厦門大學

本 科 毕 业 论 文 （设 计）

（主修专业）

非负矩阵分解在城市建设中的应用

The Application of Non-negative Matrix Factorization in Urban Construction

姓　　名：吴昕怡

学　　号：19020162203259

学　　院：数学科学学院

专　　业：信息与计算科学

年　　级：2016 级

（校内）指导老师：林鹭　　副教授

二〇二〇年五月

厦门大学本科学位论文诚信承诺书

本人呈交的学位论文是在导师指导下独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合相关法律规范及《厦门大学本科毕业论文（设计）规范》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

本人承诺辅修专业毕业论文（设计）（如有）的内容与主修专业不存在相同与相近情况。

学生声明（签名）：

年　　月　　日

致 谢

在完成本科毕业论文的这个激动人心的时刻，回眸本科的学习生活，心中不禁感慨万千。在这里。我首先要感谢我的父母。正是由于他们那无私的爱和无比的耐心和宽容才使得我能够健康而快乐地成长。在此，我向他们表达我最诚挚的敬意和最真诚的爱。

在林鹭导师的悉心指导下，我的论文才得以顺利进行。从毕业论文选题开始、论文工作的开展到论文的最后完成，都倾注了林老师大量的心血，在本论文的撰写和定稿过程中，老师提出了许多宝贵、中肯的意见。在此，特向林老师致以最诚挚的敬意和最衷心的感谢。

同时，我要感谢四年来教导我的老师们以及为我们创造良好学习氛围的校院领导们，正是这些老师孜孜不倦的教诲及辛勤的付出，我才能顺利的完成学业，才能在大学四年学到很多知识。在此，我向他们致以我最诚挚的敬意和感谢。

摘 要

城市建设作为国家民生经济的重要支撑，影响着城市里商业经济、人民生活水平、环境状况等各方面的发展。如今，世界进入信息时代，信息数据的智能化服务正在造福于各个产业，因此，探索适用于城市建设的计算机信息技术开始逐渐得到关注。

本文将非负矩阵分解用于2018年城市建筑统计年鉴，通过不同的数据预处理方式，分析出我国城市建设中存在区域性发展不均衡的重要指标以及代表城市。同时，单以排水和排污处理方面为例做非负矩阵分解算法的实现，找出影响该方面的主要指标。另外，本文还探索了将赋权归一法运用于非负矩阵分解的前期数据处理中，减少由于不等量指标数可能带来的实验误差，并分析出城市建设的八个专业类别的不平衡发展占比，从而探讨城市建设中所需要侧重的大方向。

关键词：非负矩阵分解；归一法；城市建设

Abstract

As an important support in national livelihood economy, urban construction affects the development of commercial economy, citizen's living standard, environmental conditions and many other fields in cities. Nowadays, in the era of information age, the intelligent services of information data are benefiting various industries. Hence, the exploration of computer technology which is suitable for urban construction began to get more and more attention.

In this paper, Nonnegative Matrix Factorization is applied to the statistical yearbook of urban construction in 2018. Through different data pre-processing methods, the important index of unbalanced regional development in Chinese urban construction and representative cities are analyzed. Meanwhile, the drainage and sewage treatment are taken as a single example for realization of non-negative matrix decomposition algorithm and identify key indexes that affect this aspect . Meanwhile, the nonnegative matrix decomposition of more than one hundred indicators in urban construction has been analyzed, and the ideal values of decomposition coefficient and iteration number of Nonnegative Matrix Factorization has been found. In addition, this paper also explores how to apply the weight data normalization method to the data pre-processing in Nonnegative Matrix Factorization, and analyzes the proportion of unbalanced development index in eight speciality classification of urban construction to discuss the general direction that urban construction which needs to be focused on.

Key Words: Nonnegative Matrix Factorization; Data Normalization; Urban Construction

目 录

Contents

第一章　绪论

随着中国城市建设的快速发展以及逐渐成熟，人民的生活质量和国家总体的经济发展都得到了快速地提升。城市建设的好坏，不仅影响着民生和经济发展，同时还与交通便利、公共安全、节水节能、环境保护等各个方面甚至各个产业发展息息相关。然而，随着劳动力和先进技术向我国各个城市的涌进，带来经济的快速发展的同时，迅速城市化带来的基础设施人均资源不足、水资源污染、能源紧缺等一系列“城市病”问题也成为了城市建设中的巨大挑战。

如今，城市的前期规划、中期建设和后期的优化均在稳定进步，以此来解决城市建设中产生的各种矛盾和问题。而处在信息时代，每个城市的信息智能化管理已经变成了从城市建设中尤为需要重视的一步。由此产生了智慧城市(Smart City)[1] 的概念——利用信息创新技术，智能化服务于城市建设和管理中，以提升资源利用的效率和改善城市生活的质量。近年来，随着智能城市的推广和城市信息化建设不断发展，城市各建设方面数据系统也越来越完善。然而，在计算机的数据处理能力和优势逐渐显露出来的当下，以往的人工计算或者不够高效和准确的算法面对大数据库的后期信息处理依然仍存在一些困扰 [2]。在智慧城市的大背景下，将科学的数据分析方法运用于城市建设中，无疑能有效促进信息智能服务的可持续发展、数据资源的可用性和建设工作效率的提升，提高城市各组织部门的资源共享、协同工作，促进城市的发展和建设。

对于数据处理有很多方法，为实现信息智能化、动态化和处理的科学性，通常会考虑统计计算、可视化分析、云计算技术等逐渐成熟的计算机技术 [3]，而在人工智能技术成熟以后，机器学习、神经网络、遗传算法等可能也会在城市建设的数据处理中得到广泛运用。

然而，如今城市建设的数据利用现状还处于探索和优化阶段，许多地方还需改进并且真正运用和帮助到城市建设中。除了前期的数据收集和统计不够完全，还存在部分缺值的问题以外，对于后期数据处理技术还不够成熟和普遍运用，可以说还处在不断完善数据库阶段，同时，目前对于许多数据指标还没有分级处理的标准，使得数据在前期量化和预处理时举步维艰 [1]。因此，为了向真正的“智慧城市”迈进，寻找一种适合城市建设的优质算法是当前值得探索的方向。目前已有学者应用深度学习 [4]、DBSCAN 空间聚类算法 [5] 等算法分析城市建设数据。

非负矩阵分解 (Nonnegative Matrix Factorization，简称 NMF) 是将非负矩阵进行非负约束的低秩分解，可以运用于文本分析与聚类、图像检索和复原、语音识别、信号分离、网络安全、生物医学工程和化学工程等各个方面。

本文分为六个部分：

第一部分为绪论部分，介绍了本文的研究背景和意义，阐述城市建设与计算机相结合的研究现状以及全文以非负矩阵分解为中心的研究方法和论文框架。

第二部分为算法介绍，介绍了非负矩阵分解的来源和 MU 算法的原理、五种除量纲的数据预处理的方法、两种结果分析时用的分类准则和数据来源。

第三部分为非负矩阵分解在城市建设中的应用，主要是做城市建设的算法实现和总体分析。

第四部分为赋权值归一化非负矩阵分解在城市建设中的应用，与第三部分的不同在于该部分考虑对不同数据进行赋权处理再实现 MU 算法。

第五部分是 NMF 在城市排水和污水处理中的应用，开展仅限于排水和污水处理类别的非负矩阵分解，并对其结果进行解读。

第六部分是对本文的小结和对未来有关研究的展望。

第 二章　相关算法和数据来源

2.1 非负矩阵分解

1999 年 Lee 和 Seung 在自然杂志上发表了一篇文章《Learning the parts of objects by nonnegative matrix factorization》[6]，首次提出非负矩阵分解算法，并介绍其能够运用于人脸识别和提取文本语义特征以及异于主成分分析法 (Principal Components Analysis) 和矢量量化法 (Vector Quantization) 的地方。其基本思想是对非负矩阵进行非负约束的降维分解，具体过程如下：

对于已知的所有矩阵元素非负的矩阵 $V = (V_{ij})_{n\times m}$，求特征矩阵 $W = (w_{ij})_{n\times r}$ 和系数矩阵 $H = (h_{ij})_{r\times m}$，其中 W, H 的所有元素仍为非负，使得 $V \approx WH$ 的逼近误差最小，即 $V_{ij} \approx \sum_{k=1}^{r} w_{ik}h_{kj}$，使

$$\min_{W\geqslant 0, H\geqslant 0} f(W, H), \tag{2.1}$$

本文 $f(W, H) = \frac{1}{2} \parallel V - WH \parallel_F$。$h_{ij}$ 的大小体现出 W 第 i 列对 V 第 j 列的贡献大小。

本文非负矩阵分解算法采用经典的 MU 算法 [6]：

```
W=rand(n,r);
H=rand(r,m);
for i=1:iter
    H=(H.*(W'*V))./(W'*W*H+eps);
    W=(W.*(V*H'))./(W*H*H'+eps);
end
```

其中 $iter$ 为迭代次数，r 为分解系数 [7]，eps 为机器零，目的是避免出现分母为零的情形。

2.2 数据的收集和处理

数据取自中华人民共和国住房和城乡建筑部采集的 2018 年城乡建筑统计年鉴 [8]，本文仅选取城市的相关数据，其中有九个有关城市建设的专业类别（公用设施，建设用地情况，能源，供水和节水，交通建设，排水和污水处理，园

林绿化，环境卫生，集中供热），每个专业类别都有不等量的指标，共 231 个指标，除去有相关性和具有叠加性的一些指标，剩下 143 个指标，关系如下：

表 2.1

专业类别	各类别指标
公用设施	建成区供水管道密度，人均道路面积，燃气普及率等共 15 个指标
建设用地情况	商业服务业设施用地面积，耕地面积，居住用地面积等共 9 个指标
能源	人工煤气生产能力，天然气储气能力，液化石油气销售气量等共 22 个指标
供水和节水	供水总量，节水措施投资总额，工业水重复利用量等共 20 个指标
交通建设	桥梁数，道路照明灯盏数，地下综合管廊长度等共 14 个指标
排水和污水处理	污水排放量，市政再生水利用量，污水处理厂座数等共 22 个指标
园林绿化	公园个数，绿地面积，公园面积等共 9 个指标
环境卫生	道路清扫保洁面积，生活垃圾无害化处理能力，公厕数等共 12 个指标
集中供热	蒸汽供热能力，热电厂热水供热总量，一级供热管道长度等共 20 个指标

数据包括 673 个城市，对缺值部分标以 0 处理后，即得到 673×143 的数据矩阵。

2.3 几种数据预处理方法

对于不同指标的数据，数据量纲和数量级差异很大，因此有必要对数据进行预处理。在做数据预处理的时候，去除量纲和数量级带来的误差是十分重要的一步，常见的方法是数据归一化和标准化。有效的预处理方法可将原数据转换为具有某种特征或性质的数据，使得每个指标特征对结果贡献相同，甚至可以提高算法收敛速度 [9]，因此数据预处理在算法的实现中尤为重要。设原数据矩阵为 $V = (V_{ij})_{n\times m}$，处理过后的矩阵为 $v = (v_{ij})_{n\times m}$，以下列出了几种数据归一化的方法：

1. 经典归一化：对 n 个指标的对应 m 个数据分别取平均值 $a_i, i = 1, 2, ..., n$，处理方式为原数据各数值和对应平均值的差的绝对值除以指标各项数值之和。归一化公式为

$$v_{ij} = \frac{|V_{ij} - a_i|}{\sum\limits_{k=1}^{m} V_{ik}}, \qquad a_i = \frac{1}{m}\sum_{k=1}^{m} V_{ik} \tag{2.2}$$

所求得的矩阵 v 仍为非负矩阵，且各元素数值控制在 [0,1] 之间。这种方法通常用在数值相差较大的数据中，即极差不够稳定的情况。归一化后所

得的矩阵 v 主要体现各元素数值与该数据对应指标的总体水平之间的的差异。

2. min-max 标准化：各个特征向量被去掉其原来的取值量纲, 统一被重新赋予新的数据量纲，处理方法为原数据各数值减去该数据所属专业类别的最小值的差除以极差。标准化公式为

$$v_{ij} = \frac{V_{ij} - min\{V_{ik}\}_{k=1}^{m}}{max\{V_{ik}\}_{k=1}^{m} - min\{V_{ik}\}_{k=1}^{m}} \tag{2.3}$$

所求得的矩阵 v 仍为非负矩阵，且各元素数值控制在 [0,1] 之间。这种方法适用于数值比较集中，各指标极差较稳定、不会很大的情况下。

3. z-score 标准化：给予原始数据的均值和标准差进行数据的正态标准化处理，处理后的原始数据变为符合标准正态分布的数据, 即平均值为 0、标准差为 1。数据标准化公式为

$$v_{ij} = \frac{V_{ij} - \mu_i}{\sigma_i} \tag{2.4}$$

其中 μ_i, σ_i 分别为第 i 个指标的对应的平均值和标准差。

4. 平均值归一化：样本 $V = (V_{ij})_{n\times m}$，对 n 个指标的对应 m 个数据分别取平均值 $a_i, i = 1, 2, ..., n$ ，公式为

$$v_{ij} = \frac{V_{ij}}{a_i}, \qquad a_i = \frac{1}{m}\sum_{k=1}^{m} V_{ik} \tag{2.5}$$

所求得的矩阵 v 仍为非负矩阵，各元素数值控制在 [0,1] 之间，且满足 $\sum_{k=1}^{m} v_{ik} = 1, i = 1, 2, ..., n$。这种方法是在除量纲的基础上最好的保留了数据的原始特征，比较适合需要保留原始数据状态的算法。

5. 中值归一化 (Median Normalization)：与第二种方法相似，数据被重新赋予数据量纲，但前者数据列中最大值为 1，最小值为 0，后者着重于体现数据与指标中位数的差距。数据归一化公式：

$$v_{ij} = \frac{V_{ij} - mid\{V_{ik}\}_{k=1}^{m}}{max\{V_{ik}\}_{k=1}^{m} - min\{V_{ik}\}_{k=1}^{m}} \tag{2.6}$$

其中 $mid\{V_{ik}\}_{k=1}^{m}$ 表示第 i 个指标的对应的中位数，所求得的矩阵 v 各元

素数值控制在 [-1,1] 之间。

本文采用了经典归一化、平均值归一化以及改进的中值归一化进行数据预处理。

2.4 分类准则

当实现非负矩阵分解后，我们需要对分解得到的特征矩阵 W 和系数矩阵 H 进行分析。本文中所用的实验结果分析的方法是：对于特征矩阵，主要是通过提取出其每个特征列中若干个突出项，就可以得到比较具有代表意义的特征指标；对于系数矩阵，则是利用合适的分类准则 [10]，选取出比较需要改进的城市或者比较具有特征代表性的城市进行进一步分析和解读。

设数据矩阵为 $V = (V_1, V_2, \cdots, V_m)$，通过非负矩阵分解得到的特征矩阵为 $W = (W_1, W_2, \cdots, W_r)$，系数矩阵为 $H = (h_{ij})_{r\times m}$，以下是两种分类准则：

1. 最大元准则 (MAX rule)：计算 H_j 中的最大元所对应的行数 i

$$i = \arg\min_k \{h_{kj}\}, \tag{2.7}$$

则 W_i 为 V_j 的近似特征列。

2. 最近子空间准则 (Nearest-Subspace，简称 NS)：计算回归残差矩阵 $R = (r_{ij})_{r\times m}$

$$r_{ij} = \frac{1}{2} \parallel V_j - W_i h_{ij} \parallel_2^2, \tag{2.8}$$

再计算出 R 中每一列的最小元所对应的行数 i

$$i(j) = \arg\min_k \{r_{kj}\}, \tag{2.9}$$

则 $W_{i(j)}$ 为 V_j 的近似特征列。

本文将会应用这两种方法对分解结果进行分析和解释。

第三章　非负矩阵分解在城市建设中的应用

随着我国城市建设的不断发展，建设幸福城市逐渐变得重要了。基础设施建设是人们生活的地基，推进人与自然的和谐发展是城市建设的趋势，对于我国城市建设的优化，从不局限于单个指标或方向的改善，而是需要各个部门的协调工作，同时对严重滞后的方面进行查漏补缺。将非负矩阵分解应用于城市建设中，既有利于快速筛选出城市建设中需要努力的方向，又使得相关决策具有科学性，避免盲目。

本章利用经典归一化对 2018 年城市建设的数据进行非负矩阵分解的实现：设组成城市建设的原始数据矩阵 $V=(V_{ij})_{n\times m}$，其中 $n=143,\quad m=673$，然后利用公式 (2.2) 进行归一化处理得到矩阵 $v=(v_{ij})_{n\times m}$。对于所得到的新的数据矩阵 v 体现了原数据与其对应指标的全国平均水平的距离，所以通过这种归一化处理之后，选取合适的分解系数进行 MU 算法的实现，可以提取出我国在城市建设上各地发展不够均衡的特征建设指标——通过分析分解结果，可以得出我国哪些城市的哪方面建设比较好或者比较落后，以及有哪些是大部分城市已经淘汰掉的相关产业。

3.1 各参数的选取

在本文所考虑的算法中，需要提前预设分解系数 r 和迭代次数 $iter$，由算法的一致收敛性可知，r 和 $iter$ 均是越大时分解结果越逼近数据矩阵。在实际计算中，通常取经验值来选取合适的 r 和 $iter$，本文用以下方法来确定这两个参数：

假设我们实际能接受计算机在计算该算法时，迭代次数最多为 200 次，算法所得结果最坏能接受分解系数为 50，分解的误差函数 $error$ 为

$$error=\frac{\parallel V-WH\parallel_F}{\parallel V\parallel_F} \tag{3.1}$$

我们希望实验结果的最终误差足够小来分别确定合理的参数 r 和 $iter$：

1. 在选取参数 r 时，我们设 $iter=200$，用 MATLAB 做出随着 r 值的增大，$error$ 的变化，函数变化图如下所示

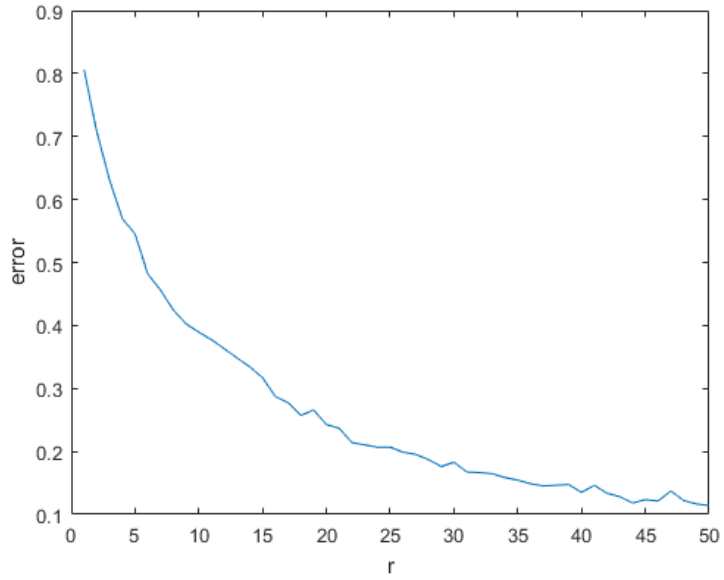

图 3.1: **分解系数** r

因此，考虑到数据取自九个城市建设的不同专业类别，并且分解系数在超过 10 后，$error$ 减小速度明显下降。因此，我们在之后的实验中设分解系数为 $r = 10$ 。

2. 在选取参数 $iter$ 时，我们设 $r = 200$ ，用 MATLAB 做出随着 $iter$ 值的增大 $error$ 的变化，函数变化图如下所示

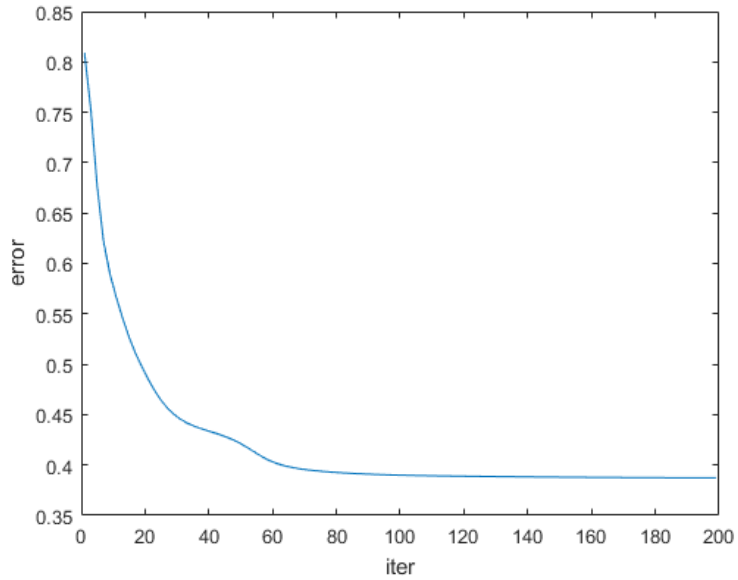

图 3.2: **迭代次数** $iter$

图中，$error$ 在迭代 100 次后基本上不再减小，因此我们在之后的实验中

设迭代次数为 $iter = 100$ 。

3.2 数值实验

通过一番对两个参数的优化后，现在令分解系数为 $r = 10$ ，迭代次数为 100 次，利用 MATLAB 实现 MU 算法。对数据矩阵进行 100 次 MU 算法计算出的误差波动如下图所示

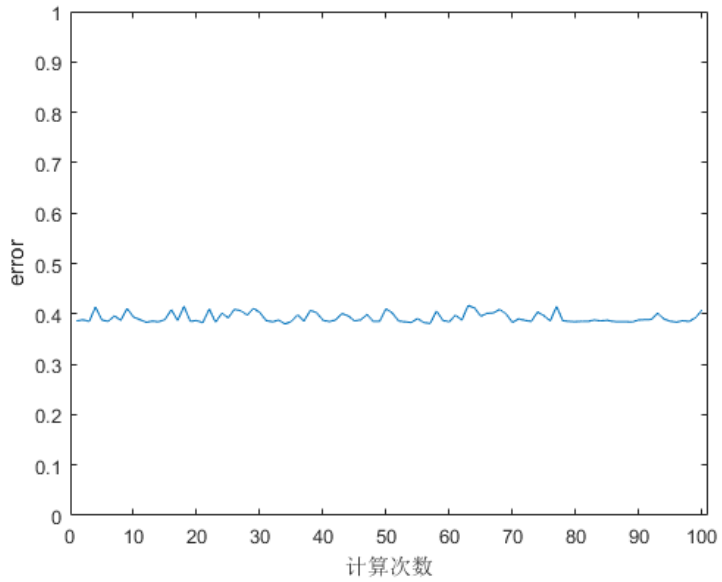

图 3.3: 100 次试验结果误差对照

由此可知算法结果的 $error$ 几乎稳定在 0.40 左右，误差数结果总方差为 9.4472×10^{-5} ，可以看出实验结果的误差比较稳定。

对于分解得到的特征矩阵 W ，每列选取 2 个突出项。以下是 3 次实验结果得到的 W 各列突出项情况：

表 3.1

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
1	人均公园绿地面积，液化石油气储气能力	供居民家庭人工煤气总量，人工煤气燃气损失量	轨道交通固定资产投资，地下综合管廊固定资产投资
2	天然气储气能力，排水污泥处理固定资产投资	锅炉房蒸汽供热能力，锅炉房蒸汽供热总量	供居民家庭人工煤气总量，人工煤气燃气损失量
3	供燃气汽车液化石油气总量，大桥及特大桥数	供水固定资产投资，轨道交通固定资产投资	供集中供热天然气总量，市政再生水管道长度

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
4	液化石油气燃气损失量，地下水供水生产能力	排水污泥处理固定资产投资，天然气储气能力	供燃气汽车液化石油气总量，液化石油气加气站座数
5	供居民家庭人工煤气总量，人工煤气燃气损失量	排水再生水利用固定资产投资，液化石油气燃气损失量	自备水计划用水用户，地下综合管廊长度
6	人工煤气生产能力，人工煤气自制气量	自备水计划用水用户，地下综合管廊长度	排水再生水利用固定资产投资，液化石油气燃气损失量
7	超计划定额用水量，节水措施投资总额	人工煤气生产能力，人工煤气自制气量	人工煤气生产能力，人工煤气自制气量
8	自备水计划用水用户，地下综合管廊长度	超计划定额用水量，节水措施投资总额	锅炉房蒸汽供热能力，锅炉房蒸汽供热总量
9	锅炉房蒸汽供热能力，锅炉房蒸汽供热总量	锅炉房热水供热总量，市政再生水管道长度	排水污泥处理固定资产投资，天然气储气能力
10	排水再生水利用固定资产投资，供集中供热天然气总量	供燃气汽车液化石油气总量，大桥及特大桥数	超计划定额用水量，节水措施投资总额

由表 3.1 可以看出，稳定且频繁出现在突出项的指标是人工煤气、天然气、液化石油气的相关指标，说明全国关于发展能源（燃气）方面的大部分指标普遍不够均衡。产生这样的原因有很多，主要是由于各城市之间消费需求和各技术成熟度的不同：各地因为生活方式和人口数量的不同，需求量就有很大的区别，而市场需求在很大程度上还会影响到技术的进步，同时，技术的成熟与否又决定了发展速度与质量，间接决定了供应量。比如，在液化石油气的储气能力上，按照最大元准则对实验结果进行计算，可以得到北京、上海、广州（分别对应分解得到的系数矩阵 H 的第 1、168、448 列）明显高于其他城市，与实际情况相符，这不仅是因为人口基数大，更是因为这些城市在液化石油气的储存技术已经足够成熟。

将数值实验结果的能源（燃气）方面的相关指标进行总结可得如下表格：

表 3.2

燃气类型	相关指标（出现次数）
人工煤气	供居民家庭总量（3），燃气损失量（3），生产能力（3），自制气量（3）
天然气	储气能力（3），供集中供热总量（2）
液化石油气	燃气损失量（3），供燃气汽车总量（3），汽车加气座数（1），储气能力（1）

从表 3.2 中可以看出天然气的相关指标占比最少，人工煤气的相关占比最高，这刚好验证了我国燃气发展结构的现状：由于天然气作为绿色能源之一, 已成为城市燃气的主要供能来源，正在逐步代替液化石油气，但由于液化石油气的供能技术成熟、灵活机动、基建成本低等优点在中国仍具有稳定市场 [11]；而人工煤气由于其极大的污染性和较高的运输成本，已逐渐退出我国城市能源市场，从而导致我国区域性发展不均衡，即大部分城市已经淘汰人工煤气，然而对于油气资源短缺的地方仍有使用人工煤气。

对于不同类型的燃气，区域性发展不均的相关专项指标不同，这取决于不同燃气侧重的服务领域。比如对于天然气，关于各城市的储气能力有很大的差别，这一是说明了我国仍有大量城市的天然气储气技术需要改善，储气能力的而建设还需要大力发展，二是从侧面体现了我国天然气的储气设施建设和气源应急能力不足，从而会影响其稳定供能；而对于液化石油气，其燃气损失量的不同说明，其液化石油气的能量利用率还有待提高。当然，为了实现国家能源的可持续发展，仅以燃气提供城市中人民生活和各行业发展是远远不够的，城市能源建设仍应大力发展更多可再生能源或低碳能源来普遍代替不可再生能源的燃烧发电。

除了能源方面，实验结果突出的节水方面也值得关注。由于我国水少人多的国情，节水技术的发展一直很被重视，主要侧重于农业灌溉技术方面的节约用水。而如今，水资源的日渐紧张和城市化的迅速发展之间的矛盾也突现出来，因此在城市规划建设时，节约用水的相关工作也开始逐渐被考虑甚至被关注。

对于三次数值实验结果，利用最近子空间准则求出 10 个特征列分别对应的特征指标有突出贡献的城市个数，统计结果如下

表 3.3

特征列	1	2	3	4	5	6	7	8	9	10
第一次城市个数分布	7	10	27	7	25	4	443	11	127	12
第二次城市个数分布	24	16	202	9	14	11	4	42	338	13
第三次城市个数分布	495	23	36	10	11	8	4	37	9	40

从表 3.3 可以看出不同指标两极分化的城市个数。比如，第一次数值实验中有 7 个城市对第 1 列特征列贡献较大，即有 7 个城市在人均公园绿地面积、液化石油气储气能力两个指标上与全国平均水平相差较远。对于统计结果来说，城市个数越少，说明对于相关指标，大部分城市的发展水平比较平均，只需要对这些个别落后城市进行技术上的引进或资源上的补足即可。比如三次实验中都是仅有 4 个城市与全国平均水平相差较远的人工煤气生产能力、人工煤气自制气量两个指标（分别对应第一次实验的第 6 列、第二次实验的第 7 列、第三次实验的第 7 列特征列），通过分解结果的系数矩阵 H 可知 4 个城市分别为邹平市（山东省）、嘉峪关市（甘肃省）、安阳市（河南省）、大连市（辽宁省），说明这四个城市还没有完全摒弃人工煤气的使用。而对于城市个数较多的情况，说明我国在这些指标上两极分化严重，导致城市与城市之间发展水平相差甚远。

同时，该数值实验仍存在非计算误差可以排除：集中供热方面的指标特征比较明显，但产生此现象的主要原因是南北差异过大而导致的，即在南方地区一般一年四季都不需要供热，因此数据存在缺值或为零较多；另外，其他数据中也存在着不少的缺值部分，除了因为地方相关的统计记录工作不到位外，还有一些近几年的新增指标，这些指标只有少数城市有数据，多数城市还没有数据。因此如果在数值实验前，这些数据均以 0 值代替，必然影响最终结果，还需要考虑改进方法。

3.3 数值实验的改进

3.3.1 改进中值归一化

对于暂时没有办法在短期内解决的部分数据缺失所带来的实验误差下，我们还是希望通过数学方法和算法的改进来减少这类误差。因此在做数据预处理的时候，考虑利用中值归一化方法来代替经典归一法。然而对于传统的中值归

一化，处理过后的数据矩阵元素有正有负，所以这里对其进行改进使得处理过后的数据矩阵各元素非负：设已知组成城市建设的原始数据矩阵 $V=(V_{ij})_{n\times m}$，进行归一化处理后得到矩阵为 $v=(v_{ij})_{n\times m}$，v 可由关系式得出：

$$v_{ij}=\frac{|V_{ij}-mid\{V_{ik}\}_{k=1}^{m}|}{max\{V_{ik}\}_{k=1}^{m}-min\{V_{ik}\}_{k=1}^{m}} \tag{3.2}$$

其中 $mid\{V_{ik}\}_{k=1}^{m}$ 表示第 i 个指标的对应的中位数，这样所求得的矩阵 v 仍为非负矩阵，且各元素数值控制在 [0,1] 之间。

对于新的数据矩阵 v，它体现了原数据与其对应指标的全国中间水平的距离，这种归一化的方法与上一种不同的在于，它不仅能体现我国在城市建设上区域性不平衡发展的特征建设指标，同时相比于常规归一化还能减少大部分为 0 的数据指标的影响，即减少前期做缺值归零处理的影响。因此通过这种归一化处理之后，选取合适的分解系数进行 MU 算法的实现后，提取出的特征建设指标能够更准确地体现的我国在城市建设上区域不平均衡方面。

3.3.2 数值实验

对于原数据矩阵，考虑到在供热方面的数据差异主要是由于南北的差异过大产生的，因此需要除去有关供热的指标，剩下一共 123 个指标，即原数据矩阵为 $V=(V_{ij})n\times m$，其中 $n=123,\quad m=673$。

在数据归一化后，我们仍然需要确定分解系数和迭代次数两个参数，选取方式同上，非负矩阵分解误差 $error$ 的变化随分解系数 r 与迭代次数 $iter$ 变化的函数图像分别如下图所示：

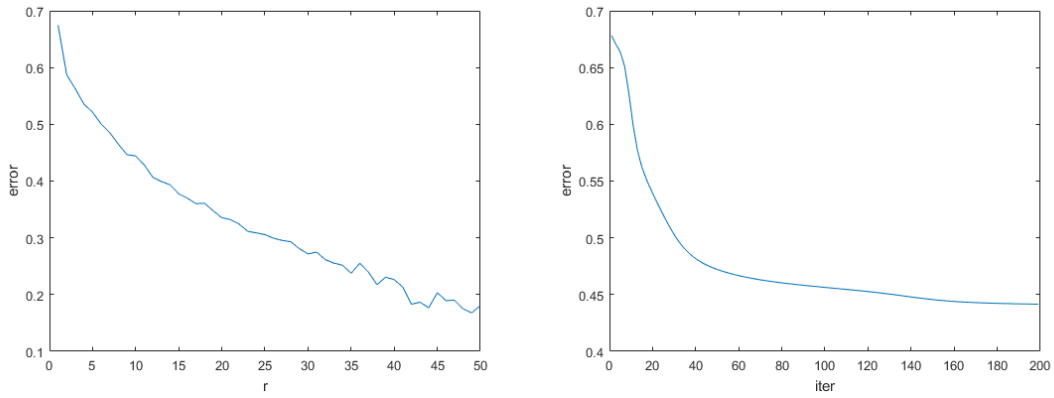

图 3.4: **参数优化曲线**

同样可合理地取分解系数为 $r=10$，迭代次数为 100 次，利用 MATLAB 实现 MU 算法。对数据矩阵进行 100 次 MU 算法计算出的误差波动如下图所示

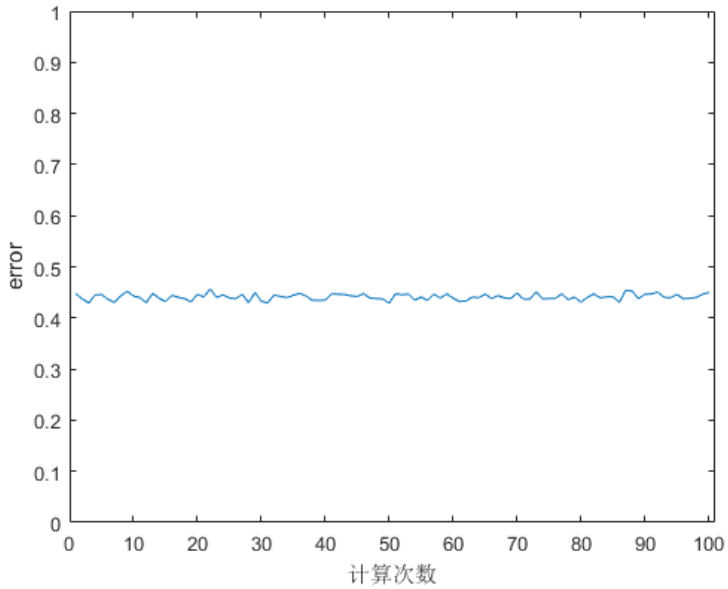

图 3.5: 实验结果误差的波动

由此可知算法结果的误差几乎稳定在 0.45 左右，误差结果总方差为 3.3502×10^{-5}。

对于分解结果中的特征矩阵 W，观察其各列选取 2 个突出项，取 3 次试验结果，统计结果如下

表 3.4

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
1	排水固定资产投资，污水处理固定资产投资	建成区路网密度，建成区排水管道密度	地下综合管廊长度，新建地下综合管廊长度
2	供燃气汽车天然气总量，天然气汽车加气站座数	市政再生水生产能力，用于再生水管道长度	人均日生活用水量，建成区供水管道密度
3	再生水利用固定投资总额，园林绿化固定投资总额	公园个数，公园面积	建成区路网密度，建成区排水管道密度
4	燃气汽车普及率，建成区路网密度	绿地与广场用地面积，耕地面积	绿地与广场用地面积，公厕数
5	绿地与广场用地面积，耕地面积	工业新水取用量，实际用水重复利用量	市政再生水生产能力，用于再生水管道长度
6	液化石油气销售总量，供居民家庭液化石油气总量	供居民家庭液化石油气总量，人行道面积	天然气汽车加气站座数，道路面积

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
7	实际用水重复利用量，工业用水重复利用量	供人工煤气管道长度，人工煤气销售总量	液化石油气销售总量，公园个数
8	地下综合管廊长度，新建地下综合管廊长度	排水固定资产投资，污泥处置固定资产投资	污水处理固定资产投资，液化石油气汽车加气站座数
9	桥梁数，污水管道长度	轨道交通固定资产投资，道路桥梁固定资产投资	人口密度，建成区路网密度
10	建成区供水管道密度，建成区排水管道密度	人口密度，绿地与广场用地面积	实际用水重复利用量，工业用水重复利用量

由表 3.4 可知，突出项里有关能源的指标数依然是占比较高的一部分，说明能源区域性发展不平衡确实是城市建设中十分需要注意的方面。同时，我们还能看到一些比较稳定出现的指标：建成区路网密度，绿地与广场用地面积，公园座数，实际用水重复利用量等，它们也是普遍性会影响城市建设的因素，这些都是与人民生活幸福度密切相关甚至有直接影响的实际指标，这不仅反映出城市建设的质量会直接影响到人民生活质量，还说明我国仍有许多城市的基础设施建设不够完善，绿化水平还有待提高等问题。

对于三次数值实验，利用最近子空间准则求出 10 个特征列分别对应的特征指标有突出贡献的城市个数，统计结果如下

表 3.5

特征列	1	2	3	4	5	6	7	8	9	10
第一次城市个数分布	1	17	1	325	53	11	29	7	8	221
第二次城市个数分布	445	1	8	66	23	5	12	1	13	99
第三次城市个数分布	13	136	201	10	1	71	12	4	199	26

从表 3.5 可分析出，建成区路网密度是大部分城市稳定出现的突出指标（分别对应第一次实验的第 4 列、第二次实验的第 1 列、第三次实验的第 3 列特征列），说明我国在城市交通基础设施建设两级分化严重，还有明显短板，需要加强建设。通过对非负矩阵分解所得的系数矩阵 H 进行分析，可将单个突出指标按照其对整体数据的影响由大到小进行排列，排名前八的相关指标和代表城市（取三个）如下所示

表 3.6

排名	指标	代表城市
1	建成区路网密度	天水市（甘肃省）；耒阳市（湖南省）；郑州市（河南省）
2	建成区排水管道密度	珠海市（广东省）；来宾市（广西省）；德惠市（吉林省）
3	建成区供水管道密度	昆玉市、可克达拉市（新疆省）；中山市（广东省）
4	燃气汽车普及率	康定市（四川省）；安宁市（云南省）；泸水市（云南省）
5	人均日生活用水量	北屯市（新疆省）；三亚市（海南省）；庄河市（辽宁省）
6	人口密度	聊城市（山东省）；神木市（陕西省）；醴陵市（湖南省）
7	绿地与广场用地面积	五指山市、琼海市（海南省）；阳江市（广东省）
8	天然气汽车加气站座数	天津市；绥化市（黑龙江省）；信阳市（河南省）

实验结果选出的相关城市是相关建设发展水平与全国中间水平相距较远的城市。其中，部分为急需侧重发展的城市，结果可以体现这些城市中十分需要侧重发展的相关指标对应的建设方面，比如甘肃省的天水市需要发展道路建设，而新疆省的昆玉市和可克达拉市需要发展供水建设方面；而另一部分为发展较好的城市，比如广东省的珠海市和中山市。这些城市不仅仅只是相关指标的对应值较高或者较低，还体现出的是比较有代表性的数据特征，可以通过研究算法分析出来的发展较好的城市来给其他区域提供某方面可借鉴的建设建议，比如海南省五指山市在绿化方面的建设就可供其他城市学习。

第 四章　赋权归一化在城市建设中的应用

本章主要是探讨非负矩阵分解在城市建设涉及的八个专业类别（公用设施，建设用地情况，能源，供水和节水，交通建设，排水和污水处理，园林绿化，环境卫生）的运用。

4.1 赋权归一化

在本文前面所用的数据归一化方法，对于各项指标都是无差别对待，但事实上根据国情、民意、发展重点、优势劣势等各方面因素，对城市建设的不同方面不同时期侧重点是不一样的，比如自 21 世纪以来，节能减排就受到了广泛的重视。另外，虽然对于各指标采取无差别计算，但在大方向中却变成了有差别，比如能源方面就共考虑了 22 个建设指标，而园林绿化仅占 8 个指标，这种不等量的指标数导致了一些重要指标特征在专业类别的总结分析中被忽视，从算法上说，就是要排除因可记录数据指标量多而影响数值实验的结果。而对各指标进行赋权，就能很有效的解决以上考虑到的问题。对于不同的指标，我们可以对其在数据分析时进行不同权重的赋值，即增大或减少其对算法结果的影响，从而达到强调某项指标或无差别计算等实验目的 [12]。

考虑到我们希望赋权值法所用的数据要尽可能保留原始特征，故这里采用平均值归一化方法，设组成城市建设的原始数据矩阵 $V=(V_{ij})_{n\times m}$，则利用公式 (2.5) 对数据矩阵进行平均值归一，可得到数据矩阵 $v=(v_{ij})_{n\times m}$ 。

设赋权值 $\{k_i\}_{i=1}^n$ ，转换的数据矩阵为 $v=(v_1,v_2,\cdots,v_n)^T$ ，其中 v_i 代表了数据矩阵 v 的第 i 个行向量，即第 i 个指标的数据向量；k_i 代表了对第 i 个指标的赋权值，则令

$$v'=\begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix}, \qquad V_i'=k_iv_i \tag{4.1}$$

可得到最终的数据矩阵 $V'=(V_{ij}')_{n\times m}$ ，然后即可进行非负矩阵分解。

4.2 数值实验

实验的原始数据仍是利用除去集中供热相关指标后的 123 个指标、673 个城市的数据矩阵 $V=(V_{ij})_{123\times 673}$，由 15 个公用设施相关指标、9 个建设用地情况相关指标、22 个能源相关指标、20 个供水和节水相关指标、14 个交通建设相关指标、22 个排水和污水处理相关指标、9 个园林绿化相关指标、12 个环境卫生相关指标这八个专业类别按顺序组成。为了让这八个专业类别组成的数据矩阵在非负矩阵分解时能无差别地提取特征，可设置权重：

$$k_i\begin{cases}\frac{1}{15}, & \text{i=1,2,...,15}\\ \frac{1}{9}, & \text{i=16,17,...,24}\\ \frac{1}{22}, & \text{i=25,26,...,46}\\ \frac{1}{20}, & \text{i=47,48,...,66}\\ \frac{1}{14}, & \text{i=67,68,...,80}\\ \frac{1}{22}, & \text{i=81,82,...,102}\\ \frac{1}{9}, & \text{i=103,104,...,111}\\ \frac{1}{12}, & \text{i=112,113,...,123}\end{cases} \tag{4.2}$$

然后对数据矩阵 V 进行赋权值归一化：

$$v_{ij}=\frac{k_iV_{ij}}{a_i},\qquad a_i=\frac{1}{m}\sum_{k=1}^{m}V_{ik} \tag{4.3}$$

得到转换后的数据矩阵 $v=(v_{ij})_{n\times m}$。首先确定分解系数和迭代次数两个参数的优化，选取方式同上，非负矩阵分解的 $error$ 变化随分解系数 r 与迭代次数 $iter$ 变化的函数图像分别如下图所示：

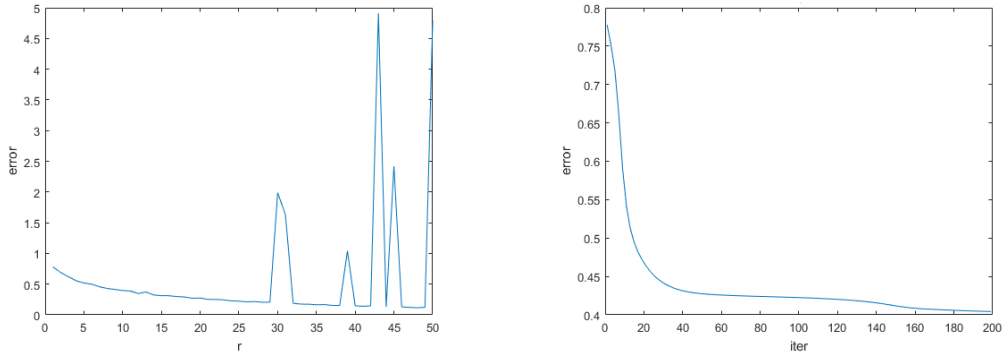

图 4.1: **参数优化曲线**

由结果可知随着分解系数的增大，误差范数甚至不够稳定，原因在于平均值归一法保留了原始数据的特征，则原来为零值的数据元素经过归一化处理后仍然为零，这在计算机的算法实现中会产生误差，且使得算法结果不稳定。

同样取分解系数为 $r = 10$，迭代次数为 100 次，利用 MATLAB 实现 MU 算法，对数据矩阵进行 100 次 MU 算法计算出的误差波动如下图所示

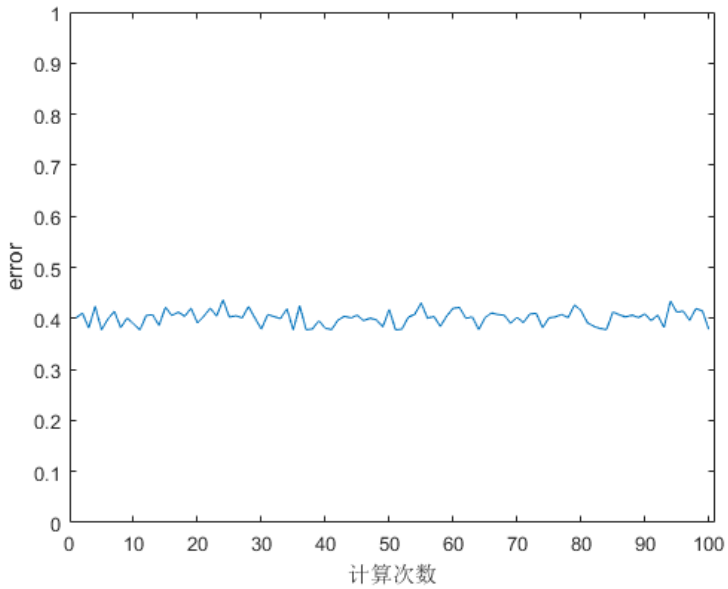

图 4.2: 实验结果误差的波动

可发现算法结果的误差几乎稳定在 0.4 左右，误差结果总方差为 2.0810×10^{-4}。

对于分解结果中的特征矩阵 W，仍然选取 2 个突出项，取 3 次误差范数小于 0.4 的试验结果，统计结果如下

表 4.1

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
1	工业用地面积，耕地面积	液化石油气燃气损失量，园林绿化固定资产投资	天然气储气能力，污泥处理排水量
2	液化石油气燃气损失量，再生水利用量	道路交通设施用地面积，耕地面积	道路交通设施用地面积，耕地面积
3	供居民家庭天然气量，天然气燃气损失量	供居民家庭天然气量，天然气燃气损失量	供居民家庭天然气量，天然气燃气损失量
4	天然气储气能力，污泥处理排水量	人均道路面积，人均公园绿地面积	超计划定额用水量，节水措施投资总额

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
5	自备水计划用水用户，地下综合管廊长度	人工煤气生产能力，人工煤气自制气量	自备水计划用水用户，地下综合管廊长度
6	供燃气汽车液化石油气量，大桥及特大桥数量	超计划定额用水量，节水措施投资总额	液化石油气燃气损失量，再生水利用量
7	超计划定额用水量，节约用水量	地下综合管廊固定资产投资，供燃气汽车液化石油气量	地下综合管廊固定资产投资，轨道交通固定资产投资
8	人工煤气生产能力，人工煤气自制气量	液化石油气燃气损失量，再生水利用量	供燃气汽车液化石油气量，大桥及特大桥数量
9	地下综合管廊固定资产投资，轨道交通固定资产投资	自备水计划用水用户，地下综合管廊长度	人均道路面积，人均公园绿地面积
10	人均道路面积，人均公园绿地面积	天然气储气能力，污泥处理排水量	人工煤气生产能力，人工煤气自制气量

将表 4.1 的结果统计并转为各专业类别的指标占比的饼状图，呈现为下图所示：

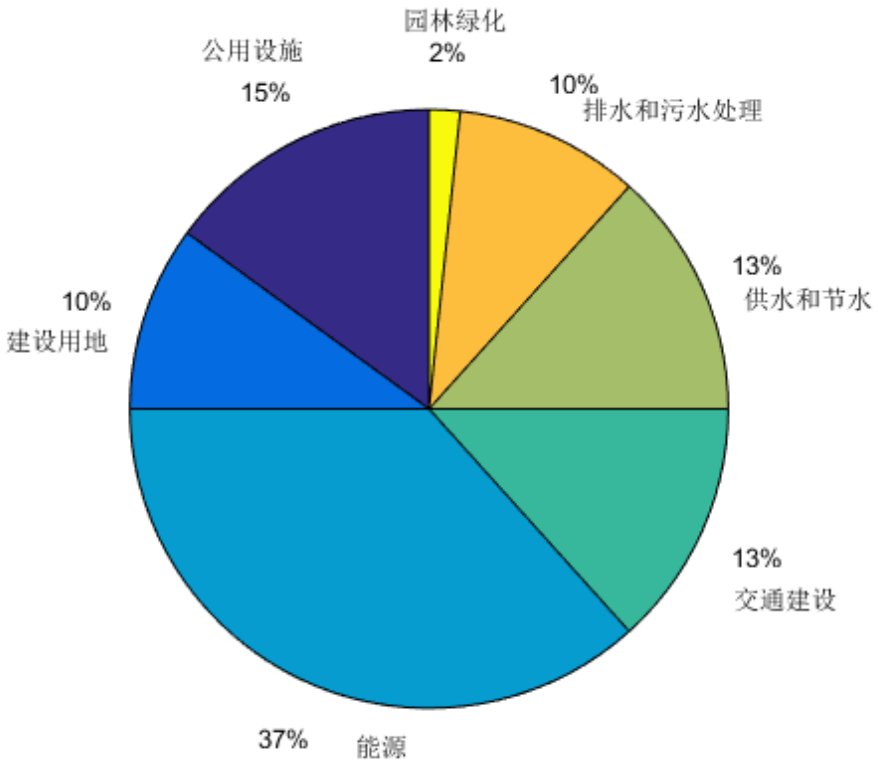

图 4.3: 指标占比

由图 4.3 可知，能源是如今城市建设最应该关注的部分，它的区域不平衡问题十分突出。同时，占比第二的公共设施说明了我国在公共基础设施建设上还需加强，虽然我国的基础设施建设投入了大量的财力物力和时间成本，但是随着人们生活质量的不断提高和城市化的迅速蔓延，人民对公共设施的需求也是呈指数型增长，要解决这一矛盾、跟上人民生活质量的脚步，仍需做大量的努力。

此外值得关注的一点是，环境卫生相关指标占比为 0，这说明这一方面在全国各地发展均衡，可以体现出我国在 2018 年及之前在环境卫生方面做出的全国范围内的努力颇有成效。

将数值实验得到的特征矩阵 W 和系数矩阵 H 进行信息整理和提取转换可得下图

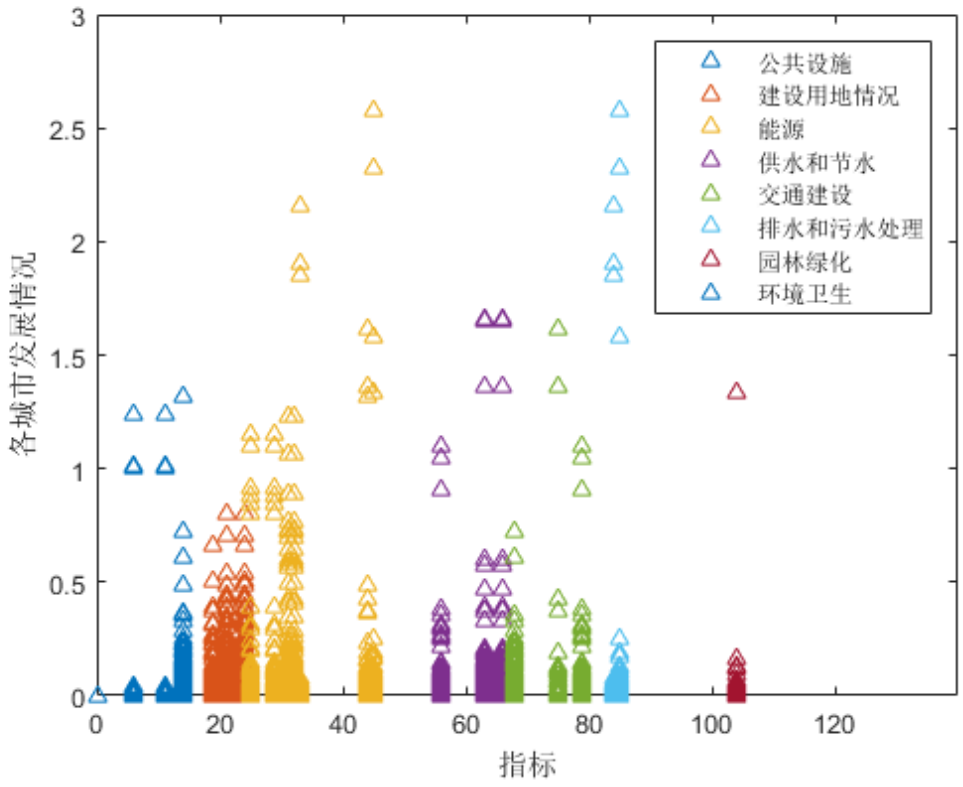

图 4.4: 各城市发展分布情况

图 4.4 体现的是非负矩阵分解提取出来的与全国平均水平相差甚远的城市发展情况，横轴是城市建设的八个专业类别包含的 123 个指标（按顺序），数轴是提取出来的代表城市与平均水平的差距。可以看到有关能源、排水和污水处理的相关指标的代表城市比较分散，即存在较多对相应的特征列有突出贡献的城市，说明在关于能源、排水和污水问题，我国存在小部分城市发展跟不上平

均水平，所以目前可以侧重于扶持这些个别城市的发展和引进相关技术；而关于建设用地在图中的点分布较为集中，说明建设用地情况在我国是除了环境卫生外最平衡的大类方向，即全国发展较为平衡，则目前可以侧重于各个城市对于技术的运用和政策的调整更多的是需要考虑根据当地的发展情况和特点来进行区域性调整。

第 五章　非负矩阵分解在城市排水和污水处理中的应用

考虑到不同专业类别的指标之间关联性不高、量级差别大等问题，在面对城市建设的某一方面数据时可以考虑做进一步数据分析。排水和污水处理作为城市建设的重要组成部分 [13]，这里单独对其相关指标对应的数据矩阵进行非负矩阵分解的实现；同时，因为排水排污方面缺值较少，对其做算法的实现得出的结果更为符合实际意义，若将来数据普遍不缺值时，可以考虑将非负矩阵分解运用于其他专业类别中分别进一步讨论。考虑排水和污水处理中 14 个指标（包括污水排放量，污水管道长度，雨水管道长度，雨污合流管道长度，建成区排水管道长度，污水处理厂座数，污水处理厂处理能力，污水处理厂数量，污水处理厂干污泥产生量，污水处理厂干污泥处置量，污水处理总量，市政再生水生产能力，市政再生水利用量，市政再生水管道长度），归一化方法仍然采用经典归一化，然后用 MATLAB 实现 MU 算法。

5.1 数值实验

这里同样要确定算法中预设参数，这里稍有不同的是，由于指标数较少，我们可以对分解系数进行算法稳定性分析。对于 14 个数据指标，设原数据矩阵为 $V = (V_{ij})n \times m$，其中 $n = 14, \quad m = 673$，令迭代次数为 80，非负矩阵分解 $error$ 的变化随分解系数 r 变化的函数图像分别如下图所示：

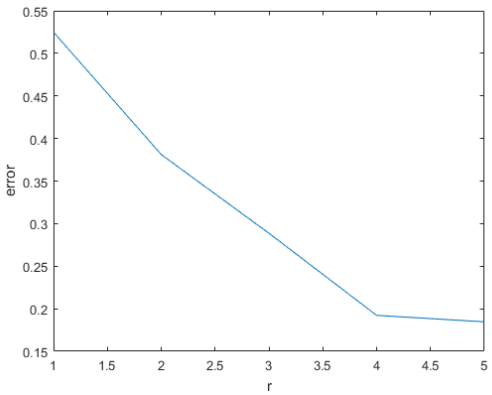

图 5.1: **分解系数** r

由此可知，在仅关心误差的情况下依然是分解系数越大误差越小。接下来做稳定性分析：对比分解系数为 $r = 2, 3, 4$ 下非负分解矩阵的实验结果，即对比

特征矩阵的特征列的稳定性。对于分解结果中的特征矩阵 W ，观察其各列选取 2 个突出项，取 3 次试验结果，统计结果如下

表 5.1 ： $r=2$

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
1	市政再生水生产能力，市政再生水管道长度	污水处理厂数量，市政再生水利用量	市政再生水生产能力，市政再生水管道长度
2	污水处理厂数量，市政再生水利用量	市政再生水生产能力，市政再生水管道长度	污水处理厂数量，市政再生水利用量

表 5.2 ： $r=3$

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
1	市政再生水生产能力，市政再生水管道长度	污水处理厂干污泥处置量，市政再生水利用量	污水处理厂干污泥处置量，市政再生水利用量
2	污水排放量，市政再生水利用量	市政再生水生产能力，市政再生水管道长度	污水管道长度，雨水管道长度
3	污水处理厂干污泥处置量，污水处理厂数量	污水排放量，污水处理厂干污泥处置量	市政再生水生产能力，市政再生水管道长度

表 5.3 ： $r=4$

特征列	第一次实验的突出项	第二次实验的突出项	第三次实验的突出项
1	污水处理厂干污泥处置量，市政再生水生产能力	雨水管道长度，市政再生水管道长度	污水处理厂干污泥处置量，市政再生水利用量
2	污水排放量，污水处理总量	污水处理厂干污泥处置量，市政再生水利用量	污水处理厂干污泥处置量，市政再生水生产能力
3	雨水管道长度，市政再生水管道长度	市政再生水生产能力，市政再生水管道长度	市政再生水生产能力，市政再生水管道长度
4	污水处理厂干污泥处置量，市政再生水利用量	污水排放量，污水处理厂数量	污水排放量，污水处理总量

5.2 结果分析

由以上表格可以看出，随着分解系数的增加，非负矩阵分解的分解结果越来越不稳定，但是误差范数越来越小，这说明我们在做非负矩阵分解的时候需要权衡结果的稳定性误差值之间的冲突。

由实验结果可知，当 $r=2$ 时，实验的分解结果比较稳定。可以看出在所分解出来的四个突出指标中，再生水的利用在污水处理中比较重要：污水经过简

单处理后，可以应用于水质没那么高的行业中，甚至于随着技术的提高，再生水的质量越来越高，可以被越来越多的行业所利用，以此来减少水资源的消耗。

对于 $r = 2$ 的三次数值实验，利用最近子空间准则求出 2 个特征列分别对应的特征指标有突出贡献的城市个数，统计结果如下

表 5.4

特征列	1	2
第一次城市个数分布	150	523
第二次城市个数分布	525	148
第三次城市个数分布	137	536

可以看出污水处理厂数量，市政再生水利用量这两个指标在城市排水和污水处理中属于区域不平衡问题较大的两个指标。

第六章　总结与展望

6.1 总结

近年来，城市人口不断增长和迅速城市化所带来的“城市病”问题之间的矛盾已经变成了我国城市建设中必须思考和解决的挑战。而在信息化时代的今天，“智慧城市”的需求越来越高。运用合理的技术处理和分析由此产生的数据库是当前城市建设的重要方向。

本文展示了如何将非负矩阵分解运用于城市建设的总体数据分析和部分相关类别数据分析（如排水和污水处理），对比了三个不同归一化方法对非负矩阵分解的实现的区别，该算法可对城市建设的数据库实行动态化分析，提取出当前需要侧重发展的特征指标。对于 2018 年城市建设年鉴的分析可以得出的结论是，能源方面是城市建设中最需要关注的重点，此外节水和公共设施的建设也应该得到改善，从民生意义上解释，就是一个直接影响水资源紧缺的现状，一个直接影响生活在城市的人民的生活质量。对城市建设的总体数据分析有利于寻找各类指标之间的关系，有助于之后的协同工作，而它的弊端在于由量级的不同和指标间的联系不大的时候数值实验误差较大，因此当追求精度或希望对某一专业类别进行数据分析时可以考虑单独对其分析的方式。而对于希望对某些城市建设方向进行侧重或从轻处理时，可以在数据预处理时进行赋权。以上方法均可以从不同意义上提取出所有指标中，因区域性发展不平衡而需要被侧重视的指标，从而有利于城市建设的均衡发展。

6.2 展望

由于一些指标数据对于城市建设比较重要，但如今仍有许多城市缺值，这给算法带来的误差较大，可以说完整的数据是算法里不可缺少的起步点，因此，政府的实时监测工作会直接决定后期的算法数据分析，这方面还需加大监测力度才能为今后的云计算等发展“智慧城市”奠定基础。同时，对于城市建设数据的前期数据预处理可以不仅停留在用统一的归一化处理，对不同类型的指标可采取不同预处理的方式，当城市建设的各类指标数据可以进行科学地分级处理时，算法的误差会减少许多。

我们正在见证城市建设的稳定发展和计算机技术的广泛普及，如今城市建设的信息化已成为必然。本文对非负矩阵分解在城市建设中的应用的探索有望

为今后“智慧城市”的发展提供帮助，为我国的城市建设出一份力。

参考文献

[1] 郁健红. 大数据在智慧城市建设中的作用和深度应用 [J]. 民舍，2020:1.

[2] 冯超. 大数据时代的城市规划 [J]. 建材与装饰，2019:88-89.

[3] 缪梦仪. 大数据可视化分析践行建筑业大数据智能 [J]. 中国建设信息化，2019:24-27.

[4] 林伯钧. 基于深度学习算法的实时手机数据分类及其对智慧城市建设的影响研究 [J]. 科技创新导报，2019:240-242.

[5] 李新延，李德仁. DBSCAN 空间聚类算法及其在城市规划中的应用 [J]. 测绘科学，2005:51-53.

[6] D.D.Lee, H.S.Seung. Learning the parts of objects by nonnegative matrix factorization[J], Nature, 1999:788–791.

[7] 钱瀚. 非负矩阵分解在医疗检测报告中的应用 [D]. 厦门大学：厦门大学数学科学学院，2014.

[8] 中华人民共和国住房和城乡建设部. 2018 年城乡建设统计年鉴 [Z]. http://www.mohurd.gov.cn/xytj/tjzljsxytjgb/index.html.

[9]Johannes Burdack, Fabian Horst, Sven Giesselbach, etal. Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning[J]. Frontiers in bioengineering and biotechnology，2020:0-260.

[10] Yifeng Li, Alioune Ngom. Classification approach based on non-negative least squares [J]. Neurocomputing. 2013, 118(11): 41-57.

[11] 陈姿屹. 城市燃气利用现状及其应用 [C]. 姜东琪. 2017 中国燃气运营与安全研讨会论文集，中国江苏常州：煤气与热力杂志社，2017:30-35.

[12] 王兴权，蔡福海，罗建国. 起重机桁架臂整体可靠性评估指标的归一化方法研究 [J]. 建筑机械，2019:72-74.

[13] 季诚云，丁玉珍. 我国城镇污水处理厂建设运行现状及存在问题分析 [J]. 建材与装修，2020:231-232.

附 录

附录 A

```
V=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\1.xlsx','1','B6:EN678'); %读入数据
V=V'; %数据矩阵V的转置
n=size(V,1); %行数n
m=size(V,2); %列数m
v=zeros(n,m); %创建零矩阵v
for i=1:n
    for j=1:m
        v(i,j)=abs(V(i,j)-mean(V(i,:)))/sum(V(i,:)); %将数据矩阵V归一化后存入矩阵v
    end
end
r=10; %分解系数r
iter=100; %迭代次数iter
H=abs(rand(r,m)); %起始系数矩阵H
W=abs(rand(n,r)); %起始特征矩阵W
for i=1:iter
    H=(H.*(W'*v))./(W'*W*H+eps);
    W=(W.*(v*H'))./(W*H*H'+eps); %迭代分解矩阵v
end
norm(v-W*H,'fro')/norm(v,'fro') %用F范数求误差
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\1W.xlsx',W); %存储特征矩阵
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\1H.xlsx',H); %存储系数矩阵
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\1v.xlsx',v');
```

附录 B

```
V=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\2中值归一化\2.xlsx','1','B6:DT678'); %读入数据
```

```
V=V'; %数据矩阵V的转置
n=size(V,1); %行数n
m=size(V,2); %列数m
v=zeros(n,m); %创建零矩阵v
for i=1:n
    for j=1:m
        v(i,j)=abs((V(i,j)-median(V(i,:))))/(max(V(i,:))-min(V(i,:))); %将数据矩阵V归一化后存入矩阵v
    end
end
r=10; %分解系数r
iter=100; %迭代次数iter
H=abs(rand(r,m)); %起始系数矩阵H
W=abs(rand(n,r)); %起始特征矩阵W
for i=1:iter
    H=(H.*(W'*v))./(W'*W*H+eps);
    W=(W.*(v*H'))./(W*H*H'+eps); %迭代分解矩阵v
end
norm(v-W*H,'fro')/norm(v,'fro') %用F范数求误差
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\2中值归一化\2W.xlsx',W); %存储特征矩阵
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\2中值归一化\2H.xlsx',H); %存储系数矩阵
```

附录 C

```
V=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\3赋权归一化\3.xlsx','1','B6:DT678'); %读入数据
V=V'; %数据矩阵V的转置
n=size(V,1); %行数n
m=size(V,2); %列数m
K=[15,9,22,20,14,22,9,12]; %不同专业类别的指标数
k=zeros(n,1);
l=0;
for i=1:8
    for j=1:K(i)
        k(j+l)=1/K(i); %录入指标权重
    end
```

```
    l=l+K(i);
end
a=zeros(1,n);
for i=1:n
    a(i)=mean(V(i,:)); %求各指标平均数
end
v=zeros(n,m);
for i=1:n
    for j=1:m
        v(i,j)=k(i)*V(i,j)/a(i); %将数据矩阵V赋权归一化后存入矩阵v
    end
end
r=10; %分解系数r
iter=100; %迭代次数iter
H=abs(rand(r,m)); %起始系数矩阵H
W=abs(rand(n,r)); %起始特征矩阵W
for i=1:iter
    H=(H.*(W'*v))./(W'*W*H+eps);
    W=(W.*(v*H'))./(W*H*H'+eps); %迭代分解矩阵v
end
norm(v-W*H,'fro')/norm(v,'fro') %用F范数求误差
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\3赋权归一化\3W.xlsx',W); %存储特征矩阵
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\3赋权归一化\3H.xlsx',H); %存储系数矩阵
```

附录 D

```
r=1:50; %分解系数取值可接受范围
iter=200; %迭代次数为200
error=zeros(1,size(r,2)); %创建误差error
for i=1:size(r,2)
    H=abs(rand(r(i),m));
    W=abs(rand(n,r(i)));
    for j=1:iter
        H=(H.*(W'*v))./(W'*W*H+eps);
        W=(W.*(v*H'))./(W*H*H'+eps); %迭代求分解矩阵
    end
```

```
    error(i)=norm(v-W*H,'fro')/norm(v,'fro'); %求对应误差的F范数
end
plot(r,error); %画图
xlabel('r'); %标x轴参数
ylabel('error'); %标y轴参数
```

附录 E

```
r=10; %取r值为10
iter=1:2:200; %迭代次数取值可接受范围
error=zeros(1,size(iter,2)); %创建误差error
h=abs(rand(r,m)); %取定起始系数矩阵
w=abs(rand(n,r)); %取定起始特征矩阵
for i=1:size(iter,2)
    H=h;
    W=w;
    for j=1:i
        H=(H.*(W'*v))./(W'*W*H+eps);
        W=(W.*(v*H'))./(W*H*H'+eps); %迭代求分解矩阵
    end
    error(i)=norm(v-W*H,'fro')/norm(v,'fro'); %求对应误差的F范数
end
plot(iter,error); %画图
xlabel('iter'); %标x轴参数
ylabel('error'); %标y轴参数
```

附录 F

```
u=1:100; %计算次数
uu=zeros(1,100);
for j=1:100
    H=abs(rand(r,m));
    W=abs(rand(n,r));
    for i=1:iter
        H=(H.*(W'*v))./(W'*W*H+eps);
        W=(W.*(v*H'))./(W*H*H'+eps); %迭代求分解矩阵
    end
    uu(j)=norm(v-W*H,'fro')/norm(v,'fro'); %用F范数求误差
```

```
end
var(uu)
plot(u,uu); %画图
xlabel('计算次数');
ylabel('error');
axis([0 101 0 1]); %控制坐标轴范围
```

附录 G

```
H=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\第二次重新\1H(2).xlsx','Sheet1',
读入系数矩阵
v=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\第二次重新\1v.xlsx','Sheet1','A1
读入数据
W=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\第二次重新\1W(2).xlsx','Sheet1',
读入特征矩阵
v=v';
n=size(v,1); %行数n
m=size(v,2); %列数m
r=size(H,1); %分解系数r
R=zeros(r,m); %创建零矩阵R
for i=1:r
    for j=1:m
        R(i,j)=(norm(v(:,j)-W(:,i)*H(i,j),2))^2/2; %求回归残差
    end
end
I=zeros(1,m); %创建零矩阵I
for j=1:m
    [i,I(j)]=min(R(:,j)); %求R中每一列最小元对应行数
end
t=zeros(1,r); %创建零矩阵t
for i=1:r
    for j=1:m
        if I(j)==i
            t(i)=t(i)+1; %求I中对应不同特征列的个数
        end
    end
end
```

```
xlswrite('D:\课件\2020第二学期\毕业论文\数值实验\1归一化实验\第二次重新\1I(2).xlsx',I); %
存储矩阵I
```

附录 H

```
H=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\3赋权归一化\3H.xlsx','Sheet4','A1:YW30'); %
读入3个分解得到的系数矩阵
W=xlsread('D:\课件\2020第二学期\毕业论文\数值实验\3赋权归一化\3W.xlsx','Sheet4','A1:AD123'); %
读入3个分解得到的系数矩阵
r=10;
m=size(H,2);
n=size(W,1);
x=0;
y=x;
for i=1:3
    for j=1:r
        [a,b]=sort(W(:,(i-1)*r+j)'); %将W每一列进行排序
        for k=1:m
            x=[x,b(n-1:n)]; %存储m次最大的两个值
        end
    end
end
for i=1:3*r
    for j=1:m
        for k=1:2
            y=[y,H(i,j)]; %存储2次H各系数
        end
    end
end
x1=0;x2=0;x3=0;x4=0;x5=0;x6=0;x7=0;x8=0;
y1=0;y2=0;y3=0;y4=0;y5=0;y6=0;y7=0;y8=0;
for i=1:size(x,2)
    if x(i)<=15
        x1=[x1,x(i)];y1=[y1,y(i)]; %分开存储
    else
        if x(i)<=24
            x2=[x2,x(i)];y2=[y2,y(i)];
        else
```

```
            if x(i)<=46
                x3=[x3,x(i)];y3=[y3,y(i)];
            else
                if x(i)<=66
                    x4=[x4,x(i)];y4=[y4,y(i)];
                else
                    if x(i)<=80
                        x5=[x5,x(i)];y5=[y5,y(i)];
                    else
                        if x(i)<=102
                            x6=[x6,x(i)];y6=[y6,y(i)];
                        else
                            if x(i)<=111
                                x7=[x7,x(i)];y7=[y7,y(i)];
                            else x8=[x8,x(i)];y8=[y8,y(i)];
                            end
                        end
                    end
                end
            end
        end
    end
end
plot(x1,y1,'^'); %画图
hold on;
plot(x2,y2,'^');hold on;
plot(x3,y3,'^');hold on;
plot(x4,y4,'^');hold on;
plot(x5,y5,'^');hold on;
plot(x6,y6,'^');hold on;
plot(x7,y7,'^');hold on;
plot(x8,y8,'^');hold on;
xlabel('指标');
ylabel('各城市发展情况');
legend('公共设施',1,'建设用地情况',2,'能源',3,'供水和节水',4,'交通建设',5,'排水和污水处理',6,'园林绿化',7,'环境卫生',8);
axis([0 140 0 3]); %控制坐标轴范围
```