

面向 GDP 数据时空多维属性的可视分析方法

周志光^{1,2,3)}, 徐杨炳¹⁾, 刘芳¹⁾, 陈伟锋¹⁾, 陶煜波²⁾, 林海²⁾, 苏为华^{3,4)*}

¹⁾ (浙江财经大学信息学院 杭州 310018)

²⁾ (浙江大学 CAD&CG 国家重点实验室 杭州 310058)

³⁾ (浙江财经大学应用经济学博士后流动站 杭州 310018)

⁴⁾ (浙江工商大学统计与数学学院 杭州 310018)

(swh@zufe.edu.cn)

摘 要: 国内生产总值是经济统计的核心指标, 文中提出面向国内生产总值数据时空多维属性的可视分析方法. 利用多维标度法对不同时空维度多维属性进行降维, 支持用户交互选择时空维度和参与相似度计算的多维属性; 利用环状饼图展示产业结构的比例关系, 设计垂直指针定位优势产业并动态表达产业结构的变化趋势; 利用圆状堆叠图呈现各个产业产值及占比的时序变化信息, 堆叠图的绘制顺序进一步提示产业结构的更替特征. 设计内圆、中环、外环的可视化模式, 实现面向国内生产总值数据时空多维属性的可视分析系统, 利用方便而快捷的交互手段关联上述可视化方法, 为用户提供快速探索经济统计数据中产业结构时空特征模式的工具. 最后通过大量的可视化效果及用户体验结果, 进一步验证了文中方法的有效性和实用性.

关键词: 国内生产总值; 可视分析; 产业结构; 多维标度; 时空多维属性

中图法分类号: TP391.41

Visual Analysis of GDP with Spatiotemporal Multidimensional Features

Zhou Zhiguang^{1,2,3)}, Xu Yangbing¹⁾, Liu Fang¹⁾, Chen Weifeng¹⁾, Tao Yubo²⁾, Lin Hai²⁾, and Su Weihua^{3,4)*}

¹⁾ (School of Information, Zhejiang University of Finance and Economics, Hangzhou 310018)

²⁾ (State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310058)

³⁾ (Postdoctoral Research Station of Applied Economics, Zhejiang University of Finance and Economics, Hangzhou 310018)

⁴⁾ (School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018)

Abstract: The gross domestic product (GDP), a significant part of economic statistics, is collected to analyze the development of national economic. In this paper, we propose a visual analytics method to visualize GDP with time-space multidimensional features. MDS is employed to reduce the dimensionality of original multi-dimensional data, and the users are allowed to select different perspectives and dimensions to calculate the similarity matrix. We use the circular pie chart to present the relationship of various industries, especially with a vertical pointer designed to locate the major industry. The angle between the vertical pointer and the center line further indicates the development trend of industrial structures. We also take advantage of a centric stream map to reveal the proportion of different industries, and shows the change of the major industry using the drawing order. The user interface of the proposed system is composed of three parts including the inner, middle and outer areas. Sev-

收稿日期: 2015-08-27; 修回日期: 2015-12-07. 基金项目: 国家自然科学基金(61303133, 61472354); 全国统计科学研究重大项目(2015LD03); 浙江省科技厅公益技术应用研究计划项目(2014C31057); 浙江省自然科学基金(LQ14F020007); 教育部博士后科学基金(2015M571846); 浙江大学 CAD&CG 国家重点实验室开放课题(A1417). 周志光(1983—), 男, 博士, 副教授, 主要研究方向为数据可视化与可视分析; 徐杨炳(1993—), 男, 在校学生; 刘芳(1976—), 女, 博士, 讲师, 主要研究方向为高维数据可视化; 陈伟锋(1983—), 男, 博士, 讲师, 主要研究方向为数据可视化; 陶煜波(1980—), 男, 博士, 副教授, 主要研究方向为数据可视化; 林海(1965—), 男, 博士, 教授, 博士生导师, 主要研究方向为图形学与可视化; 苏为华(1963—), 男, 博士, 教授, 博士生导师, 论文通讯作者, 主要研究方向为统计经济学、应用经济学.

eral intuitive interactions are also provided to help users quickly explore the time-space features of the industrial structure in the economic statistic data. A large number of experimental results as well as credible user studies further demonstrate the effectiveness and practicability of the proposed system.

Key words: gross domestic product (GDP); visual analytics; industrial structure; multidimensional scaling; spatiotemporal multidimensional features

经济和社会发展信息数据的统计与分析是国家科学决策和现代化管理的重要基础性工作,是认识国情、研究问题、制定政策的重要依据。随着中国经济与社会的高速发展及信息采集过程的规范化,经济统计数据呈现显著的时空多维特征。统计属性范畴包括国民经济核算、固定资产投资、财政金融、人口与自然资源、能源与环境、教育科技、文化体育、医药卫生等各个领域和行业的数据。而且,经济统计数据的时空维度亦不断细化,大部分经济指标数据的时间维度精确至月度,空间维度则精确至县及县以上行政区,数据的规模不断增加,结构层次关系显著。面对如此规模不断增加、结构关系日趋复杂的经济统计数据,简易而有效的数据可视化及数据分析挖掘方法是帮助民众直观地了解国情,辅助学者对经济形势进行分析与研究,协助相关职能部门制定政策的有效途径。

可视化是利用人类视觉认知高通量的特点,以图形的形式呈现隐含于数据内部的模式及特征信息,是帮助用户理解复杂现象,诠释复杂数据的重要手段,其广泛应用于科学计算、数据分析、知识发现等研究领域。在统计数据分析中,可视化技术亦广泛应用,常见的统计分析软件均提供了数据可视化功能包,如经典的 Excel 软件、专业的统计产品与服务解决方案 SPSS 软件、统计分析系统 SAS、计量经济学软件包 Eviews。R 语言的诞生更是深度融合数据处理、计算分析及制图等技术,进一步强化了可视化在数据统计分析与知识挖掘中的作用。强大的软件功能包为统计数据分析过程中特定问题的解决提供了有效的支持,却也限制了数据分析及探索过程的自由度。

近年来,作为可视化研究领域的一个分支,可视分析迅速发展为一个交叉学科方向。相比于经典的可视化技术,可视分析集成了数据挖掘、模型分析等理论和方法,更关注于用户在可视化获取知识过程中所发挥的作用,即通过交互可视界面对

复杂的数据信息进行探索和分析^[1]。针对不同的数据来源及数据分析目标,国内外学者设计3种有效的可视分析工具,如基于交通数据^[2-3]、社交媒体数据^[4-5]、电子商务交易数据^[6]、城市公共服务数据^[7]等各个行业数据的可视分析工具,充分展现了可视分析技术的有效性和实用性。面向经济统计数据开展可视分析方法的设计与研究,对于观察和分析经济形势,理解和制定经济政策等方面具有重要的理论意义和实用价值。

国内生产总值(gross domestic product, GDP)是经济和社会发展信息数据的重要组成部分,是国民经济核算的核心,即衡量一个国家或地区经济状况和发展水平的重要指标。按照生产面的计算方式核算,GDP是由3个产业数据构成。第一产业包括农业、渔业等提取天然资源的经济活动;第二产业包括纺织业、重工业等将天然资源转换成制成品或半制成品的经济活动;第三产业则包括商业、金融保险、零售业等提供服务的经济活动。其中,产值信息能够反映各个地区的产业规模,占比数据则体现了产业结构特征。利用经典的柱状图、饼图、折线图可以有效地展示单一属性数据,却不能综合考虑 GDP 数据的时空多维属性,难以提供信息量丰富的可视化及分析结果,指导用户识别当前地区产业结构特征以及与其他地区经济发展模式的异同。而且,统计分析软件提供的可视化功能包依赖于用户手动提取的小样本数据,增加数据分析过程复杂度。频繁的视图切换操作,妨碍了经济特征的准确分析与快速识别。

因此,本文面向 GDP 数据设计综合考虑时空多维属性的可视分析工具,可以有效地集成用户的先验知识及特定需求,支持用户多角度、全方位地探索数据中隐含的经济发展及产业结构特征。其核心设计主要由内圆、中环和外环3部分组成。其中,内圆区域是利用多维标度法(multidimensional scaling, MDS)对 GDP 数据各个产业产值、占比等多维属性

降维的结果,直观地展示不同时空维度产业结构发展的关联特性;中环部分描述当前时空维度各个产业的占比,并且设计垂直指针表达优势产业及其与余下产业之间的联系;外环区域则描述由初始时间至当前时间各个产业结构的产值、占比变化及优势产业更替等情况。此外,引入色调映射技术对 GDP 总量进行颜色映射,有利于不同时空维度数据差异感知;并且实现简洁而直观的交互设计,方便不同时空维度的属性对应及效果对比。相比于传统的统计数据可视化功能,本文设计的面向 GDP 数据时空多维属性的可视分析工具能够帮助用户快速发现经济统计数据中产业结构的时空特征模式,大量的可视化效果及用户体验结果可进一步验证本文工具的有效性和实用性。

1 相关工作

随着经济与社会的高速发展及信息采集过程的规范化,经济统计数据质量逐渐优化,基于经济统计数据的知识挖掘与规律分析已然成为经济统计领域的研究热点。付凌晖^[8]采用 Engle-Granger 两步检验法、格兰杰因果关系检验方法对产业结构高级化与经济增长的相互作用关系进行分析,证实了经济增长与产业结构升级的高度相关性。韩颖等^[9]结合投入产出理论与对数平均迪氏指数法,利用 LMDI 法对经济统计数据分解,进而建立各个因素对产业结构变动的作用模型,能够快速地发掘各个因素对产业结构转型的抑制和促进作用。刘刊等^[10]则运用系统工程原理分析各种结构的相互联系和作用关系,利用 Cobb-Douglas 生产函数构建经济增长的产业结构的依赖模型,揭示了中国经济增长过程对各产业结构的依赖性,进一步表明产业结构、GDP 结构和 GDP 项目结构是我国经济系统中的主要结构因素。彭冲等^[11]采用面板向量自回归模型,从动态的视角探究产业结构变迁与经济波动的关系和影响,发现产业结构的合理化对经济波动具有“熨平效应”,产业结构合理化冲击、高级化冲击与经济波动的关系均表现为对称影响。薛继亮^[12]利用 Shift-Share 模型对技术选择和产业结构进行变量分析,探索技术选择与产业结构之间的联系。在利用经典的统计分析软件对经济统计数据进行分析的过程中,通常需要用户反复从初始数据中提取小型样本,手动输入命令或编写脚本语言对数据进行可视化及模式预测,

难以根据用户的需求交互地实现内在特征模式的分析与探索,存在一定的局限性。

可视分析是近年来快速发展的交叉学科方向,主要是借助于可视化的展现手段引导用户交互式地探索数据中隐含的特征信息,其流程大致包括数据处理、视觉映射、可视化设计及数据建模等步骤,如图 1^[1]所示。

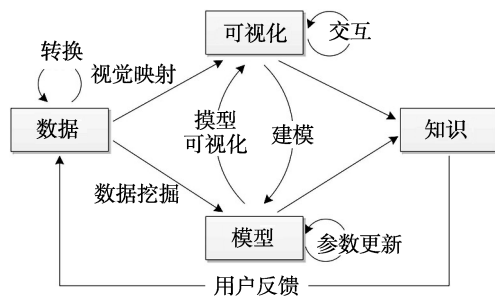


图 1 可视分析流程

针对各个行业不同应用需求产生的数据,国内外研究学者创作了大量的实用可视分析工具。Qu 等^[13]面向香港城市空气污染数据设计可视分析系统,利用极坐标及改进的平行坐标技术描述影响空气质量的因素,结合图可视化表达不同污染物之间的关联特征。Wu 等^[14]面向酒店顾客的评价信息,设计时间多尺度信息可视分析工具,能够有效地分析不同用户的特征模式。Xie 等^[6]提出淘宝交易数据可视分析系统,借助于音乐符号可视手段,帮助用户直观地探索网络交易过程中的欺诈行为。Cao 等^[15]以向日葵传播种子为隐喻,抽象化信息传播模式,能够帮助用户快速感知信息传播的过程及模式。

可视化与模型的循环验证是可视分析流程的重要环节,参数的调节及中间结果的变化可以直观地指导用户优化模型^[16]。Choo 等^[17]设计以一种基于可视分析技术的高维数据分析工具 iVisClassifier,利用线性判断分析 (linear discriminant analysis, LDA) 对高维数据进行降维,进而利用平行坐标、散点图等可视技术对降维分类结果进行展示,可由用户根据先验知识交互式地标记分类结果数据,进而引导更加准确的高维数据分类过程;Bogl 等^[18]利用可视分析方法实现时间序列模型 SARIMA 参数选择,由用户交互调整参数,系统自动记录参数,相应的图表以动画的方式渐变,辅助用户直观地比对模型,深入地理解复杂模型参数的含义,进而实现参数选择和设置。

实际数据通常具有丰富的时空属性,时空

可视分析技术是当前可视分析领域的研究热点。Ferreira 等^[19]设计了大规模城市时空数据的可视化探索工具,并以纽约出租车载客记录数据为例,支持用户从不同时空角度观察和分析数据,进而发现数据中隐含的特征模式;Wang 等^[2]则是以南京市交通数据为例,利用可视分析技术发现不同时空维度交通状况的内在关系模式;Zhang 等^[7]以城市公共服务数据为例,针对复杂的时空事件进行可视化与分析,为城市公共服务管理部门决策提供依据;Doraiswamy 等^[20]利用拓扑结构描述事件,定义标量函数自适应地发现事件并引导用户探索相关性较强的时空维度事件,进而发现与识别事件的周期、频率等模式;Turkay 等^[21]从交互可视的角度出发,对拥有多维属性的英国人口普查数据进行可视分析,进而发现多维属性的相关性及区域发展的特征模式等;Li 等^[22]面向气候变化的时空多维属性设计了 Vismate 可视化工具,可以辅助用户快速发现初始气象数据时空维度的异常特征。

GDP 统计数据不仅具有显著的时空特征,而且各个产业结构属性具有组合结构关系,产值及占比信息不同程度地反映了经济发展及产业结构特征。

2 总体流程

本文设计面向 GDP 数据时空多维特征的可视

分析系统,帮助用户快速分析不同时空维度产业结构与经济发展的关联特征模式。为清晰地表述本文可视分析工具的设计与实现过程,图2所示为该系统的总体流程图。

首先,导入初始 GDP 数据,包括不同年份、不同地区各个产业产值等相关信息;进而借助地图可视化,利用颜色表达 GDP 总量,交互拖动时间轴,帮助用户快速地发现感兴趣的空间地区;然后交互点击地图确认地区,出现面向 GDP 数据时空多维特征的可视分析窗口。其中内环区域是利用 MDS 对 GDP 数据多维属性降维可视化的结果,以帮助用户快速发现不同空间或不同时间各个节点的特征变化模式;中环区域是利用环状饼图展示当前节点各个产业占比,并在此基础上设计垂直指针定位优势产业,根据不同产业之间的占比差异,自适应计算偏移角度,提示产业结构之间的关联特征;外环区域则展示截止当前时间各个产业产值、占比的发展情况,且不同产业绘制的先后次序直观地提示了用户产业结构的更替时间点。用户可以交互拖动时间轴,观察当前地区历年来各个产业的产值、占比的发展情况,亦可以快速地发现与其他地区的差异等信息。借助于面向 GDP 数据时空多维特征可视分析系统提供的交互分析与对比功能,用户可以方便、直观地探索 GDP 数据多维属性的相关性及区域发展的特征模式。

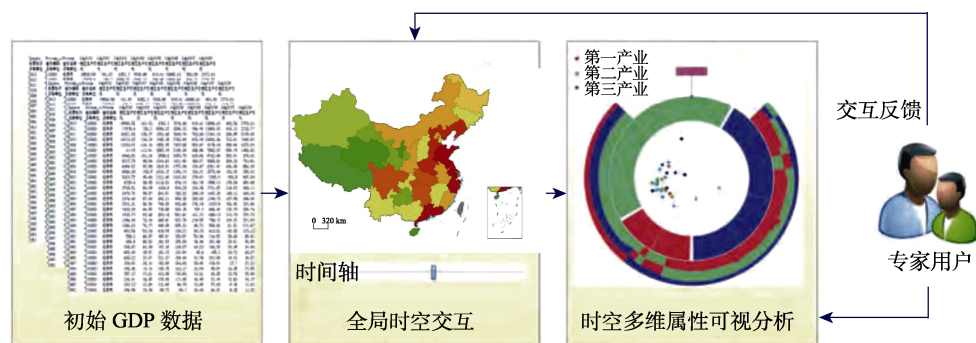


图2 本文系统总体流程图

3 GDP 数据时空多维属性的可视分析

GDP 数据是国民经济核算的重要指标,不仅具有显著的时空维度,而且其组成元素具有特定的结构关系。在利用传统的统计分析软件对时空多维的 GDP 数据进行分析的过程中,常常需要用户手动提取的小样本数据,频繁切换可视化窗口,使其不能综合考虑 GDP 数据的时空多维属性,并

提供信息量丰富的可视化效果,增加了数据分析过程的复杂度。因此,本文结合经济统计专家的应用需求,设计面向 GDP 数据时空多维属性的可视分析系统,旨在方便用户快速发现 GDP 数据中产业结构的时空特征模式。

3.1 基于 MDS 的时空多维属性降维

GDP 数据中任一节点具有多维属性,包括时间、地点、3 个产业产值以及占比等信息,如何直观

而快捷地表达与展示 GDP 数据中的时空多维属性是可视分析与设计的关键。聚类分析方法可以根据 GDP 数据的多维属性计算相似度,进而对不同时空维度节点进行聚类,以发现奇异节点及类别特征。然而,在聚类分析的过程中,由产业产值、占比等信息构建的初始多维属性被强制纳入一维谱系分类中,容易导致初始多维属性对象的关系弱化,具有一定的局限性。因此,本文利用 MDS 算法^[23]在近似的意义下,将初始多维属性对象由多维空间简化到低维空间,以寻求最佳的空间维度及空间位置,最大程度地保留初始多维属性对象之间的联系。主要步骤如下:

Step1. MDS 对象选取。为有效发现不同时空节点的相似程度,本文提供 2 类降维对象选取方式,分别是空间多维属性降维,即同一时间节点不同地区的多维属性分析;以及时间多维属性降维,即同一地区不同时间的多维属性分析。综合考虑多维属性信息,能够有效地度量各个地区产业结构的差异特点。空间多维属性降维可以帮助用户快速发现同一时间各个地区的产业结构模式差异,而时间多维属性降维可以帮助用户有效探索当前地区产业结构的时序发展特征。本文提供方便的交互接口,支持用户根据特定应用需求探索感兴趣特征模式。

Step2. 计算相似性矩阵。在降维对象确定的基础上,由用户交互指定参与相似性计算的多维属性,包括各个产业产值及占比等信息,进而定义距离函数,构建降维对象的相似性矩阵

$$\Delta = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,K} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,K} \\ \vdots & \vdots & & \vdots \\ \delta_{K,1} & \delta_{K,2} & \cdots & \delta_{K,K} \end{pmatrix} \quad (1)$$

其中, K 表示节点的数目,即参与降维的年份或地区的数目; $\delta_{i,j}$ 表示第 i 个节点和第 j 个节点的相似性,计算方式为

$$\delta_{i,j} = \sum_{t=1}^M |m_{i,t} - m_{j,t}| \quad (2)$$

其中, M 表示参与降维的多维属性的数目,由用户交互选择确定; $m_{i,t}$ 示第 i 个节点归一化的 t 属性值; $|m_{i,t} - m_{j,t}|$ 表示第 i 个节点和第 j 个节点对应 t 属性值的差异绝对值,求和即可获得节点在 M 维空间中的距离,对降维对象中任意 2 个节点均按照式(2)进行计算,即可获得相似性矩阵。

Step3. 多属性降维。利用 MDS 对多维属性降维的目标是寻找 K 个 N 维空间向量 $x_1, x_2, x_3, \dots, x_K \in \mathbb{R}^N$,使得初始 M 维空间的节点距离在 N 维空间得以有效保

持,即如

$$\|x_i - x_j\| \approx \delta_{i,j} \quad (3)$$

其中, $\|x_i - x_j\|$ 表示向量 x_i 和 x_j 的范数,通常定义为欧氏距离。因此,利用 MDS 降维则转换为一个优化问题,定义最小化目标函数为

$$f(x, \delta) = \min_{x_1, \dots, x_K} \sum_{i < j} (\|x_i - x_j\| - \delta_{i,j})^2 \quad (4)$$

进一步利用特征值分解方法求解最小化目标函数,获得最优向量 $x_1, x_2, x_3, \dots, x_K$ 。

Step4. 结果可视化。为突出各个节点的对比,直观地展示各个节点的差异,本文设置降维空间的维度 $N=2$,即在二维空间中获得各个节点的多维投影结果,极大程度地保持原始多维空间的距离程度。将投影结果展示到二维平面中,用户可以快速地发现不同时空节点的差异,进而发现产业结构发展的时空特征模式。在二维空间中,不同投影点的距离越近,表明其在综合考虑多维属性的前提下产值或结构表现为较强的相似性;反之则表现为较强的差异性。为了从视觉上区分各个节点的来源与异同,将当前时空节点的形状映射为紫色正方形标志,在空间降维分析过程中引入颜色映射 GDP 值;在时间降维分析过程中引入颜色和大小描述时间的先后顺序,以便快速识别产业结构时序发展模式。

图 3 所示为北京市时空多维属性的 MDS 降维结果。图 3a 所示为借助各个产业产值差异度量相似性。利用 MDS 获得的 1952—2013 年北京市历年来的时间节点的投影结果。其中,颜色由黄色到蓝色渐进性变化,圆点半径由小到大,均用以描述时间的变化;紫色正方形节点即当前时空节点,描述的是北京市 2013 年的多维属性投影节点。可以看出,北京市各个产业产值的变成表现出明显的线性结构,且具有显著的阶段式发展特征。图 3b 所示为借助各个产业占比差异度量相似性获得的 2013 年全国各个省份地区的投影结果,颜色信息映射为人均 GDP,粉色正方形表示的北京市的投影点距离其他投影点较远,表明北京市的产业结构具有显著的特殊性,这与北京市第一产业占比较小、第三

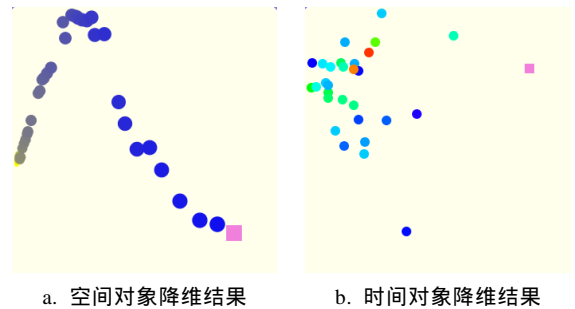


图 3 基于 MDS 的时空多维属性降维分析

产业占比较大的真实情况相契合,进一步验证了面向GDP数据时空多维属性降维算法的有效性。

3.2 产业数据占比可视化

产业数据占比的描述是GDP数据可视化的重要环节。经典的饼图是比例数据可视化的常用方法,其在较大程度上依赖于人眼视觉感知角度的大小,进而分析与判断各个属性的占比。然而,在时序GDP数据的分析与可视化的过程中,当各个产业占比差别较小时,人眼视觉难以感知产业结构占比差异,而当饼图随着时间变化不断切换时,各个产业结构占比的动态变化特征难以捕捉,存在一定的局限性。因此,本文设计垂直指针定位优势产业,并且动态地计算各个产业之间的差异,借助于垂直指针与优势产业之间的角度偏移,揭示其他产业及优势产业之间的关联特征。

在GDP数据的时空交互可视化过程中,垂直指针保持位置和角度不变,始终指向当前时空节点对应的优势产业。用户不需要通过视觉感知角度的大小关系来确定优势产业,而是直接将垂直指针指向的产业锁定为优势产业,避免了视觉感知带来的不确定性。在饼图的绘制过程中,增加各个属性的中心线,将饼图各个属性对应的角度分为均等的2部分。利用优势产业的中心线与垂直指针的角度差异描述各个产业之间的相关性,不仅能够静态数据中直观地分析各个产业占比的大小关系,而且在动态产业结构占比可视化时,中心线与垂直指针的角度连续变化更是增强了用户对时序GDP产业结构占比特征的分析与理解。具体的偏移角度度量,如

$$\alpha = \frac{p_l - p_r}{p_m} \quad (5)$$

其中, p_m 表示优势产业的占比, p_l 表示优势产业左侧相邻的产业占比; p_r 则表示优势产业右侧相邻的产业占比; α 为相对偏移角度,与优势产业占比成反比,其余产业占比的差异成正比。实际可视化过程中,垂直指针与优势产业中心线的偏移角度 $\Delta\alpha$ 的计算过程为

$$\Delta\alpha = \alpha \frac{\theta p_m}{2} = \frac{p_l - p_r}{2} \theta \quad (6)$$

其中, θ 表示显示比例数据的所有可用角度; $\theta p_m / 2$ 表示偏移的极限角度,即优势产业所占角度的一半。

借助于垂直指针能够明确地定位优势产业,用户随着指针的方向能够快速确定当前时空节点

的优势产业。优势产业中心线与垂直指针的偏移角度进一步揭示了当前时空节点产业结构的发展关系和模式。随着时间的变化,饼图各个属性的比例不断变化,优势产业不断更替,各个产业之间的关系亦发生微妙的变化。

图4所示为时空节点的产业结构占比的可视化结果。图4a所示为浙江省2002年产业结构占比可视化结果,正上方垂直的指针定位于绿色区域,表明浙江省2002年的优势产业为第二产业。垂直指针由绿色区域的中心线向蓝色区域偏移角度较大,表明第三产业占比较为接近优势产业占比,而且远超第一产业占比;图4b所示为辽宁省1960年产业结构占比可视化结果,虽然第一产业和第三产业有显著差异,但是远远小于优势产业占比,因此垂直指针相比于优势产业中心线的角度偏移较小。因此,垂直指针的引入能够帮助用户快速而准确地发现产业结构的特征模式,尤其是在面对时序变化的优势产业更替的特征模式分析过程中,具有较强的实用性。

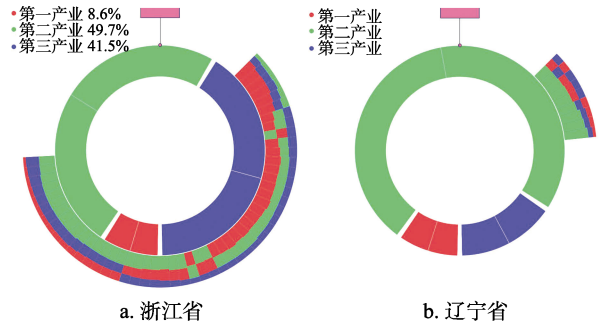


图4 不同省份产业结构占比可视化结果

3.3 产业数据堆叠可视化

GDP数据是一组时变数据,每个节点由第一、二、三产业产值构成,堆叠柱状图是展示GDP数据的有效手段。然而,GDP数据又是一组空间数据,每个节点都拥有明确的空间属性,如何有效地结合堆叠柱状图与数据本身的空间属性,是实现面向GDP数据时空多维属性可视分析的关键。在文献[20-21]的启发下,本文利用圆状堆叠图的方式对时序属性进行可视化,位于可视分析界面的外环区域,每年的属性对应均等的角度,半径的大小表示各个产业的产值或者占比大小。随着时间的变化,类似于花瓣的效果能够有效地展示各个年份的产业结构及其变化。为了更加清晰地提示优势产业的更替信息,按照产值、占比的大小关系依次对各个属性进行绘制,由于各个产业的颜色

定义不同, 所以当优势产业发生更替时, 颜色的变化能够有效地帮助用户快速感知有趣的经济现象.

增加堆叠柱状图与时空属性之间的交互功能. 结合第 4.1 节中 MDS 投影结果, 当鼠标移至各个堆叠区域时, 由堆叠区域向 MDS 投影点发出一条 Bezier 曲线段, 方便用户快速感知各个时间节点产业结构的发展差异; 对于同一时间不同地区节点的降维投影点, 用户交互点击地图窗口中的感兴趣地区, 则新建对比空间可视化窗口, 通过比较 2 个窗口的堆叠柱状图的半径及形成过程, 能够深入探索地区产业结构发展差异特征.

图 5 所示为 1952 年以来广东省和云南省的各个产业产值的堆叠可视化结果对比. 可以看出, 建国初期 2 个省份均以第一产业为优势产业, 改革开放以后, 广东省迅速转换为以第二产业为主导的发展模式, 近年来第三产业才占据上峰; 而云南省以第一产业为主导产业经历的时间比较长, 且产值相比于广东省有显著的差距, 近年来第三产业的产值也呈现显著增长的趋势. 而且, 圆状堆叠图的半径大小描述了产值信息, 广东省的产值较之云南省的产值有较大增量.

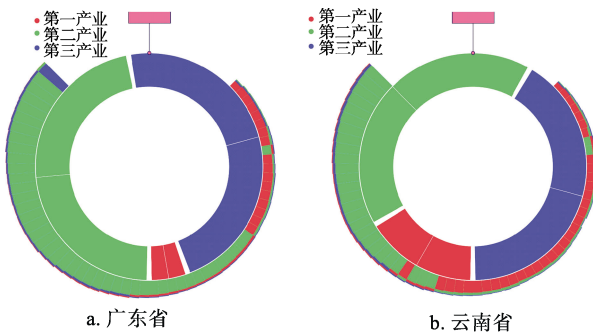


图 5 不同省份产业数据堆叠可视化结果

4 系统设计与交互

本文设计一套面向 GDP 时空多维特征可视分析的原型系统, 提供了快捷而方便的交互功能. 载入初始的 GDP 数据后, 主窗口所示为系统全局的时空交互接口, 即地图可视化窗口和时间轴. 地图各个地区颜色映射为 GDP 属性, 用户交互式地拖动时间轴, 地图各个地区颜色发生动态变化.

由于 GDP 数据呈现高动态范围且分布不均等特征, 传统的颜色直接映射方式导致大量数据映射为少量颜色, 以致用户在大部分时空节点上难以发现各个省份的 GDP 总量的差异, 存在一定的局限性. 为有效地突显各个时空区域的差异, 本文

引入色调映射技术对 GDP 数据的细节特征进行增强展示, 如

$$C_{\text{new}} = (D_{\text{max}} - D_{\text{min}}) \times \frac{\log(C + \gamma) - \log(C_{\text{min}} + \gamma)}{\log(C_{\text{max}} + \gamma) - \log(C_{\text{min}} + \gamma)} + D_{\text{min}} \quad (7)$$

其中, C_{max} 和 C_{min} 分别表示 GDP 属性的最大值和最小值; $\gamma = \eta(C_{\text{max}} - C_{\text{min}})$, 用户可以调节 η 以控制颜色映射的强度; D_{min} 和 D_{max} 是颜色映射的目标色调范围, 在本文中分别定义为蓝色调和红色调; C_{new} 则是结果色调值. 利用色调映射技术, 可以有效地增强高动态范围数据的细节特征展示.

图 6 所示为 1987 年各个省份 GDP 总量的颜色映射结果, 可以看出, 直接量化的方法数据分布不均, 局部细节特征隐藏; 色调映射可以有效地在直方图空间分配颜色. 从图 6a 可以看出, 直接量化的方式使得大部分数据集中于蓝色区域, 大部分年份对应的各个省份颜色差异不明显; 图 6b 则是采用色调映射技术对应的颜色映射结果, 数据相对均衡地分配到颜色区间中, 各个省份的颜色差异有效突显.

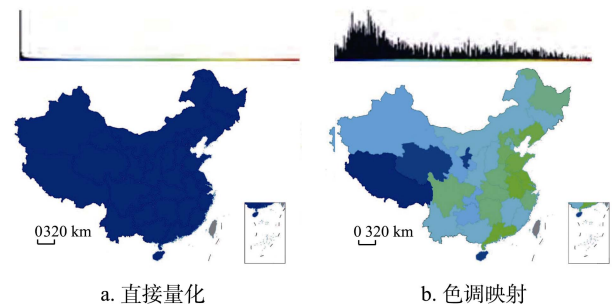


图 6 不同方式对应的地图颜色映射结果

由用户在地图窗口选中感兴趣的时空节点, 系统显示多维特征可视分析窗口, 其由内环的 MDS 投影、中环的饼图、外环的堆叠柱状图 3 部分构成. 用户可以交互选择以确定降维分析对象, 即同一时间不同地区或同一地区不同时间各个节点的差异, 同时交互选择不同的多维属性进行相似性矩阵计算, 包括各个产业产值、占比等属性. 内环区域将有效地展示指定分析对象的 MDS 降维结果, 用户可以移动鼠标交互式地查阅投影点的时空及多维属性信息. 进一步拖动时间轴, 内环区域参与 MDS 的对象发生变化, 重新投影; 中环区域饼图比例或优势产业发生变化, 垂直指针与优势产业之间的夹角随之变化; 外环堆叠柱状图的

随着时间的推移,数据展示逐年增加,不同年份产业产值及占比信息直观显示,优势产业更替一目了然.用户鼠标左键点击地图可以更换地区,亦可以右键点击增加对比地区可视分析界面,不同地区时空多维属性可视分析窗口的对比更有利于不同时空节点产业结构特征模式的分析与发现.

5 实验及结果分析

为了验证面向GDP数据时空多维属性可视分析系统的有效性和实用性,本文引入2组经济统计数据进行分析,分别是1952—2013年中国32个省份的GDP数据、1970—2013年世界212个国家的GDP数据,来源于国泰安数据服务中心.其中,不同时空节点均是由GDP总量及第一、二、三产业产值分量构成.

5.1 可视化结果

图7所示为中国典型的3个地区的产业结构可视分析结果.图7a所示为上海市产业结构图,结果表明,上海市历年来第一产业所占比重非常低,主要是以第二、三产业为主导;建国以来一直以第二产业为优势产业,直至1999年正式更替为以第

三产业为优势产业,发展至今;同时可以看出上海市历年来产值的相似性分析结果,明显提供阶段性视觉聚类区域,为经济学者分析数据与发展形势提供参考和依据.图7b所示为辽宁省产业结构图,当前投影点隐藏于历年来投影点群之中,表明产业占比特征没有显著变化;建国以来辽宁省自始至终以第二产业为主导产业,从未发生优势产业更替现象,第三产业占比呈现逐年递增趋势,将来有望替代第二产业成为优势产业.图7c所示为西藏自治区的产业结构图,明显发现数据缺失情况,即1978年以前西藏的GDP数据没有记录;而且MDS结果显示,西藏地区的产业结构发展趋势相对稳定,占比差异无明显变化,第一产业占比逐渐下滑,近年来第三产业已经成为优势产业,与该地区近年来大力发展旅游业的事实契合,进一步验证了系统的有效性.

由于世界各国在统计产业结构数据过程中,只存在第一、二产业以及其他行业数据,并没有明确强调第三产业数据.因此,本文利用GDP总量减去第一、二产业产值获得第三产业产值,进而读入世界各国历年来GDP数据,图8所示为3个典型国家的产业结构可视分析结果.图8a所示为

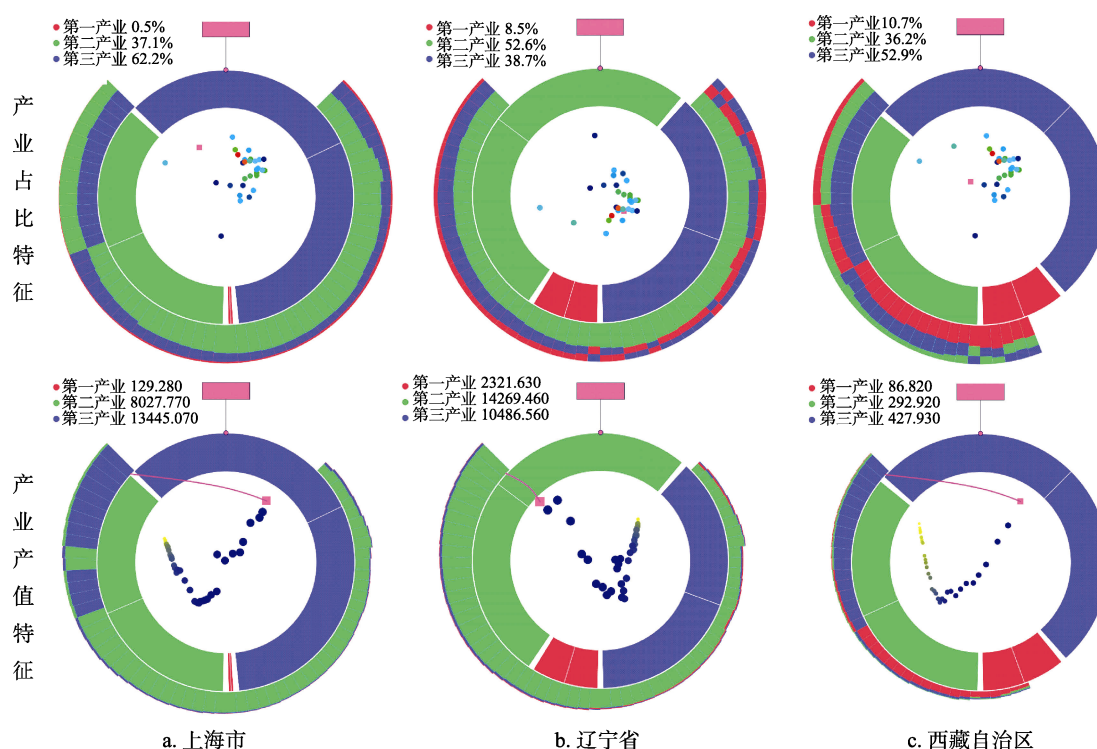


图7 中国部分地区产业结构可视分析结果

近年来中国经济发展情况, 内环区域 MDS 投影结果出现明显的聚类, 提示不同历史时期产业结构及经济发展的异同; 改革开放以后, 中国产业结构转型, 以第三产业为主导发展至今, 饼图及垂直指针进一步提示了第二产业在经济发展中占有重要比重. 图 8b 所示为印度的产业结构发展情况, 可以看出, 印度第三产业相对中国起步较早, 常年以第三产业为优势产业, 主要是源自其 IT 行业以及矿物出口和稻米等贸易行业的高速发展. 图 8c 所示为美国历年来的经济发展及产业结构信息由社会服务业、金融保险业和房地产业来推动, 历年来的第三产业高度发达, 对第一产业的依赖程度较低; 而且产业结构占比特征近年来呈现显著的聚类情

况, 提示重大历史事件对美国经济的作用与影响.

5.2 用户体验

为了进一步验证本文系统的有效性与实用性, 我们邀请了 12 名具有经济学背景的本科生和研究生参与用户体验. 在经过详尽的数据介绍和短暂的系统讲解、功能培训之后, 为用户提供统计分析软件及本文工具对中国 GDP 数据进行分析, 进而回答以下问题:

Q1. 湖北省 1976 年各个产业排名是怎样的?

Q2. 广东省优势产业结构发生更替的年份?

Q3. 天津市各个产业产值变化剧烈的年?

Q4. 上海市 1999 年产业结构特征与哪个地区相似?

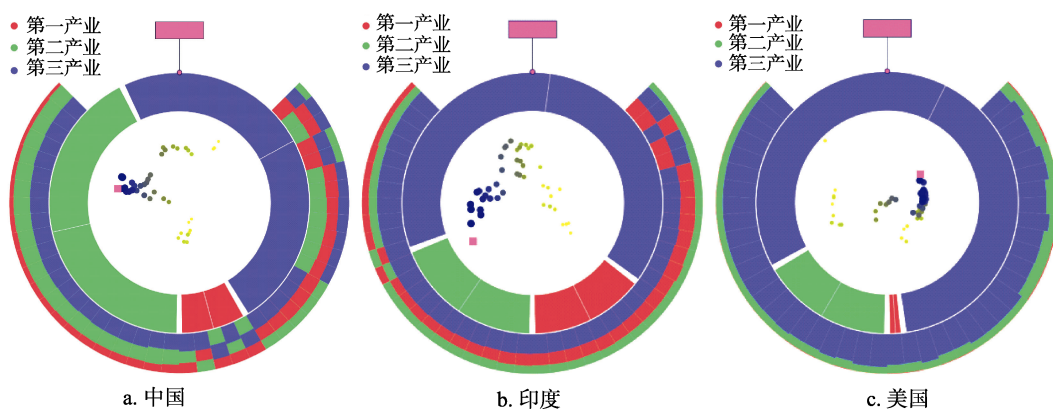


图 8 3 个典型国家产业结构可视分析结果

统计用户分析数据所用的平均时间以及回答各个问题的正确率, 结果如图 9 所示. 其中, 蓝色表示经典统计分析软件的分析结果, 红色表示本文设计的分析结果. 可以看出, 针对 Q1, 12 个用户采用 2 种分析方法均可以快速而准确地获得结果, 经过采访发现, 大部分用户并未利用统计分析软件提供的饼图去分析, 而是直接到原始数据中查询对比获得正确答案; 本文方法则是利用垂直指针及偏移角度有效获得正确答案. 然而, 查询原始数据的方法在余下问题的回答过程中却并不适用,

因此用户纷纷采用统计分析软件提供的饼图、柱状图、折线图等功能, 分析与对比不同年份、地区的 GDP 数据. 从图 9 可以看出, 利用统计分析软件提供的可视化功能在解决上述问题时存在消耗时间长、正确率低等问题. 在大致了解本文系统功能以后, 用户能够快速而准确地发现不同时空节点产业结构的特征模式, 相比于传统的数据分析软件, 用户不需要手工反复提取原始数据, 可以根据需求自由地探索数据, 在时间效率及正确率方面均有较大幅度提升.

5.3 不足之处

相比于经典的统计分析软件, 本文面向 GDP 时空多维特征可视分析工具具有丰富的交互功能, 直观的数据展示方式, 能够有效地辅助经济专家探索式地分析产业结构及特征模式, 具有一定的实用性. 但其尚存如下不足: 1) 本文借助于地图、时间轴交互, 辅以 MDS 时空多维属性投影实现 GDP 数据时空可视化, 但其时空属性融合程度不高, 将进一步参考地图与多维属性的综合可视化

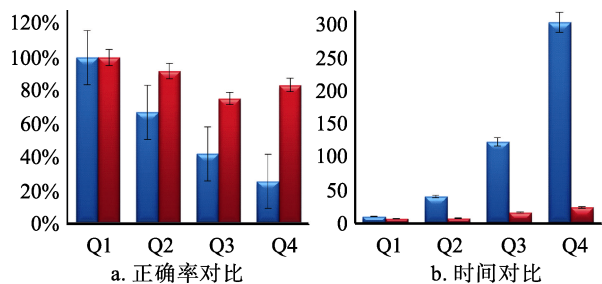


图 9 用户体验数据统计分析对比结果

界面^[24],全面地展示GDP数据的时空多维属性;2) GDP数据是经济统计数据中的重要部分,仍然存在大量的其他类型的经济统计数据,而且各个数据之间存在显著的时空层次特征,需要进一步拓展本文可视分析的功能,实现多维层次经济统计数据的可视分析;3)经济统计数据统计类型多样,在经济学领域被广泛研究,利用专业的经济统计数据进行分析软件进行数据探索常常涉及反复的数据调整和复杂的建模过程,效率低下且存在一定的不确定性。虽然本文系统能够有效地探索产业结构之间的时空关联特征,发掘经济发展模式,但是多维属性之间的关联特征及建模尚未涉及。因此,未来工作中将重点研制面向时空多维属性关系建模的可视分析方法,进一步实现经济统计数据中隐含特征模式的分析和挖掘。

6 总 结

GDP数据是国家经济统计的重要指标,本文面向GDP数据的时空多维属性设计可视分析工具,帮助用户探索式地发现不同时空节点产业结构的特征及发展模式。基于MDS的多维属性降维方法能够有效地展示不同时空节点的差异性,垂直指针的设计能够帮助用户快速感知当前时空节点产业结构比例以及优势产业及与其他产业的关联特征;圆状堆叠图能够有效地展示节点的产业结构发展模式,并直观地提示优势产业结构更替信息。由用户在全局地图绘制窗口中确定空间,交互拖动全局时间轴确定时间,各个可视化模块协同展示,为用户分析初始GDP数据提供多角度、全方位的可视接口。相比于传统的统计分析软件提供的可视化功能,本文面向GDP数据时空多维属性的可视分析工具能够有效地集成用户的先验知识及特定需求,快捷方便地探索与发现数据中隐含的模式及特征信息。大量的可视化效果及用户体验结果进一步验证了本文工具的有效性和实用性。

参考文献(References):

- [1] Keim D, Kohlhammer J, Ellis G, *et al.* Mastering the information age: solving problems with visual analytics[OL]. [2015-08-27]. <http://www.researchgate.net/>
- [2] Wang Z C, Ye T Z, Lu M, *et al.* Visual exploration of sparse traffic trajectory data[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1813-1822
- [3] Wang F, Chen W, Wu F R, *et al.* A visual reasoning approach for data-driven transport assessment on urban road[C] //Proceedings of IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2014: 103-112
- [4] Sun G D, Wu Y C, Liu S X, *et al.* EvoRiver: visual analysis of topic coepetition on social media[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1753-1762
- [5] Wu Y C, Liu S X, Yan K, *et al.* OpinionFlow: visual analysis of opinion diffusion on social media[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1763-1772
- [6] Xie C, Chen W, Huang X X, *et al.* VAET: a visual analytics approach for E-transactions time-series[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1743-1752
- [7] Zhang J W, Yanli E, Ma J, *et al.* Visual analysis of public utility service problems in a metropolis[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1843-1852
- [8] Fu Linghui. An empirical research on industry structure and economic growth[J]. Statistical Research, 2010, 27 (8): 79-81(in Chinese)
(付凌晖. 我国产业结构高级化与经济增长关系的实证研究[J]. 统计研究, 2010, 27(8): 79-81)
- [9] Han Ying, Ni Shuqian. Factor analysis of China's industrial structure adjustment[J]. Economic Theory and Business Management, 2011(12): 53-60(in Chinese)
(韩颖, 倪树茜. 我国产业结构调整的影响因素分析[J]. 经济理论与经济管理, 2011(12): 53-60)
- [10] Liu Kan, Fan Decheng, Wang Hongyu. China's economic growth and structural factors[J]. Economic Theory and Business Management, 2012, 32(8): 39-46(in Chinese)
(刘刊, 范德成, 王宏宇. 我国经济系统的结构因素与经济增长的结构依赖[J]. 经济理论与经济管理, 2012, 32(8): 39-46)
- [11] Peng Chong, Li Chunfeng, Li Yushuang. Study on the dynamic effects of industrial structure on economic fluctuation[J]. Industrial Economics Research, 2013(3): 91-100(in Chinese)
(彭冲, 李春风, 李玉双. 产业结构变迁对经济波动的动态影响研究[J]. 产业经济研究, 2013(3): 91-100)
- [12] Xue Jiliang. Technology choice and transformation and upgrading of industrial structure[J]. Industrial Economics Research, 2013(6): 29-37(in Chinese)
(薛继亮. 技术选择与产业结构转型升级[J]. 产业经济研究, 2013(6): 29-37)
- [13] Qu H M, Chan W Y, Xu A B, *et al.* Visual analysis of the air pollution problem in Hong Kong[J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6): 1408-1415
- [14] Wu Y C, Wei F, Liu S X, *et al.* OpinionSeer: interactive visualization of hotel customer feedback[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1109-1118
- [15] Cao N, Lin Y R, Sun X H, *et al.* Whisper: tracing the spatio-temporal process of information diffusion in real time[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2649-2658
- [16] Sacha D, Stoffel A, Stoffel F, *et al.* Knowledge generation model for visual analytics[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1604-1613
- [17] Choo J, Lee H, Kihm J, *et al.* iVisClassifier: an interactive vis-

- ual analytics system for classification based on supervised dimension reduction[C] //Proceedings of IEEE Symposium on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2010: 27-34
- [18] Bogl M, Aigner W, Filzmoser P, *et al.* Visual analytics for model selection in time series analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2237-2246
- [19] Ferreira N, Poco J, Vo H T, *et al.* Visual exploration of big spatio-temporal urban data: a study of New York city taxi trips[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2149-2158
- [20] Doraiswamy H, Ferreira N, Damoulas T, *et al.* Using topological analysis to support event-guided exploration in urban data[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2634-2643
- [21] Turkay C, Slingsby A, Hauser H, *et al.* Attribute signatures: dynamic visual summaries for analyzing multivariate geographical data[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2033-2042
- [22] Li J, Zhang K, Meng Z P. Vismate: interactive visual analysis of station-based observation data on climate changes[C] //Proceedings of IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2014: 133-142
- [23] Torgerson W S. Multidimensional scaling: I. theory and method[J]. Psychometrika, 1952, 17(4): 401-419
- [24] Shi C L, Fu S W, Chen Q, *et al.* VisMOOC: visualizing video clickstream data from massive open online courses[C] //Proceedings of IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2014: 277-278