# 城市空间数据分析方法理论集成研究框架

## 2.1.1 城市空间数据分析方法理论集成研究框架

数据分析是就所分析研究的内容和目的，使用与之相关的单一数据或多源数据，以统计学方法为基础，涉及应用线性代数、微积分、概率论、图论、计算机视觉、机器学习、深度学习等交叉学科的方法或算法模型解释挖掘数据、寻找模式、总结规律，从数据中发现有用的信息、得出结论以支持决策。数据分析方法在各个领域中具有广泛的应用，对城乡规划、风景园林等领域，结合地理信息系统提出有城市空间数据分析，以数据分析的方式解决城市相关问题。

数据分析是以数据为源动力，多学科方法理论综合应用的过程。因此城市空间数据分析方法理论并无固定的框架，而是以解决所研究的内容为目的，适合的选择一种或多种方法，乃至依据已有方法理论提出新的方法或算法模型。整合集成城市空间数据分析方法理论的方式则变得复杂多样。为了使得集成的方法理论易于实现且方便应用于实际的研究过程、以数据科学领域最常用的编程语言 Python 为媒介，充分利用可结合方法理论的开源代码，按照基础试验和专项探索的方式集成可操作的城市空间数据分析方法理论。

基础试验包括知识间具有潜在迭代路径的统计学、线性代数、微积分、计算机视觉、机器学习和深度学习所包含的基础核心内容。专项探索是集成可用于城市空间数据分析能够构成独立框架的方法理论，包括标记距离、权重决策、更新策略、模式生成、推理学习、时空序列、维度空间、尺度效应、强化学习和复杂网络等10个专题。在集成不同学科领域，可服务于城市空间数据分析的方法理论时，所有的知识均是可代码实现，并开展探索性试验。表1列出了所集成的方法理论，并分为一级、二级分类和计算分析内容，及探索性试验。

表 1 城市空间数据分析方法理论集成研究框架

| 一级分类 | 二级分类 | 计算分析内容 | 探索性试验或示例（Python 代码实现） |
| --- | --- | --- | --- |
| 基础试验 |  |  |  |
| 统计学 | 描述性统计 | 1. 频数（次数）分布表和直方图；2. 均值、中位数、箱型图、标准差与标准计分（z\_score）。 | 1. 示例数据的描述性统计；2. 百度地图 POI 数据的描述性统计。 |
|  | 正态分布与概率密度函数 | 1. 偏度与峰度；2. 检验数据集是否服从正态分布；3. 异常值处理；4. 给定特定值计算概率，及找到给定概率的值。 | 百度地图 POI 数据分类的正态分布与概率密度函数相关计算。 |
|  | 核密度估计 | 1. 单变量（一维数组）的核密度估计；2. 多变量（多维数组）的核密度估计。 | 开放街道地图（OpenStreetMap，OSM）点分类的核密度估计。 |
|  | 基本统计量 | 1. 标准误；2. 中心极限定理；3. t 分布 (Student’s t-distribution)；4. 统计显著性；5. 效应量；6. 置信区间。 | 示例数据的基本统计量计算。 |
|  | 相关性 | 1. 相关系数（数值数据和数值数据）；2. 相关比（数值数据和分类数据）；3. 克莱姆相关系数（分类数据和分类数据）。 | 1. 示例数据的相关性计算;2. 公共健康数据的相关性分析。 |
|  | 卡方分布与独立性检验 | 1. 卡方分布；2. （卡方）独立性检验；3.  协方差估计。 | 示例数据的卡方检验与协方差估计。 |
|  | 回归模型（含机器学习-回归） | 1. 反函数；2. 指数函数与自然对数函数；3. 回归与微分；4. 矩阵；5. 简单线性回归求解与显著性检验（F 检验）、置信区间估计、精度计算；6. 多元线性回归求解与显著性检验、置信区间估计、精度计算；7. K-近邻模型（k-nearest neighbors algorithm，k-NN）；8. 平均绝对误差（mean absolute error, MAE）和均方误差（mean squared error，MSE）;9. 多项式回归；10. 正则化;11. 梯度下降法（Gradient Descent）。 | 1. 示例数据的线性回归计算；2. 公共健康数据中经济条件数据与疾病数据的简单线性回归建模；3. 公共健康数据的多元回归、k-NN和多项式回归建模。 |
| 线性代数 |  | 1. 矩阵；2. 向量；3. 线性映射;4. 特征值和特征向量。 | bildstein\_station1 （城市点云数据）的线性变换。 |
|  |  | 主成分分析（Principal components analysis，PCA） | Landsat 遥感影像波段的 PCA 降维与 RGB 显示。 |
|  |  | 1. NDVI 指数；2. 采样精度计算。 | （建立简单工具）建立对遥感影像波段的采样工具。 |
| 微积分 |  | 1. 导数与微分；2. 积分;3. 泰勒展开式；4. 偏微分。 | 示例数据的微积分相关计算。 |
|  | 卷积与SIR空间传播模型 | 1.  一维卷积与曲线分割；2. 二维卷积与图像特征提取；3. SIR 传播模型；4. SIR 空间传播模型。 | 根据土地覆盖配置空间阻力值建立 SIR 空间传播模型。 |
| 计算机视觉 | 图像特征提取 | 1. SIFT（Scale-Invariant Feature Transform）特征检测和描述算法；2. Star 特征检测器提取图像特征；3. 特征匹配。 | 用特征检测提取 KITTI 数据集（城市影像）特征，通过特征匹配探索动态街景视觉感知。 |
|  | 超像素级分割方法 | 1. Felzenszwalb 超像素级分割方法；2. Quickshift 超像素级分割方法。 | Sentinel-2 影像超像素级分割下高空分辨率特征尺度界定。 |
| 机器学习 | 聚类 | 1. K-Means 聚类；2. 多种聚类算法比较。 | 用调研图像聚类图像主题色，分析城市环境色彩。 |
|  | 图像分类器与识别器 | 1. 视觉词袋与构建图像映射特征；2. 决策树（Decision trees）与随机森林（Random forests）；3. 信息增益（Information gain）；4. 基尼不纯度（Gini impurity）;5. 交叉验证 （cross\_val\_score）；6. 图像分类器与识别器。 | （建立简单工具）图像分类器与识别器的机器学习模型网络实验应用平台部署。 |
| 深度学习 | 多层感知机（多层神经网络） | 1. 张量（tensor）计算；2. 微积分-链式法则；3. 激活函数（activation function）；4. 前向与后向传播（propagation）；5. 自动求梯度（gradient）；6. 多层神经网络；7. runx.logx-深度学习试验管理。 | 示例数据的多层神经网络构建。 |
|  | 逻辑回归二分类与 SoftMax 回归多分类 | 1. 伯努利分布；2. 似然函数；3. 极大似然估计；4. 逻辑回归（Logistic Regression, LR）；5. SoftMax 回归多分类。 | 示例数据（图像数据集 Fashion-MNIST）的SoftMax 回归多分类建模。 |
|  | 卷积神经网络（Convolutional neural network，CNN） | 1. CNN，及其卷积层、池化层和填充；2. tensorboard 可视化；3. 可视化卷积层/卷积核；4. VGG 网络。 | 示例数据的 CNN 建模。 |
|  | 对象检测与实例分割 | 1. 对象检测；2. 对象实例分割与对象统计；3. 交并比（Intersection over Union，IoU）；4. Faster R-CNN ResNet-50 FPN 图像语义分割模型。 | 1. PennFudan 和 COCO 数据集行人检测；2. KITTI 数据集行人检测与人流量估算；3. KITTI 数据集对象实例分割与统计和关联网络结构。 |
|  | 图像分割 | 1. 开放神经网络交换 —— ONNX（Open Neural Network，Exchage）；2. Netron 网络可视化工具；3. DUC（Dense Upsampling Convolution）图像分割；4. VGG16 卷积神经网络;5. SegNet 图像分割模型；6. 图像数据增强变换；7. 图像分割模型库；8. Unet 网络模型（resnet34 编码器）图像分割。 | 1. cityscapes 数据集 DUC 图像语义分割下 城市空间要素组成，时空量度，绿视率和均衡度计算;2. sentinel-2 影像的无监督土地分类——聚类方法；3. 用 VGGNet 模型预测图像内容对象；4. ISPRS 遥感影像 SegNet 图像分割模型解译；5. NAIP 航拍影像解译（图像分割）。 |
| 其它 |  | 天空视域因子（Sky View Factor，SVF）计算 | 基于数字表面模型（Digital Surface Model，DSM） 计算 SVF。 |
| 专项探讨 |  |  |  |
| 标记距离 |  | 1. 地图分类算法;2. 模式标记特征（Pattern Signature）：     a. 类/簇大小直方图（连通域标签）;     b. 共现关系;     c. 3 层级分解 。3. 距离度量（Distance meatrics）；4. 模式发现：     a.  模式级聚——距离矩阵与层级聚类；     b. 模式搜索；     c. 模式监测；     d. 模式分割。 | 1. 对 NLCD（National Land Cover Database）美国土地覆盖类型数据集执行模式发现；2. 地表温度（Land Surface Temperature，LST）冷热点样方标记土地覆盖类型模式标记特征。 |
| 权重决策 |  | 多准则决策法，Multiple Criteria Decision-Making（MCDM）：     a. 熵值权重法；     b. 理想解法（TOPSIS）；     c. AHP、F-AHP、ARAS、 BWM （基于 GWO 元启发式算法）、DEMATEL 、 IDOCRIW （基于 GA 元启发式算法）、ELECTRE-I 、WASPAS  等 MCDM 算法。 | 1. 对 AoT  城市环境传感器测量数据中污染气体浓度数据测量点污染浓度综合指数评估（应用 TOPSIS 算法）；2. LST（降温）与绿地和建设用地矛盾的匹配评估（ WASPAS 算法评估样方单元综合评价指数）。 |
| 更新策略 |  | 1.  1维度权重决策；2. 复杂网络更新（遗传算法）；3. 2维度布局优化——空间决策支持系统；4. 元启发式算法（Meta-Heuristic Algorithm）：     a. 粒子群优化算法（Particle Swarm Optimization, PSO）；     b. 布谷鸟搜索算法（Cuckoo Search, CS）；     c. 萤火虫算法（Firefly Algorithm, FA）。 | 1. 1 维度权重决策解决不同建设内容分配有限不同土地资源条件问题；2. 2维度布局优化——空间决策支持系统复现。 |
| 模式生成 |  | 1. 聚类模式特征分析；2. 生成对抗网络（Generative Adversarial Networks，GAN） 的模式生成；3. SytleGAN；4.  条件对抗网络（Conditional Adversarial Networks，cGAN）。 | 1. 多源数据的区域统计与数据合并；2. 多源数据聚类模式组成结构；3. LST 冷热点舒适区域与地表覆盖类型的簇分布统计；4. （建立简单工具）WGAN 模型中 G 网络各层数据演化情况计算分析工具；5. StyleGAN 模型生成 NAIP 航拍影像数据；6. NAIP 遥感影像和土地覆盖类型之间的翻译转化；7. （建立简单工具）遥感影像和土地覆盖类型互译工具构建；8. 遥感影像缺失区域的补全。 |
| 推理学习 | 概率论与概率图 | 1. 概率论；2. 概率图——贝叶斯网络：     a. 基于概率模型的分类器;     b. 离散型贝叶斯网络——表征（Representation）与推理（Inference）;     c. 连续型贝叶斯网络;     d. 模型学习：参数估计；结构学习。 | 1. 示例数据概率论与概率图相关计算；2. 用生境威胁、生境类型、碳储存、生境质量和生境风险等5个生态系统服务指数关系的简化模型建立贝叶斯网络。 |
|  | 生态系统服务价值（InVEST 模型） | 1. InVEST 和 NWV 生态系统服务计算工具；2. 生态系统服务价值中的碳储存和封存、生境质量、作物授粉和作物生产（InVEST）。 | 生态系统服务价值的贝叶斯网络构建与推理。 |
| 时空序列 | 时间序列分析 | 白噪声检验、平稳性检验和周期性模式判断。 | 城市环境传感器 AoT 时序数据的白噪声检验、平稳性检验和周期性模式判断。 |
|  | 空间自相关分析 | 1. 空间权重；2. 全局空间自相关；3. 局部空间自相关；4. 地理轮廓（Geosilhouettes）。 | 公共健康数据的空间自相关分析。 |
|  | 空间动力学 | 1. 空间马尔可夫链：     a. 经典离散（时间）马尔可夫链；     b.空间马尔科夫 （Spatial Markov）；     c. 局部空间自相关马尔科夫（LISA Markov）；     d. 流动性测量。2. 空间 Kendall’s Tau：     a. 经典肯德尔等级相关系数（Kendall’s Tau）；     b. 空间Kendall’s Tau。 | 1. 犯罪率年统计量的经典和空间离散型马尔可夫链计算；2. 空气质量指数（Air quality index，AQI）的空间 Kendall’s Tau 计算。 |
|  | 不平等性和空间隔离 | 1. 不平等性（基尼系数和洛伦兹曲线）；2. 空间隔离：     a. 空间隔离指数;     b. 组间比较与作用（决定）因素成分分解和贡献度；     c. 空间隔离指数推断；     d.  局部空间隔离指数；     e. 多类空间隔离指数。 | 芝加哥城土地利用样方统计下开放空间的不平等性和空间隔离相关计算。 |
|  | 从自然语言处理到视觉模型 | 1. RNN（Recurrent Neural Networks）；2. LSTM（long short-term memory）；3. Word2Vec和Seq2Seq；4. Transformer——自然语言处理（Natural Language Processing，NLP）；5. GPT；6. Transformers 的视觉模型。 | 1. 百度全景静态图视域景观指数计算；2. 视域景观指数不同邻里尺度最优簇数选择和指数贡献度；3. 街道行业分类（POI）服务空间组成结构；4. 特征指数的 LSTM 时空预测模型。 |
| 维度空间 |  | 1. 流形学习（Manifold learning）；2. MDS、 KPCA、Isomap、LLE、t-SNE 和 UMAP 等 流形学习算法。 | 1. 街道空间多维景观指数的 MDS 降维试验；2.  街道空间单一景观指数的 KPCA 映射试验；3. 街道视域全景语义分割图的 Isomap 降维和嵌入空间的主成分分布；4. 街道 POI 行业类别空间分布结构；5. 街道空间多维景观指数的 UMAP 空间特征区段划分。 |
| 尺度效应 | 不同尺度作用结果的比较分析方法 | 1. 标记距离的尺度变化矩阵;2. 分类面积尺度变化曲线；3. 常规统计；4. 半变异（方差）函数（semi-variograms）。 | 1. NAIP 航拍影像解译（图像分割）为土地覆盖数据；2. 土地覆盖数据标记距离的尺度变化矩阵和分类面积尺度变化曲线等不同尺度作用结果比较分析；3. 用半变异函数寻找人口数量分布数据尺度效应的幅度变化关系。 |
|  | 空间相互作用模型（Spatial Interaction Modelling，SIM） | 1. SIM 族；2. 广义线性模型（GLM）与泊松对数线性回归；3. SIM 模型拟合统计。 | 绿地公园供给与人口需求关系的 SIM 参数估计。 |
| 强化学习 | 智能体模型（Agent-Based Models，ABM） | NetLogo，Mesa 和 Repast4Py 模拟系统比较。 |  |
|  | 强化学习 | 1. 表格型求解方法；2. Gymnasium 环境构建；3. 深度强化学习（Deep RL）；4. 多智能体强化学习（MARL）。 | 1. 动物的运动及其与土地覆盖类型关系的试验；2. 纳沃纳广场的人行为模拟。 |
| 复杂网络（图论） |  | 1. 图论；2. 图属性的基本度量；3. 图嵌入和图神经网络。 | 1. 构建土地利用类型的样方式复杂网络进行图属性基本度量计算；2. 公共交通复合网络构建与公园潜在服务人口压力；3. 植物调查样方物种的图嵌入； 4. 植物调查样方物种间的关联。 |

## 2.1.2 数据分类和处理

数据是数据分析的源动力，在应用方法理论进行所研究内容的数据分析前，通常需要对数据进行预处理，这包括数据的检索获取、清洗、格式转化、提取、集成，及写入数据库，从数据库中读写数据等。为了实现完成广泛类型数据的预处理任务，在城市空间数据分析方法理论集成时，尽量使用具有不同格式和内容的多源数据作为开展探索性试验的数据源，包括的数据类型主要有地理空间矢量数据和栅格数据、文本数据、影像等。根据数据内容可以分类有兴趣点、点云数据、遥感数据、航拍影像、土地利用和覆盖、地表温度、城市环境传感器测量数据、空气质量指数、数字高程、建筑高度、夜间灯光、城市影像、全景静态图、观测鸟分布、动物跟踪数据库、人口分布、GPS 轨迹数据、物种分布及丰度、公共健康数据、犯罪事件数据集，以及城市的基础数据，例如路网、公交和地铁线路站点、行政边界等。

各类数据处理完全使用 Python 编程语言实现，这与城市空间数据分析方法理论集成的 Python 代码实现保持一致，从而充分利用 Python 生态系统，构建名为 [USDA](https://richiebao.github.io/USDA_PyPI/#/) 的工具包，弹性集成数据处理技术、数据分析方法理论，建立方法理论可实现计算的工具。

表2 数据分类和预处理

| 数据分类 | 数据名称 | 数据来源 | 数据预处理内容 |
| --- | --- | --- | --- |
| 兴趣点（points of interest，POI） | 百度地图 POI 数据 | 百度地图开放平台 | 1.检索；2. CSV 转 DF 格式数据；3. DF转 GDF 地理空间格式数据；4. 数据可视化。 |
|  | 开放街道地图（OpenStreetMap，OSM） | OpenStreetMap | 1. 检索；2. 读取、转换 OSM 数据为 SHP 格式。 |
| 点云数据 | bildstein\_station1 （城市点云数据） | semantic3d | 读取和可视化。 |
|  | 伊利诺伊州 LAS 格式激光雷达数据 | 伊利诺斯州草原地质调查研究所（Illinois state geological survey - prairie research institute） | 1. 查看点云数据信息；2. 点云数据可视化；3. 由点云建立数字表面模型（Digital Surface Model，DSM）与分类栅格；4. 建筑高度提取。 |
| 遥感影像 | Landsat | 美国地质调查局（ United States Geological Survey，USGS-earthexplorer） | 1. 元数据读取；2. 波段的合成显示。 |
|  | Sentinel-2 | 哥白尼数据空间生态系统（Copernicus Data Space Ecosystem） | 1. 以 Web Mercator 方式显示 Sentinel-2 的一个波段;2.  Sentinel-2 波段合成显示；3. 元数据读取；4. 影像裁切。 |
|  | ISPRS 遥感图像数据集 | International Society for Photogrammetry and Remote Sensing | 读取与建立样本数据集。 |
| 航拍影像 | NAIP 航拍影像。（美）国家农业图像项目（National Agriculture Imagery Program，NAIP） | 美国地质调查局（ United States Geological Survey，USGS-earthexplorer | 1. 数据检索；2. 建立样本数据集。 |
| 土地覆盖 | NLCD（National Land Cover Database）美国土地覆盖类型数据集 | 美国地质调查局（United States Geological Survey，USGS）与多分辨率土地特征联盟 (Multi-Resolution Land Characteristics Consortium，MRLC) | 数据检索与建立样本数据集。 |
|  | 高精度芝加哥土地覆盖数据 | ArcGIS-Chicago Regional Land Cover | 读取与可视化。 |
|  | MCD12Q1\_v006 土地覆盖类型数据集 | 美国国家航空和航天局（National Aeronautics and Space Administration，NASA） | 1. 数据检索和地图打印；2. 批量处理和保存MCD12Q1\_v006 各年数据；3. 按采样点提取数据。 |
|  | ESA WorldCover | 欧空局（European Space Agency，ESA） |  |
| 土地利用数据 | 芝加哥城土地利用 | 芝加哥都市规划局（Chicago Metropolitan Agency for Planning, CMAP） | 1. SHP 格式转 GDF 格式与名称编码;2. 矢量 SHP 格式土地利用数据转栅格（raster）；3. 土地利用地图打印。 |
| 地表温度（Land Surface Temperature，LST） | 中分辨率成像光谱仪 (Moderate Resolution Imaging Spectroradiometer ，MODIS)的地表温度和辐射率（Emissivity ）产品（MYD21A1D v006） | 美国地质调查局（ United States Geological Survey） | 数据检索与建立样本数据集。 |
| 城市环境传感器 | Array of Things，AoT（分类测量内容为气体类、颗粒物、热环境、噪音、光、电磁场和惯性测量等） | 西北-阿贡科学与工程研究所（Northwestern-Argonne Institute for Science and Engineering，NAISE）领导 | 1. 数据清洗（数值有效区间，数据精度处理）；2. 按时间范围提取数据;3. 分类数据。 |
| 空气质量指数（Air quality index，AQI） | 空气质量指数数据集（Air Quality Index Dataset） | Kaggle | 读取与可视化。 |
| 数字高程 | 数字高程模型（digital elevation model，DEM） | 美国地质调查局（ United States Geological Survey，USGS-earthexplorer | 读取与可视化。 |
| 建筑高度（层数）数据 | 芝加哥城建筑轮廓数据，含有层数字段 | 芝加哥数据门户（Chicago Data Portal，CDP） | 读取与可视化。 |
| 夜间灯光数据 | Annual VNL V2 | 地球观测组（Earth Observation Group） |  |
| 城市影像 | KITTI 数据集（无人驾驶场景下计算机视觉算法评测数据集） | The KITTI Vision Benchmark Suite | KITTI 文件信息读取和可视化。 |
|  | PennFudan 数据集 | Penn-Fudan Database for Pedestrian Detection and Segmentation | 读取与可视化。 |
|  | COCO 数据集 | COCO | 读取与可视化。 |
|  | Cityscapes 数据集 | Cityscapes Dataset | 读取与可视化。 |
|  | ImageNet 数据集 | ImageNet | 读取与可视化。 |
| 全景静态图 | 百度全景静态图 | 百度地图开放平台 | 1. 全景图像素级语义分割；2. 语义分割图的投影变换。 |
|  | 调研图像 | 用手机应用记录调研路径 | 1. 图像数据读取与可视化；2. 读取图像 KML 调研路径数据与可视化；3. Exif（Exchangeable image file format） 可交换图像格式数据读取；4. RGB 色彩的三维图示。 |
| 观测鸟分布数据 | eBird Basic Dataset (EBD) | eBird | 读取与转化为 GDF 格式数据，及可视化。 |
| 动物跟踪数据库 | 森林公园生活实验室（Forest Park Living Lab，FPL）动物运动跟踪数据 | Movebank | 读取与可视化。 |
| 人口分布数据 | WorldPop | WorldPop | 读取与可视化。 |
| GPS 轨迹数据 | GeoLife | 微软亚洲研究院（Microsoft Research Asia） | 1. 合并 GeoLif e 数据集；2.  合并GPS轨迹和标签；3. 交通工具时速统计。 |
| 物种分布及丰度 | Countryside Survey 1978 vegetation plot data 数据集 | 英国生态与水文中心（Centre for Ecology & Hydrology） | 读取为 DF 格式数据。 |
| 其它 | 公共健康数据（公共健康指标） | 芝加哥数据门户（Chicago Data Portal，CDP） | 合并社区边界数据，转 CSV 至 GDF。 |
|  | 芝加哥市发生报告的犯罪事件数据集 | 芝加哥数据门户（Chicago Data Portal，CDP） | 时空数据（面板数据），按照给定的分组，时间长度，数值计算方法重采样数值列。 |
|  |  |  |  |
| 数据类型缩写 | DataFrame（DF）;GeoDataFrame（GDF） |  |  |