Boram Lee

**Exploratory Data Analysis**

1. Raw data  
   *\* I made new feature BVW (=NDBI/(NDVI\*NDWI)) from intuitive assumption that UHI is directly proportional to NDBI and inversely proportional to NDVI and NDWI. But Conclusively, that features did not show valuable correlation with UHI Index*

* Heatmap for Raw data with correlation
  + Auto-correlated among bands – reasonable results

A screenshot of a graph

Description automatically generated

* Pearson Correlation and Visualization with Scatter plots  
  \* Pearson correlation: -0.3597(NDVI), 0.3129 (NDBI), 0.3582 (NDWI), 0.1754 (BVW)

A blue dotted diagram with a red line

Description automatically generatedA blue dotted diagram with a red line

Description automatically generatedA diagram of a scatter plot with a red line

Description automatically generatedA diagram of a scatter plot

Description automatically generated

Why ndvi shows negative ~

Presentation

Final questions on ending part in presentation

1. Average Data with new feature from Tip3 filtering

* Heatmap for Average Data with correlation

A screen shot of a graph

Description automatically generated

* Pearson Correlation and Visualization with Scatter plots  
  \* Pearson correlation: -0.4265 (NDVI), 0.4324 (NDBI), 0.4285 (NDWI)

A blue dotted diagram with a red line

Description automatically generatedA blue dotted diagram with a red line

Description automatically generatedA blue dotted diagram with a red line

Description automatically generated

*Concluded Assumption*

1. Averaged data shows much better for training UHI Index
2. Try feature engineering such as PCA and feature selection from tree-based importance
3. Try linear regression model briefly because abs(correlation) is higher than 0.4 in NDVI, NDBI, AND NDWI.
4. Try Deep Neural Network

**Feature Engineering**

1. Principal Component Analysis

A blue and red dotted lines

Description automatically generated with medium confidence

PCA on 50 meter averaged data

1. Feature selection from tree-based importance

rf = RandomForestRegressor(n\_estimators=100, random\_state=42)

rf.fit(X, y)

feature\_importance\_rf = rf.feature\_importances\_

feature\_importance\_df = pd.DataFrame({"Feature": feature\_names, "Importance":feature\_importance\_rf})

feature\_importance\_df = feature\_importance\_df.sort\_values(by="Importance", ascending=False)

print("RandomForest Feature Importance:")

print(feature\_importance\_df.head(5))

=================================

RandomForest Feature Importance:

Feature Importance

0 B01 0.181728

3 B04 0.095628

17 Soil\_Ratio 0.068597

9 B11 0.058452

4 B05 0.051652

반경 2000m 내의 **위성 밴드 데이터들을 평균낸 값이 원본(raw) 데이터보다 UHI Index 예측 성능이 더 높은 이유**는 여러 가지로 분석할 수 있어.

**📌 1. Spatial Autocorrelation (공간 자기상관)**

• 지표면 온도(UHI Index)나 위성 밴드 값들은 **공간적으로 연관성이 매우 높은 데이터**야.

• **근처 픽셀들의 값이 유사한 패턴을 가질 가능성이 크기 때문에**, 반경 2000m 내에서 평균을 내면 **노이즈를 줄이고, 더 신뢰성 있는 대표 값을 만들 수 있어.**

• 이는 공간 데이터를 다룰 때 **Moran’s I** 같은 공간 자기상관 분석에서 흔히 나타나는 현상과 관련이 있어.

**📌 2. Noise Reduction (노이즈 감소)**

• 원본(raw) 위성 데이터는 **구름, 대기 산란, 센서 오류** 등의 영향을 받을 수 있어.

• **반경 2000m 내 평균을 내면 이상치(Outliers)나 잡음(Noise)이 완화**되면서 예측 모델이 더 일반화된 정보를 학습할 수 있음.

• 특히, XGBoost 같은 결정 트리 기반 모델은 **노이즈가 많은 데이터를 입력으로 받을 경우, 과적합(overfitting)할 가능성이 높음**.

**📌 3. Scale Consistency (스케일 일관성 증가)**

• 원본 위성 데이터는 픽셀 단위(보통 10~60m 해상도)로 제공되기 때문에, **실제 UHI Index(도시 열섬 효과)와 공간적 크기 차이가 있을 수 있음**.

• UHI Index는 **넓은 지역의 평균적인 지표면 온도를 반영하는 경우가 많기 때문에, 2000m 반경 내 평균 밴드 값이 실제 지표 환경을 더 잘 나타낼 수 있음**.

**📌 4. Feature Smoothness (특징 스무딩 효과)**

• 개별 픽셀 단위 데이터는 너무 세부적일 수 있어 **일반적인 패턴을 학습하는 데 방해**가 될 수 있음.

• 반경 2000m 내 평균을 취하면, **데이터의 변동성이 완화되어 모델이 전체적인 패턴을 더 잘 학습**할 수 있음.

• 이는 특히 **Gradient Boosting 기반 모델(XGBoost, LightGBM 등)이 지나치게 작은 변화를 학습하는 것을 방지하는 데 도움이 됨**.

**📌 5. Local Context Capture (지역적 맥락 반영)**

• 한 지점의 위성 밴드 값만 사용하는 것보다, **주변 지역의 평균 값을 함께 고려하는 것이 더 현실적인 UHI Index를 반영할 수 있음**.

• 예를 들어, **한 픽셀의 NDVI(식생 지수)가 낮다고 해서 그 지역 전체의 녹지율이 낮다고 단정할 수 없음**.

하지만 반경 2000m 내 평균을 내면 **해당 지역이 전체적으로 녹지가 많은지(혹은 적은지) 더 명확하게 파악할 수 있음**.

**✅ 결론**

**반경 2000m 내 평균을 낸 위성 밴드 데이터가 XGBoost 모델의 예측 성능을 향상시킨 이유는:**

1. **공간 자기상관**을 활용하여 UHI Index와 더 관련성 높은 데이터 제공

2. **노이즈 감소** 효과로 모델이 더 신뢰할 수 있는 정보를 학습

3. **스케일 일관성을 높여** 실제 도시 열섬 효과와 데이터 간의 차이 줄임

4. **스무딩 효과로 특징 변동성을 완화**하여 과적합 방지

5. **지역적 맥락을 반영하여** 개별 픽셀보다 더 현실적인 패턴을 학습

**📌 GWR(Geographically Weighted Regression) 모델이 공간적 패턴을 반영하는 방식**

GWR은 **일반적인 선형 회귀 모델(OLS)** 과 다르게 **각 데이터 포인트(위경도 좌표)마다 회귀 계수를 다르게 학습**하는 방식이야. 즉, **공간적으로 가까운 데이터들에 더 큰 가중치를 부여하면서 지역별로 다른 관계를 학습**하는 거지.

**✅ 1️⃣ OLS(일반 선형 회귀) vs GWR(공간 가중 회귀) 차이**

• **OLS (Ordinary Least Squares, 일반 선형 회귀)**

• 하나의 글로벌(전역) 회귀 계수를 학습함.

• y = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ + ϵ

• **모든 지역에서 동일한 β 값을 사용** → 공간적 변화 반영 ❌

• **GWR (Geographically Weighted Regression, 공간 가중 회귀)**

• 각 데이터 포인트마다 **다른 회귀 계수 β** 를 추정함.

• y(i) = β₀(i) + β₁(i)X₁(i) + β₂(i)X₂(i) + ... + βₙ(i)Xₙ(i) + ϵ

• 여기서 i는 특정한 위치(좌표)에서 추정된 값.

• **공간적으로 가까운 점일수록 더 높은 가중치 적용** → 공간적 변화 반영 ✅

**✅ 2️⃣ GWR이 공간 패턴을 반영하는 원리**

GWR은 **로컬(지역별) 회귀 분석을 수행**하는데, 이때 **공간적으로 가까운 데이터에 더 큰 가중치를 부여하는 커널을 사용**해. 즉, 특정 위치에서의 회귀 계수를 추정할 때, 그 위치에서 가까운 데이터가 더 큰 영향을 미치도록 함.

**✨ GWR의 핵심 개념**

1. **각 데이터 포인트(위경도 좌표)마다 별도의 회귀 계수(β) 계산**

→ y(i) = β₀(i) + β₁(i)X₁(i) + ... + βₙ(i)Xₙ(i) + ϵ

2. **공간적으로 가까운 데이터에 더 큰 가중치를 부여**

• 거리 가중치 함수(w(i, j))를 사용하여 **가까운 지역일수록 더 높은 가중치 적용**

• 가중치는 보통 **커널 함수(Kernel Function)** 를 사용하여 정의됨 (예: 가우시안, 비스퀘어 등)

• 예를 들어, 비스퀘어(Bisquare) 커널이면:

• 가까운 지역 데이터는 **높은 가중치**

• 일정 거리 이상 떨어진 지역 데이터는 **0 가중치** 처리

3. **각 지역별로 다르게 최적화된 회귀 계수를 학습**

• 특정 위치에서의 회귀 계수는 **그 위치 주변 데이터에 의해 결정됨**

• 즉, NDVI(식생지수)나 NDBI(건물지수)의 효과가 **지역마다 다르게 반영됨**

**✅ 3️⃣ GWR이 공간적 패턴을 반영하는 방식 (예제)**

**(1) OLS와 비교**

**📍 예제: 식생지수(NDVI)와 건물지수(NDBI)가 UHI에 미치는 영향**

• **OLS (전역 회귀)**

• UHI = β₀ + β₁(NDVI) + β₂(NDBI) + ϵ

• 모든 지역에서 NDVI의 계수(β₁)와 NDBI의 계수(β₂)가 **같은 값**을 가짐

• 즉, **도시든 시골이든 동일한 관계를 가정** → 공간적 차이 반영 ❌

• **GWR (공간 가중 회귀)**

• UHI(i) = β₀(i) + β₁(i)(NDVI) + β₂(i)(NDBI) + ϵ

• 지역마다 NDVI와 NDBI의 계수(β₁(i), β₂(i))가 다름

• **예제:**

• 도심지에서는 β₁(i) = -0.01 (NDVI가 증가할수록 UHI가 약간 감소)

• 농촌 지역에서는 β₁(i) = -0.05 (NDVI가 증가할수록 UHI 감소 효과가 큼)

• 도심지에서는 β₂(i) = 0.03 (NDBI가 증가할수록 UHI가 증가)

• 교외 지역에서는 β₂(i) = 0.01 (NDBI가 증가해도 UHI에 미치는 영향이 적음)

• 즉, **각 지역에서의 NDVI, NDBI 영향이 다르게 반영됨** → 공간적 차이 반영 ✅

**(2) GWR이 공간적 패턴을 반영한 실제 결과 해석**

**📌 위에서 출력된 GWR 결과**

| **Variable** | **Mean (평균)** | **Min (최소)** | **Max (최대)** | **해석** |
| --- | --- | --- | --- | --- |
| **X4 (NDVI)** | -0.003 | -0.659 | 0.764 | **어떤 지역에서는 NDVI가 UHI를 감소(-0.659), 어떤 지역에서는 증가(+0.764)** |
| **X6 (NDBI)** | 0.003 | -0.730 | 0.823 | **어떤 지역에서는 건물지수(NDBI)가 UHI를 증가(+0.823), 어떤 지역에서는 감소(-0.730)** |

→ 즉, **NDVI와 NDBI가 지역별로 다른 영향을 미친다는 것을 GWR이 반영하고 있음.**

**✅ 4️⃣ GWR을 왜 써야 할까?**

• **기후, 도시 구조, 지형 등의 차이로 인해 같은 변수라도 지역마다 다른 영향을 미치기 때문**

• **OLS 같은 전역 회귀는 공간적 변화를 반영하지 못함**

• **GWR은 지역별로 다른 회귀 계수를 학습하여 더 정밀한 모델을 생성 가능**

**✅ 5️⃣ 결론**

• GWR은 **지역별로 다른 회귀 계수를 학습하여 공간적 패턴을 반영**함.

• **공간적으로 가까운 데이터에 더 큰 가중치를 부여**하여 특정 지역의 특성을 잘 반영함.

• 결과적으로 **OLS보다 훨씬 더 높은 R² (설명력)를 가짐** (예: OLS R² = 0.172 → GWR R² = 0.944).

• **도심과 농촌에서 같은 변수(NDVI, NDBI)라도 다른 영향을 미칠 수 있음을 보여줌.**

• 따라서 **UHI 같은 공간적 특성이 강한 데이터에는 GWR이 훨씬 효과적**임.

혹시 더 궁금한 점 있으면 질문해줘! 😊

*Participants should consider the following in their presentation:*

*• How might local governments or urban planning decision-makers use your model?*

*• How might your model and approach be scaled to other cities?*

*• How could your model be used to address socioeconomic impact on vulnerable*

*communities?*

*• How could your model be used to address impacts on energy demand?*

*• What additional datasets might*

A graph of a window size

Description automatically generated

A graph of different sizes and numbers

Description automatically generated