陕西师範大學
SHAANXI NORMAL UNIVERSITY

大数据下的游客行为研究进展

李君轶　教授/博士

陕西师范大学旅游与环境学院

陕西省旅游信息科学重点实验室 主任

大数据带来了一场全方位的思维变革、产业变革和管理变革，它把我们的旅游科学研究、旅游实践应用的各个领域和环节都推到了一个前所未有的“大数据时代”。

- 思维变革
- 产业变革
- 管理变革

旅游科学研究

旅游实践应用

旅游者行为研究占据了现在首要的研究主题，取代了十年前的旅游吸引物/资源/产品开发和管理。(黄松山，陈钢华，2016)

挖掘旅游者行为和城市偏好已经成为近期地理信息系统研究的热门。（Zhou, X., Xu, C，2015）

Huang, S., & Chen, G. (2016). Current state of tourism research in china. *Tourism Management Perspectives*, 20, 10-18.
Zhou, X., Xu, C., & Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. Computers Environment & Urban Systems, 54, 144-153.

基于大数据的游客时空行为研究取得了众多典型的成果。几个特征：

第一是国外多，国内少。

第二是数据来源多样，包括手机信令数据、社交媒体数据（具体包括社交网络（facebook，人人网）、微博（twitter）数据、意见和评论（豆瓣，大众点评）、内容分享（youTube，Flickr，优酷等），社会化标签（Delicious），博客，BBS，论坛）、GPS数据和交通一卡通数据等。

第三是处理方式更具有创新性与技术性。

第四究研内容以**时空分布特征、旅游流空间网络结构、行为足迹模式**等为核心。

一、大数据相关研究

1.1解析大数据

Volume（数据规模大）

非结构化数据的超大规模和增长

- 占总数据量的80~90%
- 比结构化数据增长快10倍到50倍
- 是传统数据仓库的10到50倍

Velocity（数据要求处理速度快）

实时分析

- 数据输入、处理与丢掉
- 立竿见影而非事后见效

Variety（数据种类多）

大数据的异构和多样性

- 很多不同形式（文本、图像、视频、机器数据）
- 无模式或者模式不明显
- 不连贯的语法或句式

Value（数据价值密度低）

大量的不相干信息

- 对未来趋势与模式的可预测分析
- 深度复杂分析（机器学习、人工智能等）

大数据与传统数据的差异

大数据 VS 传统数据

大数据与传统数据的区别主要表现在**数据来源、数据反应速度、数据全面性**等三个方面。

Table 1. Data capture techniques with their main strength and weakness in the context of tourism and urbanism studies.

Data capture	Strength	Weakness	Example of application
Land use and census data	Applicable to many scales and over long time period	Infrastructure and service based, static view of urban dynamics	Estimate the tourism intensity of an area
Manual surveys	Capture high-level information such as motivations and reasons for stay in specific areas	Very costly and applies to limited time periods	Capture the motivation for visiting and length of stay
Near-field communication	Precise real-time mobility data	Costly Infrastructure deployment	Describe the social and spatial characteristics of space (Kostakos et al., 2008)
GPS logs	Precise mobility data	Does not scale well if deployed for the purpose of a survey alone. limited in time and participants	Cluster tourist routes (Asakura and Iryob, 2007)
Cellphone (device-based)	Timely mobility data, potentially augmented with in-situ survey	Does not scale well if deployed for the purpose of a survey alone. limited in time and participants	Context-Aware Experience Sampling to capture the experience in-situ. (Froehlich et al. 2006)
Cellphone (aggregated network-based)	Use existing infrastructure to provide real-time density and mobility data, covering multiple geographic scales (neighbourhood, city, country)	Reveal large-scale phenomena but do not explain the reasons	Real-time traffic detection (Yim, 2007)
User-generated content	Exploit publicly available data with no need for deployment or pre-existing infrastructure	Credibility of information and no systematic coverage	Reveal flows of photographers (Girardin et al. 2008)

Girardin, F. et al. "Digital Footprinting: Uncovering Tourists with User-Generated Content." Pervasive Computing, IEEE 7.4 (2008): 36-43. © 2008 Institute of Electrical and Electronics Engineers.

Girardin, F. et al. "Digital Footprinting: Uncovering Tourists with User-Generated Content." Pervasive Computing, IEEE 7.4 (2008): 36-43. © 2008 Institute of Electrical and Electronics Engineers.

数据源	优势	劣势	应用举例
土地利用和人口数据	可以运用到长时期的多种量表	基础设施和服务方面，城市动力学的静态视角	一个区域的旅游强度估算
人工调查数据	收集高级别的信息，例如停留在特定区域的动机和原因	花费大，只能运用到有限的时间阶段	游览和停留时间的动机
近距离无线通讯	精确的实时移动数据	基础设施部署贵	描述空间的社会特征和时空特征
GPS 日志	精确的移动数据	当运用到单纯的研究目的时不能很好的量化，时间和参与者限制	旅游线路集群（Asakura and Iryob,2007)
手机信令数据（基于设备）	及时的移动数据，有可能扩大到现场调查	不能很好的量化，时间和参与者受到限制	通过情景感知体验抽样调查来获取现场体验（Froehlich et al,2006)
手机信令数据（基于聚合网络）	提供实时密度和移动数据，多种地理尺度（附近，城市，乡村）	解释大规模的现象但不能解释原因	实时的交通检测（Yim,2007)
用户自生成内容	公开的开发可利用的数据，不需要部署或之前存在的基础设施	信息的可信度，没有系统覆盖范围	解释摄影者的流动（Girardin et al,2008)

1.2 相关研究

（1）技术层面的研究——数据标准化，数据存储、可视化

（2）数据分析层面——数据分析算法、数据建模、关联分析

（3）情景模拟与预测——情景模拟、预测模型

（4）管理和应用层面——舆情管理、行为预测、网络营销

- 2012年5月，香山科学会议组织了以“大数据科学与工程——一门新兴的交叉学科”为主题的会议，深入讨论了大数据的理论与工程数据研究和应用方向，指出目前最关注的是**大数据分析算法和大数据系统效率。**

- 2012年6月，中国计算机学会举办了“大数据时代，智谋未来”学术报告会，就大数据时代的**数据挖掘、体系架构理论、大数据安全、大数据平台开发与大数据现实案例**等内容进行了全面的讨论。学者们也从**大数据的概念与原理、挖掘技术、实际应用与隐私保护等角度**进行了研究

二、大数据游客行为研究

2. 1研究内容

1. 旅游网络结构研究（旅游流空间）
2. 旅游热点（兴趣点）发现
3. 游客空间行为建模与预测
4. 旅游客流监测与调控

2.1.1 城市(景区)网络空间结构

利用手机、社交媒体数据进行旅游流网络结构的研究，目前已经比较成熟。

Novel methods and tools are being developed to explore the significance of the new types of user-related spatio-temporal data. This approach helps uncover the presence and movements of tourists from **cell phone network data** and the **georeferenced photos** they generate.

经典文献：

Girardin, F. et al. “Digital Footprinting: Uncovering Tourists with User-Generated Content.” Pervasive Computing, IEEE 7.4 (2008): 36-43. © 2008 Institute of Electrical and Electronics Engineers.

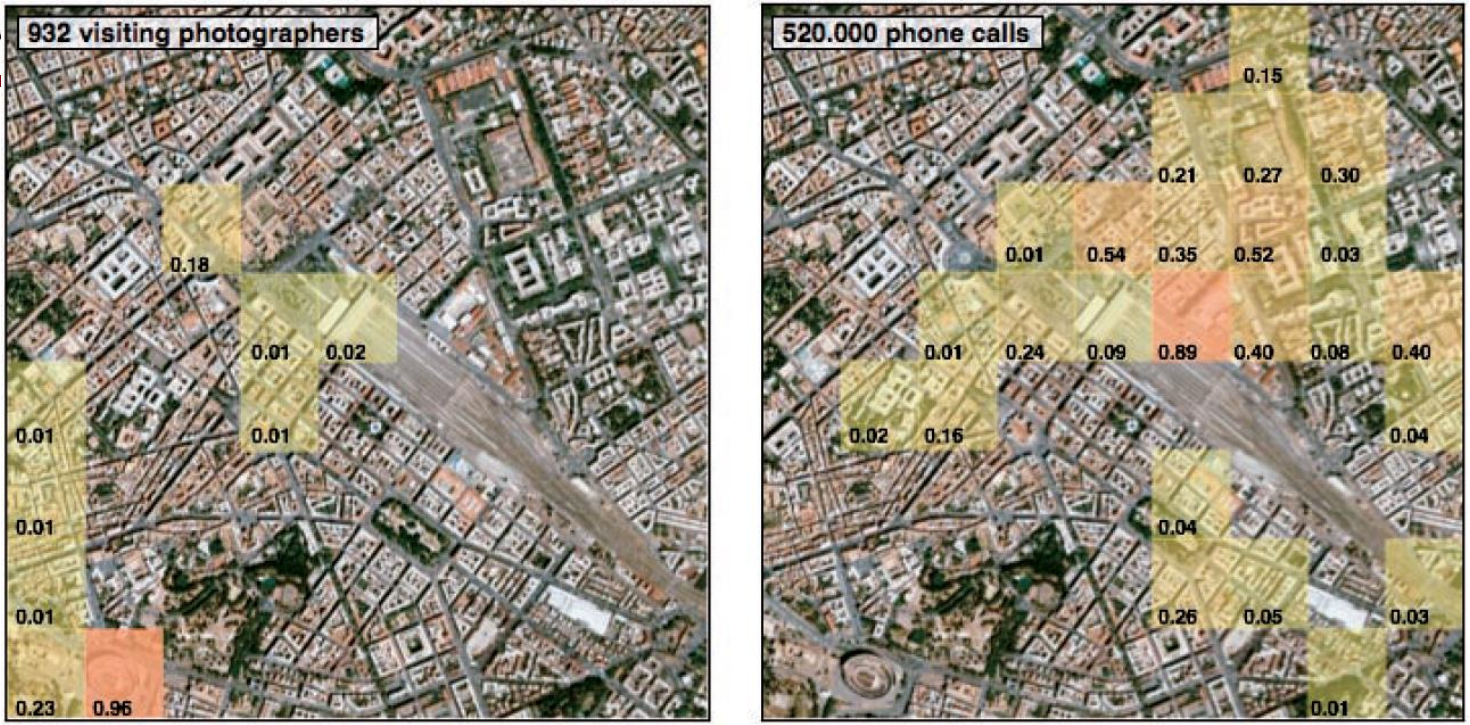
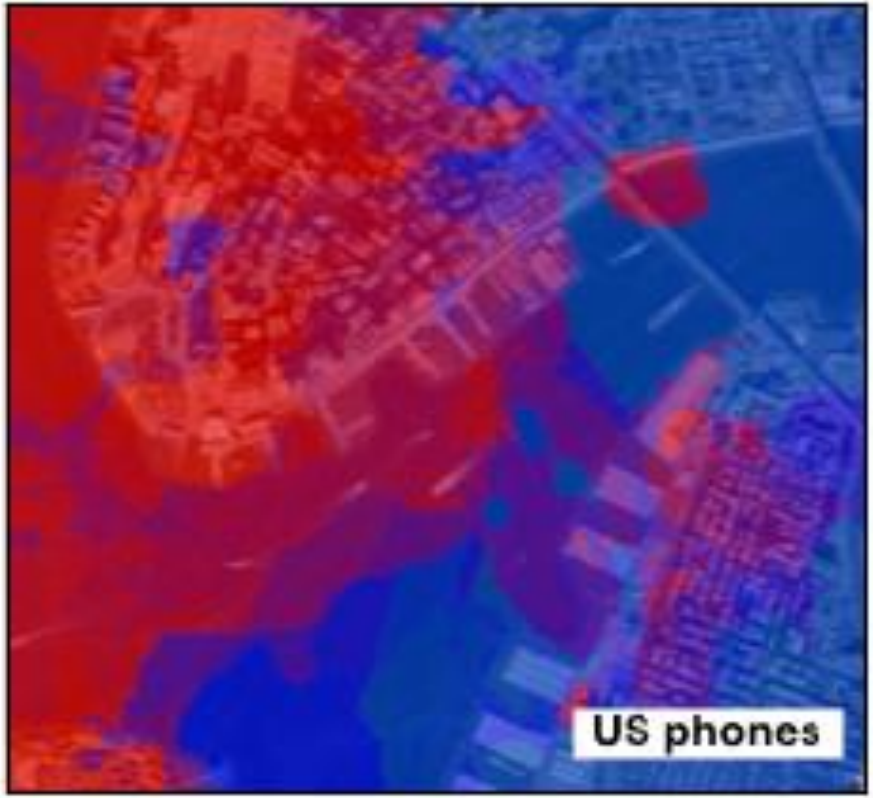


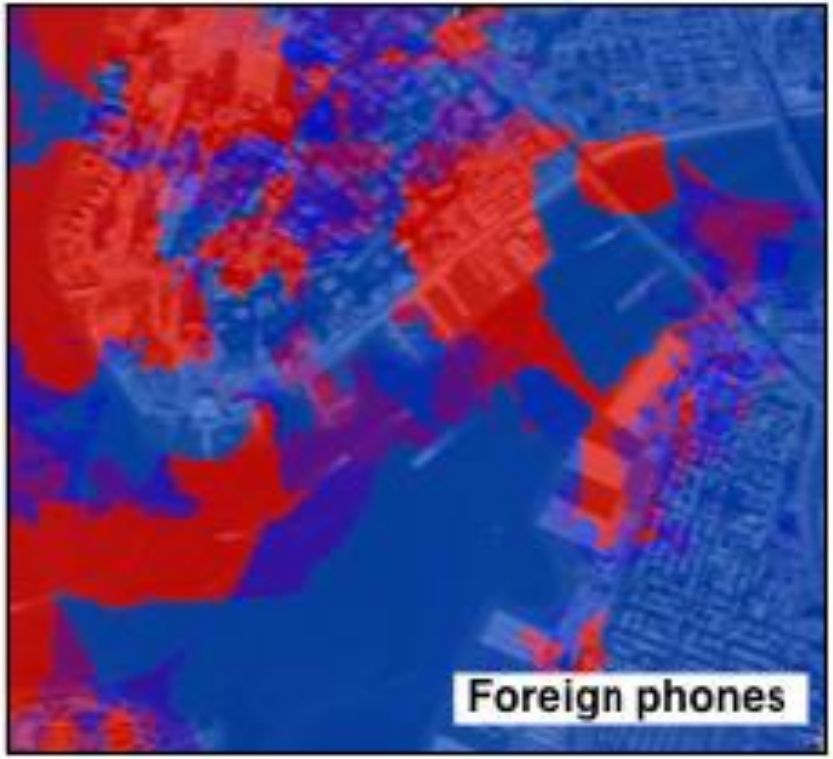
Figure . Geovisualizations of the presence of (left) 932 tourist photographers and (right) **520,000** phone calls from foreign mobile phones in the Coliseum and Piazza della Repubblica area from September to November 2006. Both types of data cover the train station area in the proximity of the Piazza della Republica. The values in each cell are normalized.

Week days

Weekend


US phones


US phones


Foreign phones

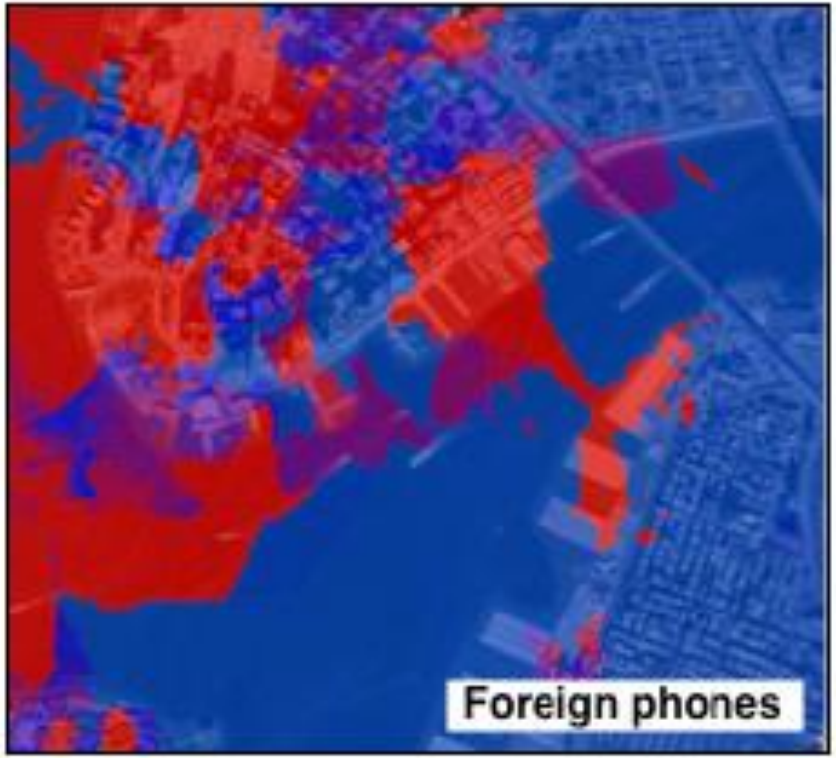
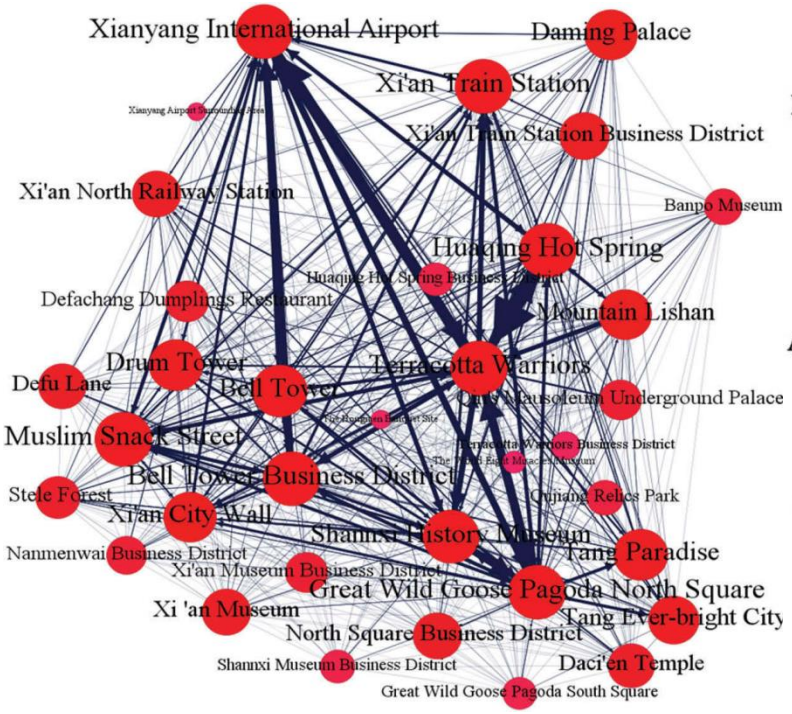
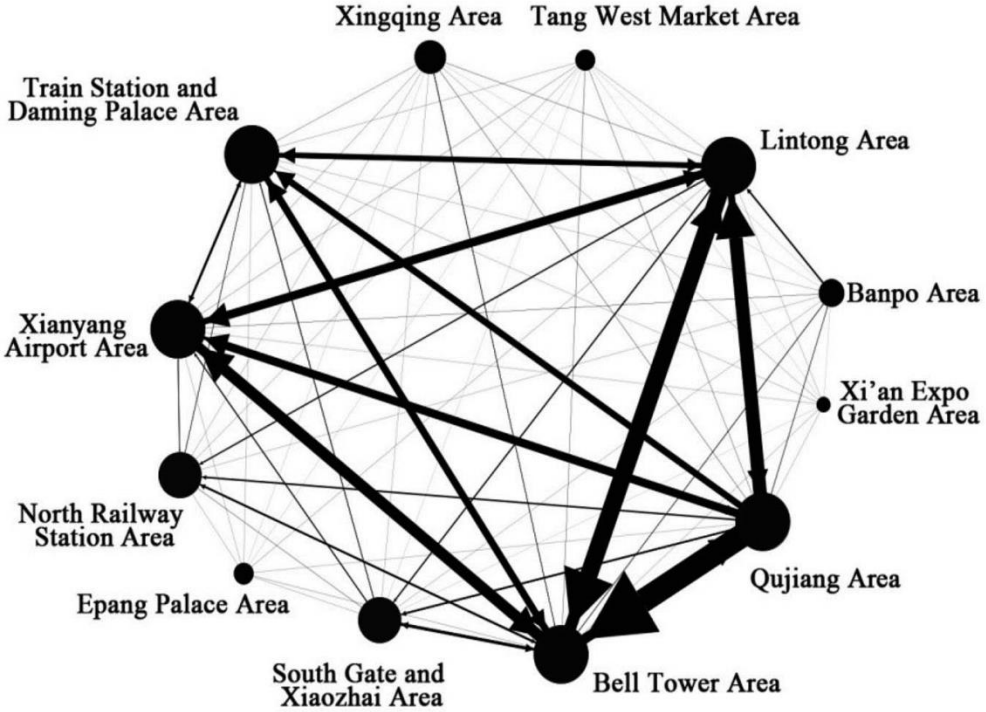

Foreign phones

Figure 5. Spatial distribution of photographers (in yellow) and photos (in red) in Summer 2006, 2007, and 2008 in Lower Manhattan and West Brooklyn.



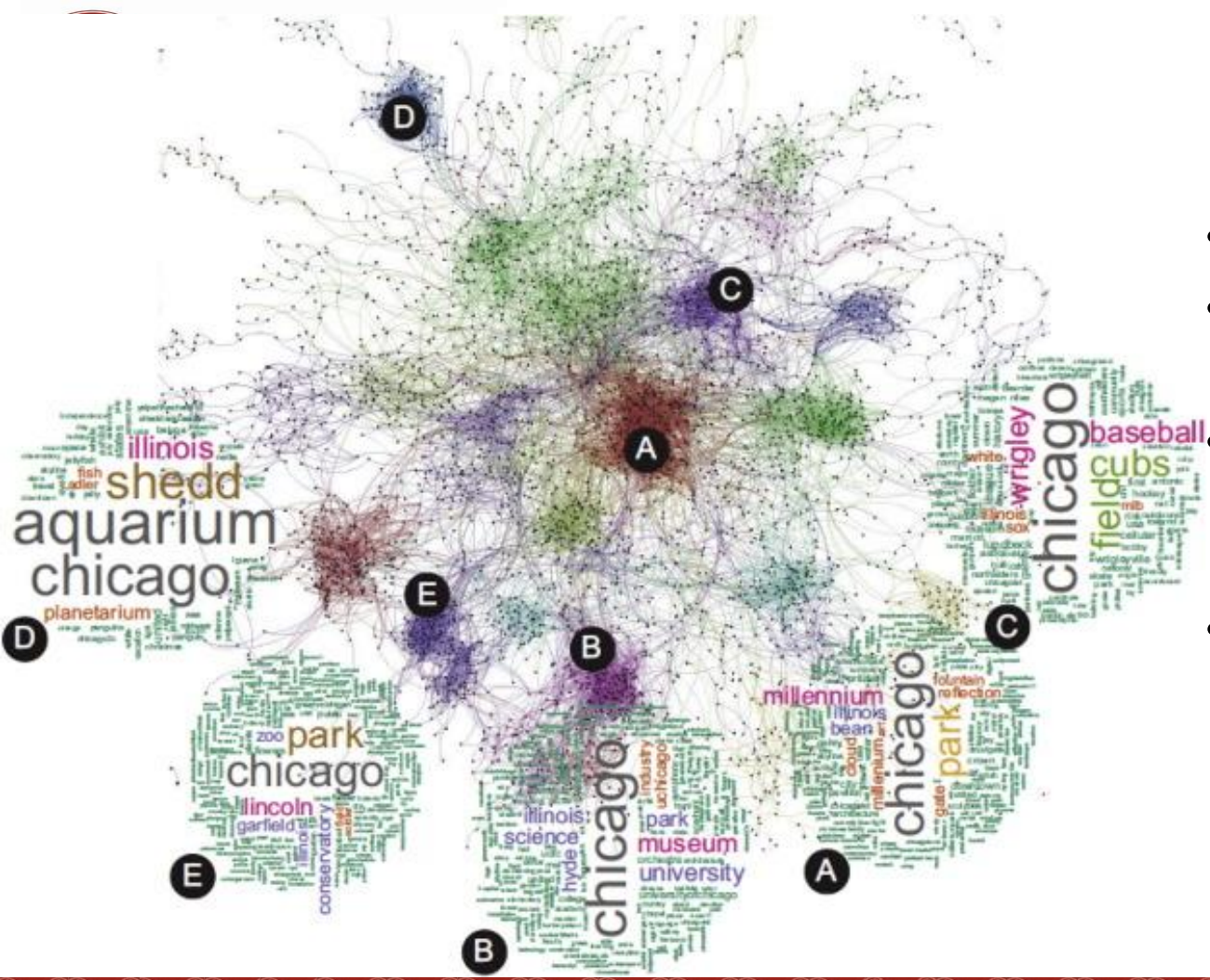


Jia Tang, Junyi Li (2016): Spatial network of urban tourist flow in Xi'an based on microblog big data, Journal of China Tourism Research, DOI:10.1080/19388160.2016.1165780

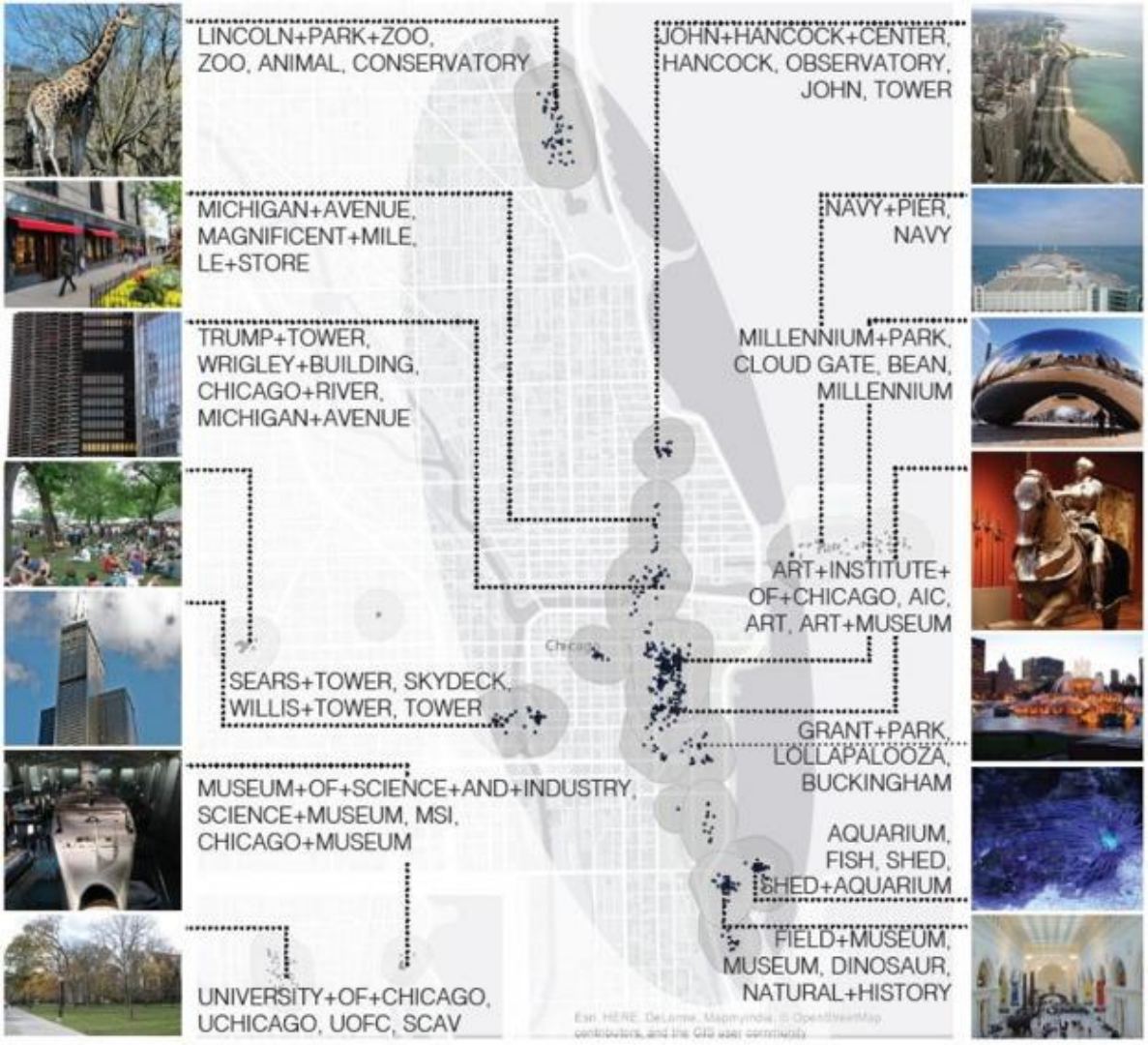
2.1.2 旅游热点发现

- 通过城市边界进行地理标签图像（基于Flickr ）搜集。收集到的数据文本标签、地理坐标位置、图片的时间标记。数据一收集便将标签识别并提取出来。收集到的与旅游目的地有关的标签都被放到一起进行标签特征计算。对不同旅游目的地根据其受欢迎程度进行排序。发现了通过与事件相关的照片中分离与旅游者相关的照片，照片的时间特征和地理特征信息是非常有用的。（Xiaolu Zhou, Chen Xu, Brandon Kimmons，2015）

Zhou, X., Xu, C., & Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers Environment & Urban Systems*, 54, 144-153.



- 芝加哥的标签社区
- （Tag communities in Chicago）.
- （相应的字符代表了检测到的社区的字母云）
- The corresponding characters represent word clouds for the detected communities.



- 在芝加哥的辨别旅游目的地空间分类（Space clusters of the identified tourist destinations in Chicago）. **Note:** Map of University of Chicago and Museum of Science and Industry are in the inset map because these two places are far away from the main attractions and are difficult to fit into the figure.

文章编号:1003-2398(2015)03--0057-07 DOI: 10.13959/j.issn.1003-2398.2015.03.009

基于微博的西安市居民夜间活动时空分布研究

陈宏飞[1], 李君轶[1], 秦 超[1], 刘 广[2], 孙九林[1,3]

(1.陕西师范大学 旅游与环境学院, 西安 710062; 2. 陕西师范大学 网络信息中心, 西安 710062;
3. 中国科学院 地理科学与资源研究所, 北京 100101)

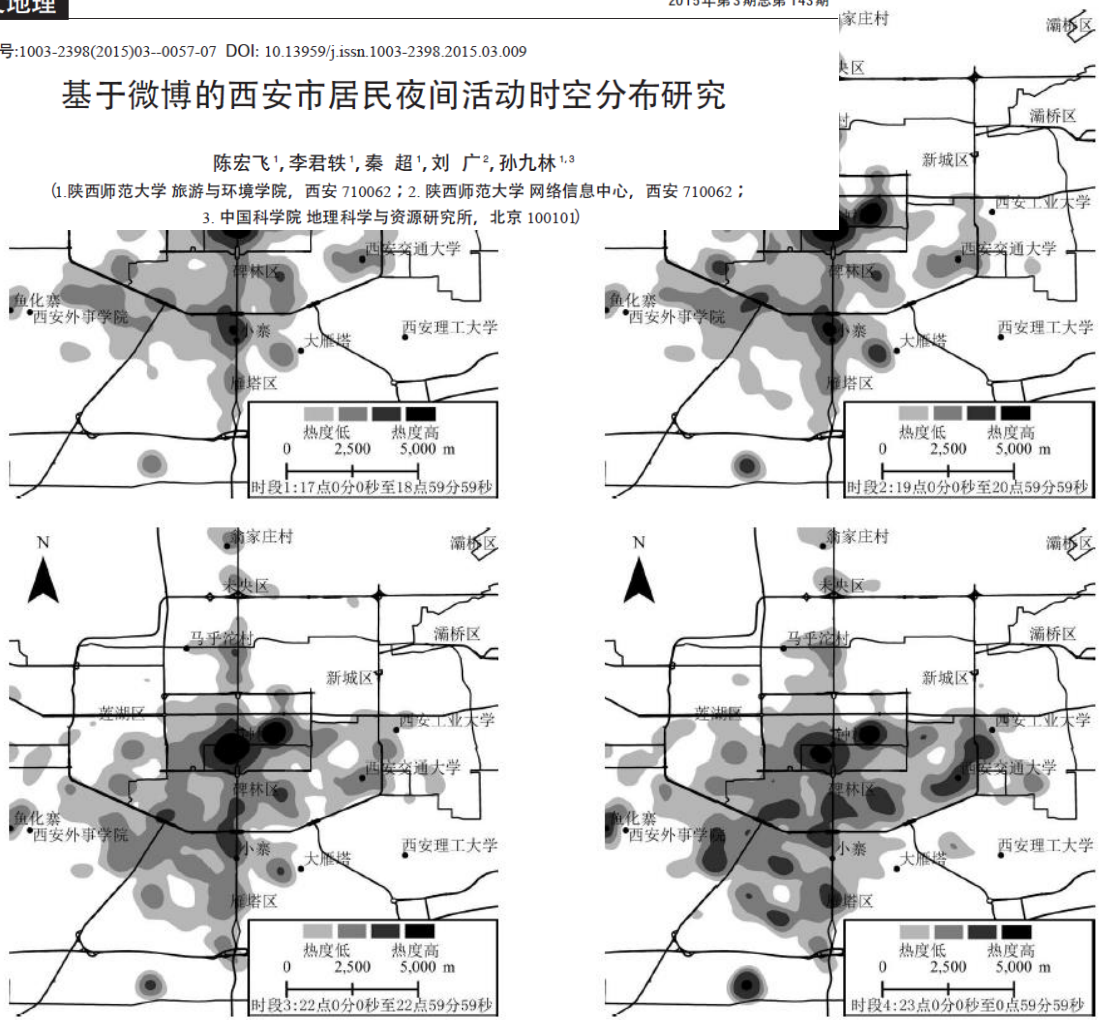
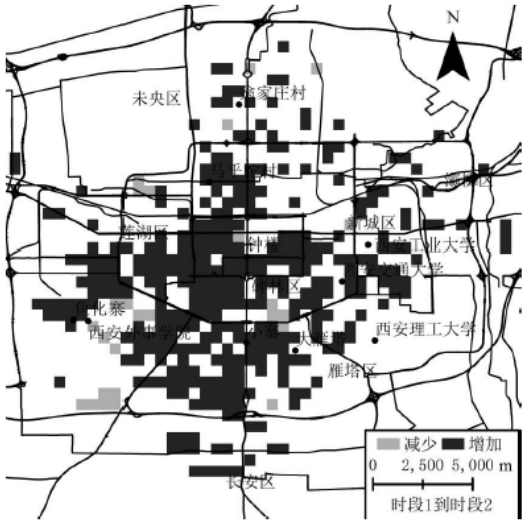
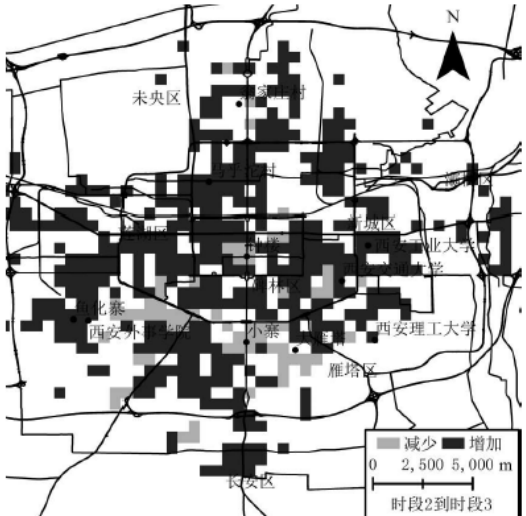


图2 分时段微博发布用户空间分布热度图

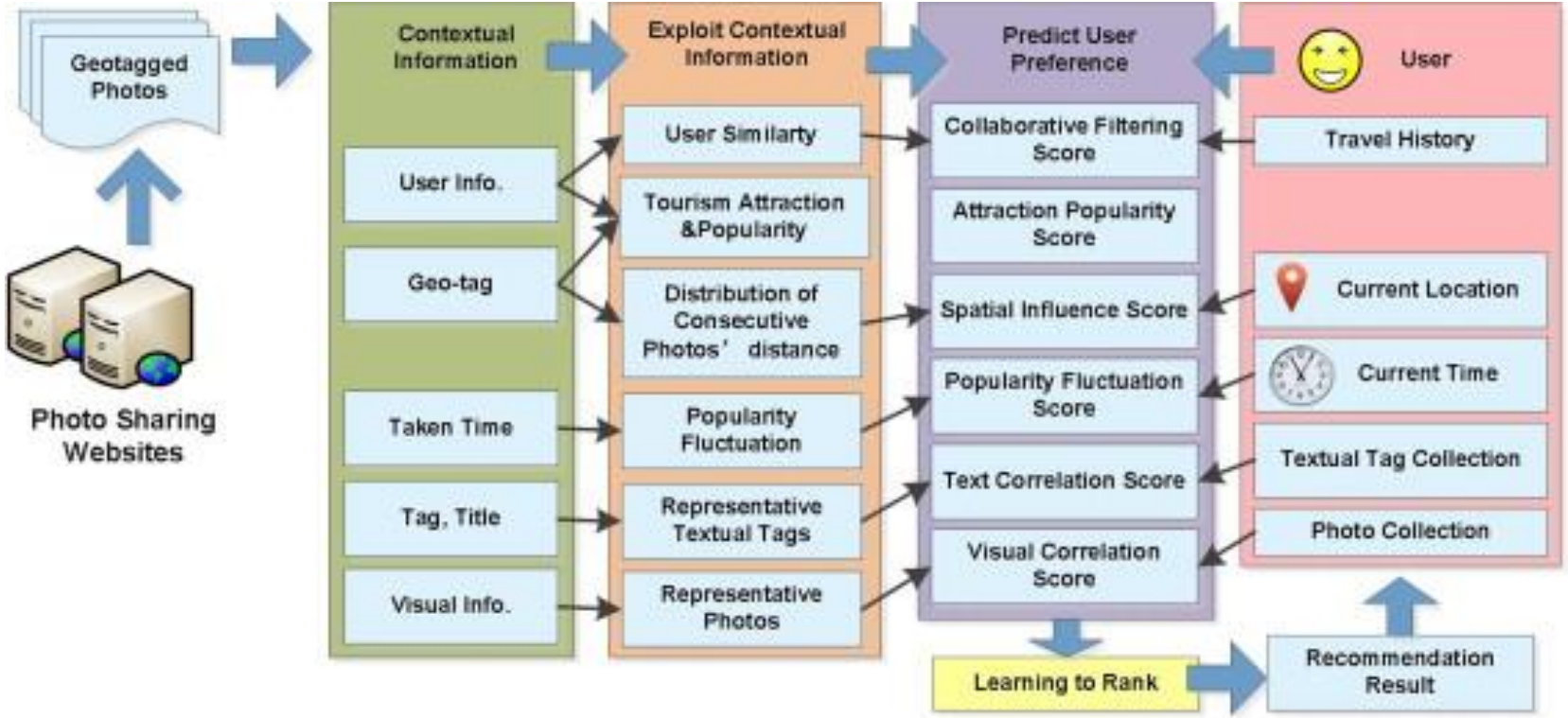




- 提出了利用带有地理标签的网络照片的不同上下文信息，然后通过学习排序来结合各种信息呈现出个性化旅游景点推荐。

Jiang, K., Yin, H., Wang, P., & Yu, N. (2013). Learning from contextual information of geo-tagged web photos to rank personalized tourism attractions. *Neurocomputing, 119*(16), 17-25.

- Algorithm overview

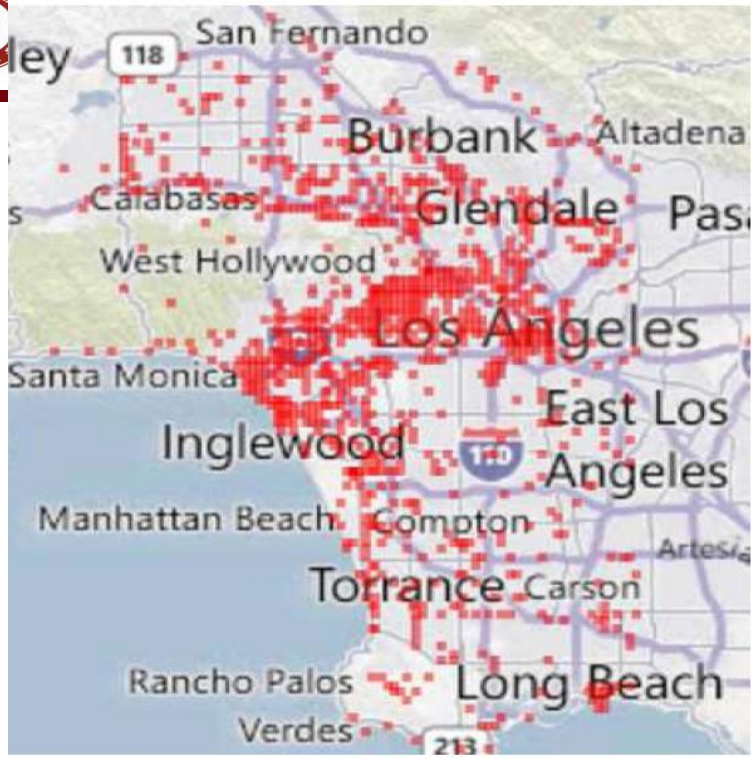


Panoramio dataset

3. 游客空间行为建模与预测

- 基于游客的空间轨迹，利用数学方法建模（常用马尔科夫链）并进行游客的空间行为预测，实现基于位置的服务和推荐（LBS）

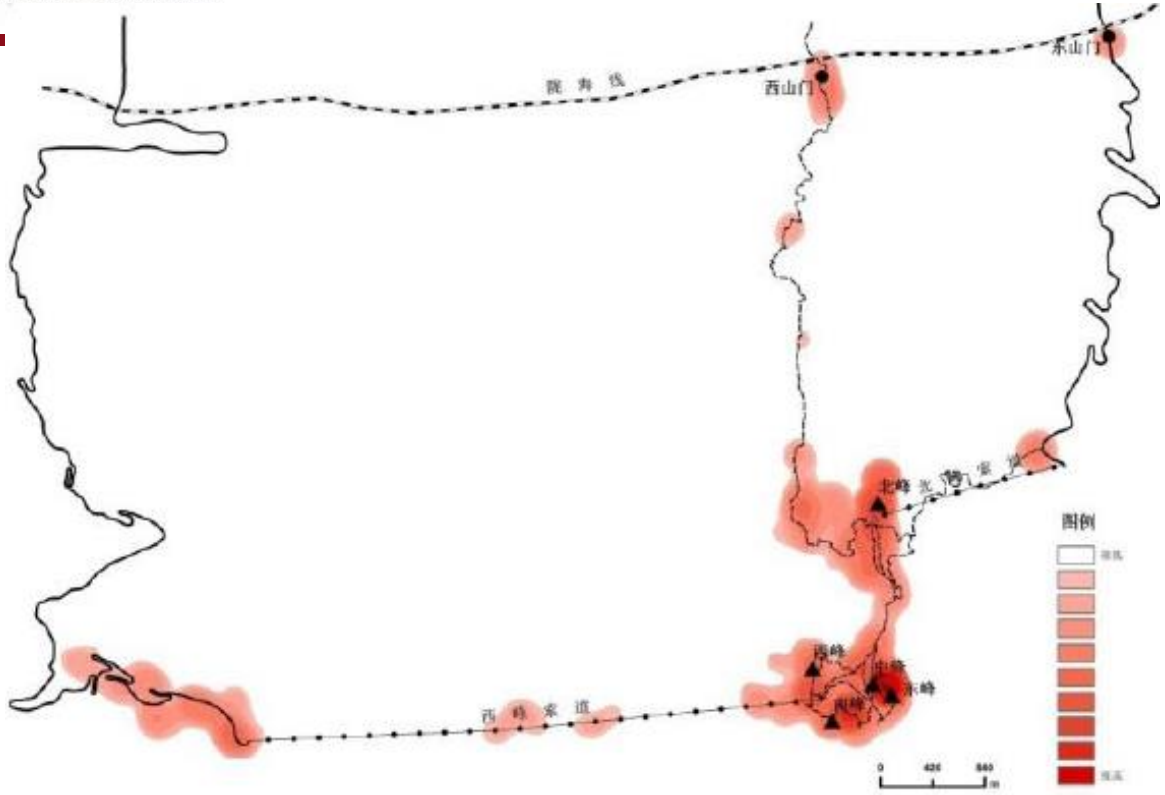
Jie Bao, Yu Zheng,Mohamed F. Mokbel. Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data，CONFERENCE PAPER ·NOVEMBER 2012 DOI: 10.1145/2424321.2424348



(a) New York users in Los Angels

(b) New York users in New York City.

Figure 1: User Location History Distributions.



（b）热点区域分布

图 4-10 微博发布点分布和热点区域空间分布

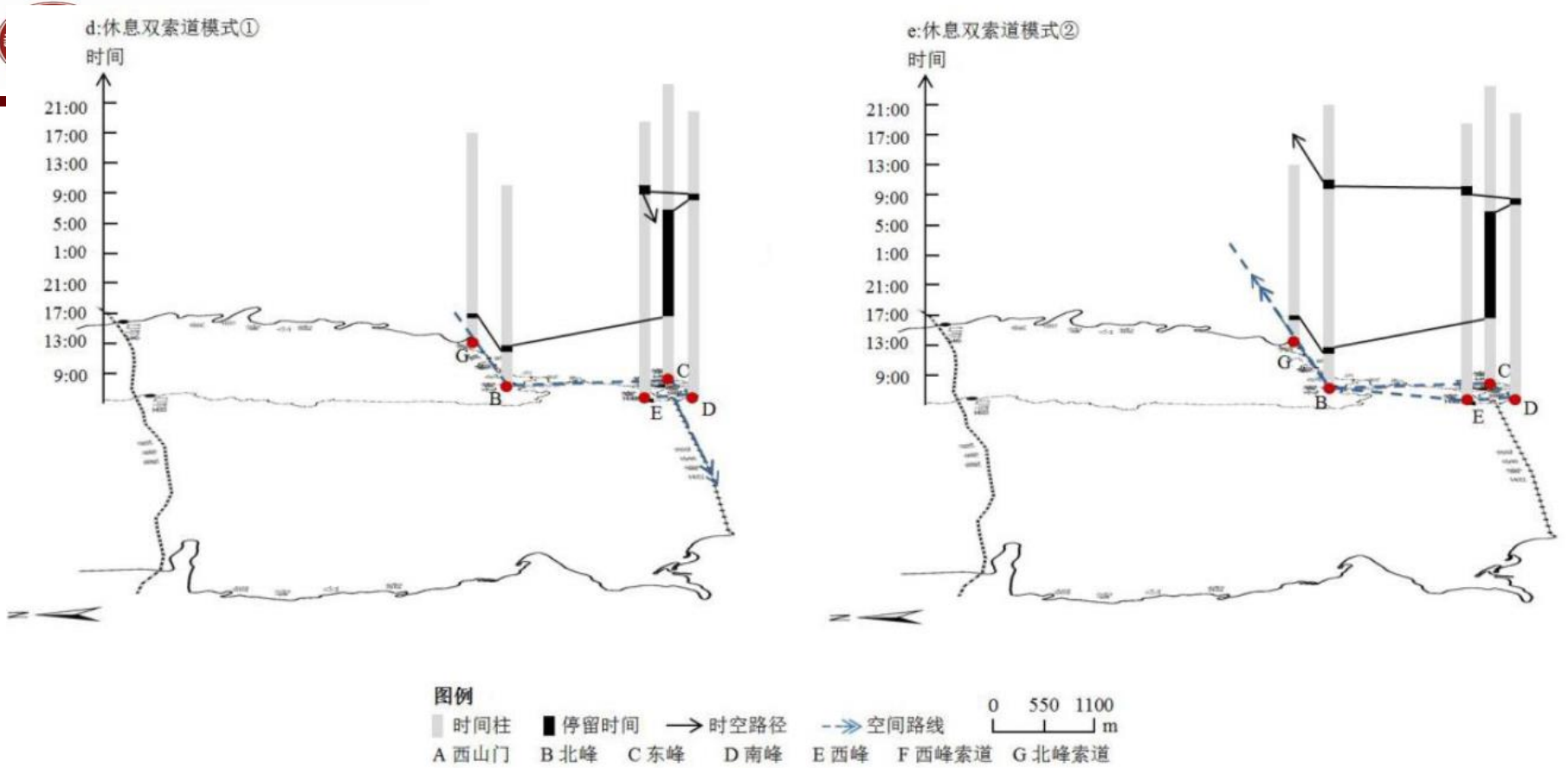


图 5-2 休息双索道模式示意图

Fig.5-2 Asleep double cableway pattern

2.1.4 旅游流时空模拟与预测

- 主要是利用实时数据（LBS, 手机信令）等模拟旅游流。并对其进行预测。

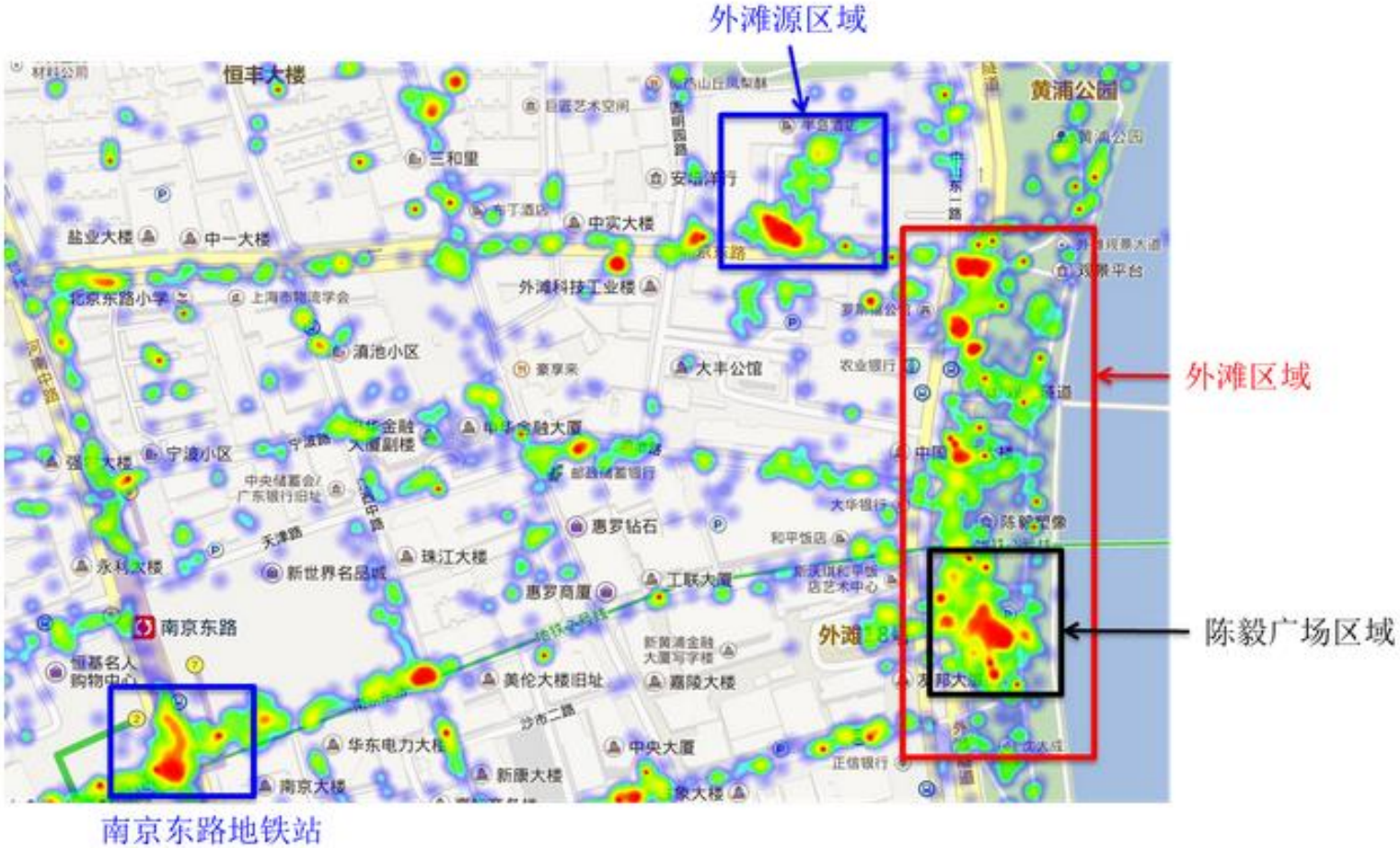
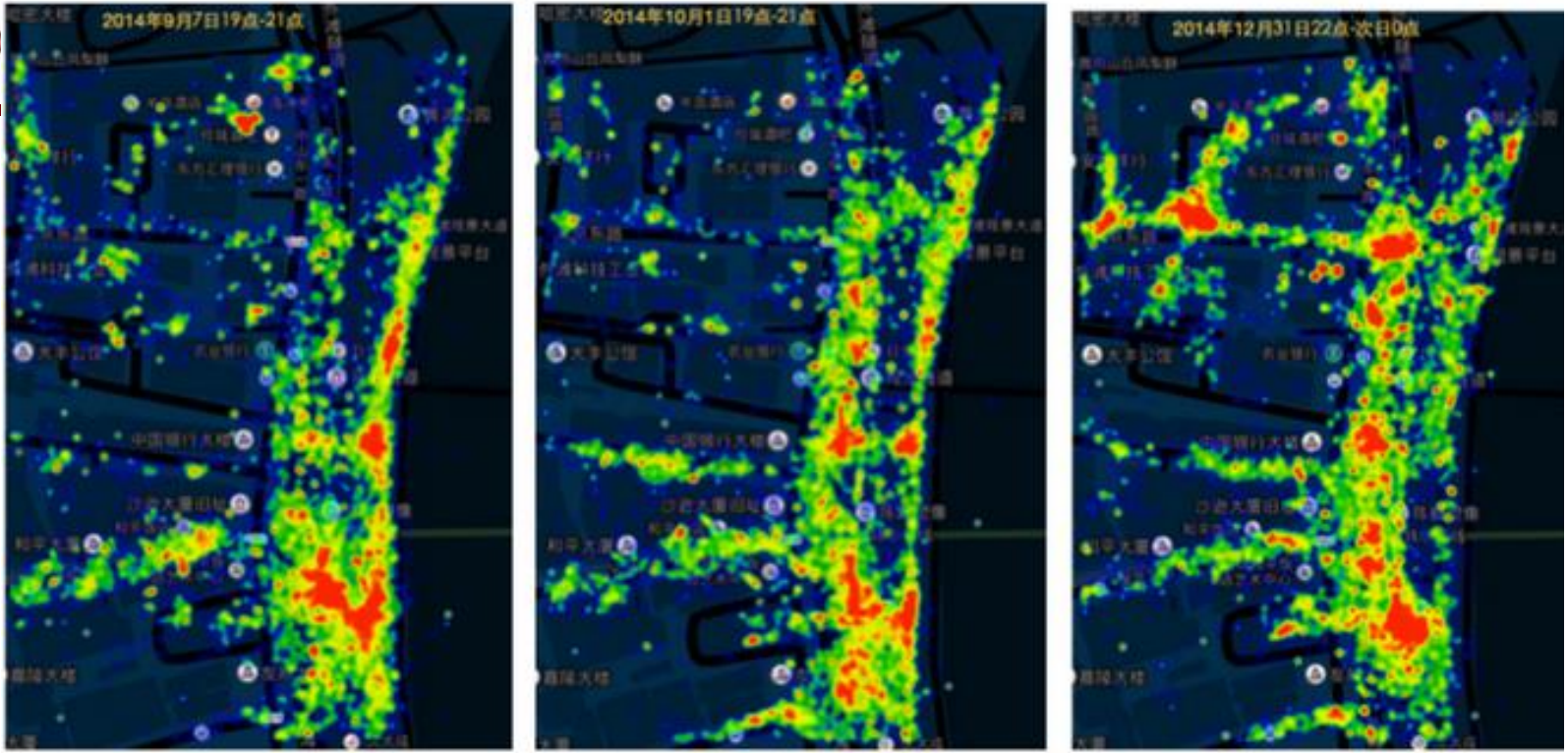
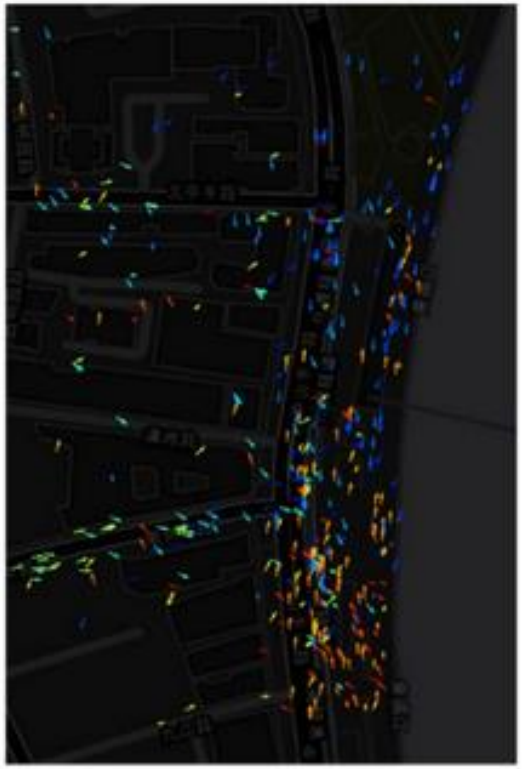


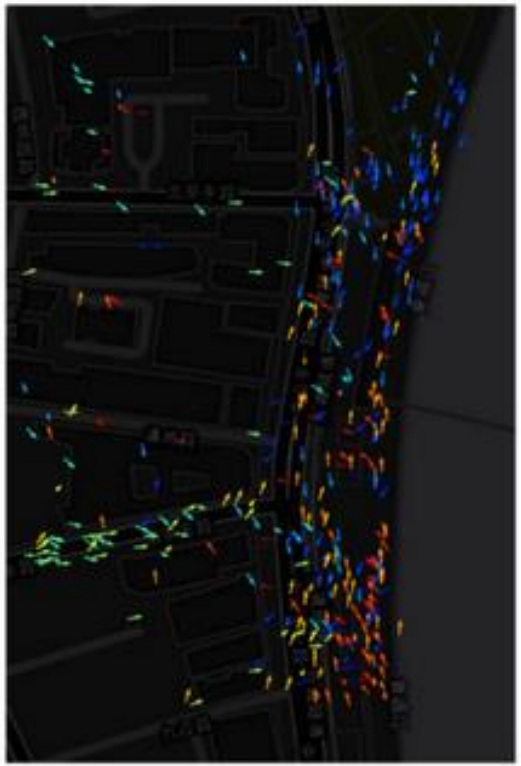
图1 2014.12.31 事发时外滩区域人群热力图

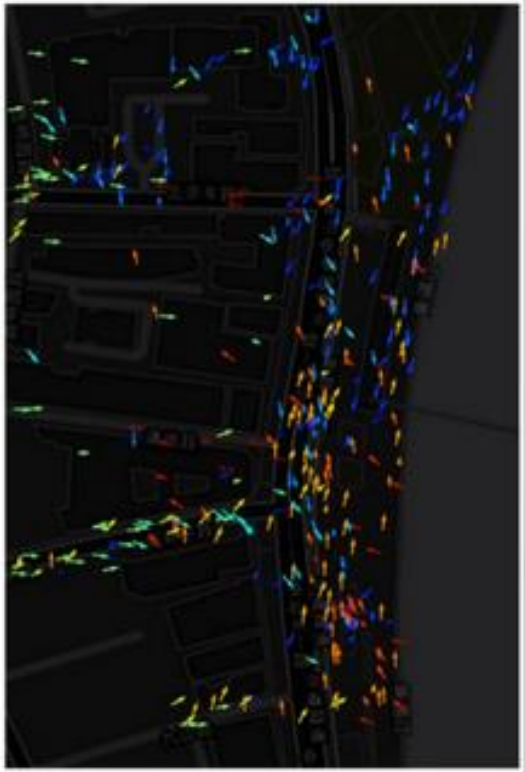


（1）中秋前夜　　（2）国庆当晚　　（3）跨年当晚

图4 外滩和外滩源区域人群分布热力图（2小时）







（1）中秋前夜　　（2）国庆当晚　　（3）跨年当晚

图5 外滩和外滩源区域人群流动方向示意图（部分采样）

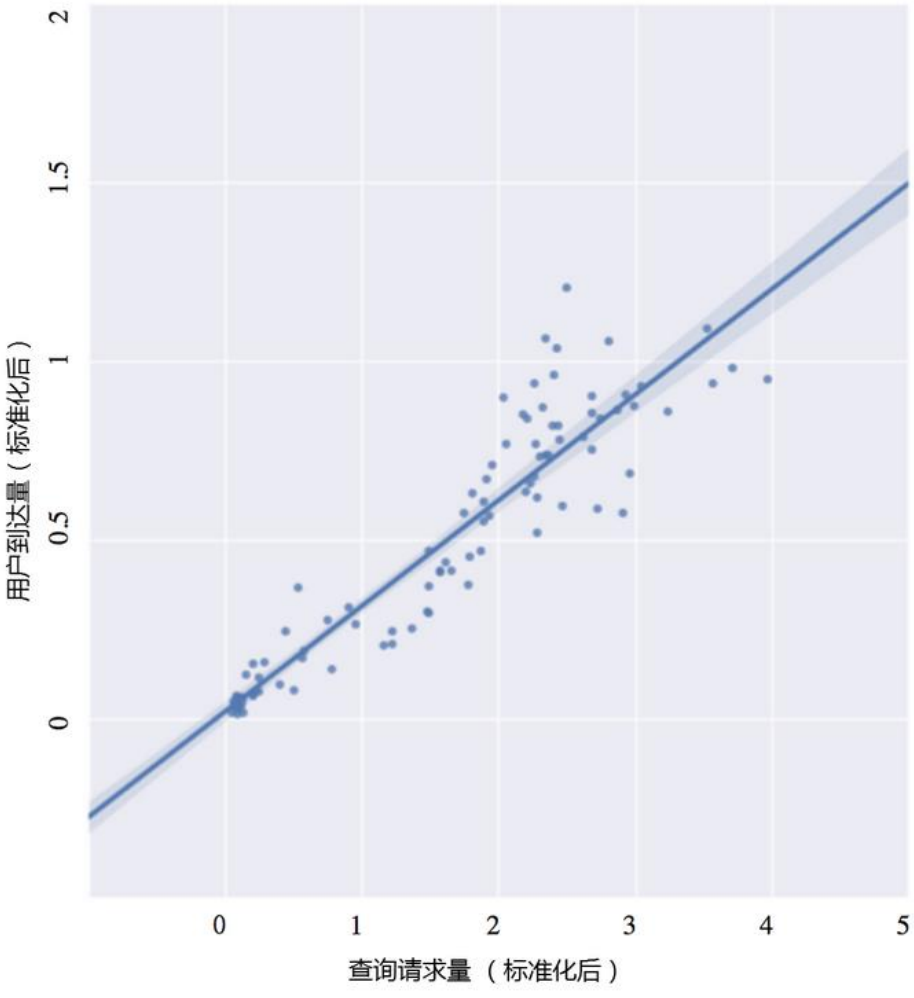
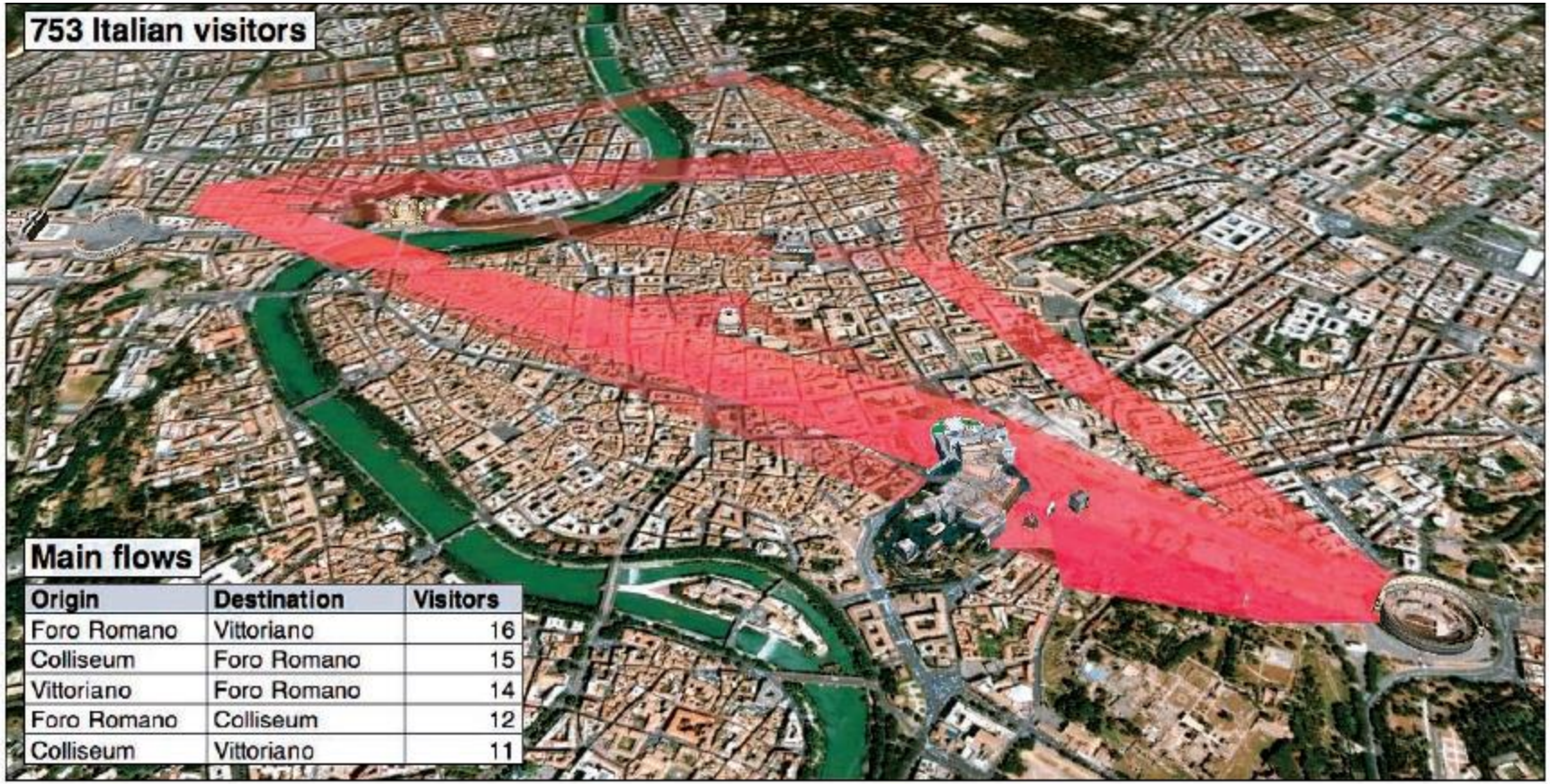
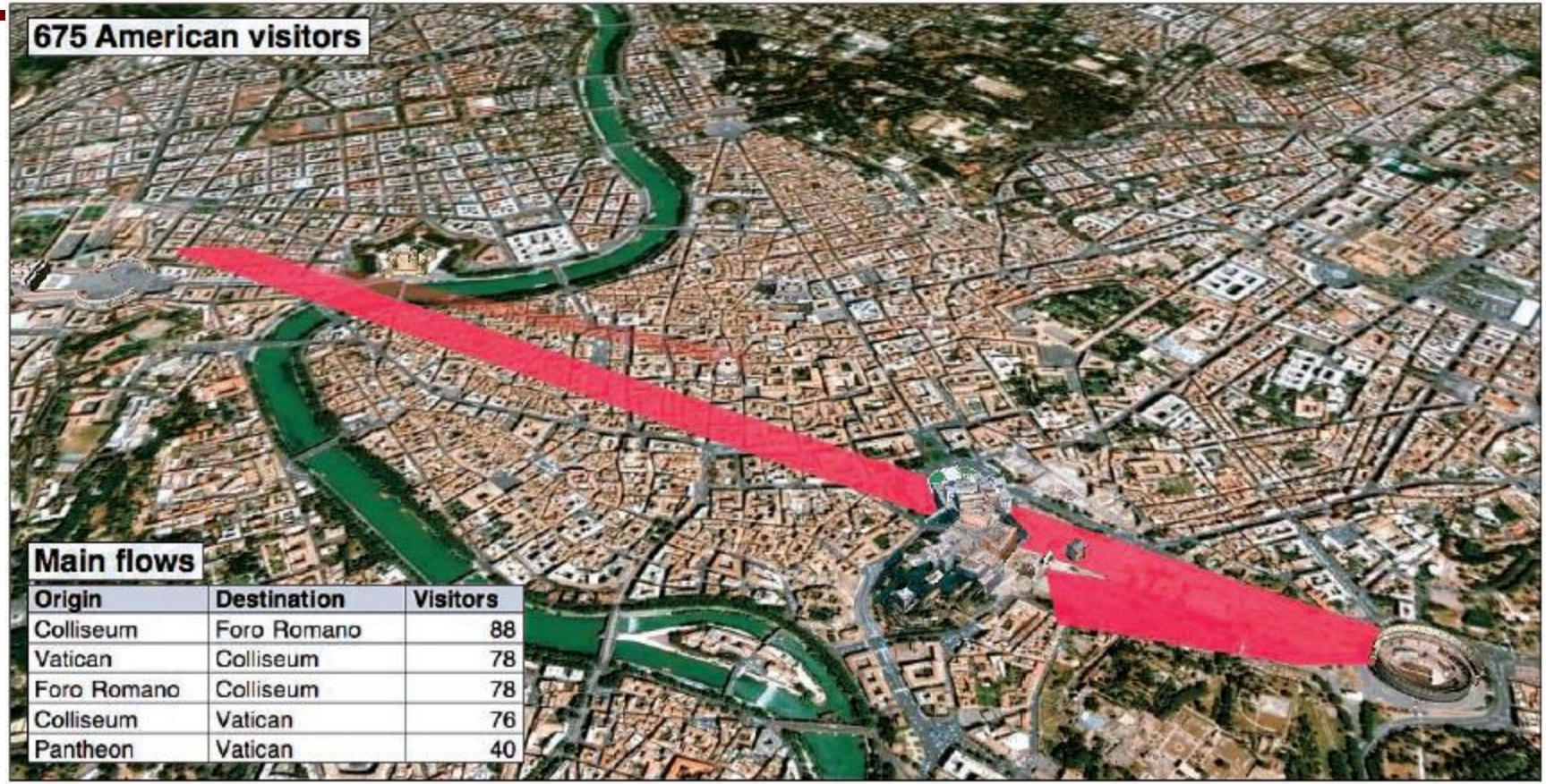


图10 外滩地图搜索请求与人员到达数量相关性分析



Main flows

Origin	Destination	Visitors
Foro Romano	Vittoriano	16
Colliseum	Foro Romano	15
Vittoriano	Foro Romano	14
Foro Romano	Colliseum	12
Colliseum	Vittoriano	11



Main flows

Origin	Destination	Visitors
Colliseum	Foro Romano	88
Vatican	Colliseum	78
Foro Romano	Colliseum	78
Colliseum	Vatican	76
Pantheon	Vatican	40

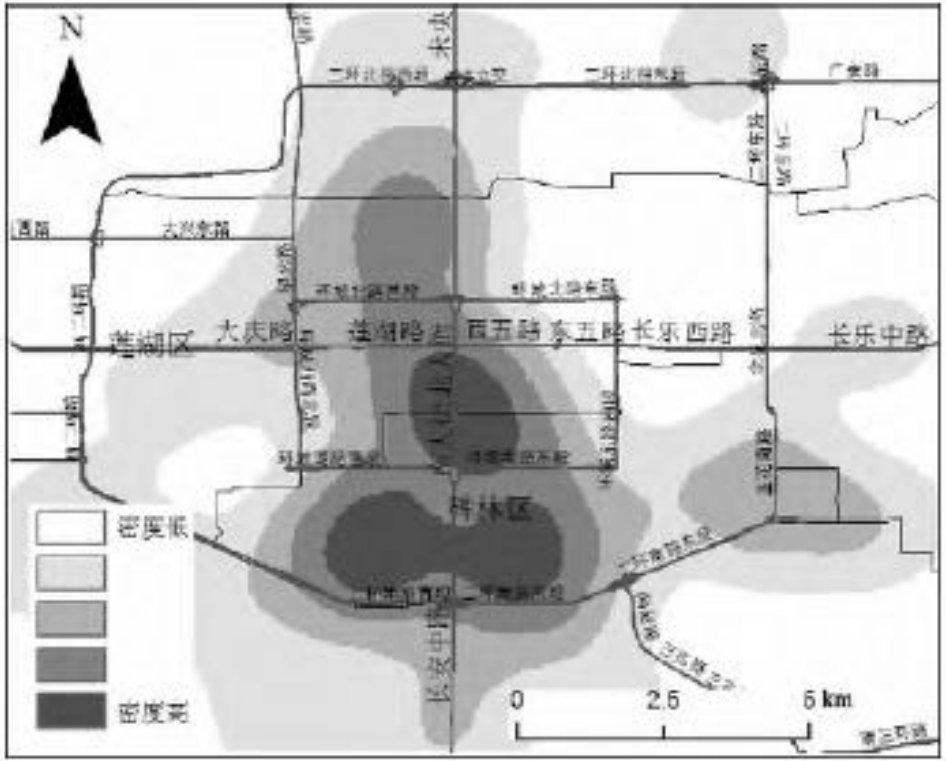


图3　城市运动公园4km以上游客空间分布

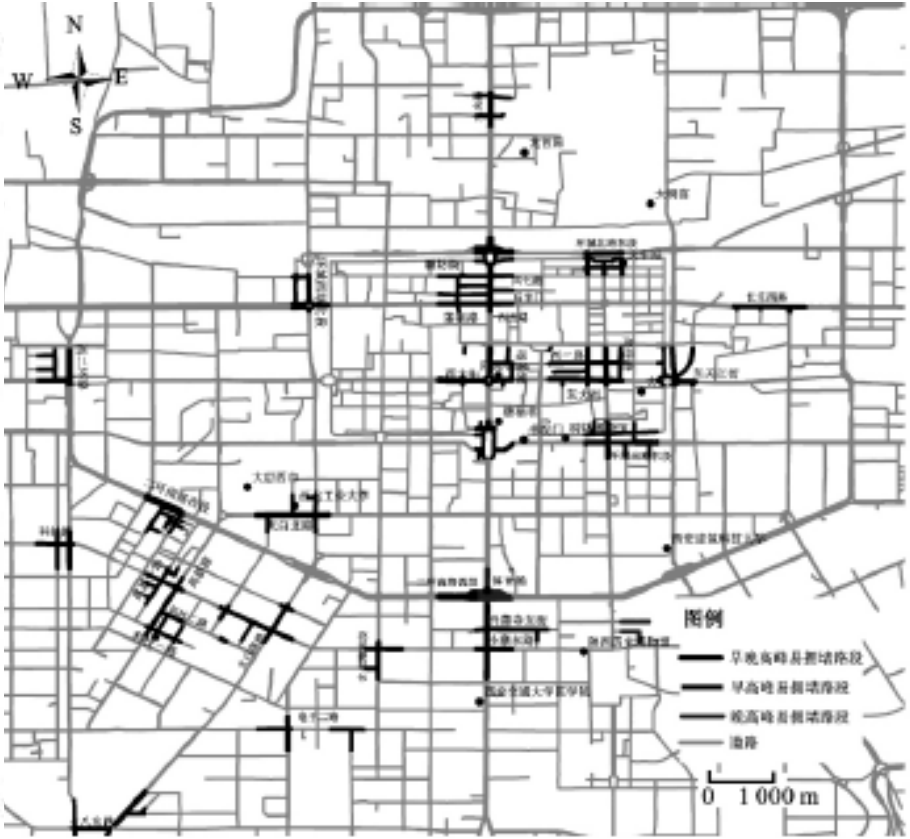
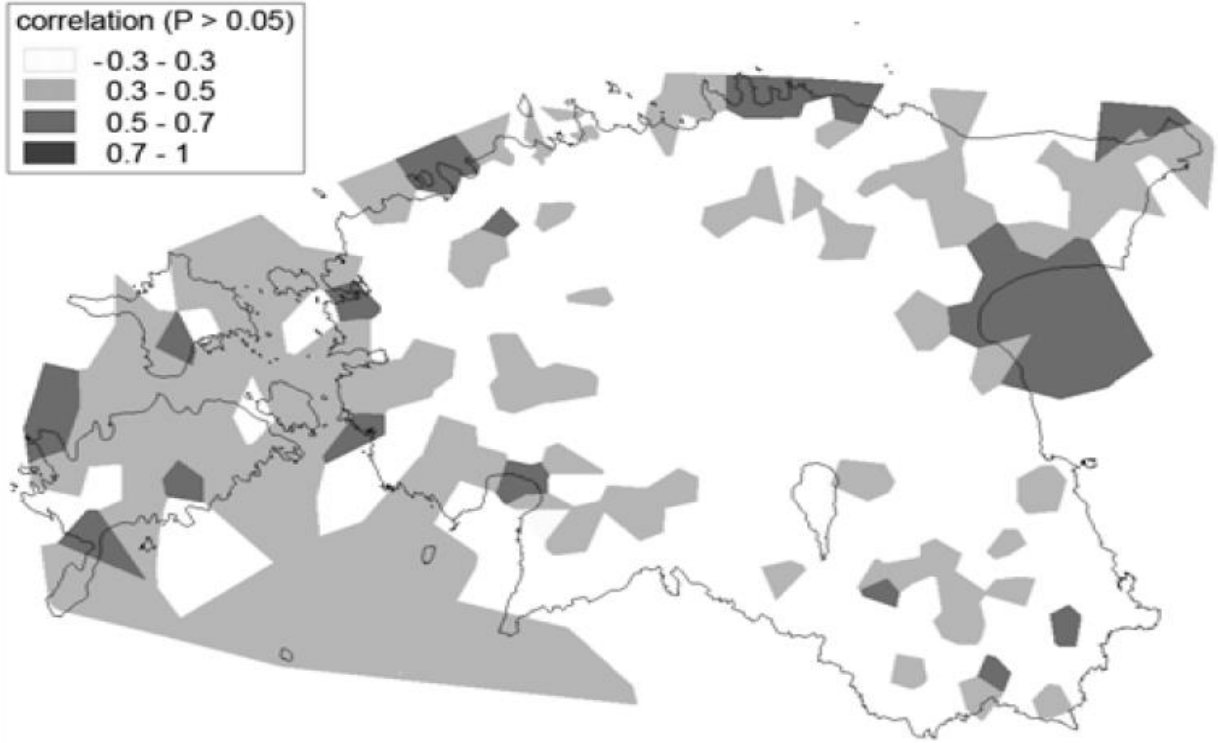


图5　西安市早晚高峰交通拥堵空间分布

2.2.5 其他

e.g. Järv等（2007）通过借助手机定位技术分析了气温与游客空间行为和目的地选择的相关性。



Correlation between tourists' locations and daily mean air temperature at the level of network cells during summer period (June-August) in 2004. Correlations are shown at significant level of $p < 0.05$

2.2 主要研究方法

- 2.2.1 可视化 (visualization)
- 其中非常重要的内容是 空间可视化
- （Spatial visualization ）
- 是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程

- **2.2.2 空间分析(Spatial analysis)**
- 利用空间分析分析空间分布、空间形成及空间演变的信息。空间分析——空间建模——模拟。
- 路径分析，叠加分析，网络分析，空间统计

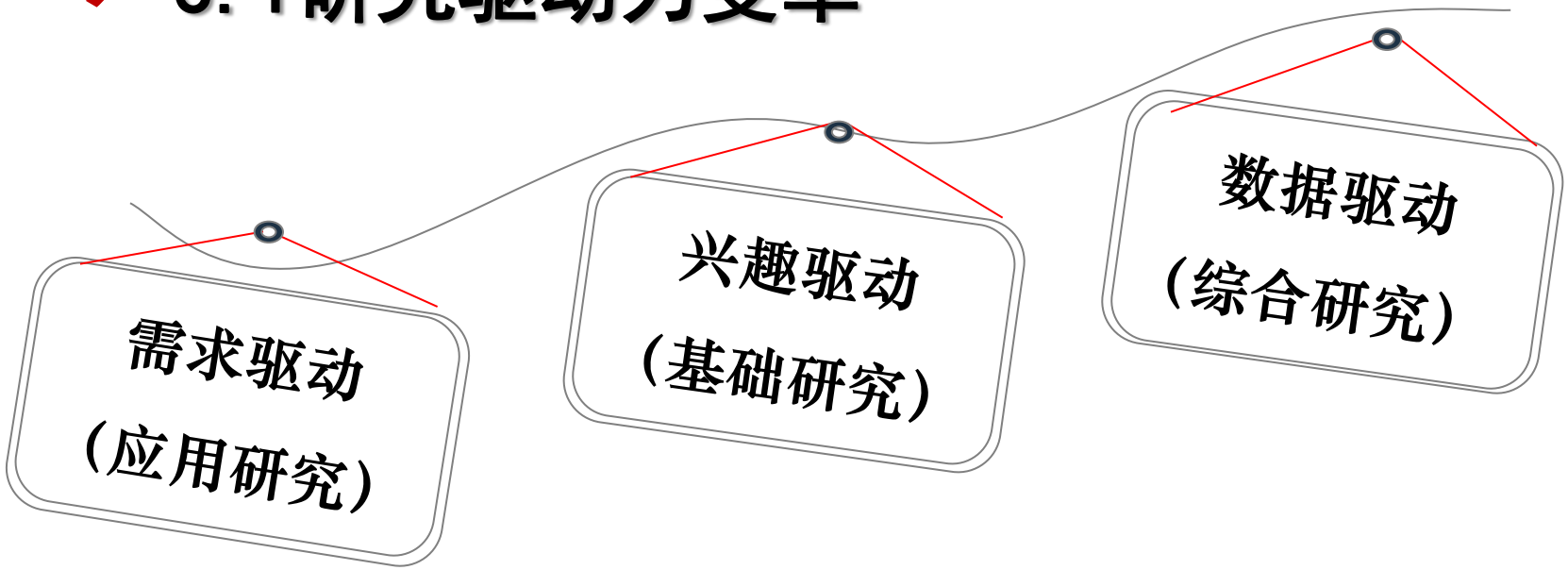
2.2.3 文本分析(Text analysis)

非结构化数据（文字、图形、符号、声频、视频）的处理，目前比较多的应用旅游行为的研究之中。

- 2.2.4 统计分析（ Statistical analysis ）
- 利用各种统计方法分析数据，研究事物之间的各种关系（相关性）。
- 预测方法，各种预测模型也会应用到研究之中。

3 展望

3.1研究驱动力变革

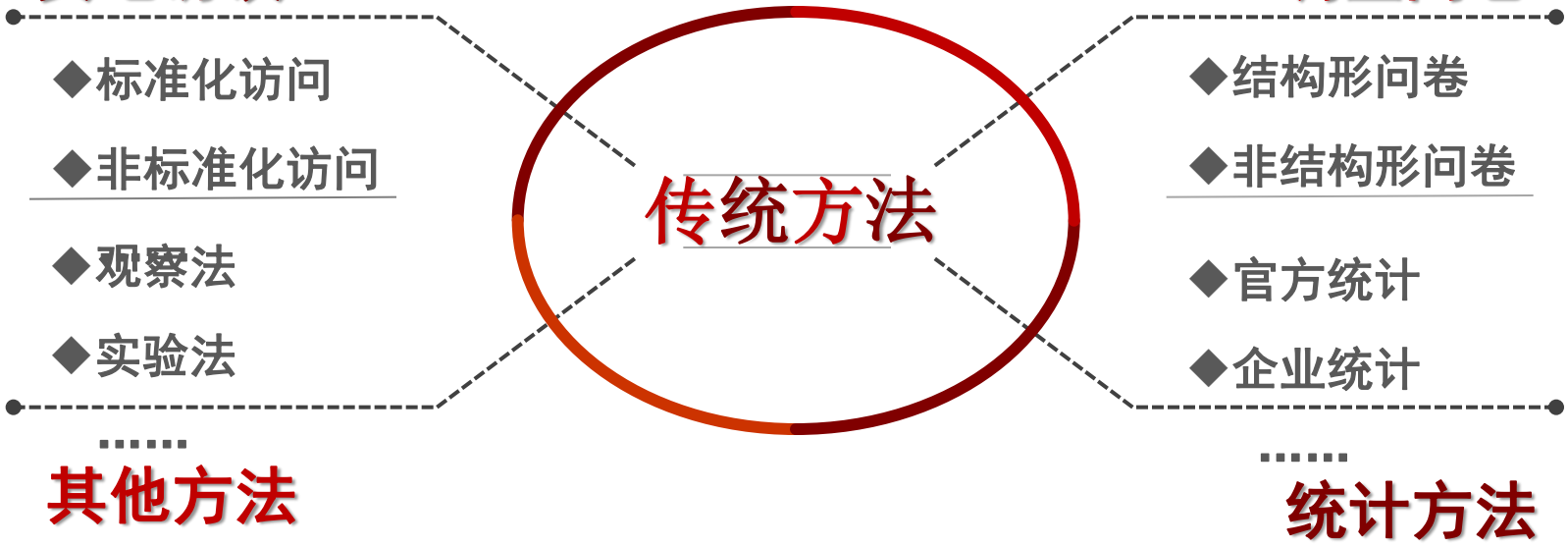


单一驱动——综合驱动

有别于需求驱动的应用研究和兴趣驱动的基础研究，大数据结构的多样化为旅游研究提供不同结构的数据源，拓宽了学者的研究思路，丰富学科研究内容。

3.2 研究方法变革

在方法上，中国旅游研究运用了更多的量化研究手段。但有些过度依赖和信奉实证主义和后实证主义。然而国际旅游研究有大量的公认的可选择的研究范式，如解释主义、建构主义和批判主义。中国现阶段旅游研究似乎慢了一步。**研究多为内容驱动而非理论驱动或知识驱动。多数的文章缺乏理论和知识贡献。（黄松山，陈钢华，2016）**


实地访谈
◆标准化访问
◆非标准化访问
◆观察法
◆实验法
.....
其他方法
传统方法
调查问卷
◆结构形问卷
◆非结构形问卷
◆官方统计
◆企业统计
.....
统计方法

新数据
新思维
新模型
新方法

有什么？——方法融合

• 新方法：

- 文本分析！
- 语义分析（语义引擎）！
- 复杂网络！
- 数据挖掘（数据挖掘算法）！
- 可视化（空间可视化，其他可视化）！
- 数据缩减！

传统方法 + 新方法

3.3 学科认识变革

- 旅游学科的认识——多学科交叉、融合
- 大数据——多学科交叉融合

单枪匹马无法做大数据

01 需要多学科交叉合作

02 需要融合多源信息建模

03 需要行程标准化的分析模块

04 需要形成功能强大的爬虫模块

3.4 局限与拓展

- 大数据目前更多的是——现象总结、规律总结，缺乏因果、机制的分析
- 大数据解决现象、规律问题
- 传统方法解决因果、机制问题

3.5未来可能的研究方向

- 1. 基于大数据的游客行为建模
- 2. 数据的标准化处理
- 3. 数据挖掘方法
- 4. 基于大数据的客流预测
- 5. 数据融合，同化
- 6. 隐私和伦理
- …… 。

谢谢

lijunyi9@snnu.edu.cn

13379263086