PORTFOLIO

SELECTED GIS & SPATIAL DATA SCIENCE PROJECTS (2022-2024)

YUANPENG CHEN | Y.Chen363@lse.ac.uk
MSc. | Geographic Data Science (Expected)
London School of Economics (LSE), London, United Kingdom

CONTENTS

01

SPATIAL DATA MODELING & VISUALIZATION (May, 2023)

Urban Expansion Simulation

OVERVIEW

Dongguan, a city that exemplifies China's rapid urbanization during the Reform and Opening-Up era, has witnessed significant shifts in its urban landscape. This study used the Logistic-Cellular Automata (CA) for two objectives:

1. To understand the historical urban expansion in Dongguan City from 2001 to 2005, offering insights into its urban development trends.
2. To evaluate the performance of the CA model in simulating urban expansion.

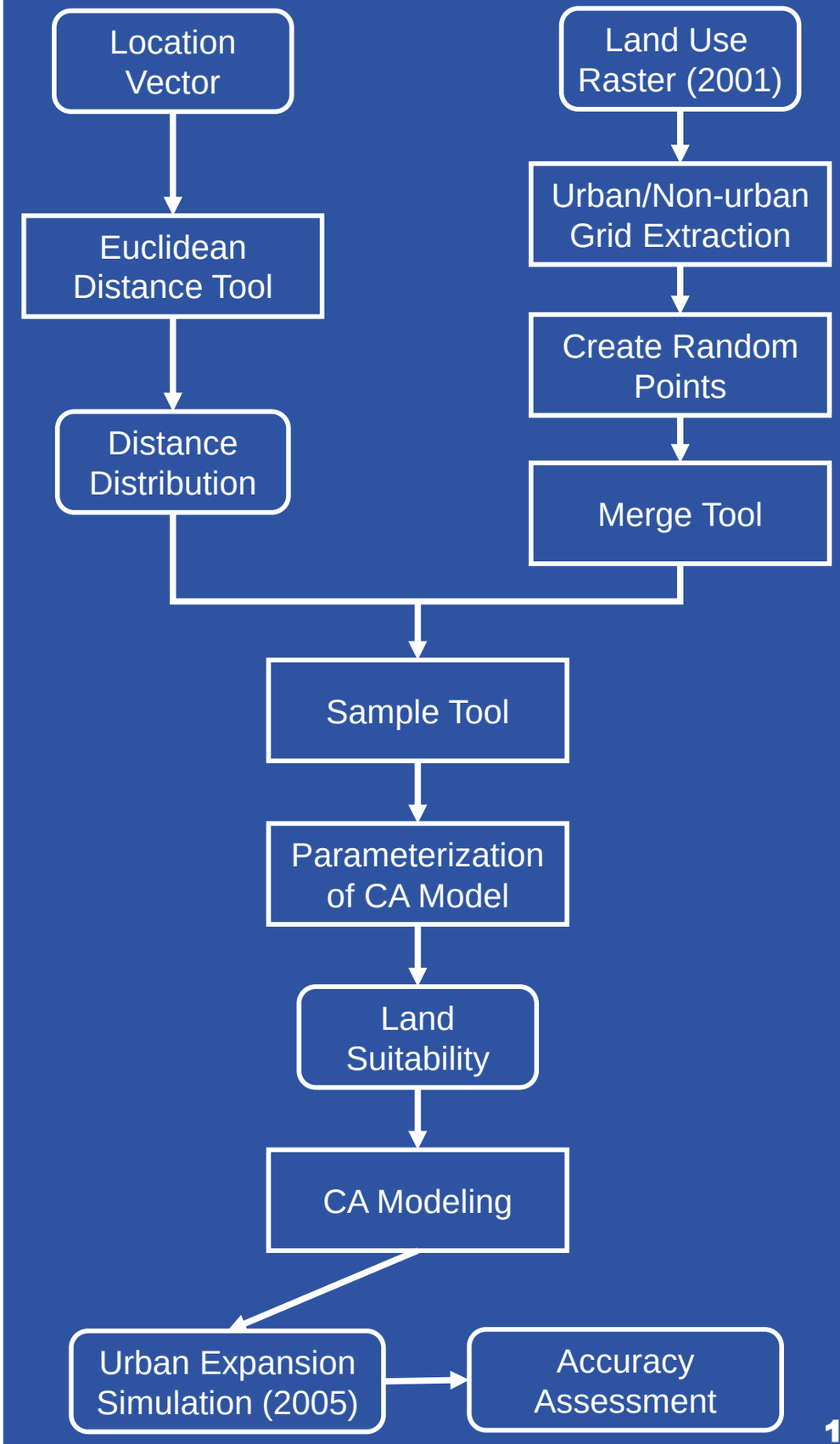
TOOLS & SKILLS

ArcMap, Python, Cellular Automata, SPSS, Logistic Regression

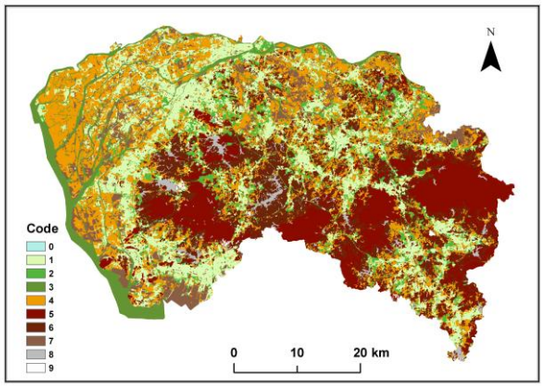
DATASET

1. Dongguan land use raster data for the years 2001 and 2005.
2. Vector data for 5 location factors:
   - City center point
   - Town center points
   - City Road lines
   - Railway lines
   - Expressway lines

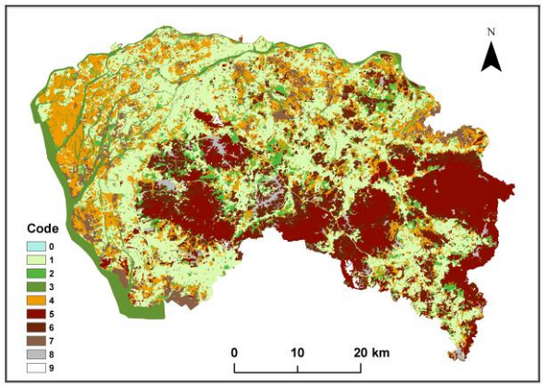
WORKFLOW



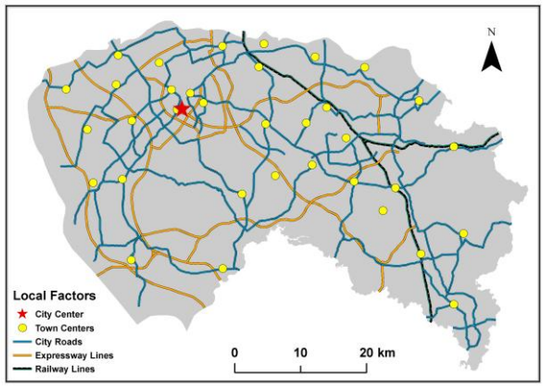
Data Set



Land Use (2001)

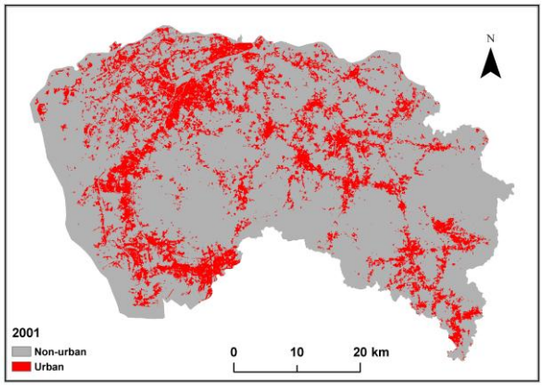


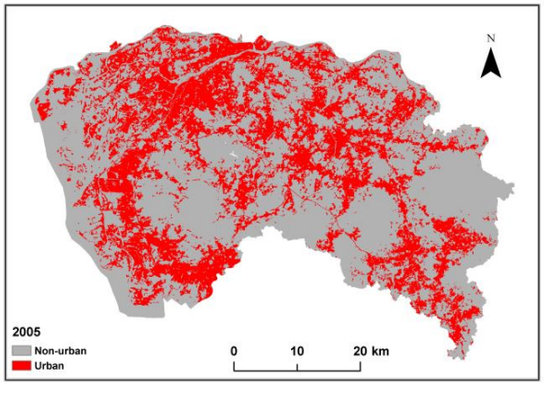
Land Use (2005)

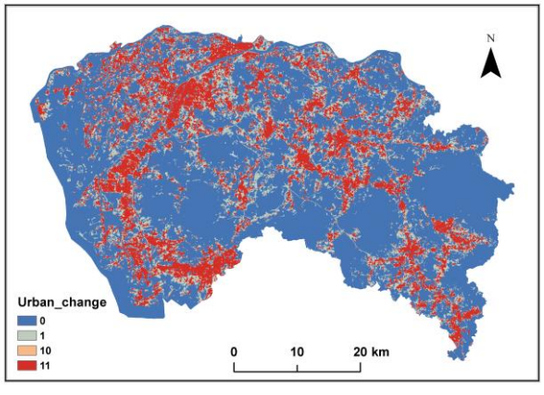


Local Vectors

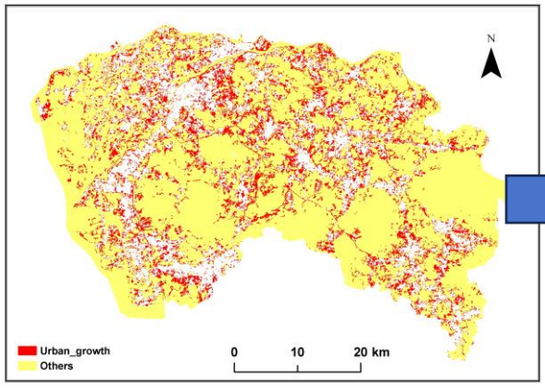
Urban/Non-urban Grids Extraction



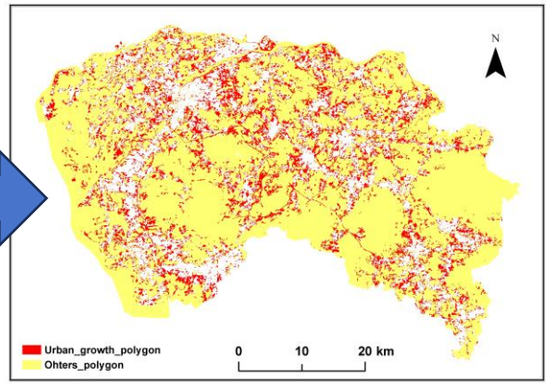




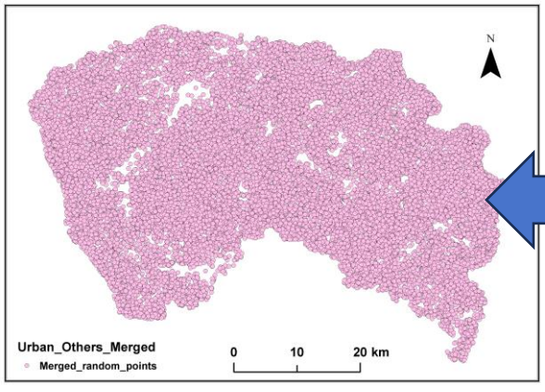
Creating Urban growth/Non-urban growth (others) random points & merge



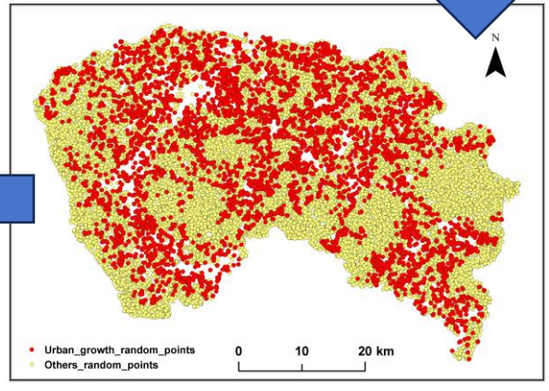
Growth/Non-growth Raster



Growth/Non-growth Polygon



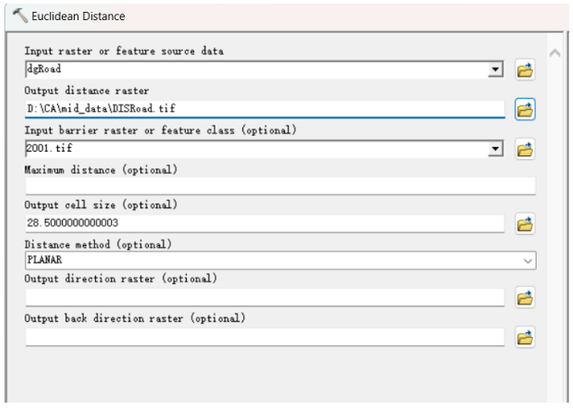
Merged Points

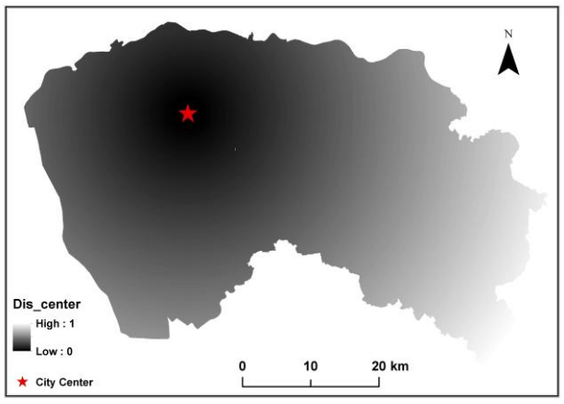


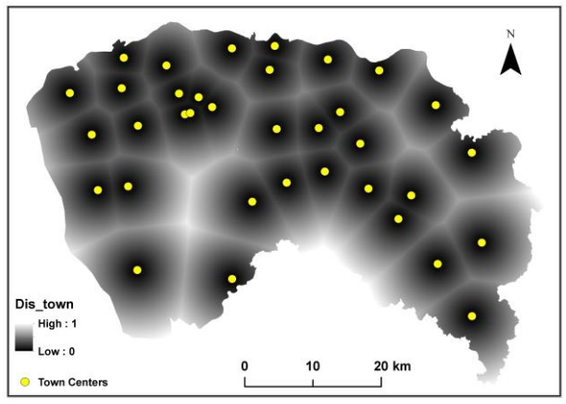
Growth/Non-growth Random Points

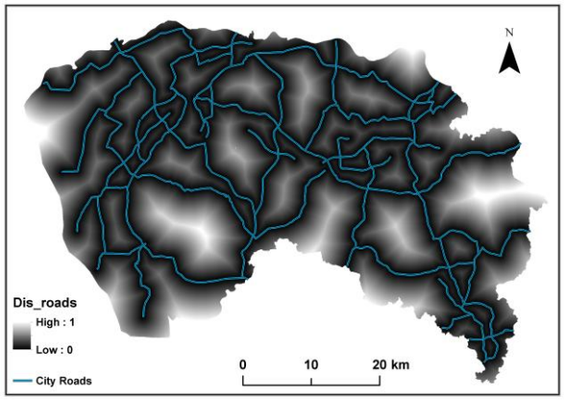
Number of Random Points:
Urban Growth (5000) ; Others (20000)

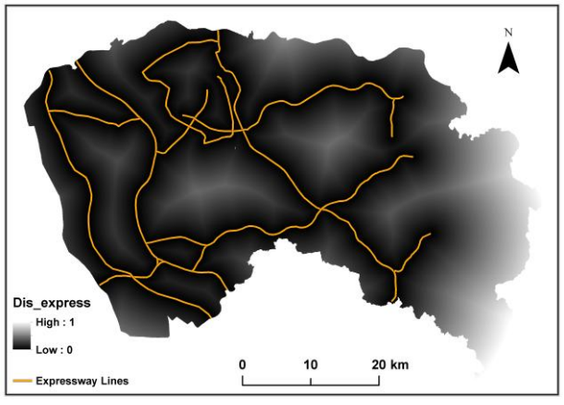
Euclidean Distances of Local Factors (Normalized)

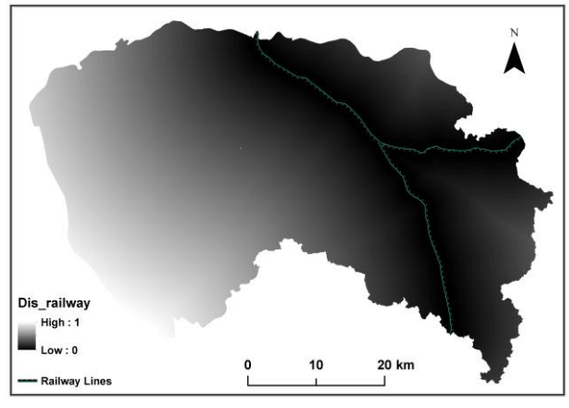




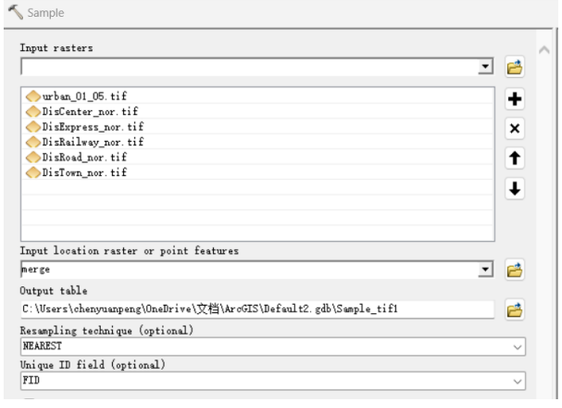






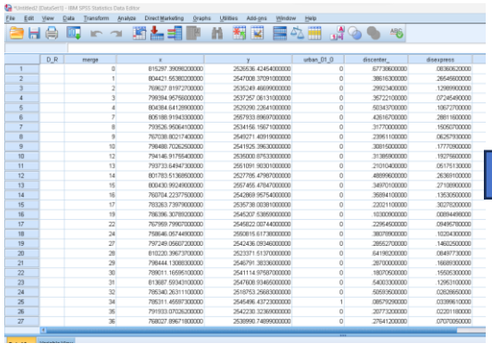


Sampling

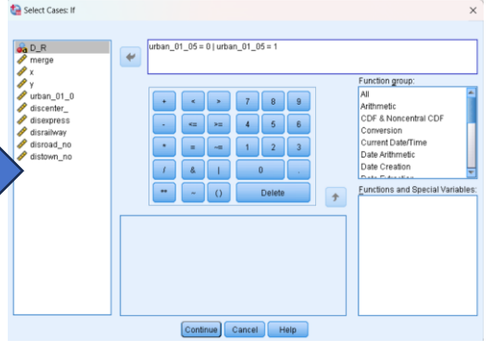


Extracting raster values for the 25,000 random points in the "Urban Change" layer (dependent variable) and the Euclidean distances of the five location factors (independent variables); exporting sampling values to a **.dbf** table.

CA Parameterization in SPSS



Load table in SPSS



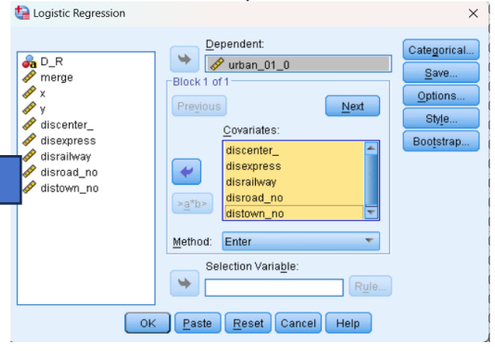
Filter: Select Class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	discenter_	-.300	.107	7.901	1	.005	.741
	disexpress	-.923	.132	48.971	1	.000	.397
	disrailway	-.387	.076	26.019	1	.000	.679
	disroad_no	-3.873	.135	820.631	1	.000	.021
	distown_no	-1.651	.125	173.957	1	.000	.192
	Constant	.274	.054	25.482	1	.000	1.315

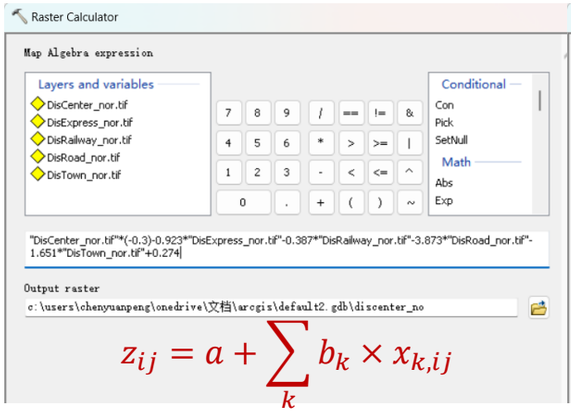
a. Variable(s) entered on step 1: discenter_, disexpress, disrailway, disroad_no, distown_no.

CA Parameters (B)

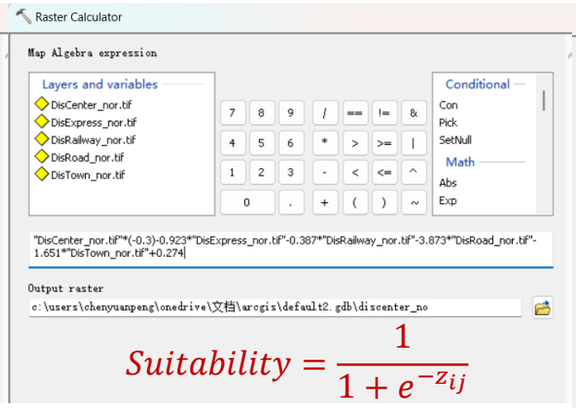


Logistic Regression

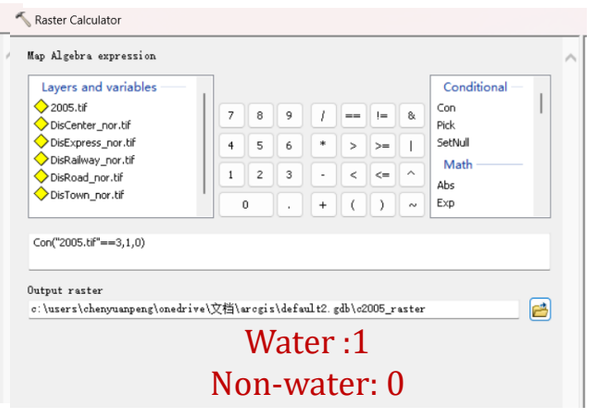
Land Suitability & Water Constraint



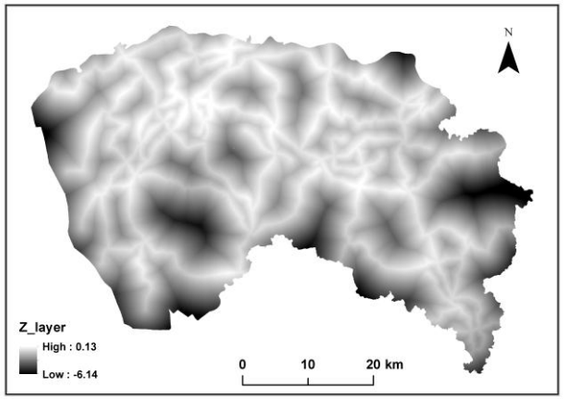
$$z_{ij} = a + \sum_k b_k \times x_{k,ij}$$

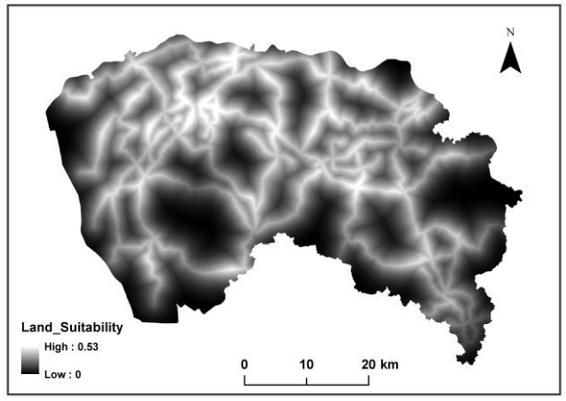


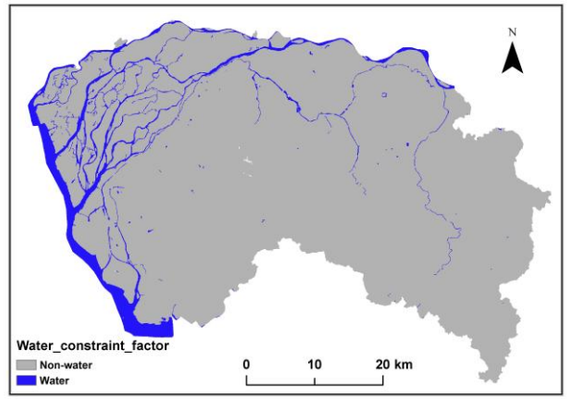
$$Suitability = \frac{1}{1 + e^{-z_{ij}}}$$



Water :1
Non-water: 0

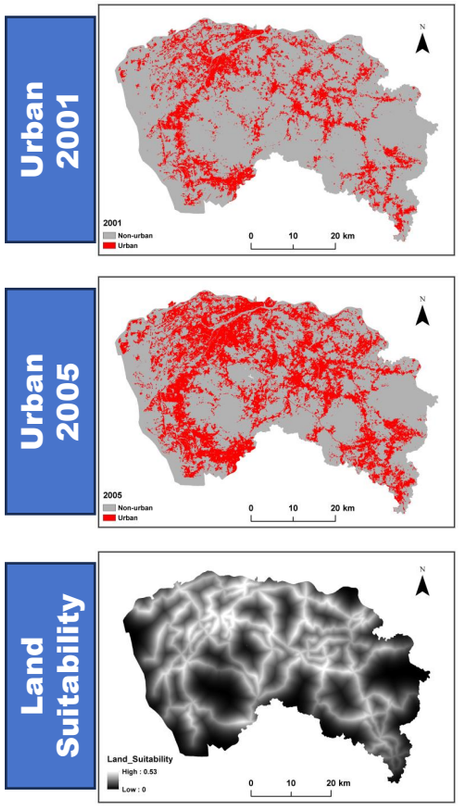






Logistic-CA Modeling

Input



CA Model

Urban Expansion Simulation

$$P_n^t = Suitability \times \Omega_n^t \times con(s_n) \times r$$

$$S^{t+1} = \begin{cases} developed & P_{ij}^t \geq threshold \\ updeveloped & P_{ij}^t < threshold \end{cases}$$

Where:

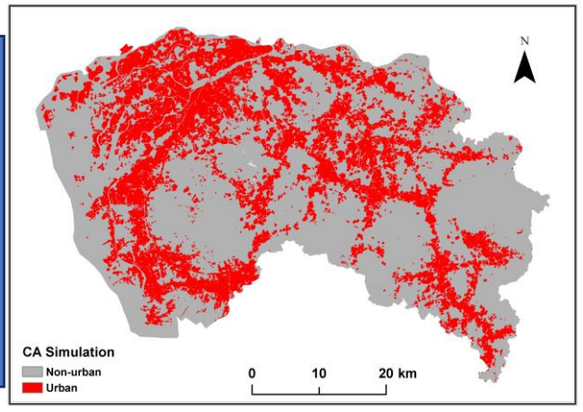
- P_n^t: Probability that the nth grid is urban land at time t;
- Ω_n^t:Proportion of urban land in the 8 surrounding grids of the nth grid.
- Con: Water constraint (0/1)
- r: Random factor by the random() in Python;
- S (binary value): Determines the urban land development status of the nth grid at time t:
  - *If P is above the threshold, the grid is urban land.*
  - *If P is below the threshold, the grid is non-urban land.*

Urban Expansion Simulation in Python

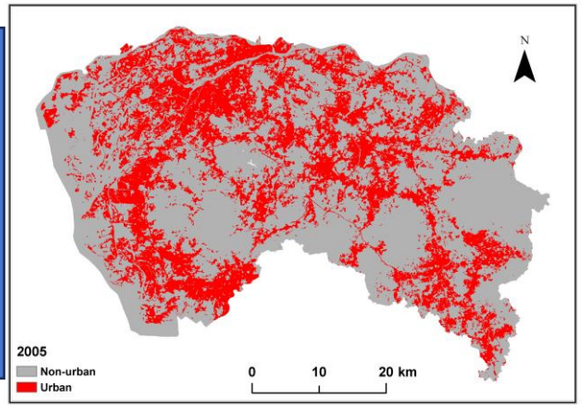
Partial Code

```
# CA modeling
pthreshold = 0.6
# Set the threshold for P at 0.6
tempdata = np.zeros((nrows, ncols), dtype=np.uint8)  # Create temporary
data

# Start a cellular automaton window, calculate neighborhood factor
while num_init < num_2005:  # Loop until the land development requirement
is met
    for i in range(nrows):
        for j in range(ncols):

# Convert double to integer
tempdata = np.int16(tempdata)
file = r'path_to_output_simulation.tif'  # Output image file path
with rasterio.open(file, 'w', driver='GTiff', height=nrows, width=ncols, count=1,
dtype=tempdata.dtype, crs=info['crs'],
                   transform=info['transform'], nodata=255) as dst:
    dst.write(tempdata, 1)  # Write to the output file
```

CA Simulation



Land Use 2005



Accuracy Assessment

Partial Code

```
# Accuracy validation
# 1. Overall accuracy
p0 = np.trace(result) / m
print("Overall Accuracy", p0)
# Calculating kappa coefficient
pe = ((sum(result[0,:])*sum(result[:,0])) + (sum(result[1,:])*sum(result[:,1])))/
(m ** 2)
kappa = (p0 - pe) / (1 - pe)
# 2. Kappa coefficient
print("Kappa", kappa)
# 3. Calculate FOM (Figure of Merit)
# Since class C cannot be determined, it is not included in this FOM
calculation
A = 0  # Number of false negatives
B = 0  # Number of true positives
D = 0  # Number of false positives
for i in range(m):
    if T[i, 0] == 1 and T[i, 1] == 0:
        A += 1
    elif T[i, 0] == 1 and T[i, 1] == 1:
        B += 1
    elif T[i, 0] == 0 and T[i, 1] == 1:
        D += 1
FOM = B / (A + B + D)
print("FOM", FOM)
```

```
Overall Accuracy 0.8047511289205442

Kappa 0.8047648224340291

FOM 0.5577572913299222
```

Conclusion

Logistic Cellular Automata (CA) provided robust urban expansion simulations. In Dongguan's case, the simulation yielded positive results with predetermined location factors and natural constraints, serving as an example for future urban expansion simulation tasks.

02

URBAN LAND USE & MACHINE LEARNING (January, 2023)

Object-oriented City Land Use Classification, and Analysis using R

OVERVIEW

The project's objective was to perform object-oriented urban land use classification of the Haizhu District in Guangzhou City based on three types of machine learning algorithms, and to explore the importance of various imagery spectral, textual, and shape attributes or characteristics in identifying the “urban village” land use category.

TOOLS & SKILLS

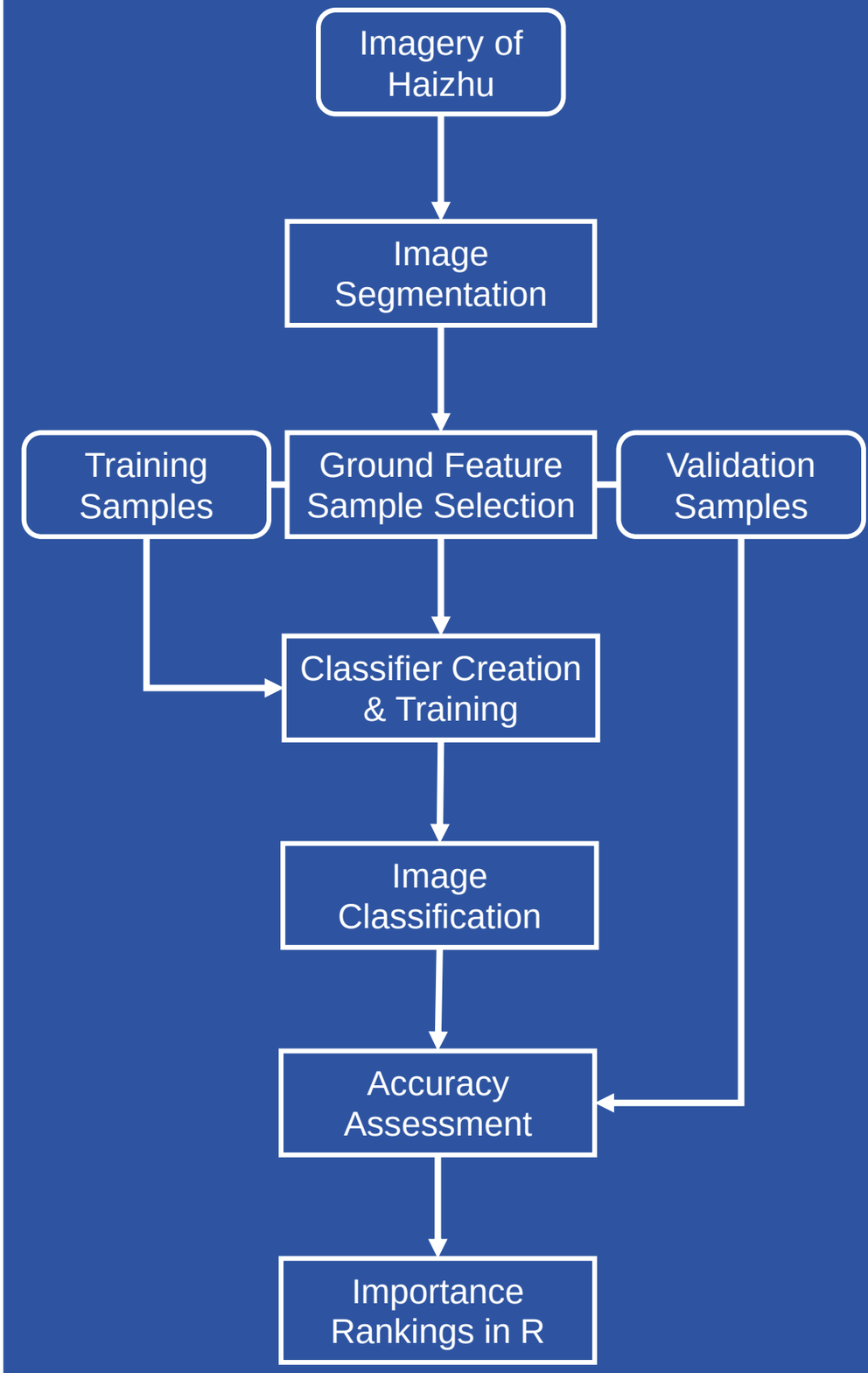
eCognition, ArcMap, R, Support Vector Machine, Decision Tree, Random Forest Regression

DATASET

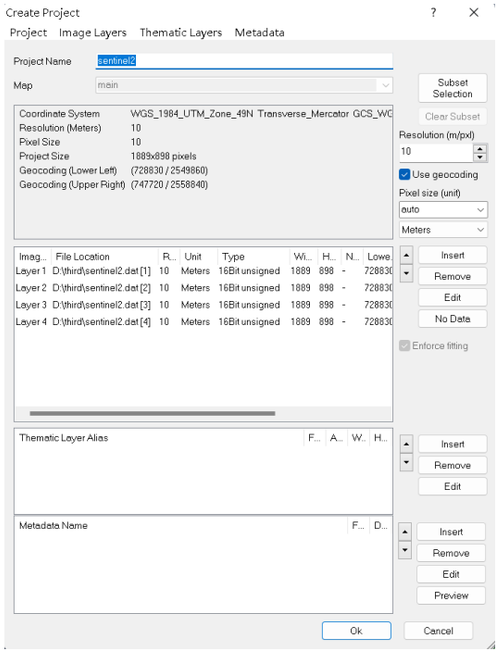
A Sentinel-2 multi-spectral satellite image of Haizhu District, Guangzhou

(Resolution: 10m)

WORKFLOW



Create Project & Load Image in eCognition



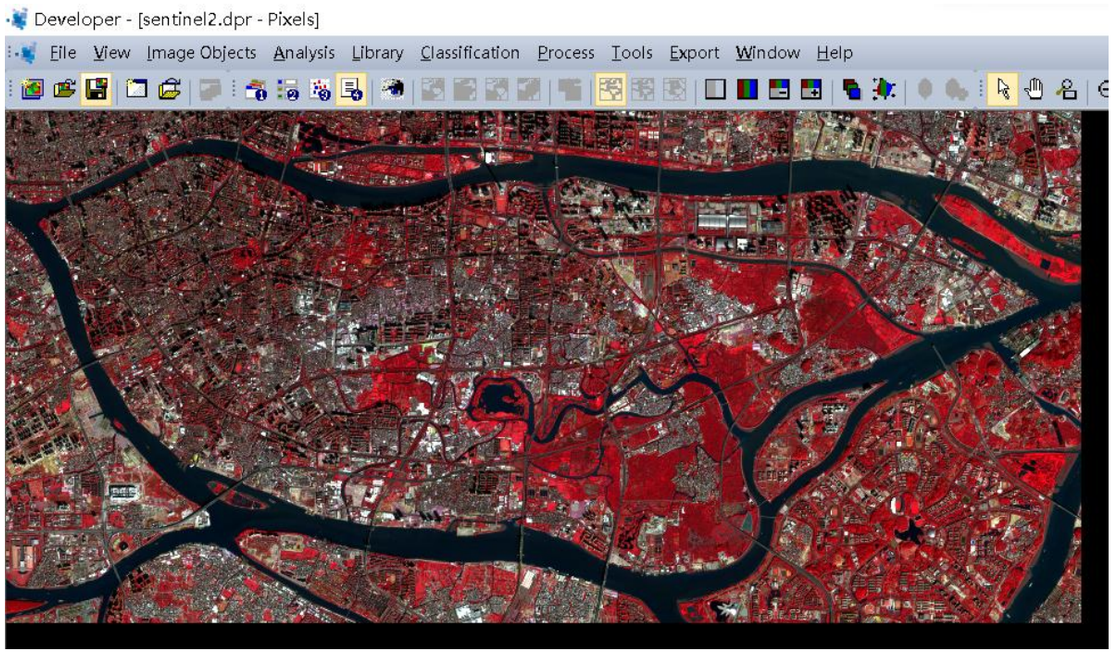
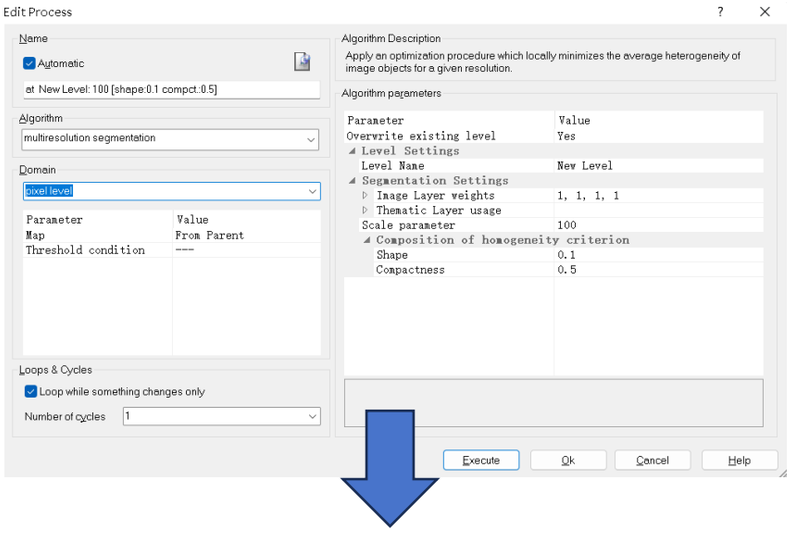
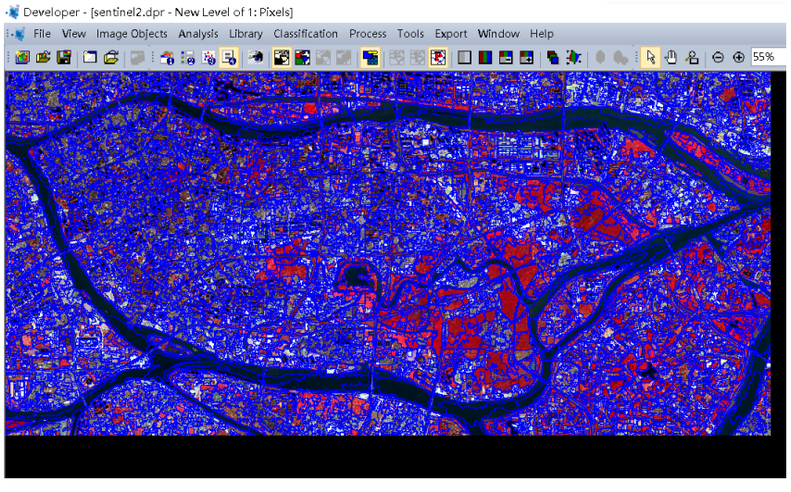
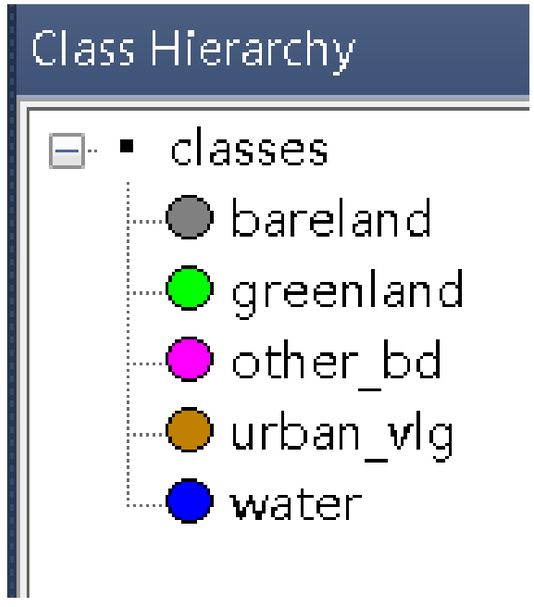


Image Segmentation (Pixel Scale Parameter: 100)



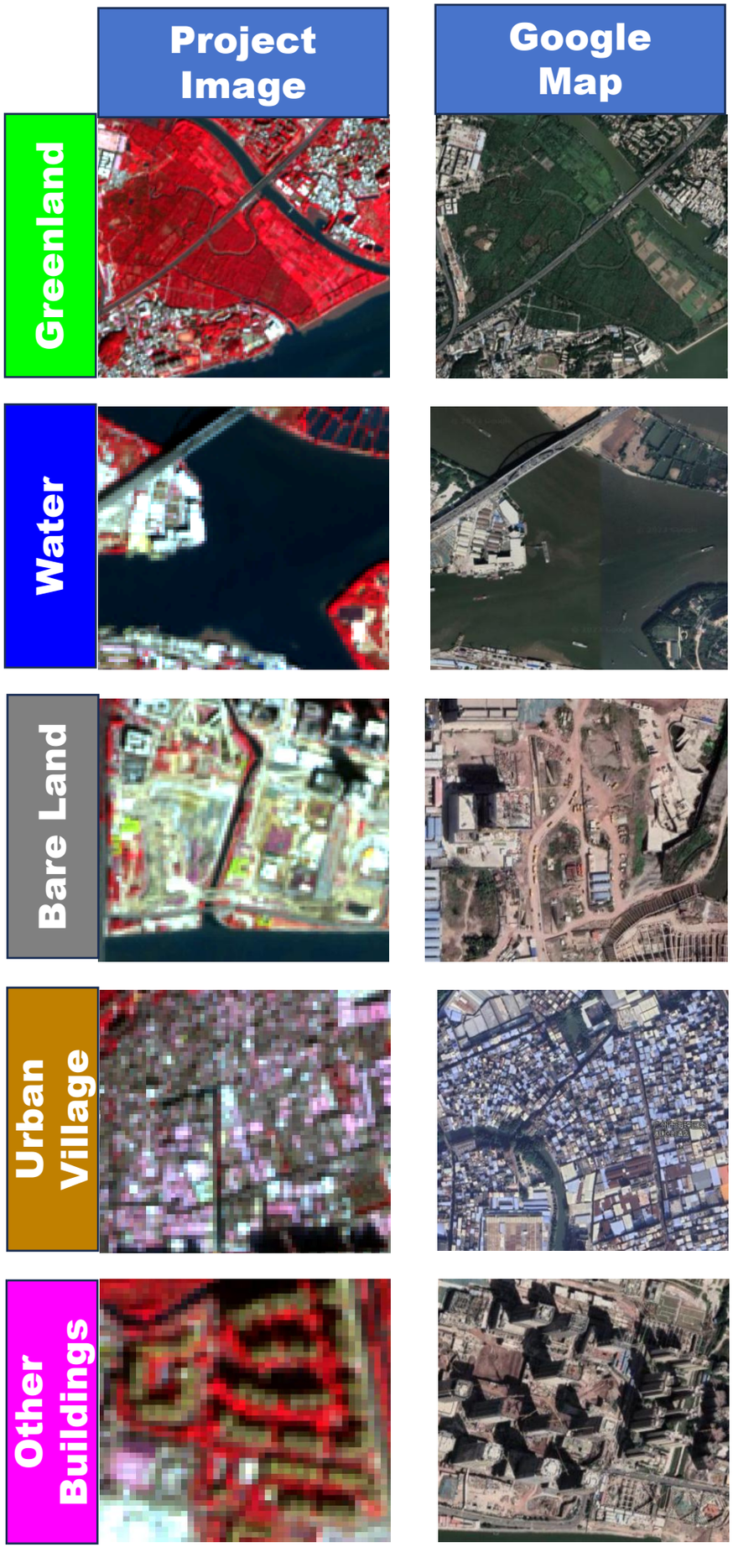


Names of Ground Features to be classified



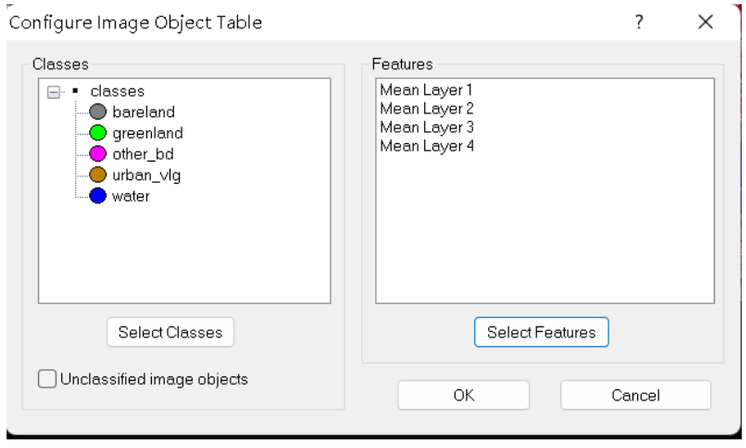
- ✓ Bare land
- ✓ Green land
- ✓ Urban village
- ✓ Buildings other than urban village
- ✓ Water

Sample Selection with the Support of Google Map (Ground Truth)

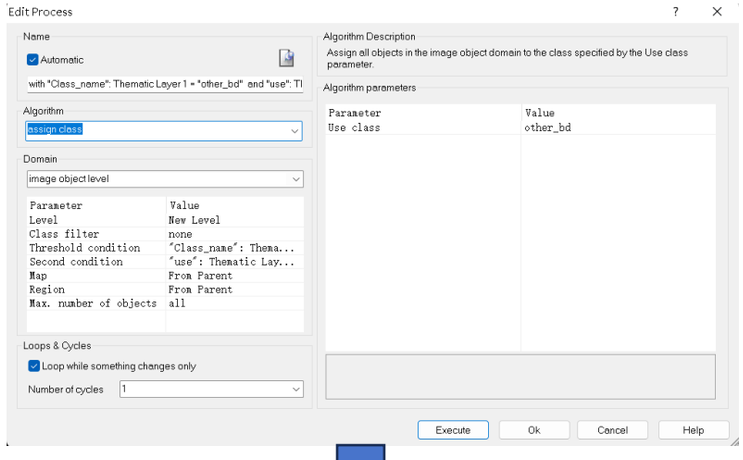


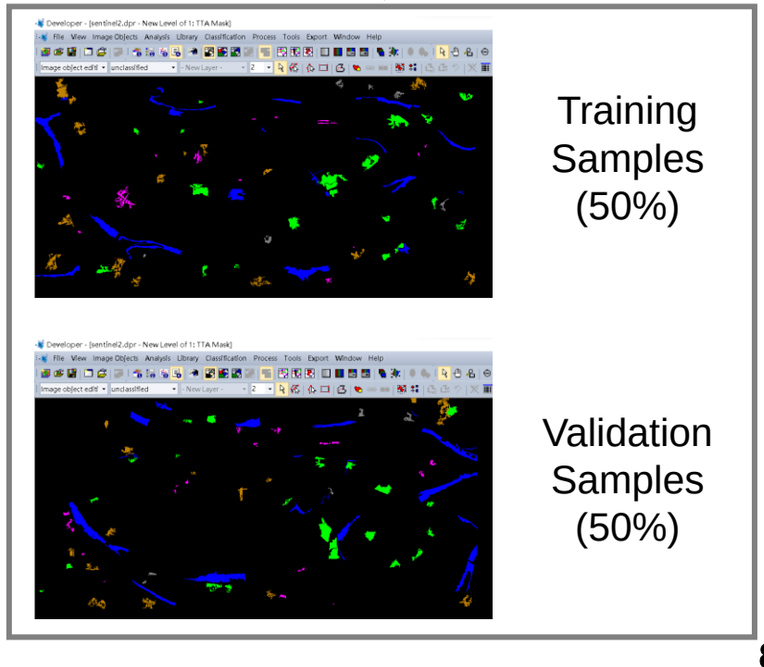
Training/Validation Samples Separation

1. Sample Objects Configuration



2. Assign Classes

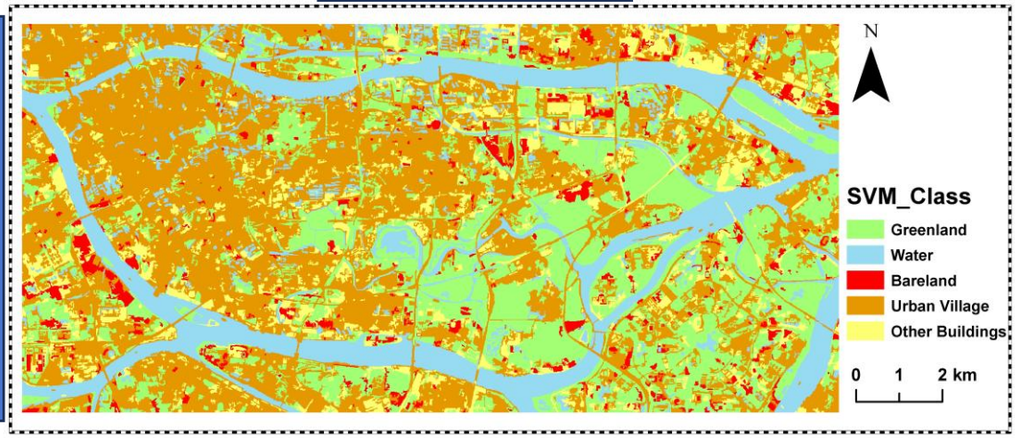




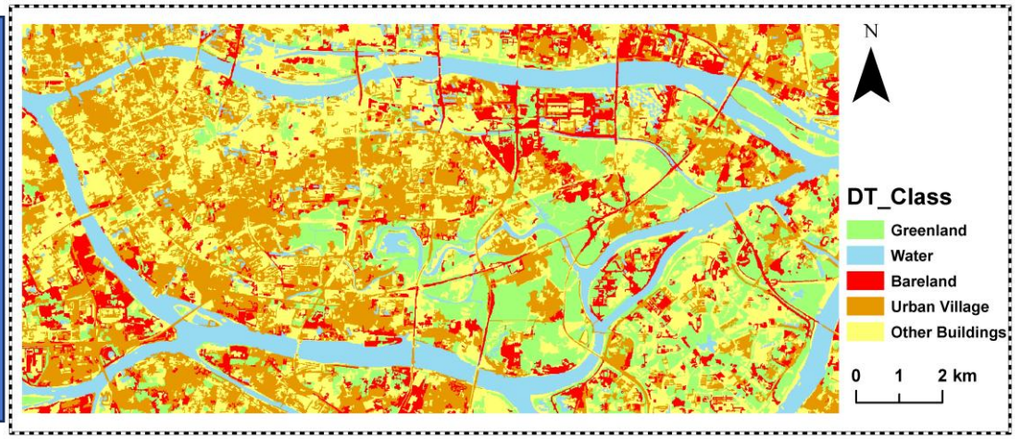
Classification & Accuracy Assessment

Classifier

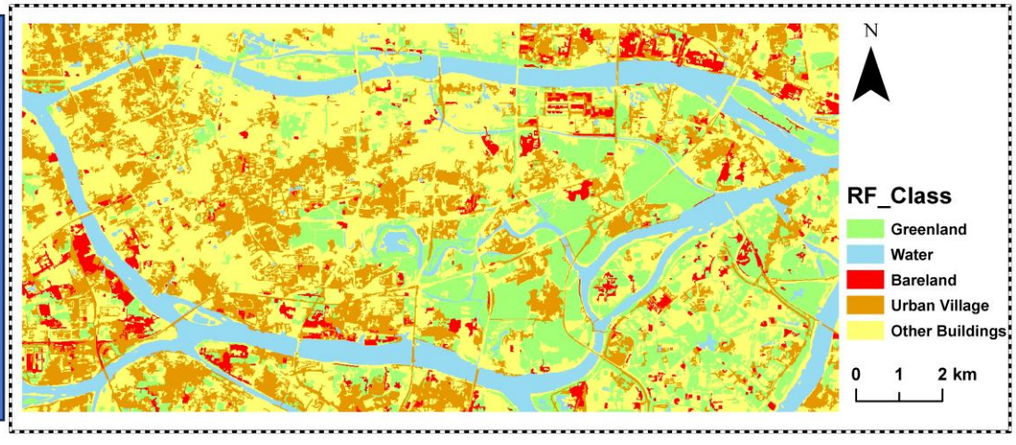
Support Vector Machine



Decision Tree

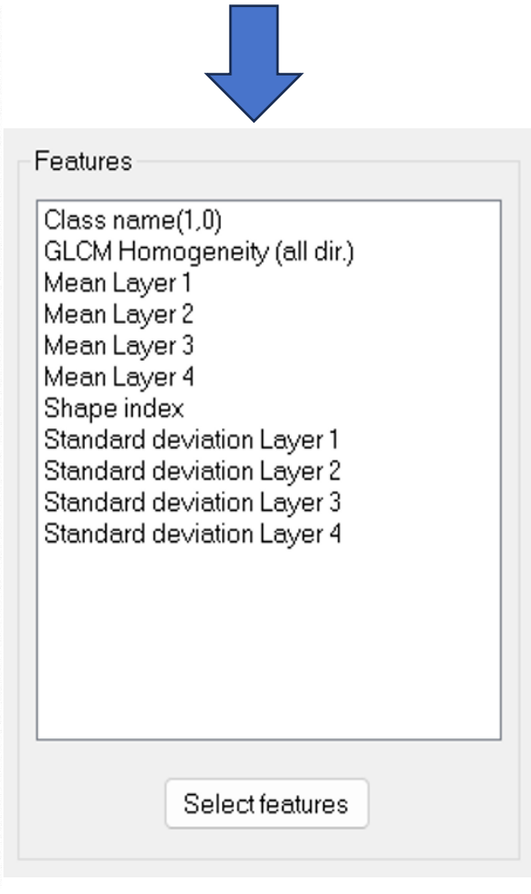


Random Forest



Selected Features for Classification:

1. Class Name
2. Band mean value
3. Band standard deviation
4. Shape Index
5. Texture Attribute



Accuracy Assessment (Overall Accuracy & Kappa)

	Overall Accuracy Objects	Overall Accuracy Pixels	Kappa Objects	Kappa Pixels
SVM	93.50%	94.95%	0.92	0.93
DT	92.20%	93.51%	0.90	0.90
RF	**97.40%**	**97.08%**	**0.96**	**0.95**

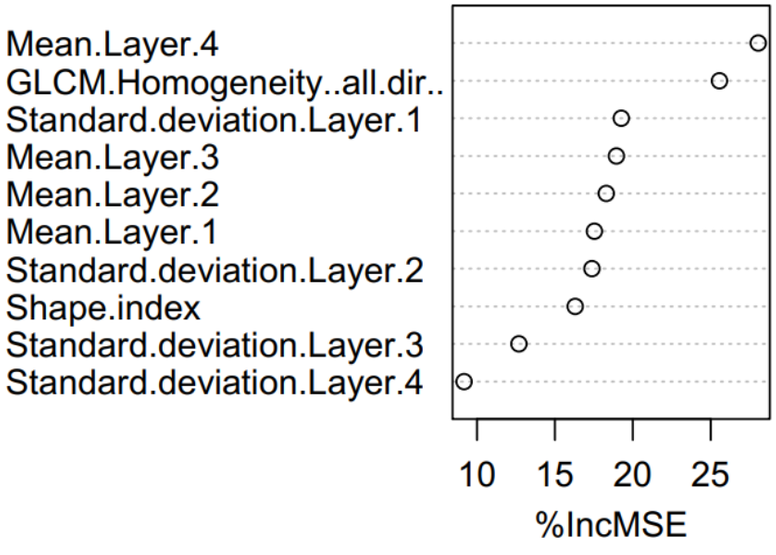
Importance Ranking in R

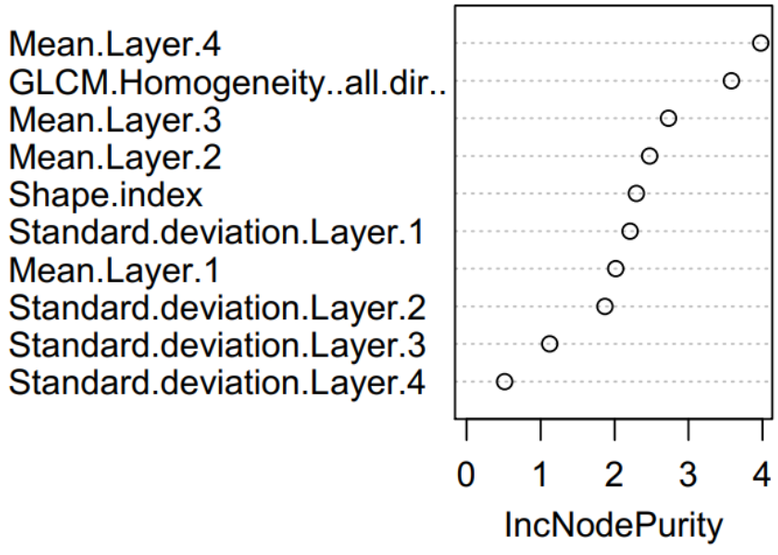
R Codes & Printed Results

```
# 1.Setting up the environment
setwd("D:/R_data")

# 2.Importing the randomForest toolkit
library(randomForest)

# 3.Reading the selected samples' feature values into the table
#   and save them to the "traindata" variable.
traindata <- read.csv("samples_R.csv")

# 4.Performing the random forest regression with 1000 number of random trees
result <- randomForest(Type~., traindata, importance=TRUE, ntree=1000)
print(result)

# 5.Sorting the features in order of importance
importance(result)

# 6.Plotting sorted results
varImpPlot(result)
```

```
> importance(result)
                                %IncMSE IncNodePurity
GLCM.Homogeneity..all.dir.. 25.555565     3.5796078
Standard.deviation.Layer.4   9.181054     0.5132136
Standard.deviation.Layer.3  12.702707     1.1239761
Standard.deviation.Layer.2  17.368765     1.8701442
Standard.deviation.Layer.1  19.255831     2.2108382
Mean.Layer.4                28.051382     3.9770616
Mean.Layer.3                18.959132     2.7315442
Mean.Layer.2                18.303619     2.4735125
Mean.Layer.1                17.547345     2.0164987
Shape.index                 16.296901     2.2931501
```

Ranking Plots





Conclusion

Overall, in object-oriented classification tasks, the texture feature of the image (GLMC Homogeneity of all Directions) and the mean value of the fourth band (near-infrared band) help the identification of "urban village" ground features more obviously; conversely, the standard deviation of the third/red band and the NIR band played a smaller role in urban village identification.

ENVIRONMENTAL MODELING & REMOTE SENSING (December, 2022)

City Cluster Water Quality Evaluation & Pollution Visualization

03

OVERVIEW

Located in southern China, the Pearl River Delta City Cluster is a crucial industrial, economic, and transportation hub facing significant water pollution challenges. This project aims to apply quantitative remote sensing modeling techniques to visualize and assess the levels of two types of water quality information: organic pollutant & suspended sediment concentration (SSC), in this critical region.

TOOLS & SKILLS

PCI Geomatica V9.0, EASI Macrolanguage Programming, Image Processing, Remote Sensing Modeling & Inversion

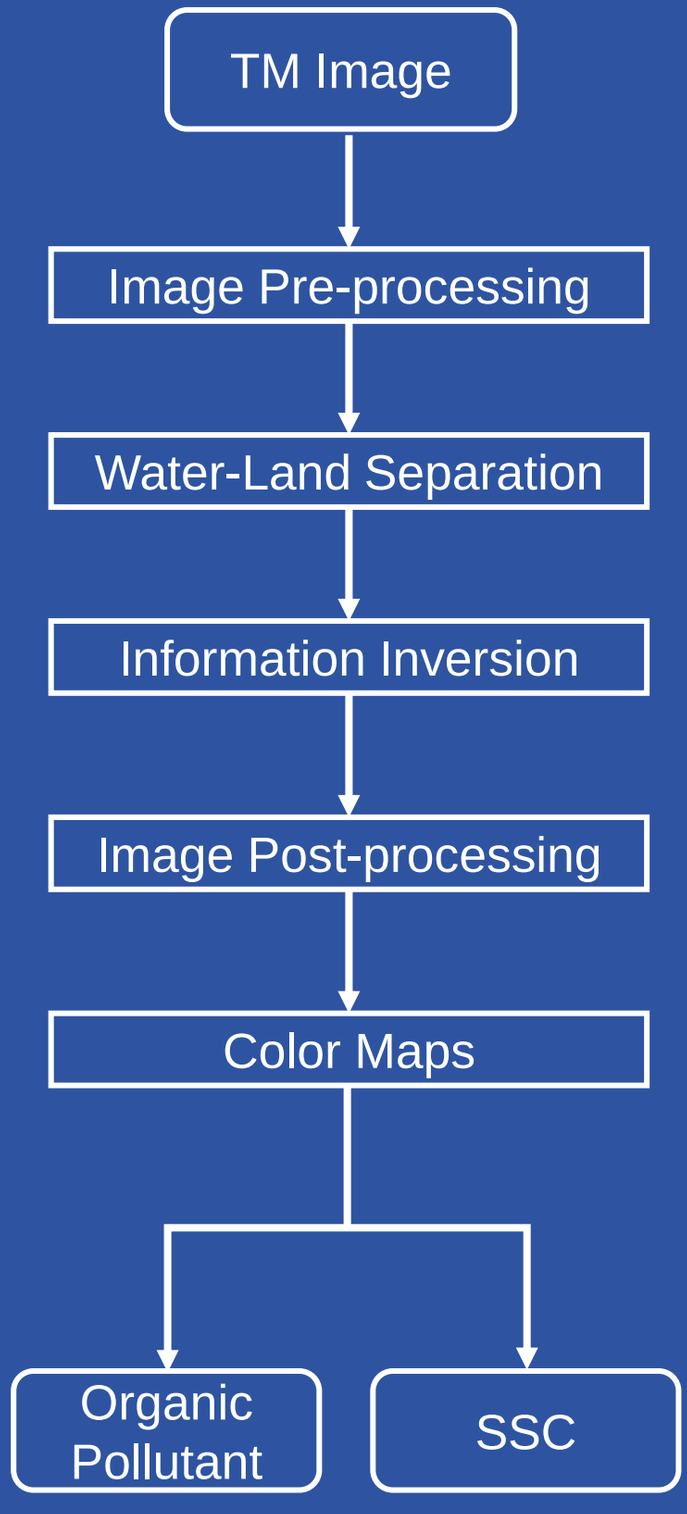
DATASET

A Landsat TM image of the Pearl River Delta Region

ACKNOWLEDGEMENT

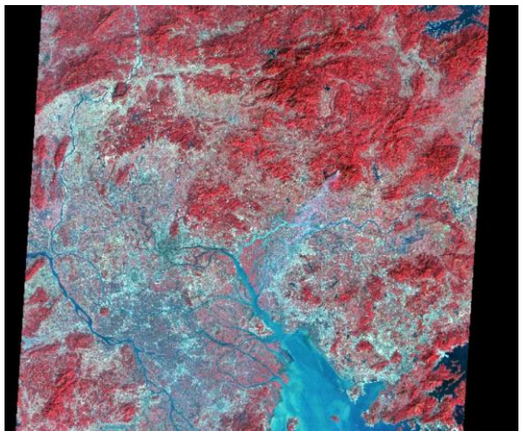
This project's reference EASI codes were provided by Professor Ruru Deng from the SYSU School of Geography and Planning. Sincere thanks to Professor Deng for his invaluable contributions to the fields of remote sensing and GIS.

WORKFLOW

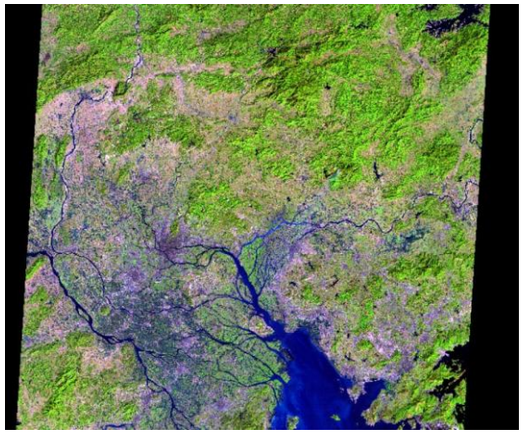


Raw Data (RGB)

Band 3,2,1



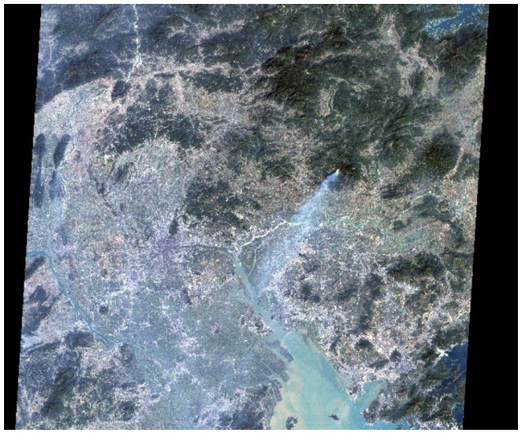
Band 4,3,2



Band 5,4,3

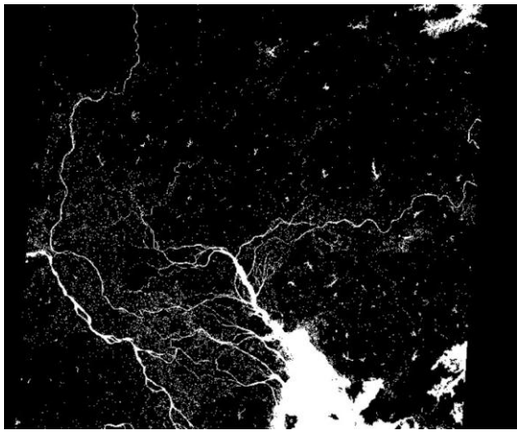
Pre-processing

1. Radiometric Calibration
2. Atmospheric Correction

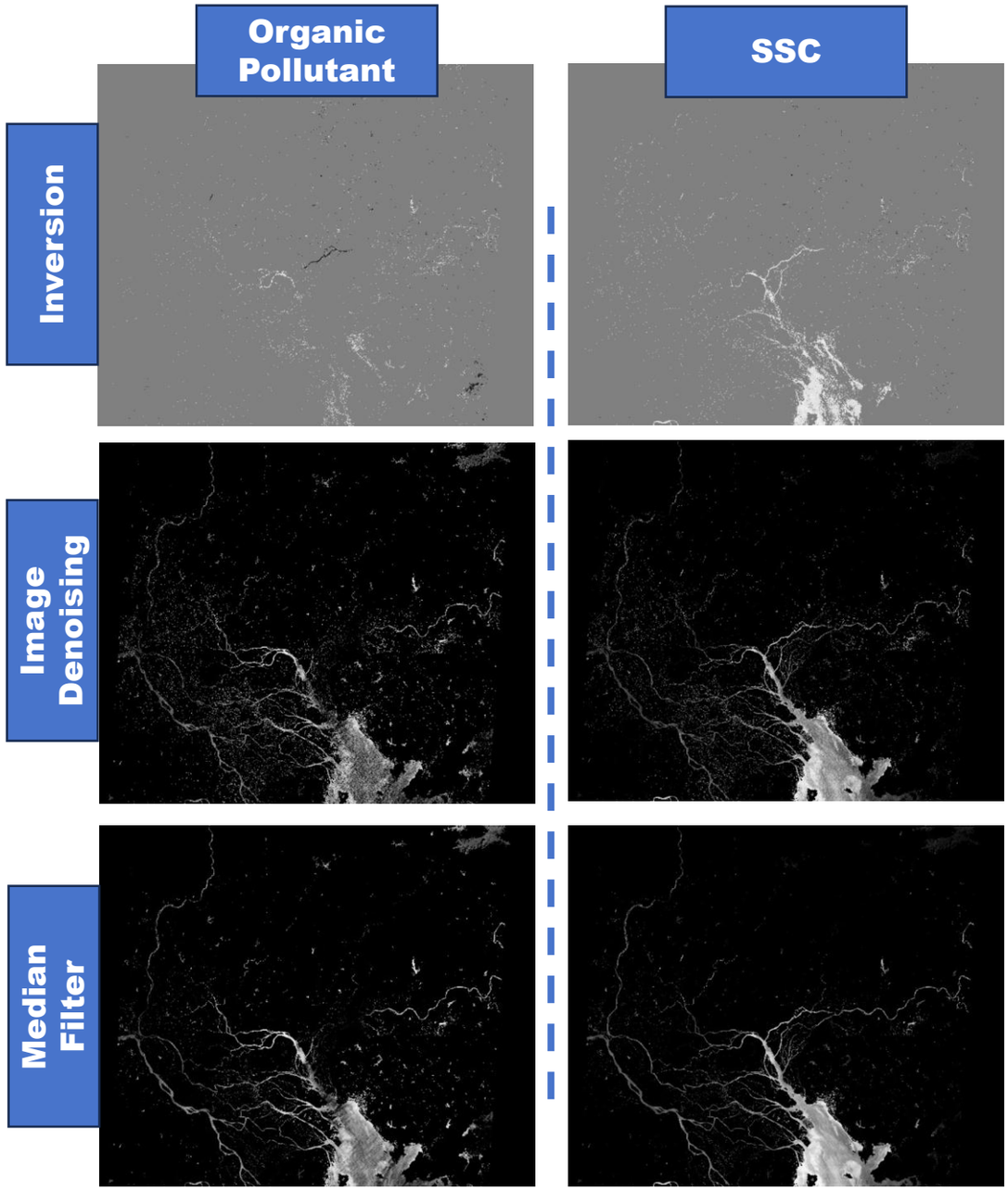


Water-Land Separation

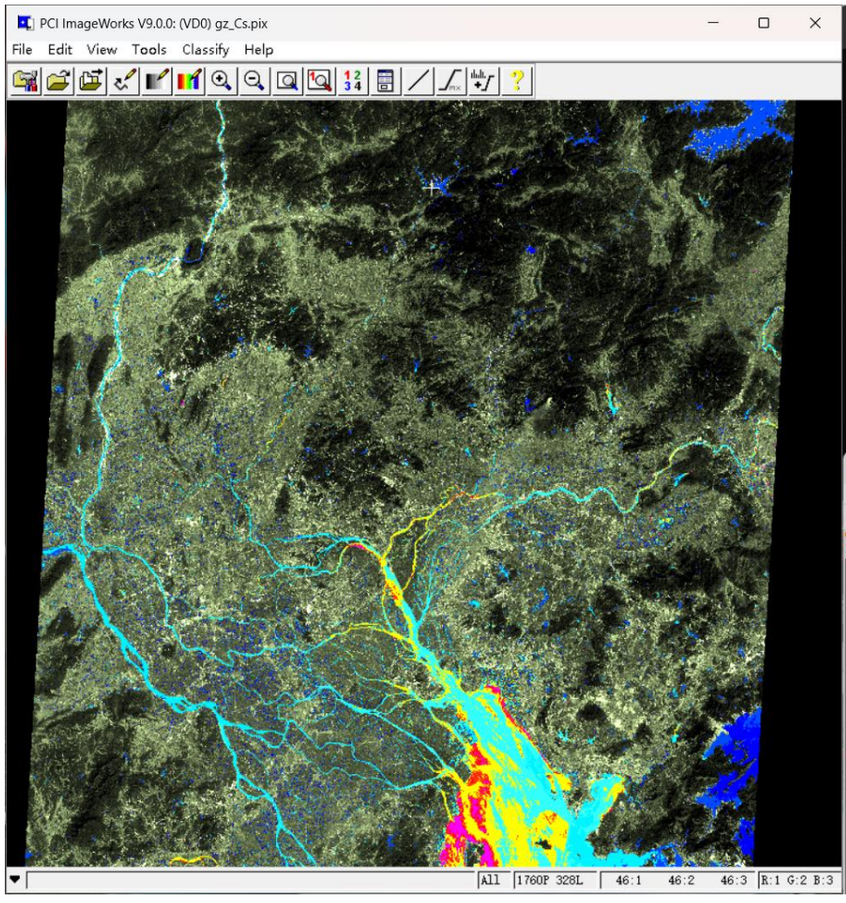
Binary Map



Inversion & Post-processing



Density Slicing & Export Color Maps




```
D:\Geomatica_V90\exe\easi.E:
 Creating segment:  1  [ 150: @@Georeferencing      ]        8 Blocks long
-------- The end of createGeo32Img.EAS.  ---
img_in:   D:\DLYG\gz.pix
img_out:   D:\DLYG\outImage\gz_tm_rc.pix
------ Transfering data form file_in of 8bit to file_ex of 32bits ! -----
-------- Transfering data to 32bits is finished ! -----
------ Doing Calibration ! ! -----
--------The End of calibrationTm.EAS ! -------
------ Doing Calibration ! ! -----
--------The End of rad_c.EAS ! -------
------ Doing air_c ! ! -----
--------The End of air_c.EAS ! -------
------ Doing water_mask  ! -----
--------The End of water_mask.EAS ! -------
------ Doing sand_pollutant.EAS  ! -----
--------The End of sand_pollutant.EAS ! -------
------ Doing mend.EAS ! -----
--------The End of mend.EAS ! -------
img_in:   D:\DLYG\outImage\gz_tm_rc.pix
---  Median filtering .
(FME     100%)
(FME     100%)
---  Average filtering .
(FAV     100%)
```

*D:\DLYG\EASI1\myMain.EAS - Notepad++

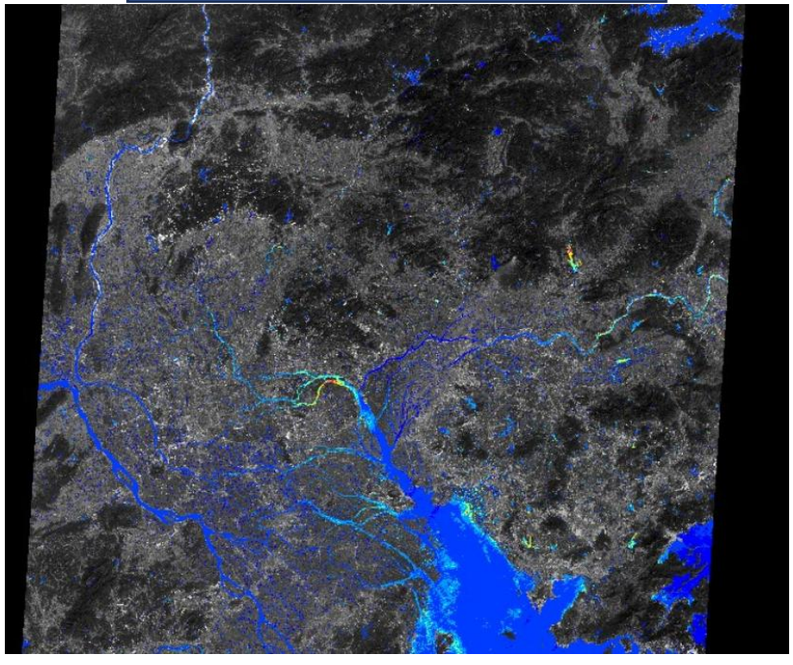
文件(F) 编辑(E) 搜索(S) 视图(V) 编码(N) 语言(L) 设置(T) 工具(O) 宏(M) 运行(R) 插件(P) 窗口(W) ?

myMain.EAS | air_c.EAS | calibrationTm.EAS | createGeo32Img.EAS | Med_ModFilter.EAS | mend.EAS

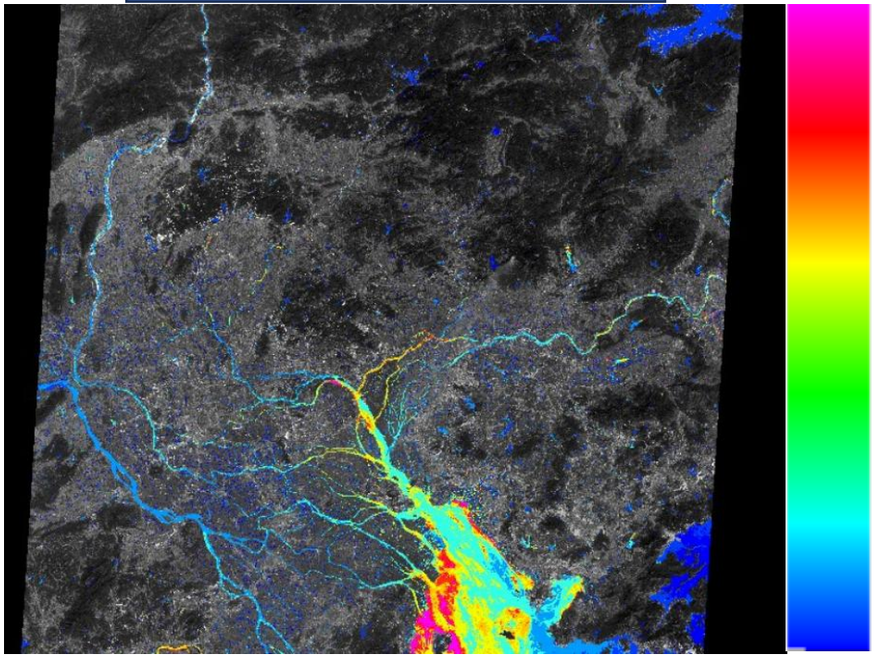
```
25  doc load image-processing modules:
26  load programPath+"\transferDataTo32.EAS"
27  load programPath+"\calibrationTm.EAS"
28  load programPath+"\rad_c.EAS"
29  load programPath+"\air_c.EAS"
30  load programPath+"\water_mask.EAS"
31  load programPath+"\sand_pollutant.EAS"
32  load programPath+"\mend.EAS"
33  load programPath+"\tif_export_colour.EAS"
34  load programPath+"\Med_ModFilter.EAS"
35
36  imgOrg=pathOrg+"\gz.pix"
37  imgIn= pathOut+"\gz_tm_rc.pix"
38  imgRc= pathOut+"\gz_tm_rc.pix"
```

Exported Maps

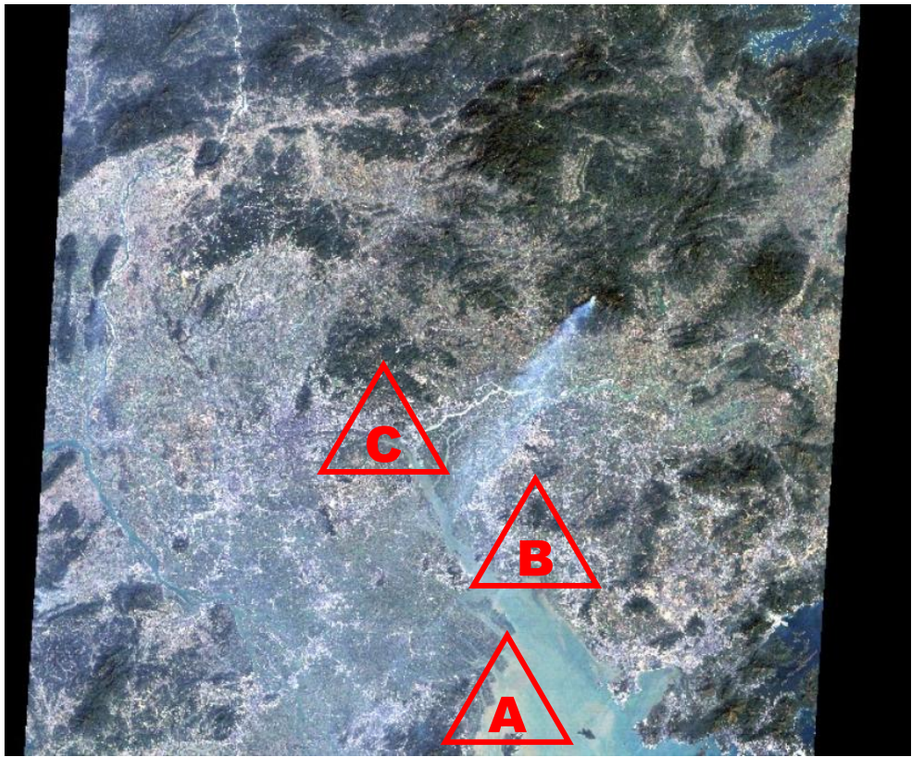
Organic Pollutant



SSC

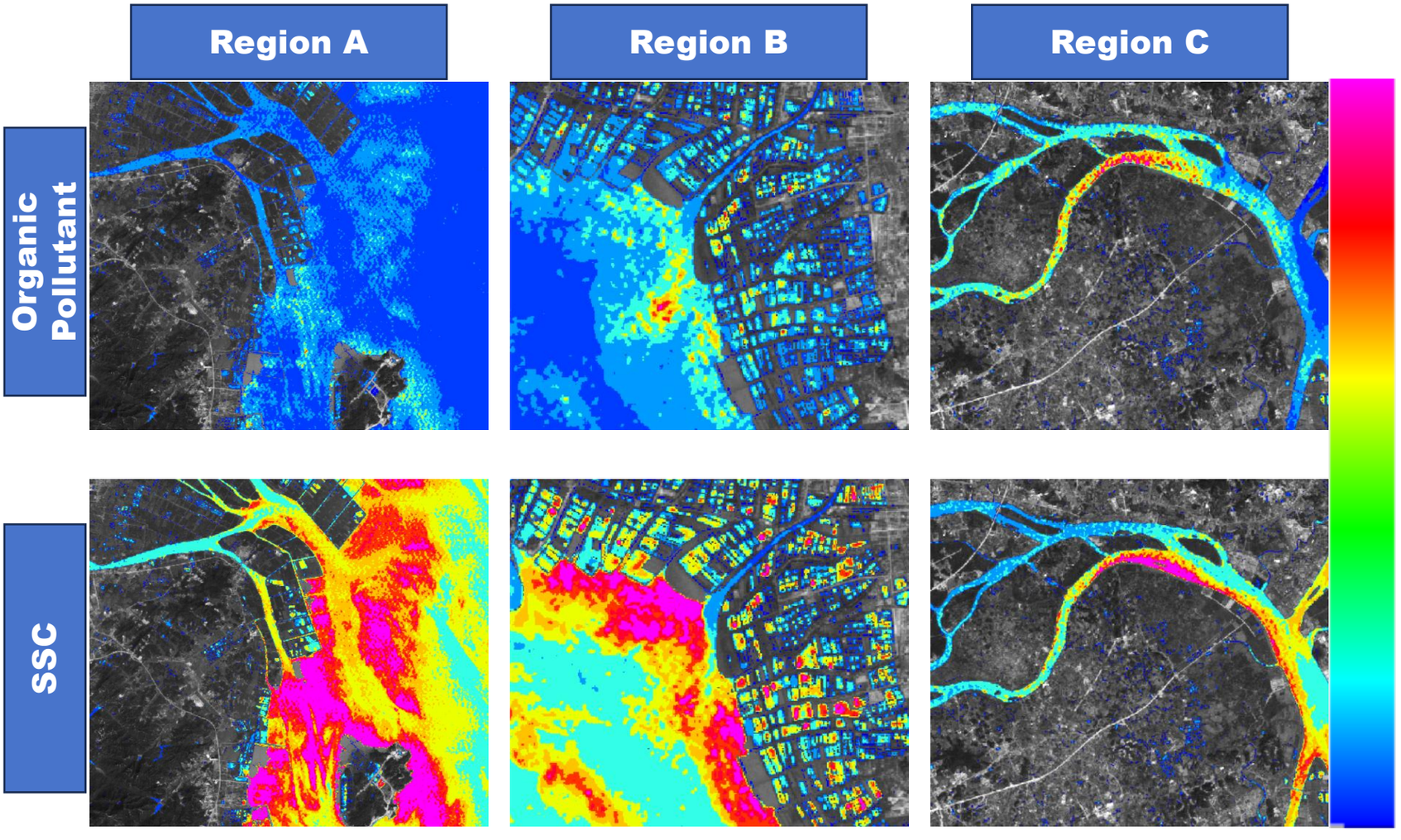


Case Assessment (Regions with High Pollution Levels)



- ✓ **Region A:** Lindingyang Bay, Pearl River Estuary
- ✓ **Region B:** Dongbao River, Shenzhen
- ✓ **Region C:** Pearl River, Guangzhou

Conclusion: Regions of city centers, industrial zones, major traffic routes, and river basin outlets are typically areas with **higher** levels of water pollution.



SPATIAL ANALYSIS (October, 2022)

Point Pattern Analysis of City Public Facilities

04

OVERVIEW

The project aimed to conduct spatial analysis of the public facility points in Jiaxing City, Zhejiang Province. The purposes included:

1. To investigate basic spatial characteristics of facility points, such as boundary contours and central tendencies.
2. To employ distance-based statistical point pattern analysis methods and functions to assess the spatial clustering and distribution patterns of these points.

TOOLS & SKILLS

Jupyter Notebook, Python Spatial Analysis Library, Statistical Point Pattern Analysis

DATASET

Points of Interest (POI): Public Facility Locations of Jiaxin, Zhejiang Province

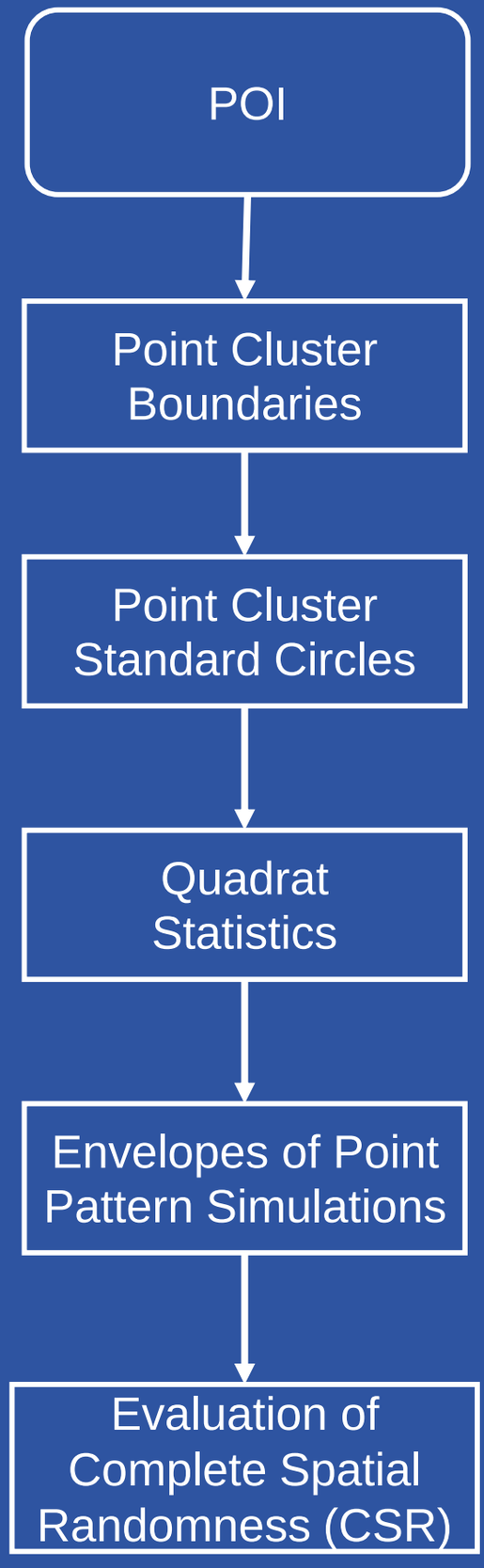
REFERENCES

[1]. Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, 256-274.

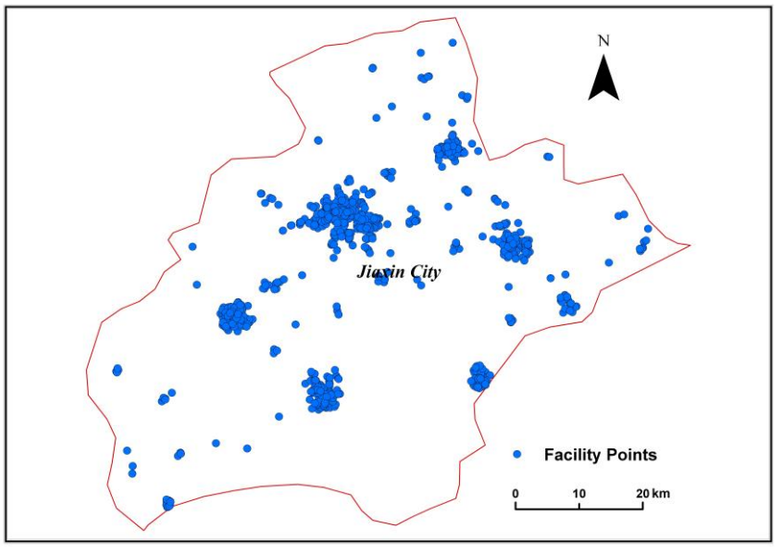
[2]. Rey, S., & Kang, W. (n.d.). Distance based statistical method for planar point patterns. Retrieved from

https://pysal.org/notebooks/explore/pointpats/distance_statistics.html

WORKFLOW



POI Data



Spatial layout in ArcMap

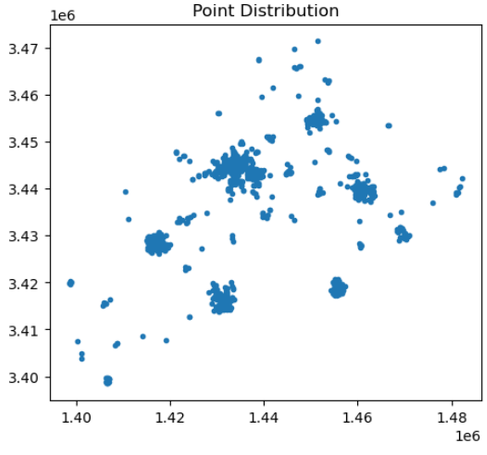
```
# open the shapefile
f = 'Jiaxin.shp'
fo = ps.io.open(f)
pp = PointPattern(np.asarray([pnt for pnt in fo]))
fo.close()
```

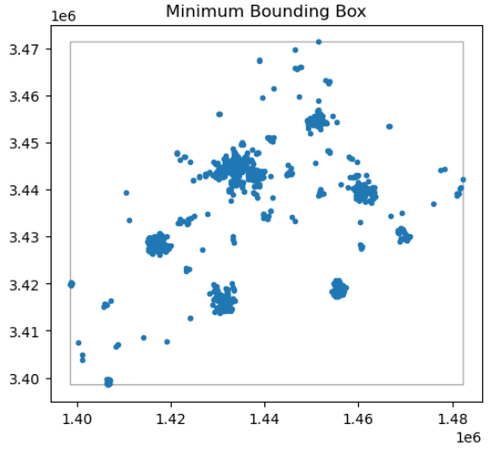
```
# Attributes of PySAL Point Patterns
pp.summary()
```

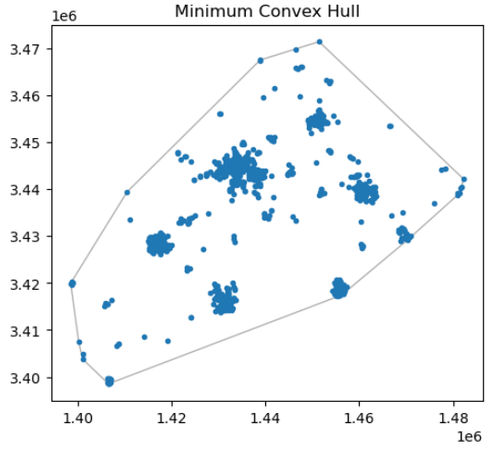
```
Point Pattern
961 points
Bounding rectangle [(1398590.0667995564,3398527.4751835745), (1482190.276057229,3471371.925912306)]
Area of window: 6089811324.18219
Intensity estimate for window: 1.5780456057545496e-07
              x             y
0  1.432403e+06  3.439677e+06
1  1.455499e+06  3.417574e+06
2  1.460882e+06  3.427831e+06
3  1.406834e+06  3.399587e+06
4  1.430687e+06  3.417733e+06
```

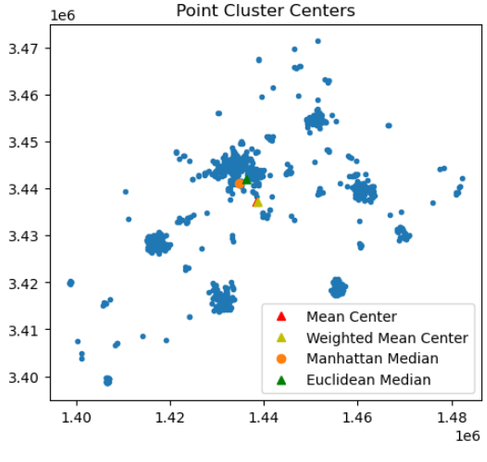
Load shapefile in JupyterLab & Print basic information of point cluster

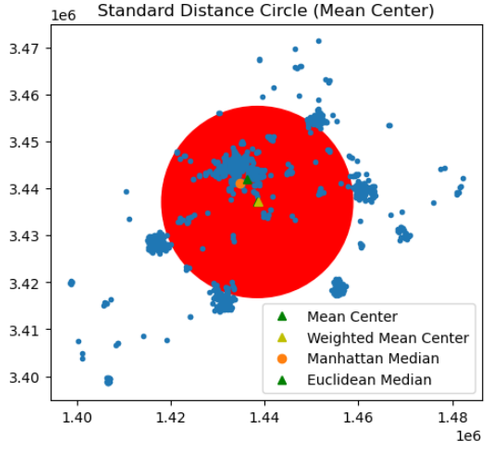
Point Cluster Boundaries, Centers, and Standard Circles

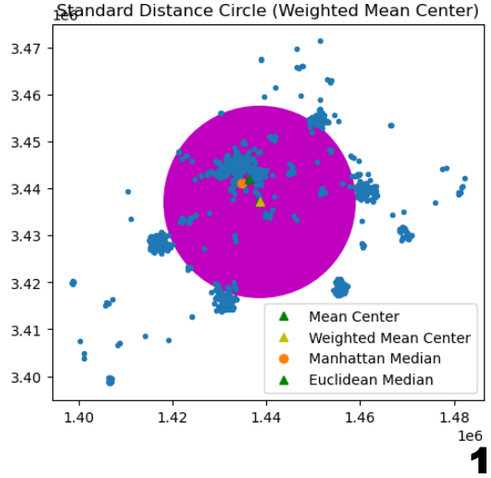




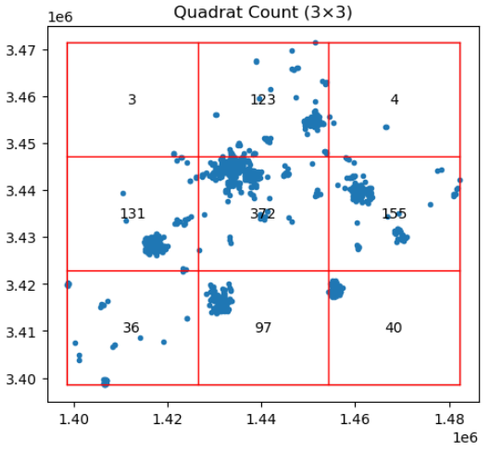


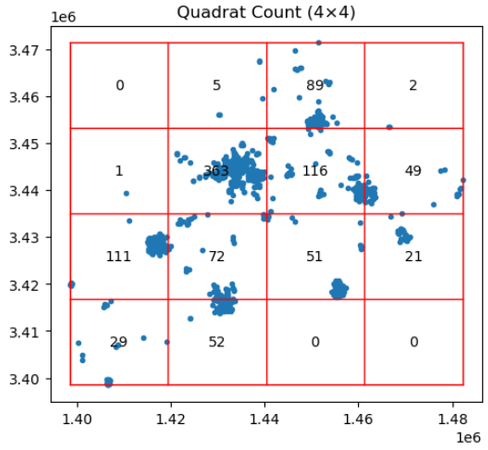


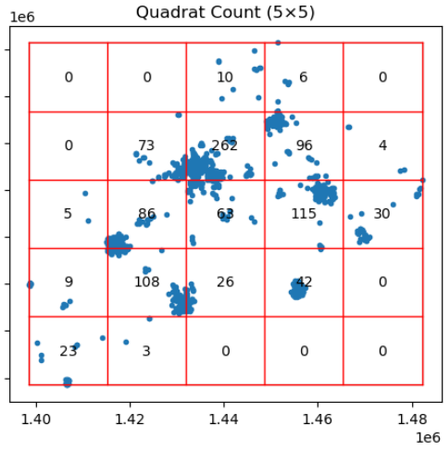




Quadrat Statistics







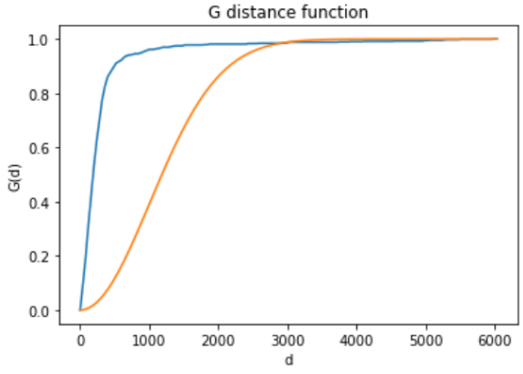
Supplementary Statistics

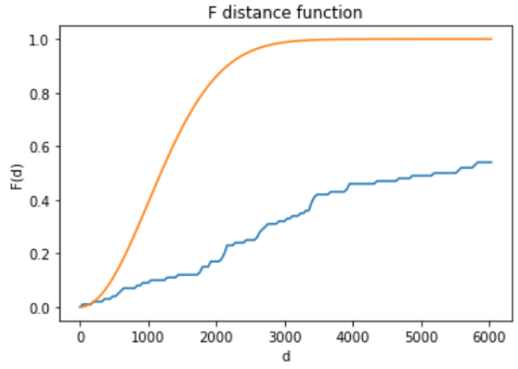
```
#According to chi-squared, degree of freedom and p-value,
# determine whether the underlying process is  CSR or not
print("chi-squared test statistic: ",q_r.chi2)
print("degree of freedom: ",q_r.df)
print("analytical p value: ",q_r.chi2_pvalue)
```

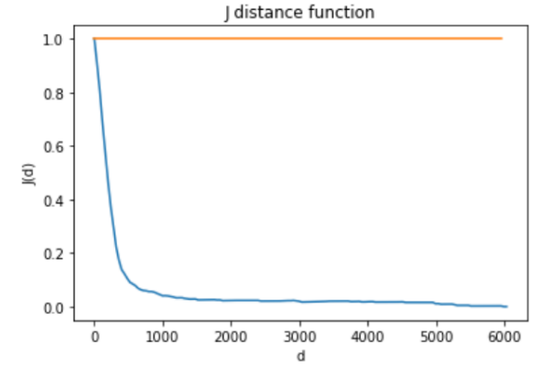
```
chi-squared test statistic:  2253.8543184183145
degree of freedom:  24
analytical p value:  0.0
```

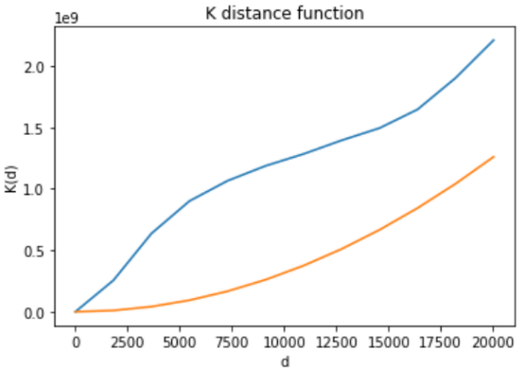
When the p-value is close to 0, it indicates a higher statistical significance of the result. In the context of point pattern analysis, such as with complete spatial randomness (CSR), a very low p-value suggests that the distribution of point clusters does not follow a random pattern but instead exhibits a significant clustering pattern.

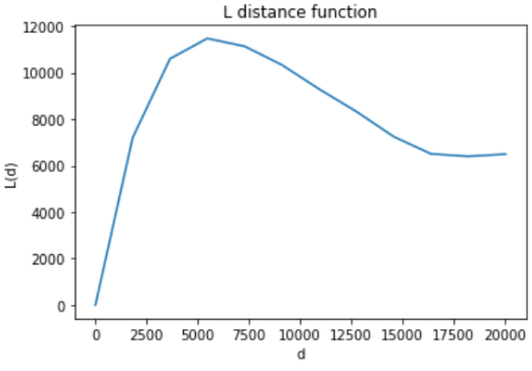
Nearest Neighbor Statistical Functions





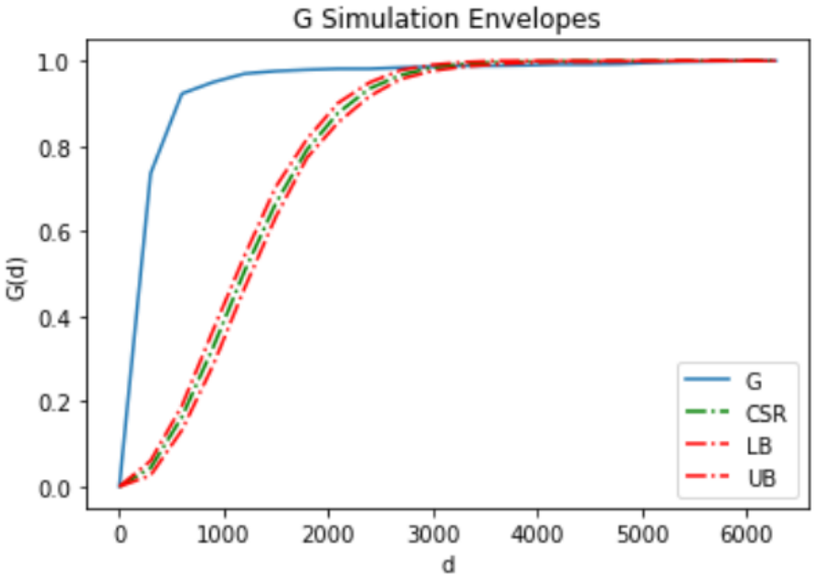




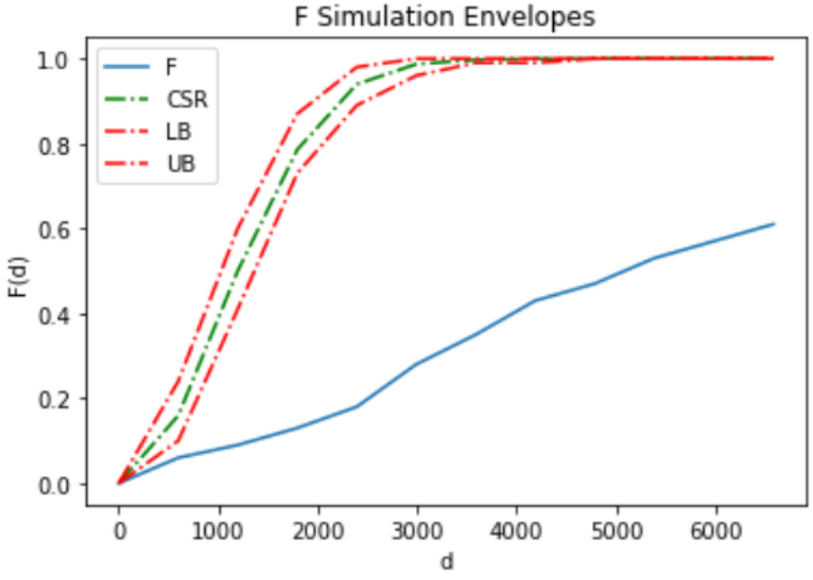


The results from all functions suggest that the point group's spatial distribution pattern is characterized by a clustering pattern.

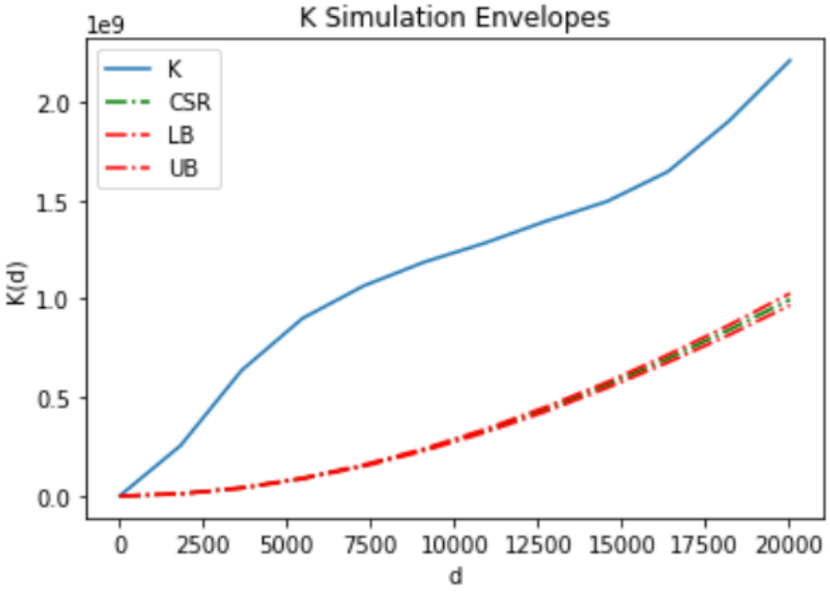
Simulation Envelopes & CSR Evaluation



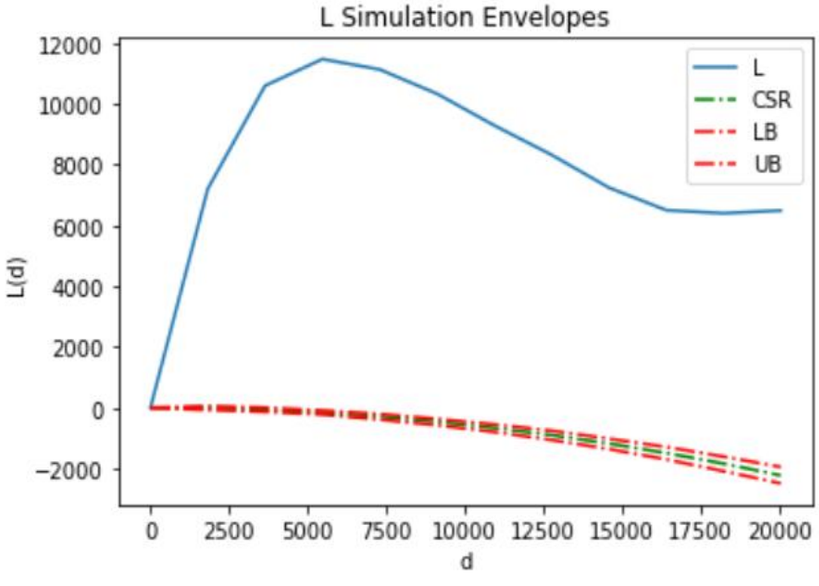
Nearest Neighbor Distance Distribution Simulation: Mostly over the expected confidence interval, generally suggesting a clustering pattern rather than CSR.



Empty Space Simulation: Consistently below the expected confidence interval, indicating a clustering pattern instead of CSR.



Ripley's K Simulation: Consistently over the expected confidence interval, indicating a clustering pattern instead of CSR.



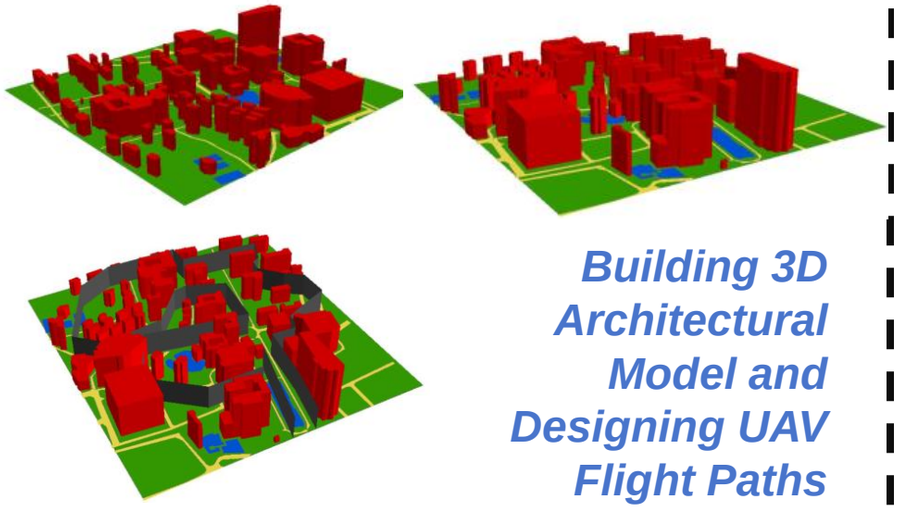
Besag's L Simulation: Consistently over the expected confidence interval, indicating a clustering pattern instead of CSR.

Conclusion: Jiaxin's public facility points exhibit a "clustering distribution" in the spatial pattern.

05 OTHERS

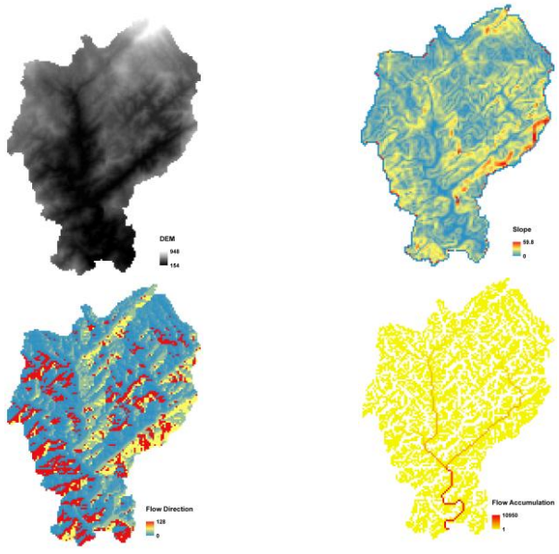
A. OTHER GIS WORKS

3D Modeling (November, 2022)

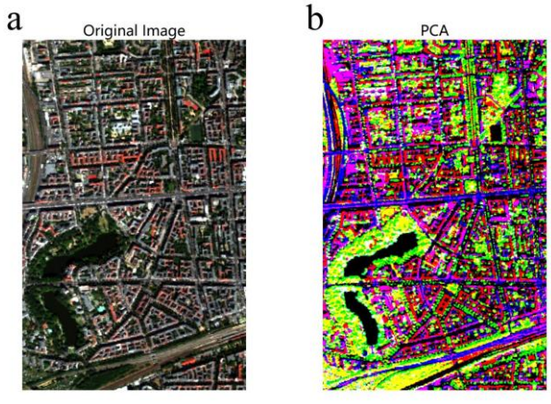


Building 3D Architectural Model and Designing UAV Flight Paths

Hydrological Analysis (January, 2023)



Hyperspectral Image Processing (May, 2023)



Hyperspectral Image Dimension via Principal Component Analysis

Cloud Programming (July, 2023)

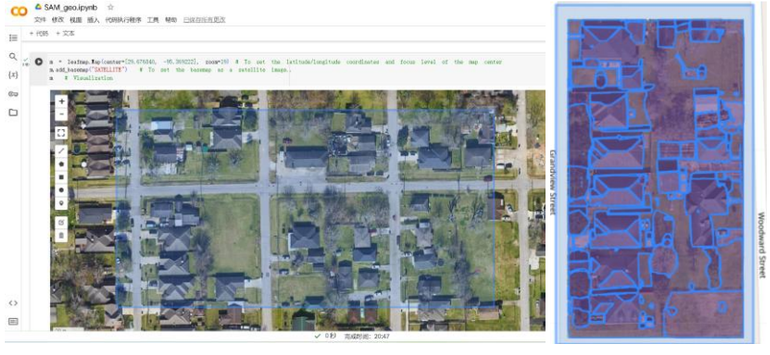
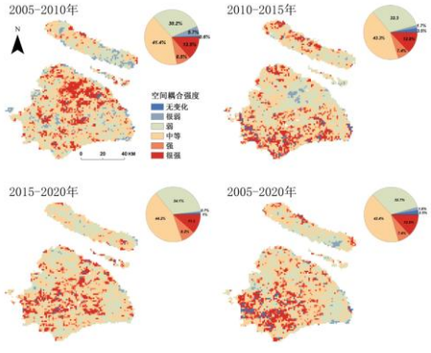
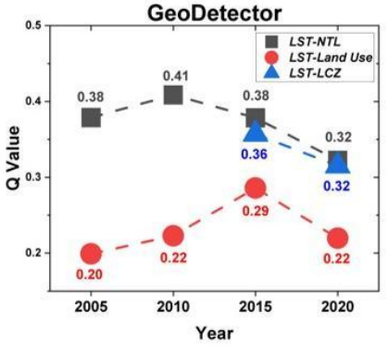


Image Segmentation Based on Segment Anything Model (SAM)

Urban Climate Research (Team Project | July-September, 2023)





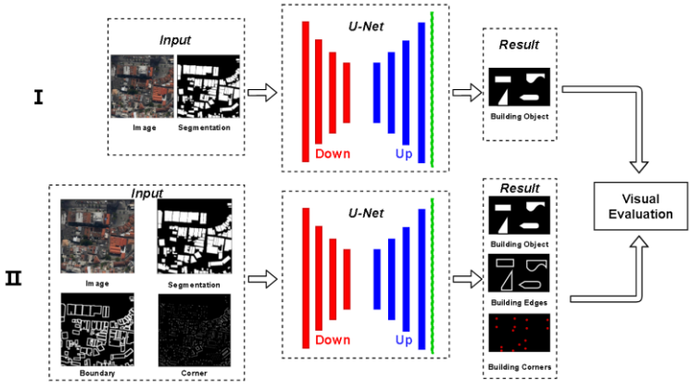
Spatiotemporal Analysis of Urban Heat Island in Shanghai, 2005-2020

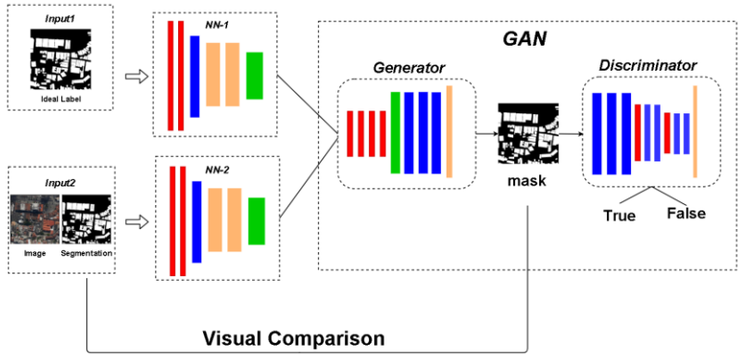
Contribution:

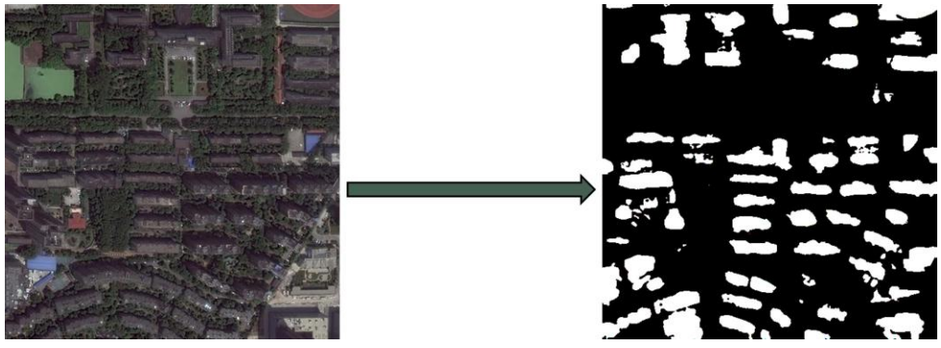
1. Data collection & processing
2. Spatial Coupling Index calculation
3. Geodetector simulation

A. OTHER GIS WORKS

Deep Learning & Computer Version(December, 2022)

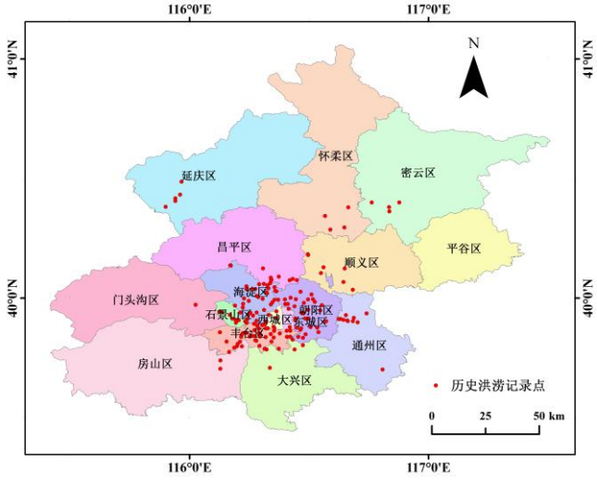


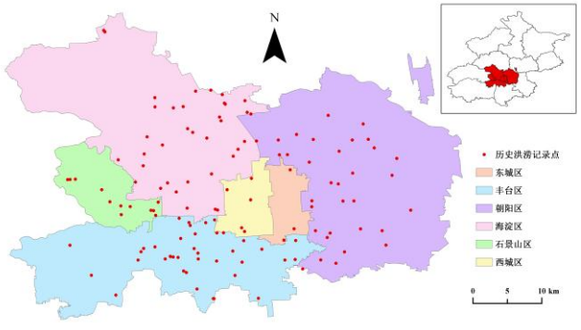


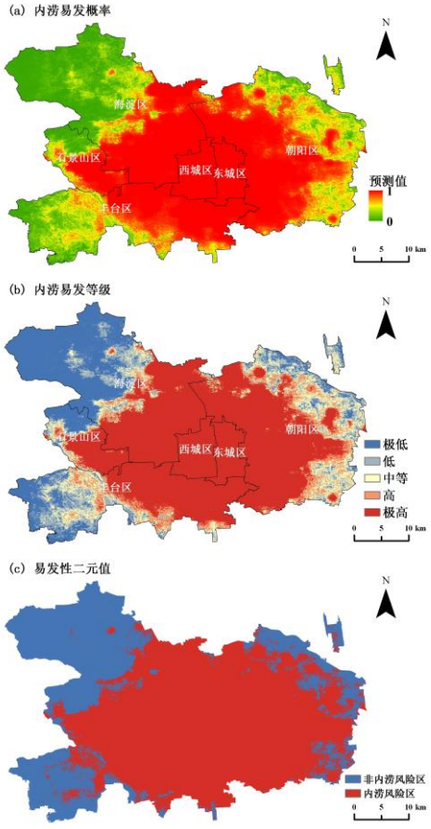


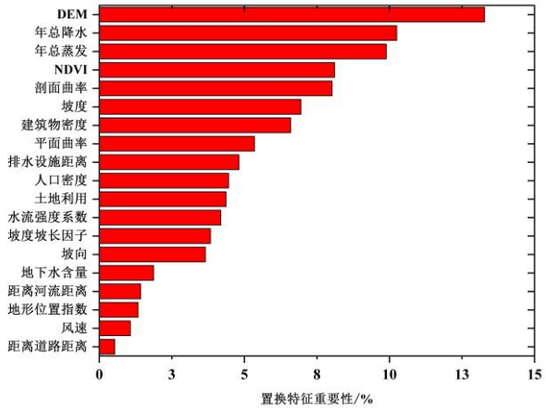
Building Segmentation and Regularization Based on Vertex and Edge Prediction

Machine Learning & Spatial Data Modeling (December, 2023 to May, 2024)









Analysis of Urban Flood Susceptibility Based on a Positive-Unlabelled SemiSupervised Machine Learning Algorithm

B. WRITING

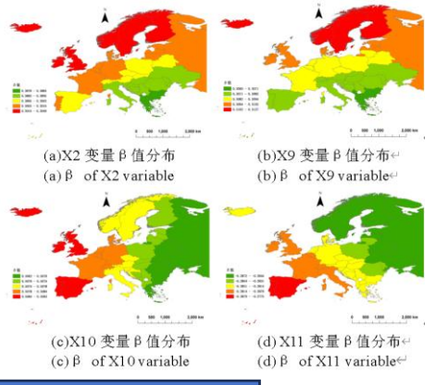
Analysis of Service Industry Development and the Relevant Indicators in European Countries

Chen Yuanpeng

School of Geography and planning, Sun Yat-Sen University, Guangzhou 510000, China

Abstract

Objective: Following the development of the commercial sector, the service industry, as a new type of comprehensive industry structure serving the general public, plays a significant role in various fields such as production, consumption, healthcare, finance, education, transportation, agriculture, forestry, animal husbandry, sports and entertainment, and scientific research in modern society. As the industry continues to develop and diversify, modern service industries have also derived numerous emerging industries from the most basic production and exchange of goods, meeting the diverse needs of the general public, expanding economic interaction between individuals and countries, and easing problems of social division of labor and its allocation. As there are many European countries and regions with varying levels of development, and their corresponding service industries also show differences in development. At the same time, the factors and degrees of influence on the development of service industries in various countries also vary. Therefore, it is of research value to evaluate the development level of service industries in different countries, and based on geographical location factors, to explore the size of the impact of different factors on the development of service industries in different regions. This paper will use relevant methods to achieve the above objectives.



(a)X2 变量β值分布 (a)β of X2 variable

(b)X9 变量β值分布 (b)β of X9 variable

(c)X10 变量β值分布 (c)β of X10 variable

(d)X11 变量β值分布 (d)β of X11 variable

Spatial Analysis

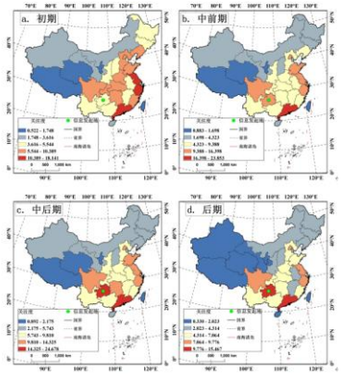
The Impact of Geographical Distance on Dissemination of Network Information: Analysis based on changes in spatial agglomeration and regression statistical methods

Chen Yuanpeng

(School of Geography and planning, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract: The study aims to investigate the influence of geographical distance on network information dissemination, as controversial viewpoints concerning the "death of distance" or "end of geography" have attracted significant attention because of the rapid development of Internet technology. By implementing and combining spatial and statistical methods, the underlying mechanism and indicators can be initially revealed. The study selects relevant hot Internet topics, TV shows and their corresponding Baidu search indices within China. Then an attention influence model is employed to estimate the Internet attention index, the attention levels of a specific topic or show. Furthermore, by introducing the global Moran's I index, analysis upon the spatial pattern of attention index can be achieved. Additionally, socio-economic and cultural indicators are incorporated, and a random forest regression model is utilized to assess the significance of geo-distance, domain counts and the indicators above on, the dissemination of network information. The findings indicate that geographical distance continues to influence network information dissemination, and its impacts diminish over time. The attention level of network information exhibited to be influenced by population age structure and culture factors. The study confirms the importance of distance as a essential factor in geographical research, including the era of Internet information.

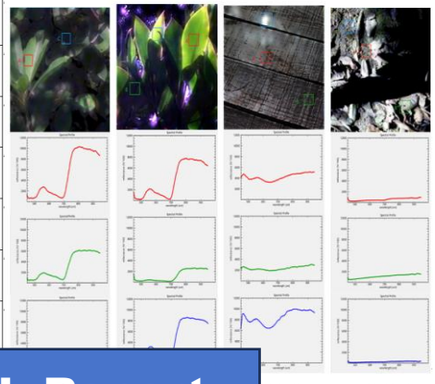
Keywords: Geographical distance, Internet attention index, Spatiotemporal changes, Spatial analysis, Random Forest regression



Geospatial Big Data

遥感解译标志表格

遥感影像	土地利用类型（编号+名称）	影像解译标志描述
	01-耕地	带有一定的植被特征，而且具有明显的平行状纹理
	03-林地	具有一定高度，而且整体上亮度较低，纹理特征较为粗糙
	04-草地	整体形态比较平整，而且亮度较林地、耕地等植被明显更高
	07 住宅用地	由密集的屋顶组成，具有明显的村庄居民地形态
	10-交通运输用地	具有明显的道路特征
	11-水利集水利设施用地	海洋部分水体整体上呈蓝色，少数湿地和湖泊水体呈黑色
	12-	



(d)枯叶泥土

Fieldwork Report

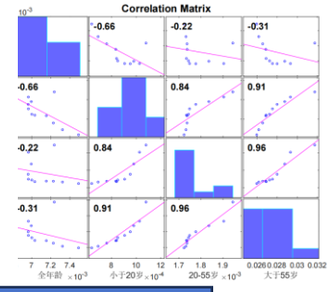
Examining Trends in China's Mortality: A Statistical Analysis Using T-Tests and Regression Models

Yuanpeng Chen

(School of Geography and Planning, Sun Yat-sen University, GuangZhou,510006)

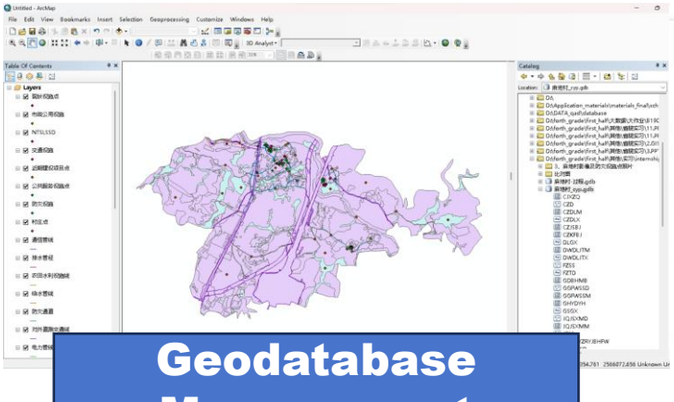
Abstract: As the country with the largest population in the world, China has a huge annual death toll and complex internal factors. There is a certain relationship between the number of deaths and mortality under the influence of different age groups, different causes of death, and different health indicators. Therefore, it is of certain significance to adopt appropriate methods to analyze the internal change mechanism of the dead population. This study collected and collated the death toll data of all age groups, influencing factors and gender in China from 2009 to 2019, as well as the index data to evaluate the national health status, and analyzed the correlation between mortality and gender, age and influencing factors through single sample t-test and regression analysis. The results showed that there was a certain relationship between mortality and overall mortality under different gender, different age groups and different influencing factors, and the linear relationship between mortality and related influencing factors could be constructed by regr

Keywords: mortali

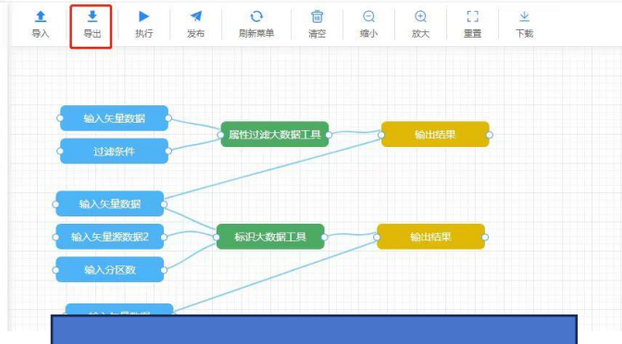


Statistics

C. INTERNSHIP WORK SAMPLES

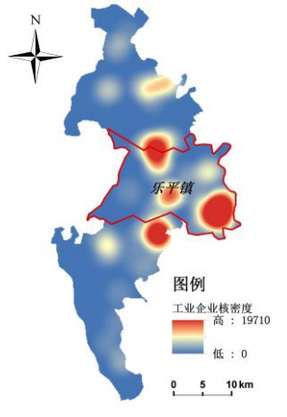


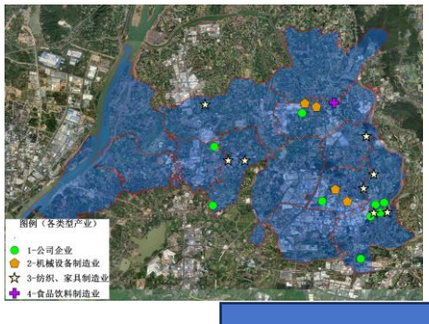
Geodatabase Management

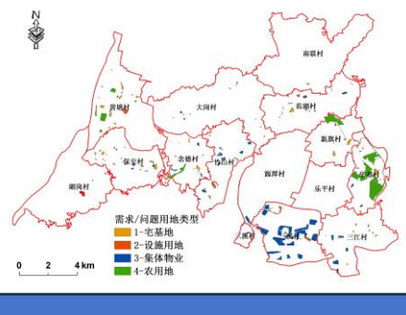


Web Modeling

Spatial Analysis







Urban & Rural Planning Maps


THANK YOU!