实 验 报 告

实验名称： 智慧城市

专业班级： 地理信息科学2018-3

学 院： 测绘与空间信息学院

姓 名： 江盟 学 号： 201801020812

指导老师： 刘文宝

2021年12月8日

**山东科技大学**

**一、实验目的：**

1. 通过实验，掌握收集时空大数据的方法；
2. 通过实验课程，了解《智慧城市》过程中通过**时空大数据**分析和机器学习发掘数据背后隐含知识的常用方法；

**二、实验内容：**

**我国省/市/县级城镇化发展水平时空变化**

要求：模仿文献1研究我国省级城镇化发展水平时空变化。时间范围可以是2007-2015，也可以是2007-至今

计算城镇化水平的步骤：

1. 确定城镇化指标，并收集相关数据
2. 无量纲化处理
3. 数据标准化
4. 计算各成分熵
5. 各成分的熵权
6. 各成分的综合评价得分，即城镇化水平
7. 城镇化水平时空变化

文献1：廖中举，张志英. 省际新型城镇化发展水平测度与比较. 管理决策, 2020, 20:168-171.

1. **实验报告的内容：**  
   1）数据简介

国家数据网https://data.stats.gov.cn/easyquery.html

国家统计局http://www.stats.gov.cn/tjsj/ndsj/

包括数据来源，数据的空间范围，空间分辨率（空间尺度），时间范围，时间分辨率

1. 数据预处理处理

对于寻找到的缺失值一般来说可以有以下3种处理方式，直接删除、保留和寻找替代值。

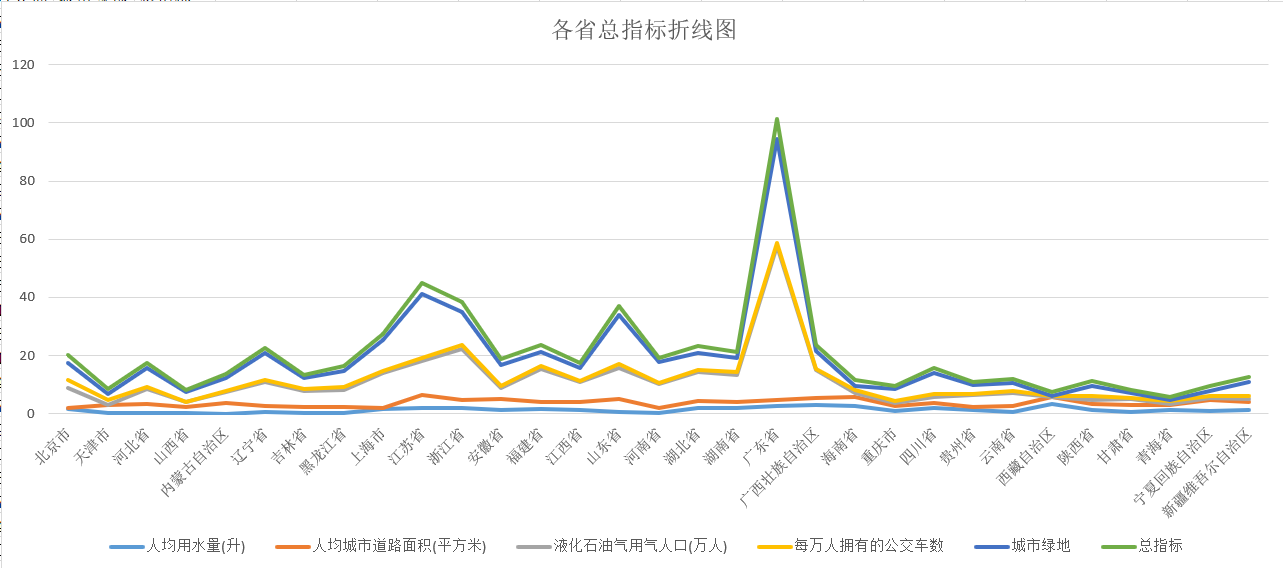
直接删除：直接删除的优点是删除以后整个数据集都变得完美了，都是有完整记录的数据，缺点是缺少了部分样本可能导致整体结果的偏差。对于有大量缺失值的在衡量利弊的情况下建议就直接删除了吧，缺失了大量关键数据的样本集统计起来也没有什么意义。

保留：保留缺失值，优点是保证了样本的完整，缺点是你得知道为什么要保留，保留它的意义是什么，是什么原因导致了值的缺失，是系统的原因还是人为的原因，这种保留建立在缺失单个数据的情况下，且缺失值是有明确意义的。

寻找替代值：如用均值、众数、中位数等代替缺失值，优点是简单且有依据，缺点是可能会使缺失值失去其本身的含义。对于寻找替代值的除了[统计学](https://www.zhihu.com/search?q=%E7%BB%9F%E8%AE%A1%E5%AD%A6&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b:,:72890012%7d)中常用的描述数据的值以外，还可以人为地去赋予缺失值一个具体的值。

1. 数据分析的原理

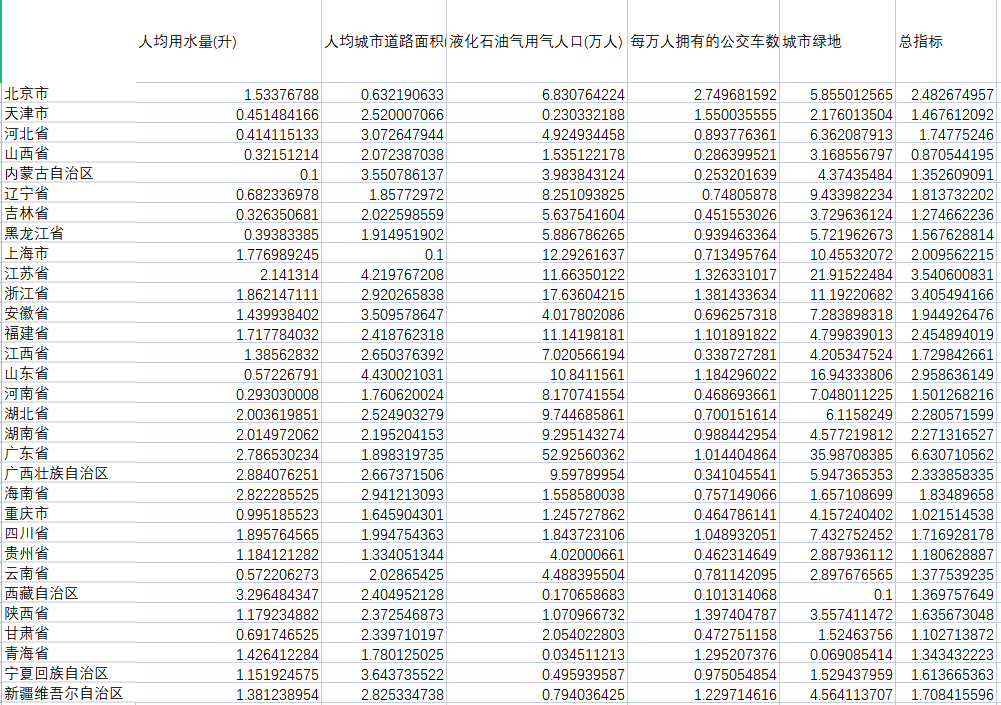
熵值法原理： 熵的概念源于热力学，是对系统状态不确定性的一种度量。在信息论中，信息是系统有序程度的一种度量。而熵是系统无序程度的一种度量，两者绝对值相等，但符号相反。根据此性质，可以利用评价中各方案的固有信息，通过熵值法得到各个指标的信息熵，信息熵越小，信息的无序度越低，其信息的效用值越大，指标的权重越大。

1. 数据分析的结果及对结果分析

图一 各省城镇化总指标图



图二 归一化后部分各省城镇化



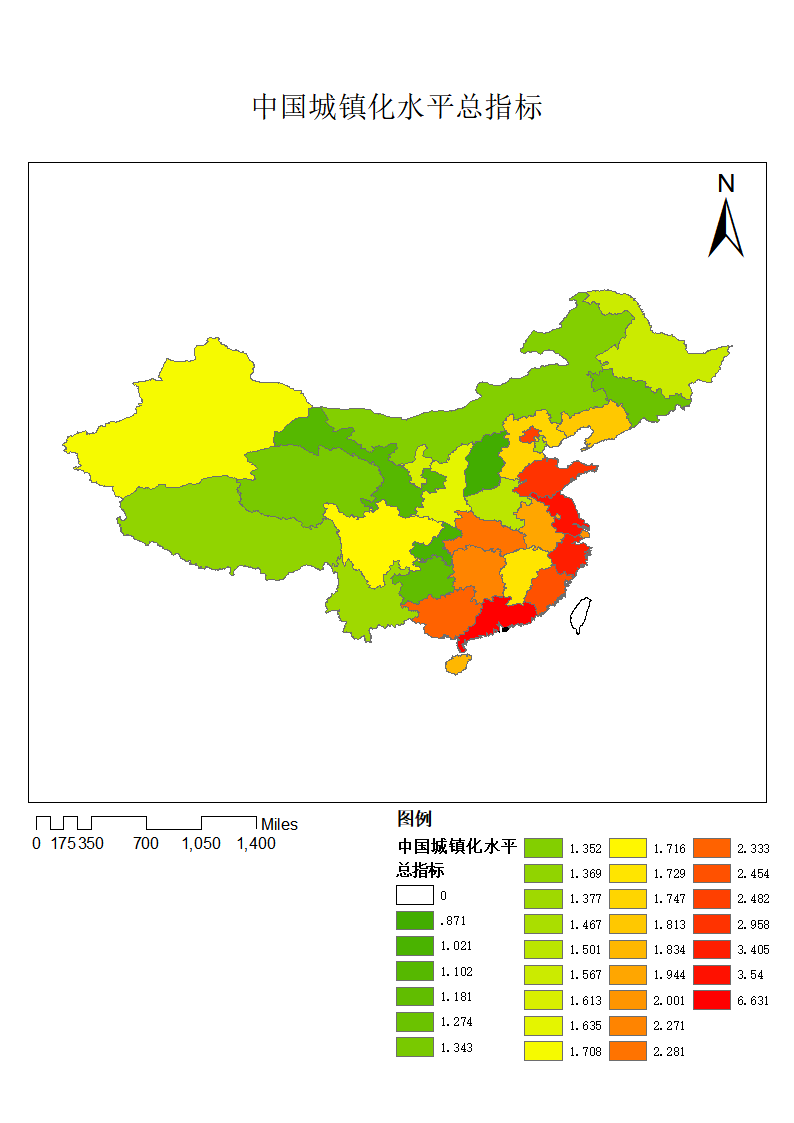
图三 各城市各子系统指标及总指标

1. 省际新型城镇化水平评价

新型城镇化水平综合评价的主要思路是先确定各指标层内每个评价指标的权重，再按照标准化后的指标值加 权平均得到2011—2019年五个维度的评价结果以及新型 城镇化水平综合评价结果。新型城镇化水平建设综合指数和人均用水量、人均城市道路面积、液化石油气用气人口、每万人拥有的公交车数、城市绿地五个子系统的指数。

对每一年的新型城镇化水平分析发现，我国30个省份2011—2019年间，在人均用水量、人均城市道路面积、液化石油气用气人口、每万人拥有的公交车数、城市绿地等5个方面呈现出持 续增长的趋势。

1. 结果图



图三 中国城镇化发展水平总指标分布图