基于多维时空数据可视化的传染病模式分析

金思辰, 陶煜波*, 严宇宇, 戴浩然

(浙江大学 CAD&CG 国家重点实验室 杭州 310058)
(taoyubo@cad.zju.edu.cn)

摘 要: 近年来, 诸如非典型性肺炎、甲型 H1N1、手足口病等传染病在社会各界引起了广泛关注, 传染病的暴发往往具有季节性、空间性和相关性. 为此提出了传染病可视分析系统, 以直观分析传染病的时空模式, 交互挖掘不同疾病、地区之间的关联性和相似性. 时间模式通过时序折线图、季节变化堆叠图可视化不同传染病的长期趋势、季节消长规律, 同时通过构建时序平行坐标系, 交互利用地区分布对比图进行异常值分析. 空间模式通过地图分类图以及交叉对比热力图来反应疾病在不同省份的空间分布规律、空间聚类结果以及相似性分析. 通过对 39 种法定传染病的整体、个别趋势以及其中异常值的分析结果, 表明所提系统能够综合考虑传染病数据的多维时空特性, 可有效帮助用户挖掘传染病传播的时空模式, 快速寻找传染病暴发时间节点和空间分布转移事件, 从而更好地进行预防、把控和分析.

关键词: 多变量时空数据; 传染病模式; 可视分析
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2019.17653

Infectious Disease Patterns Analysis Based on Visualization of Multidimensional Space-Time Data

Jin Sichen, Tao Yubo*, Yan Yuyu, and Dai Haoran

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

Abstract: In recent years, the outbreak of infectious diseases, such as SARS (severe acute respiratory syndromes), influenza A (H1N1), and hand-foot-and-mouth disease, has caused a widespread concern in our society. The outbreak of infectious diseases is often seasonal, spatial, and related. To intuitively analyze the spatial-temporal patterns of infectious diseases, interactively mining regional associations and similarities between different diseases, this paper presents a visual analysis system for infectious diseases. The seasonal and annual patterns of infectious diseases are visualized by stack graphs and line charts, and abnormal events can be extracted from a temporal parallel coordinate and regional distribution comparison bars. The spatial distribution patterns of diseases are encoded in the choropleth map to analyze the spatial clustering results and their similarities. According to the analysis of the overall, individual trends and outliers of 39 statutory infectious diseases, it is shown that the system can comprehensively explore the multi-dimensional temporal and spatial characteristics of infectious diseases and effectively help users to discover the implicit temporal and spatial patterns, and thus better prevent, control and analyze the infectious diseases.

收稿日期: 2018-11-28; 修回日期: 2018-12-12. 基金项目: 国家自然科学基金(61472354, 61672452); NSFC-广东省人民政府大数据科学研究中心项目(U1611263); 金思辰(1996—), 女, 在校学生; 陶煜波(1980—), 男, 博士, 副教授, CCF 会员, 论文通讯作者, 主要研究方向为数据可视化、可视分析; 严宇宇(1991—), 男, 博士, 主要研究方向为数据可视化、机器学习; 戴浩然(1997—), 男, 在校学生.

Key words: multivariable space-temporal data; patterns of infectious disease; visual analysis

1 简 介

随着社会卫生设施以及医疗水平的改善和提高，很多曾经引起全球性大规模暴发的传染性疾病已经得到有效的监测和防控，但是，仍有一些传染病的流行，危害着人们的生命和健康. 20 世纪 80 年代，艾滋病开始肆虐全球，至今仍然对人类的健康构成极大的威胁. 长期以来，分析传染病的传播模式一直是各国政府和有关专家关注的课题. 因此，通过传染病模式分析能够对疫情实行监控，以直观的方式为卫生行政部门在资源分配、疾病监测等方面提供科学参考.

通过可视化的方法对传染病进行分析可以追溯到 19 世纪，英国医生约翰·斯诺(John Snow, 1813—1858)使用地图标记，成功地发现了霍乱的扩散渠道及致病水源，进而阻止了 1854 年伦敦霍乱疫情的蔓延，成为地图数据可视化的典型应用. 如今可视化在空间流行疾病及公共卫生健康领域得到广泛应用，传染病时空模式分析的研究焦点也逐步转向可视分析工具的应用. 随着传染病数据的迅速增长，大量工作围绕以大数据分析为基础的传染病监测预警展开. 例如，胡雪芸等[1]通过全局空间自相关、局部自相关分析、时间序列预测分析及时空扫描分析方法，将可视化技术应用于空间流行疾病领域，对全国肺结核疾病的时空聚集情况进行分析及可视化，帮助分析疾病在时空上的聚集情况；别芹芹等[2]用地理信息系统(GIS)的方法和工具，对 2008—2009 年中国疾病预防控制信息系统收集得到的手足口病监测数据进行空间可视化分析和统计计算，得到了我国手足口病 2008—2009 年疫情的时空分布及动态变化特征. 但是大部分工作主要围绕单个特定疾病的监测和预警展开，缺少能够通过交互方式，帮助用户对比分析不同传染病之间的时空相似性、挖掘不同传染病时空模式、提取其聚类、异常值信息的多维数据协同分析可视化系统.

因此，本文提出了传染病多维时空数据协同可视化系统，通过关联多个视图，提供丰富的交互操作，支持用户分别从特定疾病分析、多种传染病比较，从时间和空间等不同角度进行传染模式、聚类以及异常值分析. 对于传染病时间模式的分析，主要通过交互生成时序折线图，以对比查看特定疾病病种、疾病类型的暴发周期和长期趋势；构建季节变化堆叠图查看疾病的累积季节消长模式；引入 *t*-SNE 算法将原始传染病多维时空属性数据降维后，构建时序平行坐标，交互利用空间分布对比条形图进行异常值分析，同时利用聚类结果优化视图. 对于传染病空间模式分析，利用地图分类图显示地区发病数，针对传染病时空聚类结果进一步提供交叉对比热力图，以帮助用户对比任意省份、疾病之间是否彼此相似，检测它们之间是否存在聚类结果中的相关性. 本文分析了国家公共卫生科学数据中心提供的自 2004—2015 年 39 种法定传染病数据，通过分析中国(不含港澳台，下同)各省份传染病的发病模式，验证本文所提系统的有效性和实用性.

本文的主要贡献如下：

- 提出了一个交互式可视分析系统，帮助用户分析不同疾病在时空上的分布.
- 设计了交叉对比热力图，辅助用户查看聚类结果中的细节信息.
- 设计了时序平行坐标异常值检测视图，帮助用户对比异常出现的细节信息.
- 通过案例分析对本文系统的可用性进行了评估.

2 相关工作

本文分析的疾病数据属于多变量时空数据，采用多个视图来展示多个变量的信息.

2.1 多变量时变数据的可视化

多变量数据是实际应用中常见的数据类型. 很多可视化方法可以用来展示多变量数据，例如，柱状图、平行坐标、散点矩阵图、雷达图和像素图. 一些工作在尝试将这些基本的可视化方法相结合的方式来展示多变量数据，例如，Keim 等[3]将传统的柱状图和 *x*-*y* 图结合，提出了层次的像素柱状图，可以展示包含类别型和数值型的多变量数据. 为了解决平行坐标展示变量相关性的问题. Zhou 等[4]使用索引点来表示高维数据中的平面来可视化平行坐标中局部变量的相关性. 还有一些研究尝试使用降维的方法来可视化多变量数据，例如，PCA[5]和 MDS[6]. 但是，数据中的类别可能在某个子空间内是分开的，在整个高维数据空间内反而

无法分开. Yuan 等[7]提出了 Dimension Projection Tree/Matrix，树中每个节点分别包含数据样本和维度的降维矩阵散点图；用户可以选择部分数据样本或者部分维度属性，生成 1 个子节点，来探索发现一个合适的子空间. 除此之外，还有些研究尝试使用图形(glyph)来展示多变量数据. Cao 等[8]对传统的图形进行了改进，提出了 Z-Glyph 来帮助用户进行异常分析；其中，正常的数据被表示成常规的形状(如直线和圆). 当异常数据偏离常规形状时，它会很容易被用户发现. Liao 等[9]对散点图进行了改进，使其能展示数据降维之后的具体信息；通过对散点图中的数据进行聚类，然后使用图形展示每个类的具体信息.

当时间作为多变量数据中一个重要属性，可视化需要兼顾数据本身属性和数据集的顺序性. Kwan[10]描述了几种基于 GIS 的、用于处理时空数据的三维地理空间可视化方法. Guo 等[11]提出了 VIS-STAMP，它可以通过聚类、排序和可视化帮助人们探索并理解多变量时空数据中的复杂模式. 系统包含多个视图：地图、平行坐标、矩阵图，以此方便分析数据中的多个变量. 周志光等[12]设计了地图视图、多维属性视图、层次聚类视图、时序平行坐标视图来协同展示多变量时空数据. 与之前的工作的不同之处在于，本文的数据包含了时间和空间维度，并且在每个时间，空间维度又包含多个疾病的相应属性. 因此，本文通过降维、聚类的方式对数据进行预处理，然后通过多个可视化视图联动的方式展示各个疾病在时空上的分布.

2.2　健康信息可视化

在健康信息可视化方面，有些工作关注单个病人健康信息的可视化，如 Lifelines[13]对临床病人历史记录的可视化，并使用时间线表示病人记录的概述. Rind 等[14]展示了交互式可视化系统 VisuExplore 的设计研究，VisuExplore 提供交互技术和可视化方法帮助分析慢性疾病病人. 还有一些工作关注于分析一群病人，以此协助临床研究人员评估和改善患者的护理质量. PatternFinder[15]是一个根据值和时间间隔进行查询的工具，并对查询得到的结果以 ball-and-chain 的形式进行可视化；它可以找到多变量数据中的时序模式，用于搜索病人的历史记录、web 日志、新闻报道、犯罪活动. Lifelines2[16]基于事件发生的时间将每个病人的记录表示成水平的时间线，使用 temporal summary 显示每一段时间内事件类型的分布. LifeFlow[17]拓展了 Lifelines2，以序列树的形式总结和表示一段时间内所有的时间序列. 在疾病时空分布可视分析方面，胡雪芸等[1]对单个疾病进行空间聚集分析、时间序列分析及时空扫描统计分析，并通过地图和折线图分别展示疾病在空间和时间上的分布. 曲玉冰等[18]研发了一套基于空间信息技术和大数据技术的登革热风险评估系统，可以对疾病进行时空聚集探测、影响因素提取、风险评估、大数据分析和可视化. 与之前的工作的不同之处在于，本文的可视分析系统可同时对多个疾病进行分析比较；而之前疾病时空分布方面的工作仅仅将可视化作为展示的方式，本文的系统需要用户根据可视化的信息交互式地分析，探索数据中包含的模式.

3　需求分析

传染病的发病模式既有时空连续性、又存在异常暴发点. 不同省份以及疾病之间存在着一定的内在关联性和相似性，如手足口病在 2005 年大规模暴发，而在此之前并无病例出现；人禽流感多暴发于冬季气候恶劣之时等等. 面对时空数据，用户无法直观获取数据中潜藏的空间信息和时间序列信息，同时多维属性增加了模式提取的困难，不利于用户获取特定层面的规律信息.

传染病分析主要包括时空规律、聚类和异常值分析. 首先，挖掘传染病时间规律包括分析总体发病趋势以及特定疾病的季节规律，帮助检测新出现和重新出现的疾病. 传染病的暴发和消退趋势以年为周期存在季度发病高峰性，在总的时间尺度上存在长期趋势，因此要把握传染病的疫情形势和防控能力，寻找出现暴发现象和消退的时间点；空间规律包括传染病的地区分布规律性、随时间的扩散趋势，同时帮助用户寻找具有相似发病模式的地区. 分析传染病的地区分布有助于帮助用户控制传染源、切断传播途径，确定传染病的空间聚集程度、扩散趋势；同时，受到地理、气候、人口、饮食习惯等内在因素的影响，各种疾病在不同省份的流行特征存在一定的模式，分析地域发病模式能够有效地预测传染病扩散趋势，为疾病监测、资源分配等方面提供科学参考；聚类分析包括帮助用户寻找具有相似传染模式的疾病. 由于传染源、传播媒介、病毒活跃时间等的内在联系，

可能导致不同传染病之间具有相似的流行特征，把握这种联系可以帮助用户分析新出现传染病可能的传播方式以及传染源；异常值分析包括帮助用户快速发现出现地区分布转移、变化的时间节点，并交互地对比查看该事件出现的时间点的地区分布变化，帮助及时发现疾病在时间、空间上的异常聚集，提供对疾病暴发进行快速反应的监测方法.

基于以上需求分析，本文总结了以下任务，并设计面向多变量时空数据的协同可视化系统:

T1 分析传染病的时序规律性.

T2 分析传染病的空间分布、演化规律.

T3 对不同传染病进行聚类分析.

T4 传染病异常值分析.

4 系统概述

根据以上任务，本文提出的系统主要由 3 个模块组成，数据处理模块、可视化模块、交互模块. 根据这 3 个模块本文提出了可视分析系统流程，如图 1 所示.

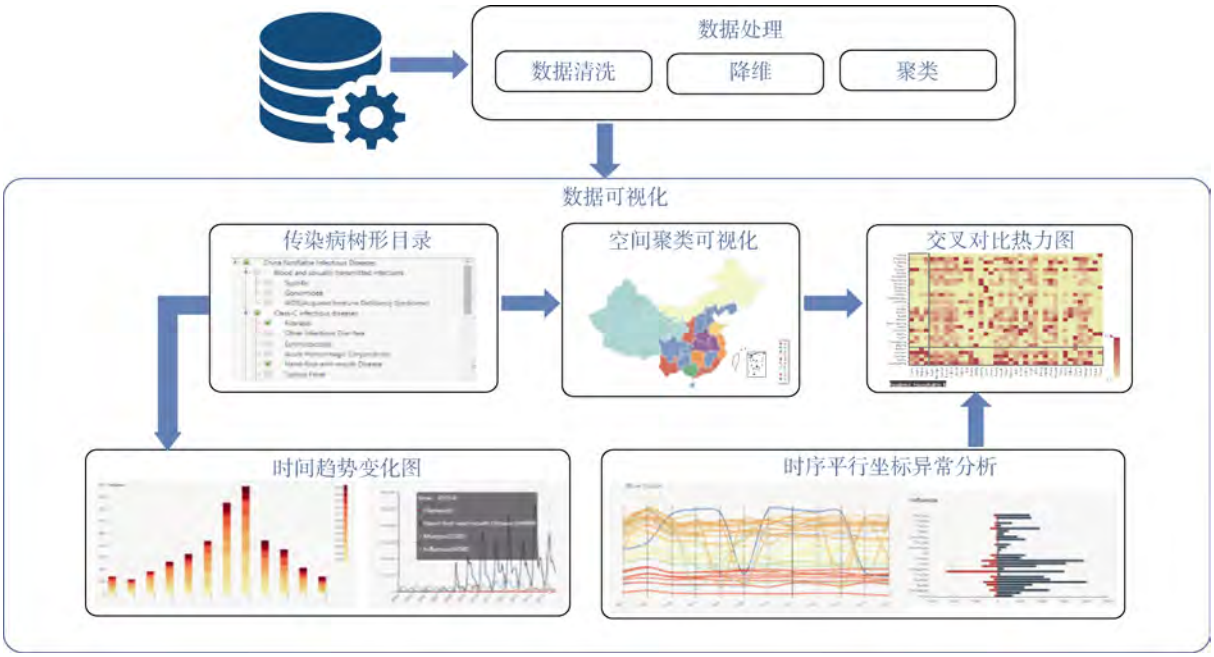


图 1 可视分析系统流程

首先对数据进行清洗、降维和聚类操作，便于进一步结合可视化图表分析传染病时空演化、聚类模式(T1,T2,T3). 然后，通过对数据可视化，以及不同视图之间的交互对比查看内部可视细节分析传染病异常值出现节点(T4)以及时空演化过程(T1,T2). 通过总体趋势帮助用户概览多维数据的时空变化规律以及总体的统计信息(T1,T2)，并交互利用细节分析对比不同疾病的地区分布结构、相同疾病的地区分布转移和变化的异同(T3,T4).

5 可视化设计

基于以上分析，本文提出了多变量时空数据的协同可视化系统，通过多个视图可视化传染病时空、聚类以及降维信息，同时利用不同视图之间的联动进行异常和时空变化规律分析.

本文系统主要包括空间视图、时间趋势变化图、时空协同分析视图，以下将对各个视图的可视化设计细节做进一步描述.

5.1 空间视图

本文根据地理学第一定律空间相关性，地物之间的相关性与距离有关. 通常，距离越近，地物间相关性越大；距离越远，地物间相异性越大. 地图视图可以直观地反映空间属性，便于用户分析相邻地区之间的疾病分布情况、新出现疾病的传播方向，以及疾病聚集的热点地区.

地图视图包括 3 个图层，如图 2c 所示，分别为整体、个别疾病发病数的地图分类图以及疾病的地区聚类信息，用户可以根据需求切换. 其中地区分类图层以省份为单位，利用颜色深浅反应发病数、发病率高低，同时交互拖动时间轴，可显示不同疾病病种、疾病类别、所有疾病发病数在各省分布随

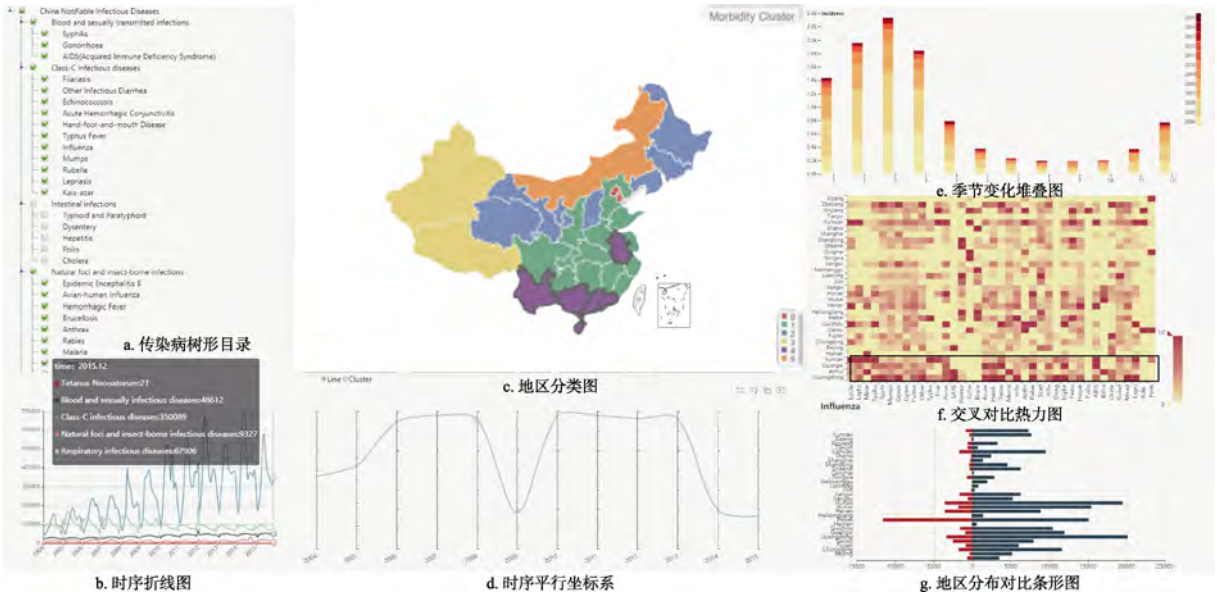


图 2　系统界面概览

时间的变化情况，以便用户寻找不同传染病高发区，以及传染病沿时间的区域扩散形式，有效地解读传染病在不同地域(如沿海、内陆)的传播趋势.

为了寻找 2004-2015 年间具有相同传染病发病趋势的地区，通过层次聚类算法结合地图分类图显示不同层次的地区聚类结果.

由于不同疾病具有不同的地区分布特征，在聚类过程中难以统一确定类别数目，对于地区分布差异较大的疾病过少的类别不能有效地体现类别内部的特征；而对于地区分布差异较小的疾病过多的类别，突出了个体差异却不能体现地区之间的相似性，因此，本文提出利用层次聚类自顶向下的方法，针对每一种疾病得到一系列不同层次的聚类结果. 用户可以交互选择聚类层次，查看相应层次的聚类结果.

基于 39 种法定传染病在 2004 年 1 月—2015 年 12 月的 12 年，按月份和地区发病数统计数据，针对每种疾病构建省份时间的 31×144(12 年×12 个月)维矩阵，基于两两向量之间余弦距离采用层次聚类的方法，得到每个层次的聚类结果.

5.2　时间趋势变化图

5.2.1　时序折线图

如图 2b 所示，时序折线图以时间为横轴，发病数为纵轴，显示 2004 年 1 月—2015 年 12 月特定传染病发病的时间变化规律，通过折线图波动情况可以直观反映不同疾病病种、疾病类别、总体疾病的变化趋势，观察疾病暴发的年度峰值、季节峰值，以及波动周期，把握传染病的控制形势，概览传染病的时序变化规律. 同时通过折线图和其他视图的交互，可以对比不同疾病在时间轴上的走势. 为满足用户进行不同疾病之间准确数值对比的需求，图表提供了鼠标事件查看具体信息，显示当前时间点图中几种疾病的发病数.

5.2.2　季节变化堆叠图

如图 2e 所示，季节变化堆叠图以月份为横轴，统计这 12 年相同月份的发病数，并利用柱形堆叠图展示统计结果，通过横向比较柱状图的高度变化可以得到疾病暴发的高峰季节. 利用堆叠图而不是普通柱状图的原因是，保留了年份的信息的同时，可以增强季度变化趋势的视觉感知，使累积的季度变化状态更加明显. 为了提供准确的数值信息、增加图表的可读性，堆叠图利用鼠标事件交互显示具体数值信息.

5.3　时空协同分析视图

5.3.1　时序平行坐标异常分析视图

本文设计了时空协同可视化视图来分析原始传染病数据中时空协同变化规律，使用降维后的数据构建时序平行坐标系，同时切换视图显示疾病聚类结果，直观地呈现用户感兴趣的特征时空演化和聚类结果，并利用正负条形图交互显示异变节点的空间分布对比.

(1) 时序平行坐标系. 如图 1 所示时序平行坐标，为了协同呈现传染病数据中的多变量时空属性，本文设计实现时序平行坐标，以年份和省份为统计单位. 首先利用 t-SNE[19]算法将传染病的多维空间属性以年为单位降解到一维特征空间，再

按照年份的时间顺序依次排列在平行坐标轴上，平行坐标每一个坐标轴上的 *t*-SNE 坐标依次显示了该年不同传染病对象在空间分布上的差异，投影点距离越相近，相应的传染病具有越相似的空间和季节分布；将不同坐标轴上属于同一传染病对象的坐标点连接起来，曲线的趋势变化能够有效地表达传染病的空间季节分布的时序变化特征，并帮助用户快速发现一些突变的异常点．这些时间节点表示传染病空间季节分布产生了变化，是值得重点关注的节点；针对这些节点，用户可以利用时间轴框选工具，查看特定传染病的其前后地区分布对比情况．

同时基于平行坐标系降维数值构建疾病的 12 维时空向量，利用 KL(Kullback-Leibler)对称距离表示两两向量之间的距离，通过 *K*-Means 聚类分析疾病相似性，结合平行坐标显示聚类结果，将聚类中心 12 维的坐标投影到平行坐标轴，利用线宽直观体现该类别疾病数目，如图 2d 所示．用户可以根据需要切换平行坐标系为聚类视图，隐藏具体疾病信息，显示类别聚类结果，同时通过框选特定疾病利用热力图显示所选传染病之间的地区相似性对比．

(2) 地区分布对比条形图．如图 2g 所示，针对平行坐标系中所反映出来的空间季节突变的节点，系统提供用户交互查看其异变原因的正负条形对比视图，利用正负轴分别显示此节点出现前后一年该疾病在地区分布上的差异，帮助寻找造成传染病空间分布差异的干扰因素．

5.3.2 交叉对比热力图

如图 2f 所示，利用热力图借助色彩变化可视化特定省份特定疾病发病数占该种疾病总发病数的比例数据以及统计发病数．以列表示省份，行表示疾病病种，为保持图表整洁统一，横轴仅显示疾病名称前 5 个字母，用户可以通过交互查看疾病全称，以单元格色调显示各省此种病发病占比的高低．热力图可帮助用户交叉检查多元数据，在多个变量之间显示其中差异，结合前述地区和疾病聚类结果，揭示其中的相似、差异结构，对比任意省份、疾病之间是否彼此相似，检测它们之间是否存在聚类结果中的相关性．通过和地图以及平行坐标之间的交互协同分析疾病、地区聚类结果．

为了增加热力图的可读性，用户可以通过颜色标度图例控件改变颜色映射取值范围，以适应具体疾病的发病数范围；同时热力图会根据用户的选择自动调整绘制顺序，将用户选中的疾病和地区依次邻近排列，方便用户对比具体信息．

6 案例分析

基于本文所提出的系统，分析了国家公共卫生科学数据中心发布的自 2004—2015 年 39 种法定传染病数据，挖掘中国(不含港澳台，下同)各省份 39 种传染病的发病模式，分别从传染病的整体、局部时空变化、降维、聚类结果方面进行分析．

6.1 传染病整体趋势

6.1.1 时间序列分析

(1) 长期趋势．2004—2015 年，我国法定传染病报告病例数表现为以年为周期的周期性波动，并且波动幅度呈现增大的趋势，暴发时间基本稳定在每年夏季 5—8 月，并且峰值有所增大；而在每年 1—2 月传染病流行程度明显降低，如图 3a 所示．而不同类型疾病也呈现出不同的流行趋势，如图 3b~3d 所示，主要特征体现为血源及性传播传染病呈逐渐上升趋势，并在 2012 年暴发(图 3b)；丙类传染病则呈现持续走高的趋势，暴发集中于夏季且峰值逐年上升(图 3c)；新生儿破伤风(图 3d)呈持续下降趋势，这主要是因为我国城乡新法接生技术的应用和推广，使得该病的发病率逐渐降低．

(2) 季节消长．我国传染病流行高峰多出现在夏季 5—8 月(图 4f)，但是传染病季节性流行高峰出现时间、持续时间与不同疾病类型相关，主要表现为血源及性传播传染病全年发病无明显高峰期，1—2 月发病相对较少(图 4a)；丙类传染病发病主高峰主要集中在春夏季 5—7 月，11 月出现小高峰，典型的有手足口病，3 月开始逐渐增多，随着 5 月春夏之交，气温攀升，湿度增加，病菌滋生，5 月底至 6 月中旬达到高峰(图 4b)；自然疫源及虫媒传染病则没有总体的季节性规律，不同病种之间有所差异，如人禽流感多发于春季(图 4c)，主要原因是流感病毒对温度比较敏感，夏季时在环境中很快会死亡，而冬季存活时间很长，同时由于冬春季禽舍注重保暖，通风强度不够，就容易导致病毒的传播和扩散；而乙脑主要流行于夏秋季(图 3d)，80%~90%乙脑的病例集中在 7—9 月，这主要是由乙脑的传播方式所决定的，乙脑主要通过蚊子叮咬而传播，而蚊虫多活跃于夏季；呼吸道传染病的高发于春季(图 4e)，这是因为季节交替，时暖时寒，人体内环境很难以与外界环境相适应，对病毒抵抗

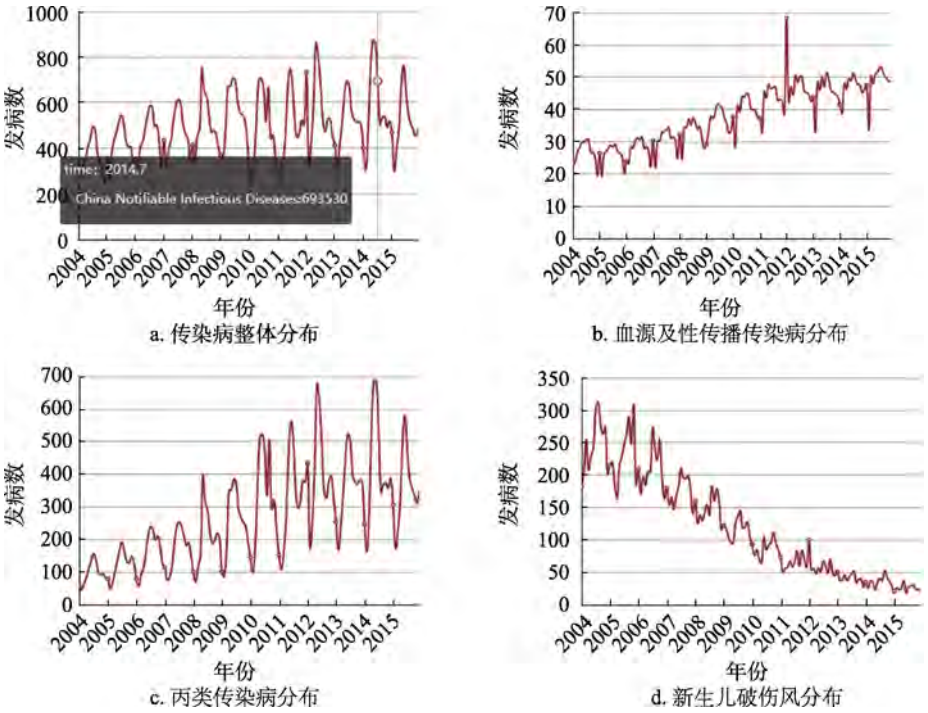


图 3　传染病整体长期趋势

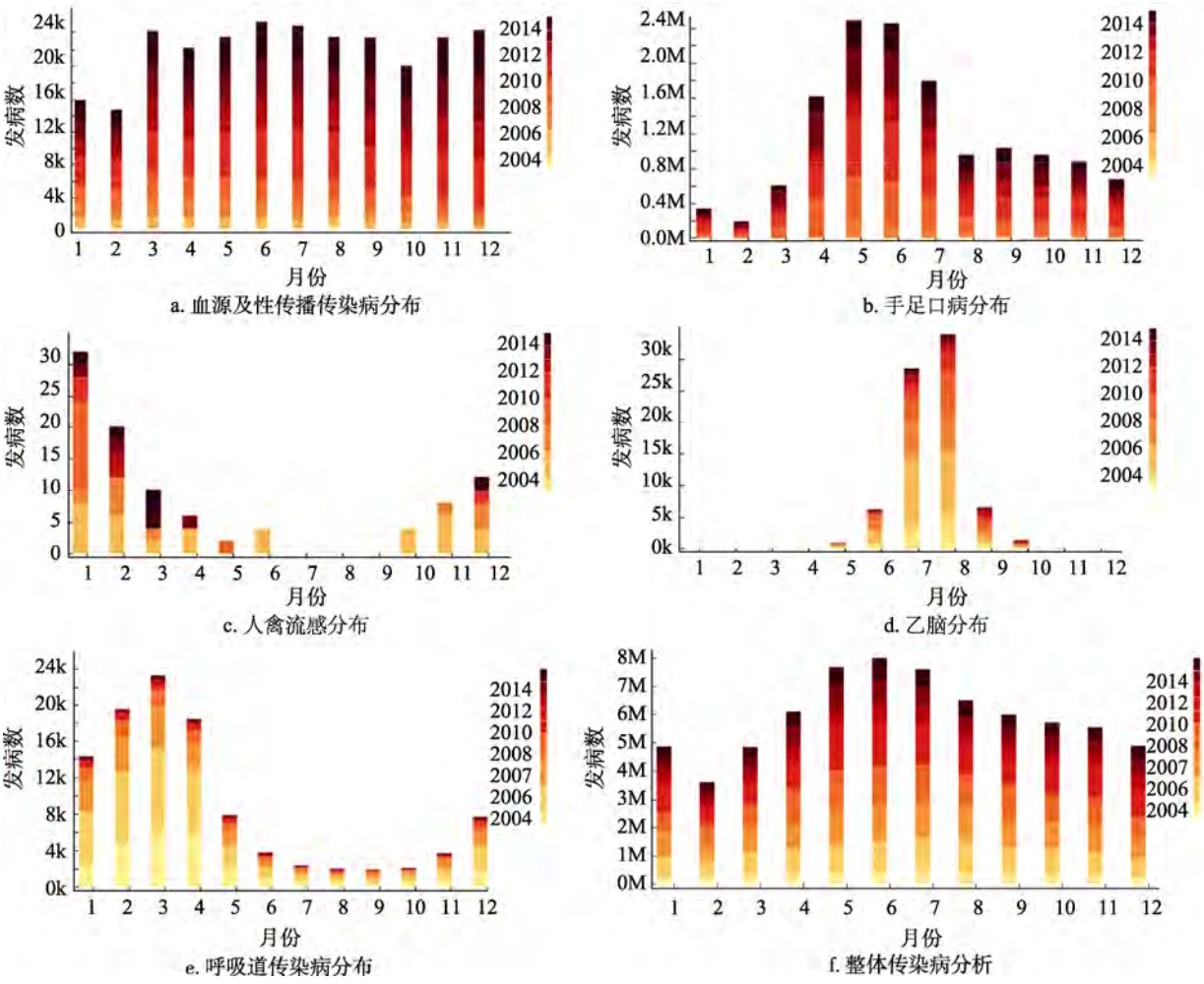


图 4　传染病季节消长

力较低，给病菌乘虚而入的机会，因此特别容易引起流脑、流感等呼吸道传染病.

6.1.2 空间分布分析

(1) 热点分析. 通过查看 39 种法定传染病的地区分布图及分析疾病聚集地，识别 12 年间 39 种疾病统计发病数的高值地区和低值地区. 由图 5a 可知，传染病暴发重灾区以广东和浙江 2 个省最为显著. 这主要由于 2 个省的气候和地理位置具有一定的相似性，夏季高温，沿海湿润，提供了传染病多发的环境因素. 同时考虑到人口因素产生的影响，对比查看总体发病率的地区分布情况图(图 5c)可以发现，发病率较高的地区为新疆、北京和浙江；由于人口基数较小，北京和新疆发病数较少但是发病率较高；传染病高发地区主要集中在东南沿海以及新疆地区.

通过交互分析不同年份的发病率地区分布图，如图 6 所示，可以发现随着时间变化热点区域有所改变，2004—2009 年，天津、浙江、北京为疾病高发区；随着时间变化，2009 年新增西藏、新疆为热点区域；到 2013 年西藏发病率降低，广西、广东、海南发病率上升，一直持续到 2015 年.

(2) 空间聚类分析. 基于发病数地区聚类结果显示我国具有相似传染病模式的省份，如图 5c 所示，大致可以分为西部发病率较高、病种特有的新疆、西藏地区；气候寒冷干燥，多集中暴发丙类传染病的东三省和内陆青海、甘肃等;发病时间、发病种类同增同减的中部地区；多发肠道传染病的南部沿海亚热带、热带地区.

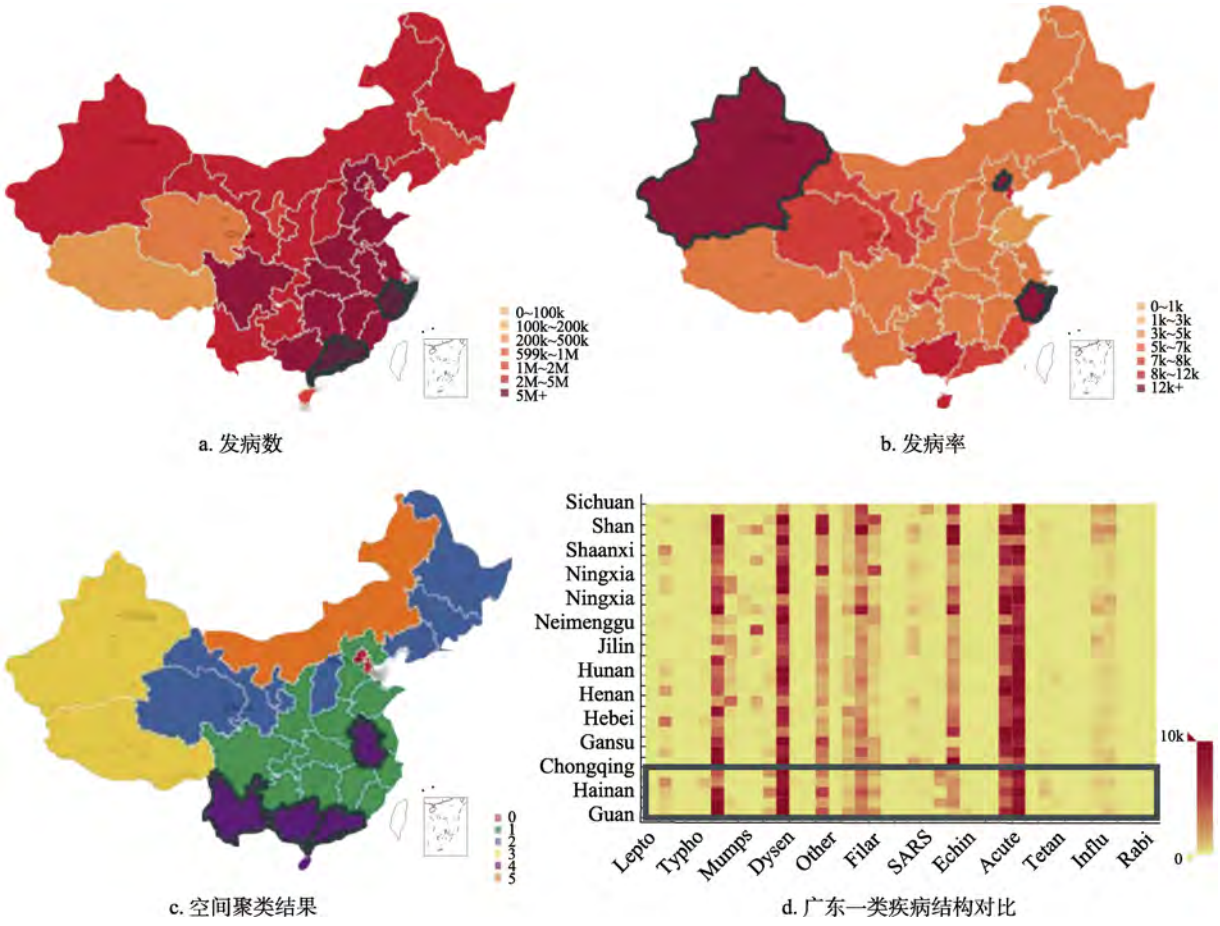


图 5 传染病整体地图分类图及聚类结果分析

交互查看热力图中各省统计发病数的分布情况，可以发现广东一类(图 5d)包括云南、广东、广西、海南、安徽具有相似的发病结构，其暴发数目较少的疾病种类基本一致，同时均多发手足口病(hand-foot-and-mouth disease, Hand)和肺结核(pulmonary tuberculosis, Pulmo)，同比其他地区在流行性腮腺炎(mumps, Mumps)、肝炎(hepatitis, Hepat)、梅毒(syphilis, Syphi)具有基本一致的发病数量.

总体而言，地区聚类结果在很大程度上受到区域地理位置和气候因素的影响，距离相近的地区传染病暴发、消退时间同步，气候相似的地区多发的病种相似.

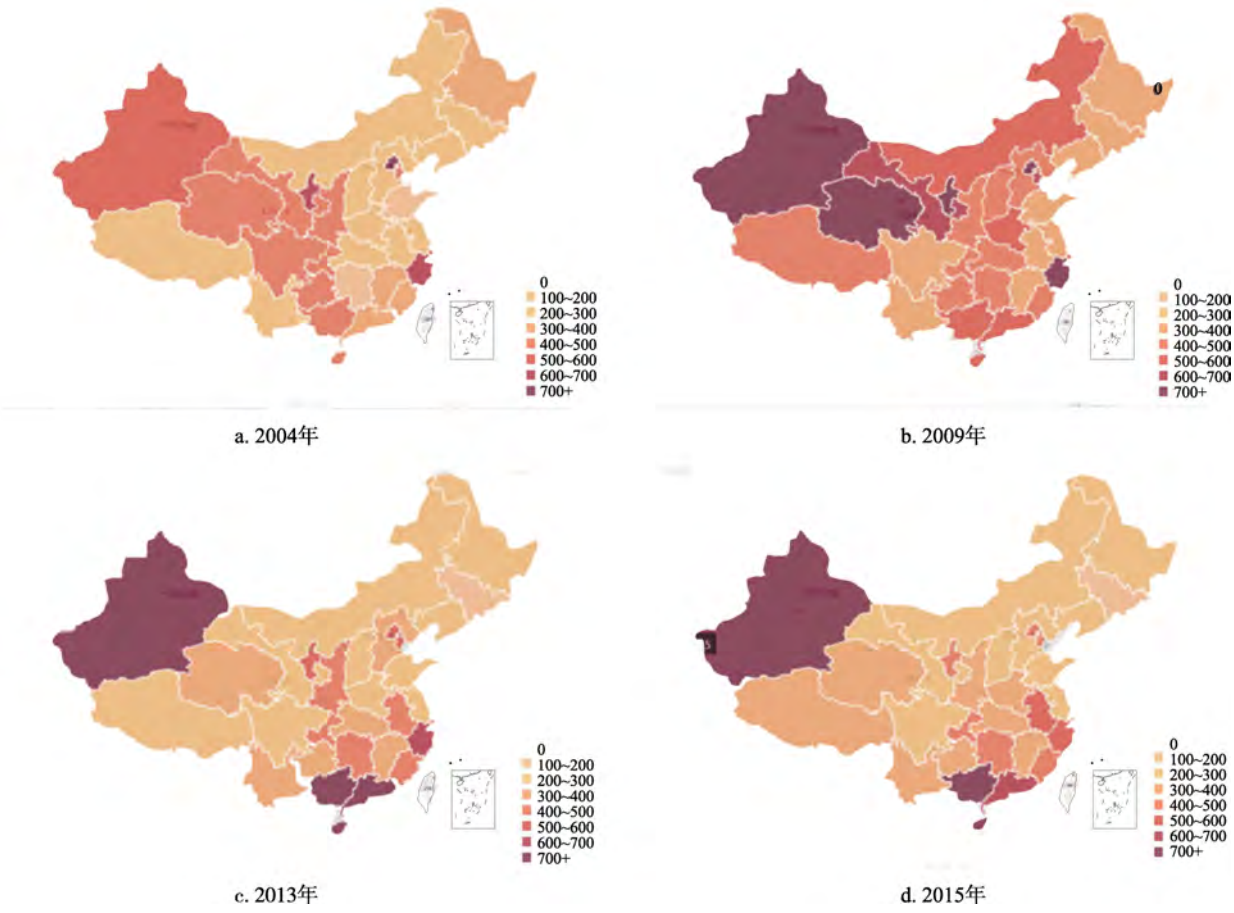


图 6　不同年份传染病整体发病数分布地图

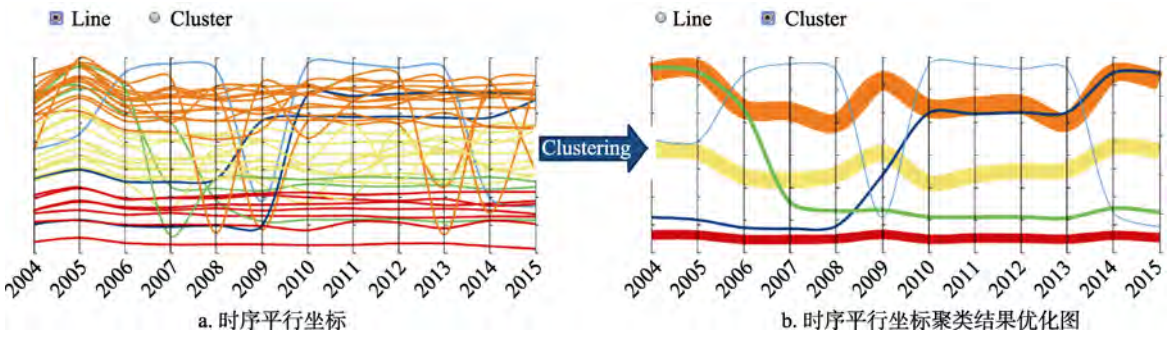


图 7　传染病时序平行坐标

6.1.3　传染病聚类分析

根据时序平行坐标的聚类结果显示，整体传染病根据空间分布和时间趋势变化可以分为 6 类，如图 7 b 所示. 本文选择其中属于一类的几种疾病：乙脑 (epidemic encephalitis B, Epide)、钩体病 (leptospirosis, Lepto)、疟疾(malaria, Malar)、伤寒副伤寒(typhoid and paratyphoid, Typho)，如图 8 所示，查看其具体时空分布结构，验证聚类结果的可靠性.

时间分布如图 9a 所示，观察发现这几种疾病高发季节集中在夏季 7—8 月，且 4 种疾病在 2004

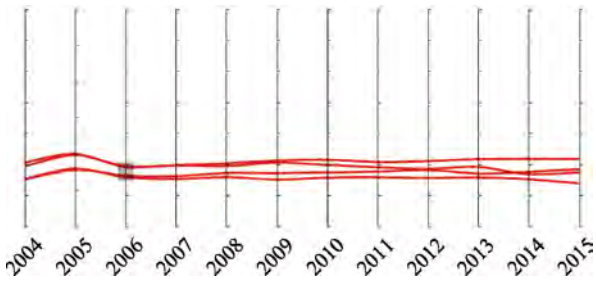


图 8　乙脑、钩体病、疟疾、伤寒和副伤寒时序平行坐标

—2015 年总体上呈下降趋势，其中乙脑和疟疾发病数都在 2007 年大幅上升，总体而言这几种疾病具有相似的时间变化趋势.

空间分布如图 9b 所示，这几种疾病都在云南多发，其次是贵州、广西、四川等地，其发病数较少的省份基本一致，大多为西北部地区.

同时根据医学分类知，这 4 种疾病中除伤寒副伤寒，乙脑、钩体病、疟疾均属自然疫源及虫媒传染病类，可见聚类结果不仅能够寻找不同传染病之间相似的时空分布，同时也符合医学分类.

6.2 传染病个体趋势

基于本文的系统可以通过时空等多角度有效分析单个传染病的传染模式，以登革热为例，介绍案例分析结果.

6.2.1 时 间

(1)长期趋势. 登革热(dengue fever, DF)是一种由登革病毒引起的，通过伊蚊传播的急性传染病，其发病特别主要表现为发病率较高且相对集中，同时传播势头迅猛. 受该病传播媒介和病毒特点的影响，该病目前主要流行于多高温潮湿天气的亚热带和热带地区.

通过观察登革热发病数在时序平行坐标的波动情况如图 10 所示，可以发现登革热在 2007 和 2013 年均出现小幅波动，同时对比折线图(图 11a)发现这 2 年均出现小范围内暴发，而在 2014 年出现剧烈波动，折线图显示登革热在 2014 年暴发，对比查看地区分布条形图(图 10)可以发现此次暴发集中于广东省.

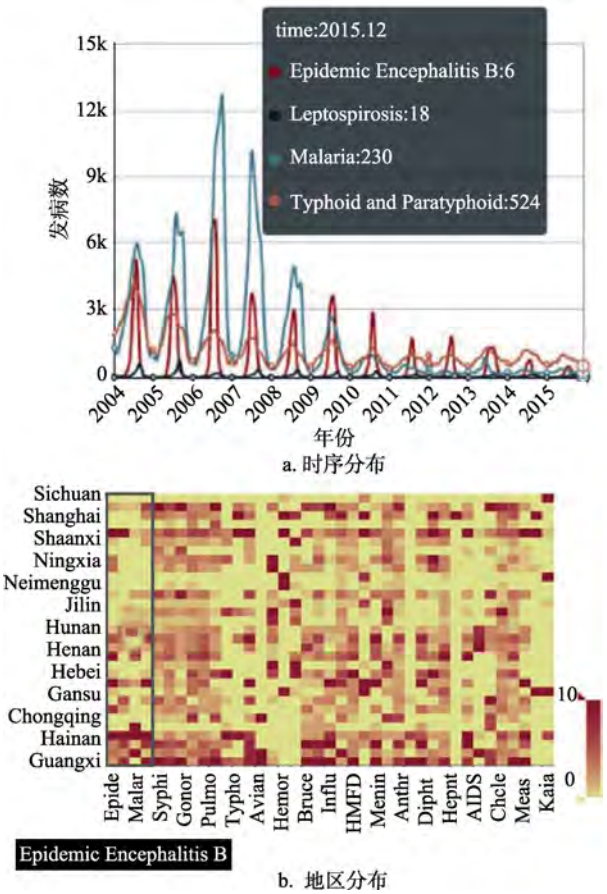


图 9 乙脑、钩体病、疟疾、伤寒和副伤寒聚类结果分析

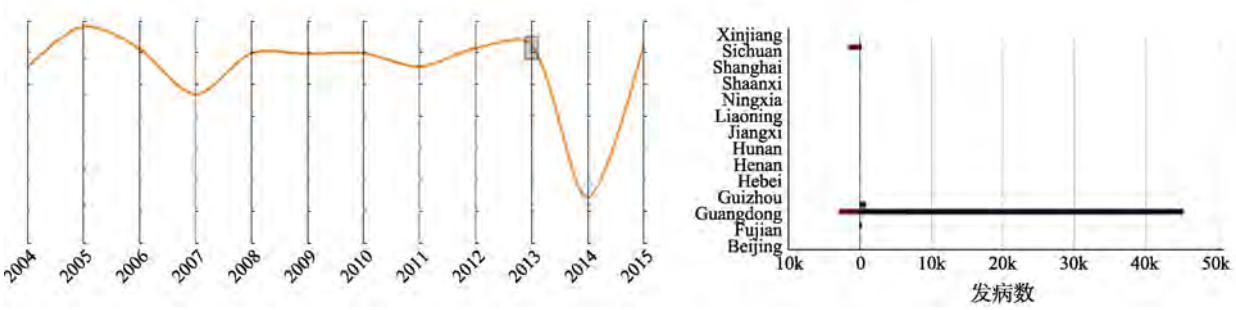


图 10 登革热时序平行坐标异常分析

总体而言，在 2004—2012 年的大部分时间里，该疾病在中国只有过零星的暴发. 但在 2013—2014 年在中国南方和沿海地带开始出现日益严重的疫情，2013 年 7 月发病人数为 175 例，8 月为 836 例，随后的几个月是登革热发病的高峰时期，9 月发病人数达到 1 480 例，10 月达到 1 665 例，11 月达到 395 例，再后面进入冬季明显回落(12 月达到 9 例). 同时卫计委在 2014 年 2 月 13 日 发布的《2013 年度全国法定传染病疫情情况》里也说明 2013 年登革热的发病相比 2012 年有所上升. 进入 2014 年，一直到 6 月发病人数都在 50 例以下. 从 7 月开始的 283 例，到 8 月的 1 936 例，紧接着就在 9 月一跃到近 2 万例，并且 10 月的病例仍然高居不下，形势相当的严峻，国家卫生和计划生育委员会、广东省政府都先后发布登革热疫情防控紧急通知. 而 11 月发病数就下降到 1 737 例，可见各地的防蚊灭蚊工作都取得

了有效的成果. 而在之后 2015 年及早预防，使疫情上升势头得以控制，使疫情得到了很好的控制.

(2) 季节消长. 革热的季节统计数据显示其传播具有明显的阶段性和季节特征(图 11 b)，并且根据其传播过程可划分为 3 个阶段：初期发展阶段为 1—6 月，此阶段发病数较少；中期暴发阶段为 6—10 月，此阶段登革热病例显著增多，表现出暴发性增长特点；后期消亡阶段为 11 月以后，进入冬季发病数明显回落.

6.2.2　空　间

(1) 热点分析. 登革热暴发热点地区主要集中在云南、广东、广西中国南方和沿海地带，其中以广东最甚，占全国发病数的 88%，如图 11 c 和图 11d 所示. 广东省地处亚热带，蚊媒孳生条件良好，同时具有人口密度高、外来人口多、流动性大等特点，在发病初期容易由输入型病例引起本地流行，加之人口密度高，扩散较快，因此广东省是全国的登革热高发区. 2004—2010 年以来疫情较为稳定，在 2006 和 2010 年出现小规模暴发，但是在 2012 年之后，广东省登革热疫情逐渐加剧，在 2013 和 2014 年发病数大幅上涨，到 2014 年累计报告病例已达 45 189 例，占广东登革热总发病数的 87%，疫情十分严重，是广东省近 12 年来最严峻的登革热疫情.

(2) 空间聚类分析. 由于各地登革热暴发季节消长趋势基本一致，集中于 5—10 月，入冬后发病势头缓和；在分析登革热的空间聚类结果时，重点讨论各省年发病数变化情况和相似性.

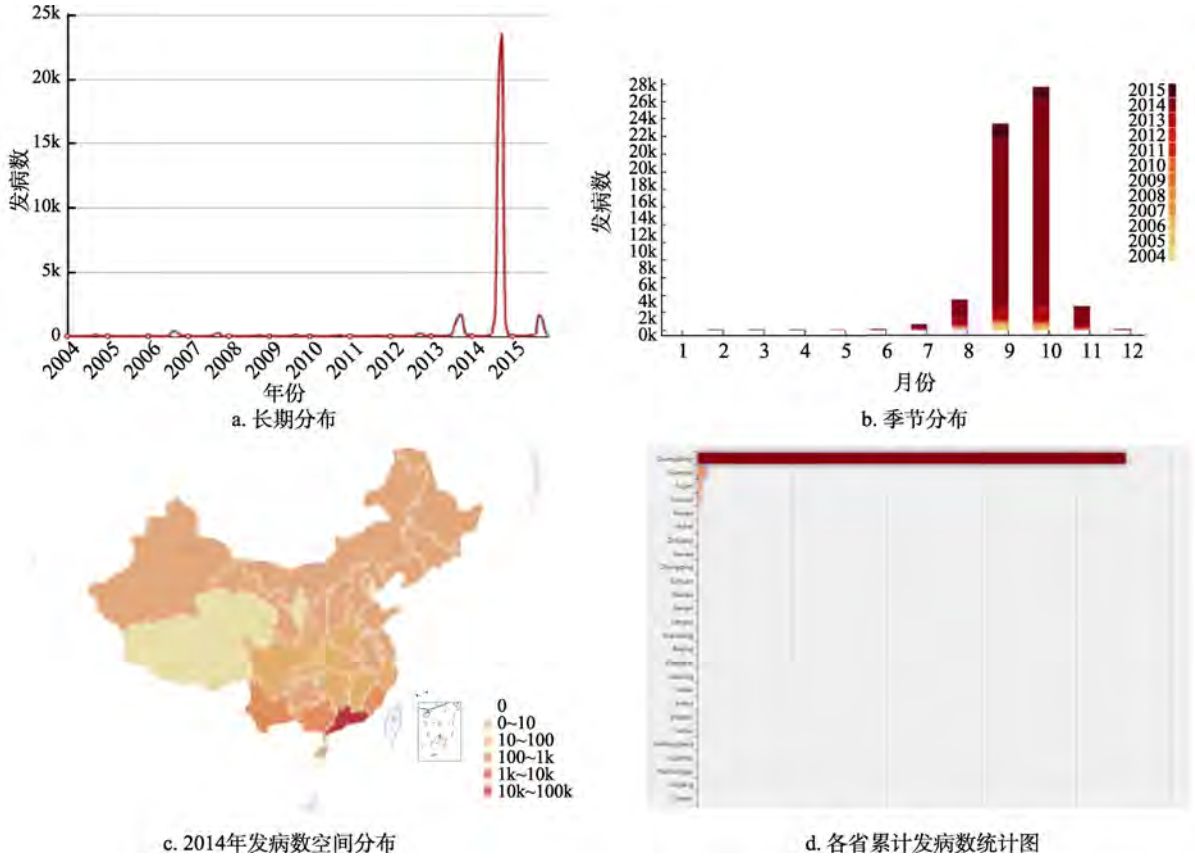
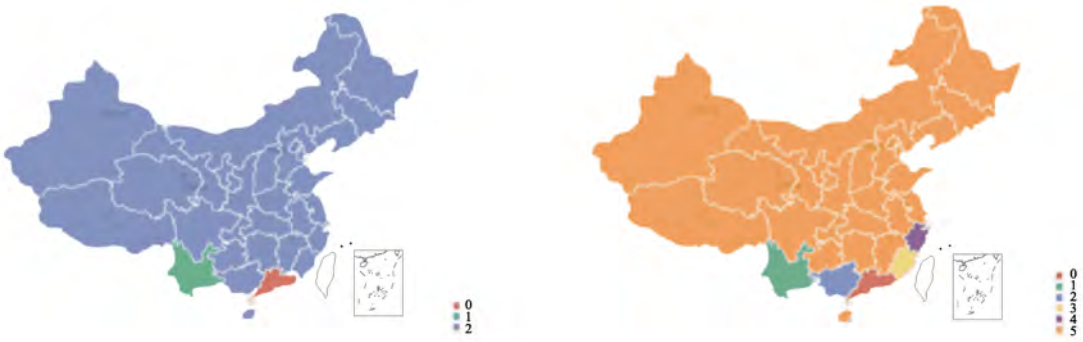


图 11　登革热发病模式细节特征分析

由图 11 d 所示登革热发病数统计结果可以发现，主要发病集中于广东，累计发病数高达 51 994 例；其次为云南 3 705 例，广西 900 例，福建 717 例，浙江 477 例；而其他省发病数则维持在较低水平，累计发病数在 100 例以下. 对比空间聚类结果可以看到，广东省由于较高的发病数自成一类，而随着聚类数目增加，云南、广西、福建、浙江也分别自成一类，而其他省份都处于暴发数目较少的一类如图 12 所示.

6.2.3　时空协同聚类和异常值分析

(1) 风疹空间分布转移. 风疹(measles)是由风疹病毒通过呼吸道和直接接触传播引起的急性病毒性传染病，主要通过感染者打喷嚏或者咳嗽形成的空气飞沫加以传播，家庭内有高度传播性.



a. 层次聚类第 1 层结果　　b. 层次聚类第 4 层结果

图 12　登革热层次聚类可视化结果

通过观察风疹发病数在时序平行坐标的波动情况如图 13 所示，发现其时空分布在 2011—2013 年出现变化和转移；结合图 14 所示时序折线图和图 15 b 和图 15 d 所示地区分布条形图分析，其在 2011—2013 年未出现异常暴发情况，发病数相较往年维持在较低水平，然而热点地区发生转移，2011 年高发病数地区为新疆、四川；而 2012 年新增高发病地区云南、广东，而且 2 个省在 2011 年发病数均处于中下水平，同时四川发病数大幅下降；2013 年伴随风疹发病数上升，新增高发病地区浙江、安徽、广西、山东.

(2) 流行性感冒疫潮. 流行性感冒(influenza)，简称流感，由流感病毒引发，主要通过人与人之间的接触、与被污染物品的接触以及空气中的飞沫传播，具有传染性强、传播速度快的特点，可以迅速在大范围人群中扩散并流行；流感流行常见于冬、春季. 流感病毒包括 3 种类型，即甲型、乙型及丙型，其中以甲型流感病毒感染比较常见.

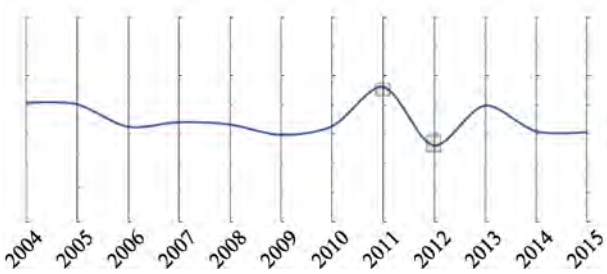


图 13　风疹时序平行坐标

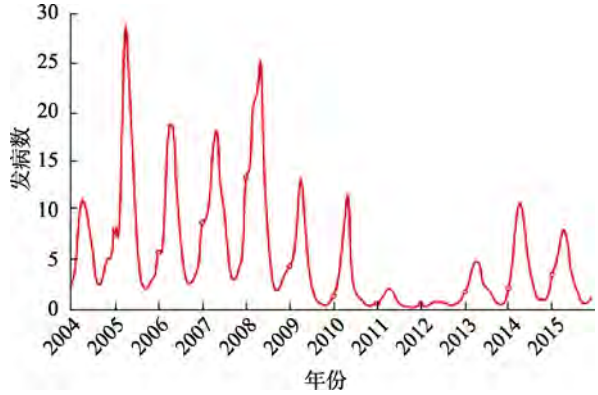


图 14　风疹时序折线图

通过观察流感发病数在时序平行坐标的波动情况，如图 16 所示，发现其时空分布在 2008 年和 2013 年 2 个时间点均出现大幅波动，结合时序折线图(图 17)和地区分布条形图(图 18b 和图 18d)分析，可以发现同比往年流感在 2009 年、2014 年出现 2 次大的疫潮，而热点地区的数量也随疾病的暴发大幅增加.

7　总　结

本文提出了基于多维时空可视化的传染病模式分析系统，针对传染病的时空变化趋势分别设计了传染病年度变化、季节消长的时间可视化分析视图、地区分布和层次聚类结果的空间视图；引入了 *t*-SNE 算法对原始传染病数据进行降维处理，利用时序平行坐标协同地区分布条形图进行时空协同聚类和异常值分析；通过交叉对比热力图帮助用户对比任意省份、疾病之间相似程度，检测它们之间是否存在聚类结果中的相关性，便于进一步分析聚类结果中的细节特征；同时利用空间层次聚类针对不同疾病选择不同层次的空间聚类结果进行可视化. 基于不同视图之间丰富的交互能够有效帮助用户协同进行传染病的时空模式，地区相似性分析以及异常值检测.

通过对国家公共卫生科学数据公布的 2004—2015 年间 39 种法定传染病数据的分析显示，

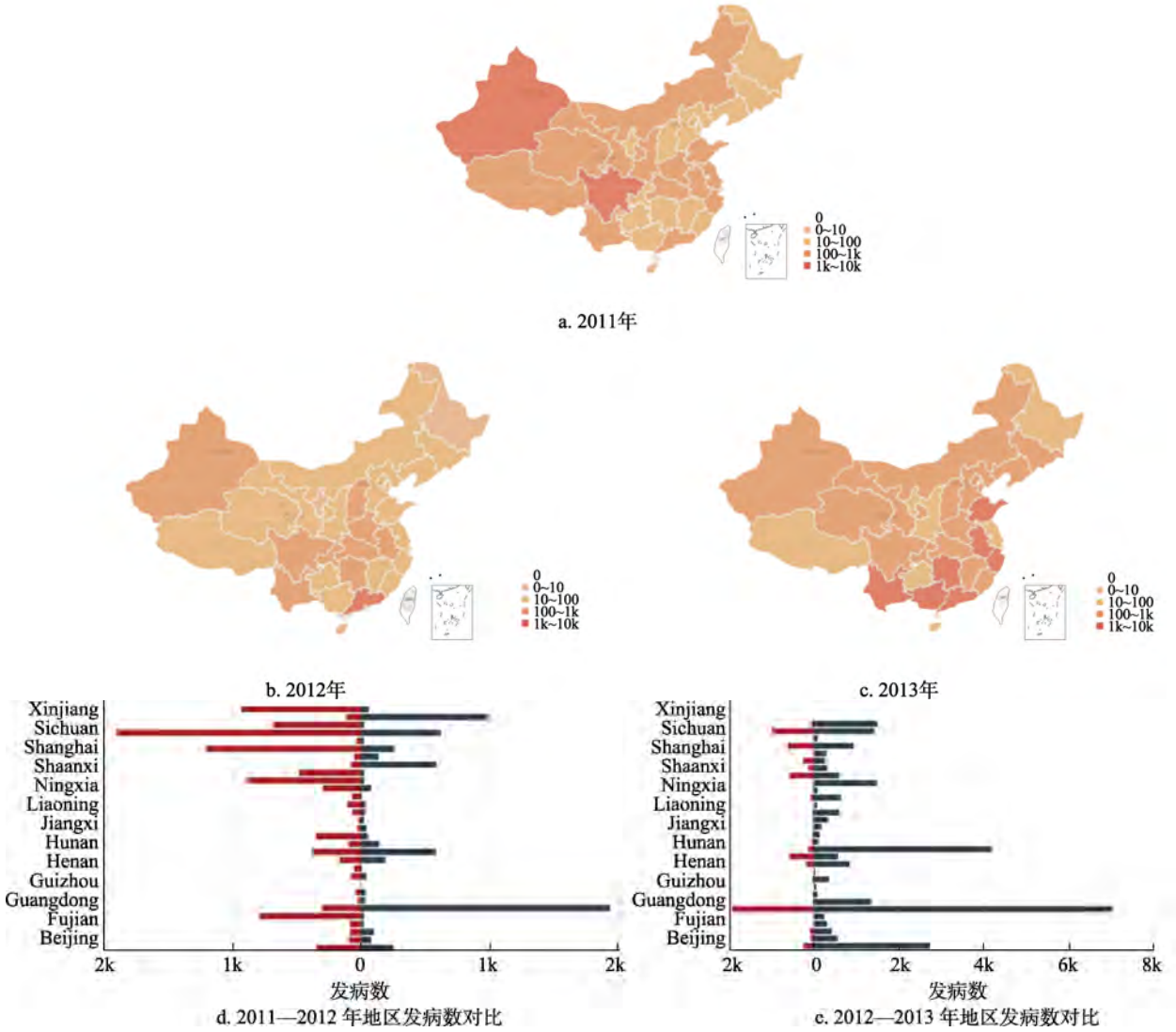


图 15　风疹 2011—2013 年空间分布转移

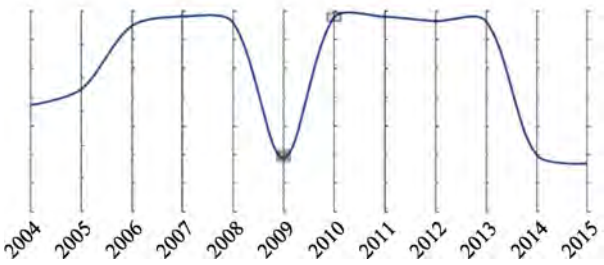


图 16　流行性感冒时序平行坐标

显本系统能够帮助用户有效、快速地分析多种传染病时空传染模式，并进一步查看发病细节特征，具有良好的实用性和扩展性．但通过总结实际应用中存在的问题发现，它仍存在以下不足之处：虽然本系统能够有效地帮助用户寻找具有相似发病模式的传染病和地区，但缺少对于聚类结果中特定类别特征的提取模型，只能在一定程度上对疫情实行监控，以直观的方式在疾病监测、资源分配等方面为卫生行政部门提供科学参考，缺少量化的预测模型．在未来的工作中，将进一步针对聚类结果、结合区域、疾病的特征信息建立特定发病模式的预测模型．

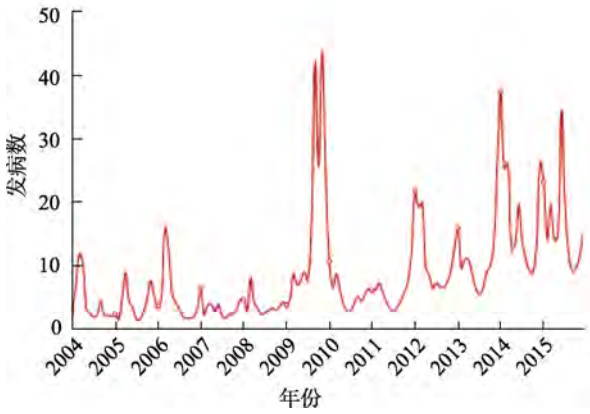


图 17　流行性感冒时序折线图

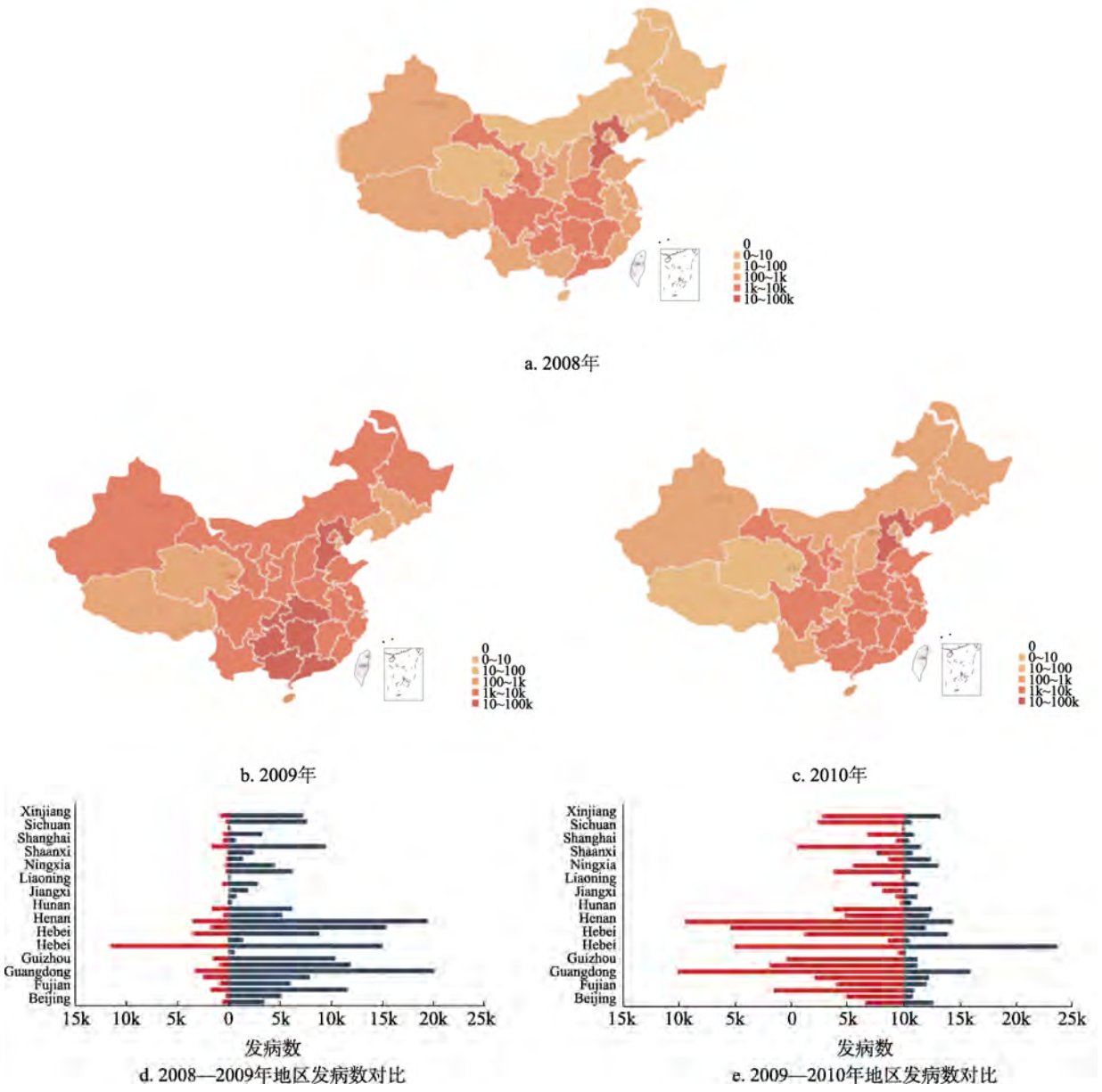


图 18 流行性感冒 2008—2010 年空间分布转移

参考文献(References):

[1] Hu Xueyun, He Zongyi, Miao Jing. Space-time clustering analysis and visualization based on the data of tuberculosis [J]. Bulletin of Surveying and Mapping, 2015(11): 106-111(in Chinese)
(胡雪芸, 何宗宜, 苗 静. 疾病数据的时空聚集分析及可视化[J]. 测绘通报, 2015(11): 106-111)

[2] Bie Qinqin, Qiu Dongsheng, Hu Hui, *et al*. Institute of geographical sciences and natural resources research[J]. Journal of Geo-Information Science, 2010, 12(3): 380-384(in Chinese)
(别芹芹, 邱冬生, 胡 辉, 等. 我国手足口病时空分布特征的 GIS 分析[J]. 地球信息科学学报, 2010, 12(3): 380-384)

[3] Keim D A, Hao M C, Dayal U. Hierarchical pixel bar charts[J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8(3): 255-269

[4] Zhou L, Weiskopf D. Indexed-points parallel coordinates visualization of multivariate correlations[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(6): 1997-2010

[5] Jolliffe I. International encyclopedia of statistical science [M]. Heidelberg: Springer, 2011: 1094-1096

[6] Borg I, Groenen P. Modern multidimensional scaling: theory and applications[J]. Journal of Educational Measurement, 2003, 40(3): 277-280

[7] Yuan X R, Ren D H, Wang Z C, *et al*. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data[J]. IEEE Transactions on Visualization and Computer Graphics, 2013 (12): 2625-2633

[8] Cao N, Lin Y R, Gotz D, *et al*. Z-Glyph: Visualizing outliers in multivariate data[J]. Information Visualization, 2018, 17(1): 22-40

[9] Liao H S, Wu Y C, Chen L, *et al*. Cluster-based visual abstraction for multivariate scatterplots[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(9): 2531-2545

[10] Kwan M P. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set[J]. Transportation Research Part C: Emerging Technologies, 2000, 8(1-6): 185-203

[11] Guo D S, Chen J, MacEachren A M, *et al*. A visualization system for space-time and multivariate patterns (vis-stamp)[J]. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(6): 1461-1474

[12] Zhou Zhiguang, Sun Chang, Le Dandan, *et al*. Collaborative visual analytics of multi-dimensional and spatio-temporal data[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(12): 2245-2255(in Chinese)
(周志光, 孙　畅, 乐丹丹, 等. 多维时空数据协同可视分析方法[J]. 计算机辅助设计与图形学学报, 2017, 29(12): 2245-2255)

[13] Plaisant C, Mushlin R, Snyder A, *et al*. The craft of information visualization[M]. San Francisco: Morgan Kaufmann, 2003: 308-312

[14] Rind A, Aigner W, Miksch S, *et al*. Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation[C] //Proceedings of Symposium of the Austrian HCI and Usability Engineering Group. Heidelberg: Springer, 2011: 301-320

[15] Fails J A, Karlson A, Shahamat L, *et al*. A visual interface for multivariate temporal data: finding patterns of events across multiple histories[C] //Proceedings of the IEEE Symposium on Visual Analytics Science And Technology. Los Alamitos: IEEE Computer Society Press, 2006: 167-174

[16] Wang T D, Plaisant C, Shneiderman B, *et al*. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison[J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1049-1056

[17] Wongsuphasawat K, Guerra Gómez J A, Plaisant C, *et al*. LifeFlow: visualizing an overview of event sequences[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2011: 1747-1756

[18] Qu Yubing, Li Rendong, Zhuang Dafang, *et al*. Application of space information technology and big data technology in dengue fever risk assessment system[J]. Chinese Journal of Disease Control & Prevention, 2017, 21(11): 1165-1169+1174 (in Chinese)
(曲玉冰, 李仁东, 庄大方, 等. 空间信息技术和大数据技术在登革热风险评估系统中的应用[J]. 中华疾病控制杂志, 2017, 21(11): 1165-1169+1174)

[19] van der Maaten L, Hinton G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605