# **GWR分析**

目的：**用于聚类的指标筛选**——食品加工厂的空间分布主要受何种**城市形态因素**的影响？

相关性分析：去除共线性变量

计算VIF（方差膨胀因子，Variance Inflation Factor），去除VIF过高的变量，

比如consumption\_count 与 retail\_count，

网格筛选：由于数据量较少，为了保证回归模型的稳定性，只针对有加工厂落位/与之邻近1km范围内的网格展开分析

将剩余变量进行**标准化**后，进行OLS线性回归分析，从OLS回归结果来看：

| **字段名** | **含义解释** |  |
| --- | --- | --- |
| **Variable** | 回归模型中使用的自变量名 |  |
| **Standardized Coef (SC)** | 标准化回归系数 (SC)，表示变量与目标变量之间的标准化后的关联强度 |  |
| **p-value** | 系数显著性检验的p值，小于0.05通常被认为统计显著 |  |
| **Adjusted R²** | 模型整体调整决定系数，表示模型对因变量的整体解释能力 |  |

1 【整体结果】：

OLS Regression Results ============================================================================== Dep. Variable: y **R-squared: 0.080**

Model: OLS **Adj. R-squared: 0.052** Method: Least Squares F-statistic: 2.870 Date: Wed, 04 Jun 2025 **Prob (F-statistic): 0.00126**

2 【单变量分析】输出 Standardized Coef (SC)（标准化回归系数）+ p值 + R² ：

Variable  **Standardized Coef (SC) p-value** Adjusted R²

0 Intercept -8.196568e-17 1.000000 0.052124

**整体解释力较低，仅约5.2%变异能被模型中变量解释，说明这些变量可能并不能很好地预测当前工厂数量的空间分布，可能存在其他更关键的变量尚未考虑。**

| **Adjusted R² 范围** | **解释** | **在城市/地理研究中的常见评价** |
| --- | --- | --- |
| **> 0.5** | 很高，模型具有很强解释力 | 很罕见，通常出现在封闭系统建模中 |
| **0.3 – 0.5** | 中等偏上，可接受 | 在城市尺度能达到这个水平就很不错了 |
| **0.1 – 0.3** | 中等，具有一定解释力 | 是城市空间模型中最常见的区间 |
| **0.05 – 0.1** | 弱，但仍可能具有分析价值 | 若能显著，说明变量方向和逻辑有意义 |
| **< 0.05** | 非常弱，解释能力有限 | 变量选择可能不足，或目标变量太复杂 |

目前 OLS 的 **Adjusted R² ≈ 0.052**（5.2%），最佳组合也只提高到 **约 0.059**（5.9%）：

* **解释能力偏弱**，但不意味着无用；
* 在 **城市空间研究、地理分布预测、工厂选址** 等现实复杂场景中，这类模型**很难出现高 R²**，因为：
  + 工厂选址受不可量化因素影响极大（政策、历史、个体意愿）；
  + 你目前只用了土地、道路、服务类“可观测”物理变量；
  + 而没有用经济、法律、政策等“不可观测”变量。

1 farming\_area\_norm 5.113939e-02 0.396674

2 aquaculture\_area\_norm 5.757688e-03 0.920678

3 pasture\_area\_norm -1.072621e-02 0.876367

4 energy\_area\_norm -5.166964e-02 0.308780

5 transportation\_area\_norm 1.061984e-01 0.039176

6 retail\_count\_norm 7.949510e-02 0.241484

7 public\_services\_norm 4.767986e-02 0.513195

8 road\_length\_norm -2.091998e-01 0.023375

9 built\_area\_norm 2.047125e-01 0.028325

10 entropy\_production 9.583607e-02 0.169940

11 entropy\_landuse 1.743733e-01 0.003400

**显著变量总结**：

* **交通设施面积**（SC = 0.106，p < 0.05）：  
  工厂通常靠近交通设施，符合一般生产性用地布局规律。
* **道路长度**（SC = -0.209，p < 0.05）：  
  道路过多区域反而工厂较少，这可能表示道路密集区偏向居住或商业区，而非工业区。
* **建筑面积**（SC = 0.205，p < 0.05）：  
  建筑面积大的地方工厂数量更多，说明工业区建设强度较高。
* **土地利用多样性**（SC = 0.174，p < 0.01）：  
  土地利用多样性较高的地区吸引更多工厂落户，可能是多功能区或综合工业区。

❌ **不显著变量**：

* 农业面积、水产面积、牧场面积、能源面积、零售、公服、生产类型多样性均未表现出显著的直接影响

3 【多变量分析】 自动进行变量组合筛选，并比较每个组合的表现指标

| **字段名** | **说明和含义** | | | |
| --- | --- | --- | --- | --- |
| **Variables** | **使用的变量组合（多个变量组合在一起用于解释工厂数量）** | | | |
| **Adjusted R²** | **调整后的决定系数，反映该组合对工厂数量整体的解释能力** | | | |
| **AIC** | **赤池信息准则（Akaike Information Criterion）用来比较模型好坏（越低越好）** | | | |
| **排名** | **变量组合** | **Adjusted R²** | **AIC** | **说明与建议** |
| **1** | **energy\_area, transportation\_area, retail\_count** | **0.0594** | **1050.16** | **包含能源面积、交通面积、零售点数** |
| **2** | **transportation\_area, retail\_count, road\_length** | **0.0593** | **1049.21** | **包含交通面积、零售点数、道路长度** |
| **3** | **farming\_area, energy\_area, transportation\_area, retail\_count** | **0.0587** | **1051.42** | **包含农业面积、能源面积、交通面积、零售点数** |

 **交通设施面积**几乎出现在所有最佳组合中，表明它对工厂选址的解释能力稳定而强烈。

 **零售点数量**也频繁出现，说明工厂位置与商业区位的关联较为显著。

 **能源面积**和**道路长度**也频繁出现，表明工业区位置受基础设施建设影响明显。

**保留交通设施面积、零售点数、道路长度**作为下一步建模的核心变量（因其最频繁出现于最佳组合中）**表明这些变量对工厂分布起主导解释作用**

显著性检验（Moran’s I）衡量空间自相关程度

（判断某个变量在空间上是否**“聚集”**或“成团”分布，而不是随机散布的）

结果：Moran’s I=0.05＞0说明

当前的目标变量 factory\_count\_norm 在空间上具有显著的正向空间自相关性。（即“聚集”）

p\_sim=0.001< 0.01，非常显著，说明这种空间聚集结构并非偶然

工厂数量在空间上存在**显著的局部聚集现象**，而不是完全随机；某些局部确实存在工厂聚集“热点区域”，意味着 **OLS 的“全局模型”可能无法捕捉这种差异**；

**用地理加权回归（GWR）是合理且有必要的**

GWR建模分析 **得出每个网格的局部系数、局部 R²、整体 GWR R² 与 AICc**

=== 模型指标对比 ===

GWR AICc : -317.5485

OLS AIC: -439.5348

**整体GWR R2 : 0.3612**

**OLS R2 : 0.0800**

普通回归模型只能解释工厂空间分布的约 **8%** 的变化

GWR 可以解释约 **36%** 的变化，模型解释力提升了**超过4倍，**大幅提升解释能力

**说明**：GWR 显著优于 OLS，变量与目标（工厂数量）之间的关系在空间上是**显著异质的**，工厂分布的驱动因素具有明显的空间异质性。用全局模型（OLS）低估了局部变化。

工厂选址和空间分布不能简单用整体平均规律来解释，必须考虑**空间局部环境因素**。

*GWR 的 AICc 比 OLS 高（更差），这似乎矛盾？*

*实际上这是一个经典情形：****GWR提高了解释力（R² ↑），但模型复杂度也大大增加；****这并不表示 GWR 不好，只说明它“更复杂”*

尽管 GWR 更复杂（AICc 更高），但它提供了更有意义、更精准的空间解释能力。

**工厂空间分布受空间异质因素影响显著，全局OLS模型无法有效捕捉这种空间差异，GWR模型大幅改善了模型表现。尤其是交通面积、道路长度、建筑面积和土地多样性具有显著且空间差异的影响。**

采用了**自适应带宽**（adaptive bandwidth）：GWR 带宽选择结果 (带宽=73.0)，表明中尺度的局部关系最有效

**说明：每个网格单元进行GWR模型拟合时，采用了空间上最近的约73个邻近单元的数据进行局部回归。**每个点的模型系数是通过 **它周围73个点的数据“加权拟合”出来的**；**这体现了数据中捕捉空间异质性的适宜尺度。**

说明数据中**空间异质性是中等尺度的特征。**局部模型能显著提高解释力（R²达36.1%），显然比全局OLS（8%）表现更佳。

GWR 模型中每个变量的局部回归系数可视化

| **功能** | **意义** |
| --- | --- |
| **局部系数地图** | **揭示变量在不同地理区域对工厂分布的影响方向与强度** |
| **局部R²地图（可再添加）** | **发现GWR在哪些区域表现最好，哪里不适合局部建模** |

# **聚类分析**

**1. 数据准备：提取图像的 RGB 特征，并融合 GWR 加权得到的城市语义向量**  
- 图像数据来源于 patch 周边的城市形态渲染图  
- 城市语义向量包括地理加权回归得到的权重指标5项，用于增强聚类的空间语义表达  
  
**2. 特征处理：使用卷积神经网络（CNN）提取图像特征向量512**  
- 使用Rest-Net预训练模型（如、VGG）提取高维图像嵌入，压缩至5维  
- 对语义向量进行标准化处理后拼接进图像特征，拼接后的特征为10维度  
图形用户界面

AI 生成的内容可能不正确。

**以上两个步骤可以参考该技术流程的画法**

**3. 最优聚类设置：使用平均轮廓系数法（Silhouette Score）评估最优聚类数 K**  
- 设置不同的组合参数

 降维维度 ∈ {2, 3, 4, 5, 10}

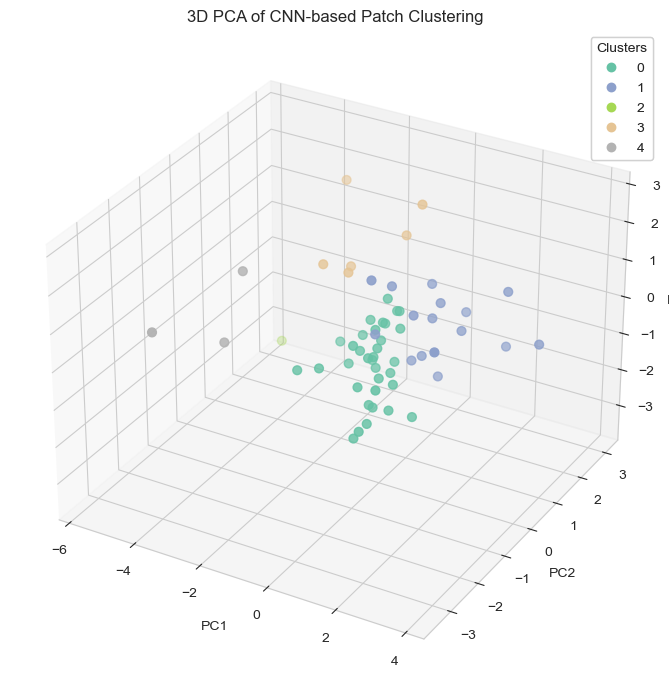
 聚类算法 ∈ {KMeans, SpectralClustering, AgglomerativeClustering}

 聚类数量 K ∈ [3, 14]  
- 选择得分最高或变化平稳的 K 作为最终聚类数  
图表, 折线图

AI 生成的内容可能不正确。

1. 高值：绿色线（Agglomerative）在 PCA=2 和 PCA=3 下，K=3 时 silhouette 得分最高（超过 0.6 和 0.4），远高于其他方法。所以首选Agglomerative + PCA=2 or 3 + K=3

2. 稳定值：Agglomerative + PCA=2：K=3~6 范围内都在 0.3 以上，这表明该组合具有良好的鲁棒性，不是“偶然峰值”

最终我们选择  
**4. 降维与可视化：使用主成分分析（PCA）对高维特征进行映射**  
- 降维后绘制散点图，查看不同聚类的空间分布  
- 分析各主成分的含义（如城市形态 vs. 功能语义）  


PC1 PC2 PC3

feature\_0 0.136513 **0.485102** 0.174019 **图像**

feature\_1 0.080137 0.165901 0.108662

feature\_2 0.244245 -0.129147 -0.076064

feature\_3 -0.181713 0.085610 **0.553074 图像**

feature\_4 -0.010743 -0.332683 -0.047339

feature\_5 **-0.441609** -0.458096 0.305153 **'beta\_trans'交通点的相关性**

feature\_6 -0.264411 **0.538318** 0.207096 **'beta\_retai'零售点的相关性**

feature\_7 -0.434027 0.314089 -0.458275

feature\_8 **0.526216** 0.061083 -0.223204 **'beta\_built'建筑密度相关性**

feature\_9 0.389792 0.021105 **0.498147 'beta\_entro'用地混合度相关性**

如上为3维映射的结果，我觉得太过于复杂了，所以转向2维

**5. 聚类分析：应用 KMeans 等聚类算法对图像+语义向量进行聚类**  
- 保存聚类标签，评估每类的代表性与空间分布

- 将聚类标签映射回原始 patch 图像与 GeoDataFrame 中

图表, 散点图

AI 生成的内容可能不正确。

PC1 PC2

feature\_0 0.136513 0.485102

feature\_1 0.080137 0.165901

feature\_2 0.244245 -0.129147

feature\_3 -0.181713 0.085610

feature\_4 -0.010743 -0.332683

feature\_5 -0.441609 -0.458096

feature\_6 -0.264411 0.538318

feature\_7 -0.434027 0.314089

feature\_8 0.526216 0.061083

feature\_9 0.389792 0.021105

主成分PC1：feature\_8建筑密度，feature\_5交通影响，feature\_7图像特征

则PC1 表达了从 **“高建筑密度” → “高交通影响”** 的一个城市形态语义轴

分布上，**右侧是建筑密度强的区域**，左侧是受交通影响更大的区域。

主成分PC2: feature\_6零售点相关性，feature\_0，7，4都是图像特征

PC2 表达了从 **“零售/商业混合功能区” → “非商业/偏居住”** 的区分。

分布上，**上方**为功能混合或零售密度高的空间，**下方**为低零售相关性的区域。

所以如何理解这几个Cluster

图片包含 图示

AI 生成的内容可能不正确。

**Cluster 0** 均位于中间区域，更偏向于依赖交通（比如能够看到路网），加工建筑往往与黑色附属建筑成组出现，可能是一些传统的工业区或者物流集散地，常与道路邻接，可能是临街厂房。

**Cluster 1 与0相同均位于中间区域，**Cluter1则更多受到建筑密度的影响。场地相对开放

**Cluster 2**只有一个，乡村型或郊区型加工场景，偏向环境嵌入式的小规模加工（这是一个养蜂场）与Cluster 4一样属于“极端样本”

**Cluster 3** 零售相关性高，可能是零售驱动型的城市块

**Cluster 4** 周边要么是大量绿地要么是规则化大格网，商业程度和建筑密度都极低，城市边缘或独立厂房，存在少数 outlier 情况，可能是典型的空间“极端样本”  
【可以直接拿出质心的样本来写结论！】  
**6. 聚类结果输出与解释**  
- 可视化聚类分布、类别图像示例、以及与城市变量的关系

【最后一张图上应该显示我用来聚类的这五个指标：交通设施，道路，建筑，土地混合度，以及零售点（商业）】

地图

AI 生成的内容可能不正确。