynamic Spatial Weight Matrix and Localized Space-Time Autoregressive Integrated Moving Average for Network Modeling[J]. *Geographical Analysis*, 2014, 46(1): 75-97

[37] Du Z, Wu S, Zhang F, et al. Extending Geographically and Temporally Weighted Regression to Account for Both Spatiotemporal Heterogeneity and Seasonal Variations in Coastal Seas[J]. *Ecological Informatics*, 2018, 43: 185-199

[38] Huang B, Wu B, Barry M. Geographically and Temporally Weighted Regression for Modeling Spatiotemporal Variation in House Prices[J]. *International Journal of Geographical Information Science*, 2010, 24(3): 383-401

[39] Xia D, Wang B, Li H, et al. A Distributed Spatial-Temporal Weighted Model on MapReduce for Short-term Traffic Flow Forecasting[J]. *Neurocomputing*, 2016, 179: 246-263

[40] Cai P, Wang Y, Lu G, et al. A Spatiotemporal Correlative K-nearest Neighbor Model for Short-term Traffic Multistep Forecasting[J]. *Transportation Research Part C: Emerging Technologies*, 2016, 62: 21-34

[41] Jia T, Yan P. Predicting Citywide Road Traffic Flow Using Deep Spatiotemporal Neural Networks[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2020, DOI: 10.1109/TITS.2020.2979634

[42] Yu B, Yin H, Zhu Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting[C]. The 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018

[43] Zhang J, Zheng Y, Qi D, et al. Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks[J]. *Artificial Intelligence*, 2018, 259: 147-166

[44] Zhang Y, Cheng T, Ren Y, et al. A Novel Residual Graph Convolution Deep Learning Model for Short-term Network-Based Traffic Forecasting[J]. *International Journal of Geographical Information Science*, 2020, 34(5): 969-995

[45] Kong L, Xia M, Liu X Y, et al. Data Loss and Reconstruction in Wireless Sensor Networks[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2014, 25(11): 2 818-2 828

[46] Qin M, Du Z, Zhang F, et al. A Matrix Completion-Based Multi-view Learning Method for Imputing Missing Values in Buoy Monitoring Data[J]. *Information Sciences*, 2019, 487: 18-30

[47] Cheng S, Lu F. A Two-step Method for Missing Spatio-Temporal Data Reconstruction[J]. *ISPRS International Journal of Geo-Information*, 2017, 6(7): 187

[48] Cheng S, Peng P, Lu F. A Lightweight Ensemble Spatiotemporal Interpolation Model for Geospatial Data[J]. *International Journal of Geographical Information Science*, 2020, DOI: 10.1080/13658816.2020.1725016

[49] Cheng S, Lu F, Peng P. Short-term Traffic Forecasting by Mining the Non-stationarity of Spatiotemporal Patterns[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2020, DOI: 10.1109/TITS.2020.2991781

[50] Yue Y, Yeh A G O. Spatiotemporal Traffic-Flow Dependency and Short-term Traffic Forecasting[J]. *Environment and Planning B: Planning and Design*, 2008, 35(5): 762-771

[51] Lu F, Liu K, Duan Y, et al. Modeling the Heterogeneous Traffic Correlations in Urban Road Systems Using Traffic-Enhanced Community Detection Approach[J]. *Physica A: Statistical Mechanics and Its Applications*, 2018, 501: 227-237

[52] Cheng S, Lu F, Peng P, et al. Short-term Traffic Forecasting: An Adaptive ST-KNN Model That Considers Spatial Heterogeneity[J]. *Computers, Environment and Urban Systems*, 2018, 71: 186-198

[53] Cheng S, Lu F. Short-term Traffic Forecasting: A Dynamic ST-KNN Model Considering Spatial Heterogeneity and Temporal Non-stationarity[C]. The Workshops of the EDBT/ICDT 2018 Joint Conference, Vienna, Austria, 2018

[54] Song C, Shi X, Wang J. Spatiotemporally Varying Coefficients (STVC) Model: A Bayesian Local Regression to Detect Spatial and Temporal Nonstationarity in Variables Relationships[J]. *Annals of GIS*, 2020, 26(3): 277-291

[55] Cheng S, Lu F, Peng P, et al. Multi-Task and Multi-View Learning Based on Particle Swarm Opti-

mization for Short-term Traffic Forecasting [J]. *Knowledge-Based Systems*, 2019, 180: 116-132

[56] Cheng S, Lu F, Peng P, et al. A Spatiotemporal Multi-View-Based Learning Method for Short-term Traffic Forecasting [J]. *ISPRS International Journal of Geo-Information*, 2018, 7(6): 218

[57] He J, Lawrence R. A Graph-Based Framework for Multi-Task Multi-View Learning[C]. The 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011

[58] Wang Jinfeng, Xu Chengdong. Geodetector: Principle and Prospective[J]. *Acta Geographica Sinica*, 2017, 72(1): 116-134(王劲峰,徐成东. 地理探测器:原理与展望[J]. 地理学报, 2017, 72(1): 116-134)

[59] Ge Y, Jin Y, Stein A, et al. Principles and Methods of Scaling Geospatial Earth Science Data[J]. *Earth-Science Reviews*, 2019, 197: 102897

[60] Song Changqing, Cheng Changxiu, Yang Xiaofan, et al. Understanding Geographic Coupling and Achieving Geographic Integration [J]. *Acta Geographica Sinica*, 2020, 75(1): 3-13(宋长青,程昌秀,杨晓帆,等. 理解地理"耦合"实现地理"集成"[J]. 地理学报, 2020, 75(1): 3-13)

Review of Interpolation, Reconstruction and Prediction Methods for Heterogeneous and Sparsely Distributed Geospatial Data

CHENG Shifen[1,2]　*PENG Peng*[1,2]　*ZHANG Hengcai*[1,2]　*LU Feng*[1,2]

1　State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

2　University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Spatiotemporal data mining is the core research topic of geographic information science. In the era of big data, the explosive growth of geographic spatiotemporal data puts forward an urgent demand for spatiotemporal knowledge discovery, which promotes the continuous development of spatiotemporal data mining technology. However, the universal heterogeneity and sparse distribution characteristics of spatiotemporal big data restrict the realization of spatiotemporal data mining algorithms, and significantly affect the description and analysis capabilities of natural and social complex systems. Thus, this paper focuses on the series of bottlenecks faced in the expression and application of heterogeneous and sparsely distributed spatiotemporal data. We systematically summarized the research status and existing problems of several key spatiotemporal mining tasks including missing spatiotemporal data interpolation, sparse spatiotemporal data reconstruction, and spatiotemporal state prediction, condensed four key scientific problems, and gave four corresponding solutions. The proposed methods are expected to enrich the method system in the field of spatiotemporal data mining and improve the quality and application value of spatiotemporal data modeling.

Key words: spatiotemporal autocorrelation; spatiotemporal heterogeneity; spatiotemporal interpolation; spatiotemporal prediction; multi-task and multi-view learning

First author: CHENG Shifen, PhD, Postdoctoral fellow. His research interest is spatiotemporal data mining. E-mail: chengsf@lreis.ac.cn

Corresponding author: LU Feng, PhD, professor. E-mail: luf@lreis.ac.cn

Foundation support: The National Natural Science Foundation of China(41631177, 41771436);the China Postdoctoral Science Foundation (2019M660774, 2020T130644).

引文格式:CHENG Shifen, PENG Peng, ZHANG Hengcai, et al.Review of Interpolation, Reconstruction and Prediction Methods for Heterogeneous and Sparsely Distributed Geospatial Data[J].Geomatics and Information Science of Wuhan University, 2020, 45(12):1919-1929. DOI:10.13203/j.whugis20200488(程诗奋,彭澎,张恒才,等.异质稀疏分布时空数据插值、重构与预测方法探讨[J].武汉大学学报·信息科学版,2020,45(12):1919-1929.DOI:10.13203/j.whugis20200488)