EPFL

Semester project

Data-driven Analysis of City Scale Human-building Interaction

Yihan Wang

MSc Civil Engineering

Supervisor:

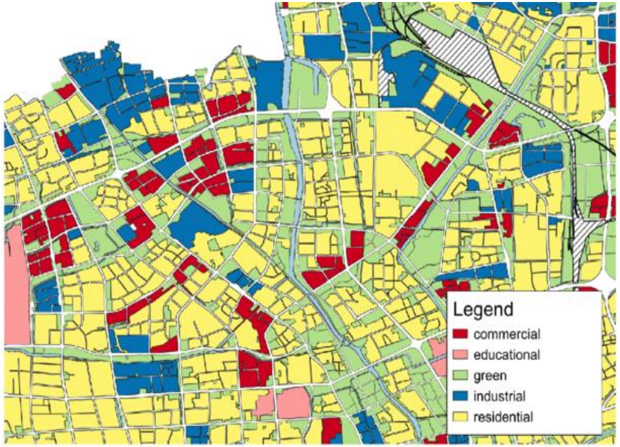
Andrew Sonta, Kanaha Shoji

24th Jan 2024

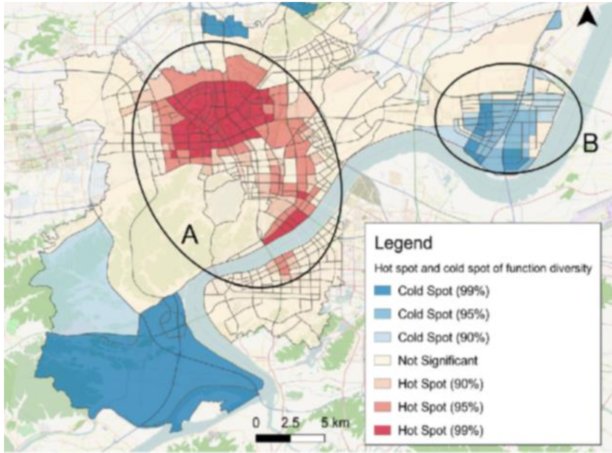
École polytechnique fédérale de Lausanne

Background : Urban planning and human living experience

West Lake, Hangzhou, China



Hangzhou urban planning map[1]



Hangzhou function diversity hot and cold spot map[1]

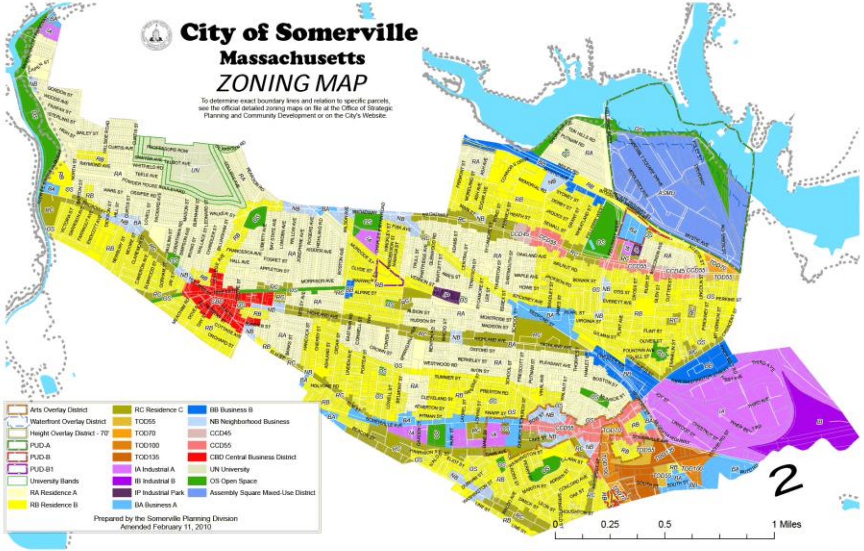
Extraordinary view existence ≠ Good living experience !

1. Qiu, Yining, et al. "Understanding the urban life pattern of young people from delivery data." *Computational Urban Science* 1 (2021): 1-16.

Background

More challenging urban design exists…

- Euclidean zoning (areas with specific rules for building use)



Jeopardizing vibrant and socially resilient communities[1]

Question: What kind of urban design is considered benefiting internal individual residents and society?

Objective: Understanding human's interaction with urban forms

2. Jacobs, J. (1961). The Death and Life of Great American Cities. New York.

Data-driven Analysis of City Scale Human-building Interaction

OUTLINE

Data Exploration

- Pedestrian Counts
- Urban Forms
- Other Features

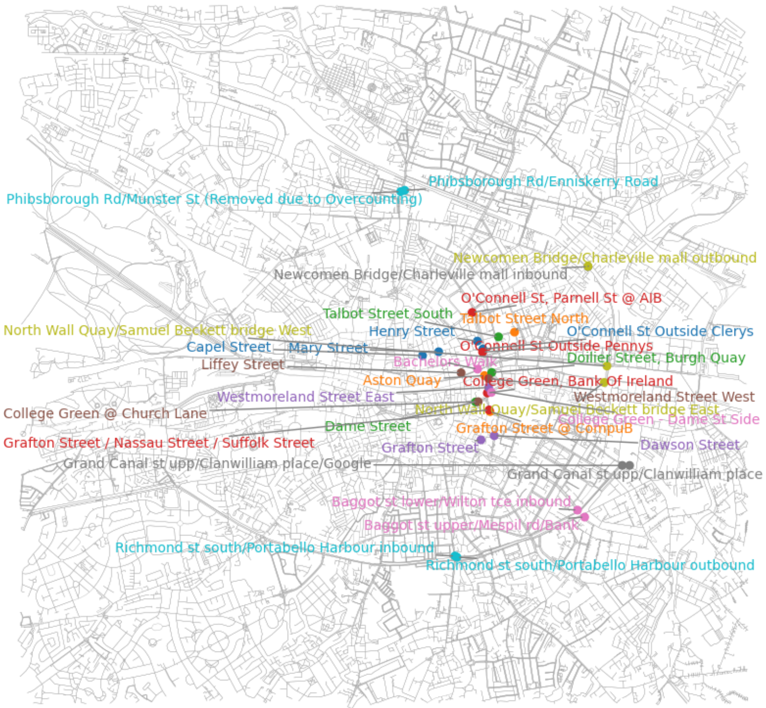
Model Regression & Interpretation

Yihan Wang

Data – Pedestrian Counts : Sensor Locations

Dublin

Sensor Locations on Street Network in a region in Dublin

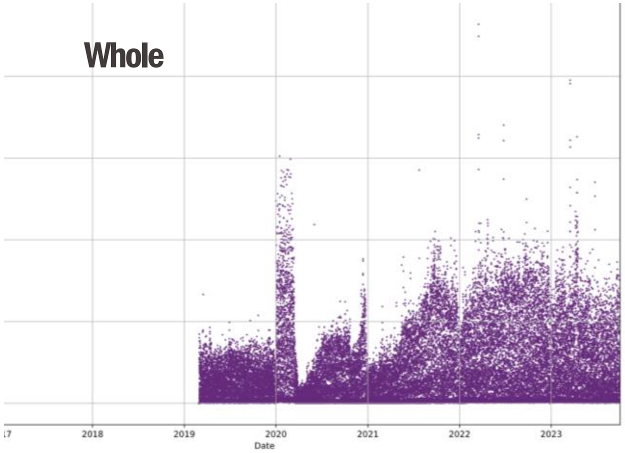


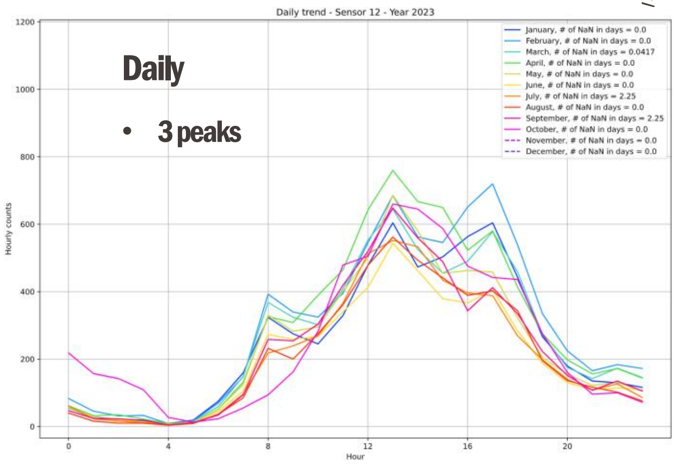
- 34 pedestrian counters
- Densely distributed around central region

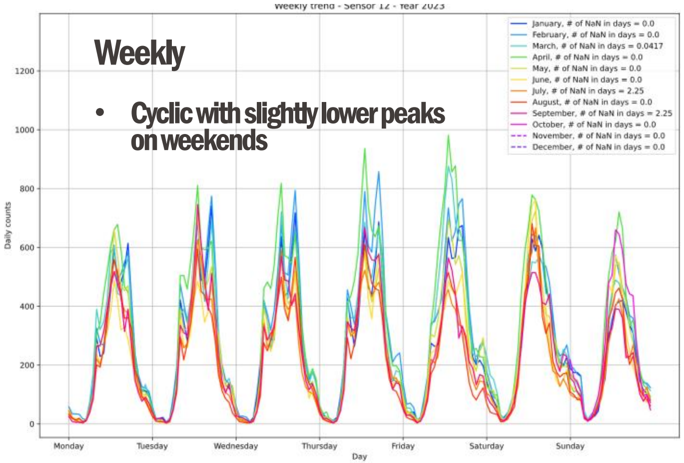
Data – Pedestrian Counts : Hourly Pedestrian Flow

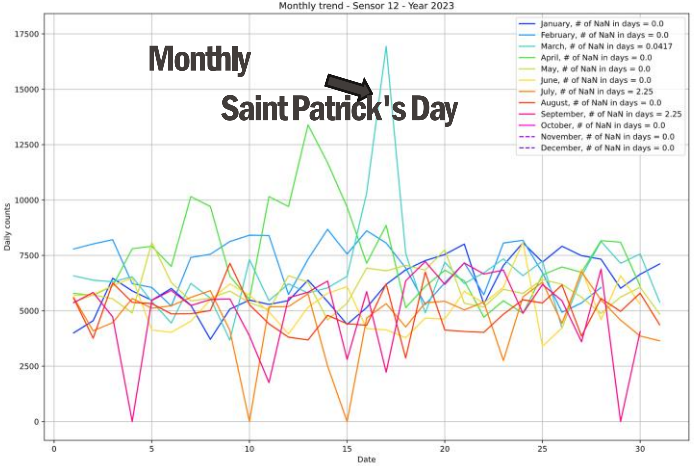
Sensor ID: 12

College Green, Bank of Ireland









O U T L I N E

Data Exploration

- Pedestrian Counts
- **Urban Forms**
- Other Features

Model Regression & Interpretation

Data - Urban Forms: From Counter Nearest Node Polygon

Sensor ID: 12

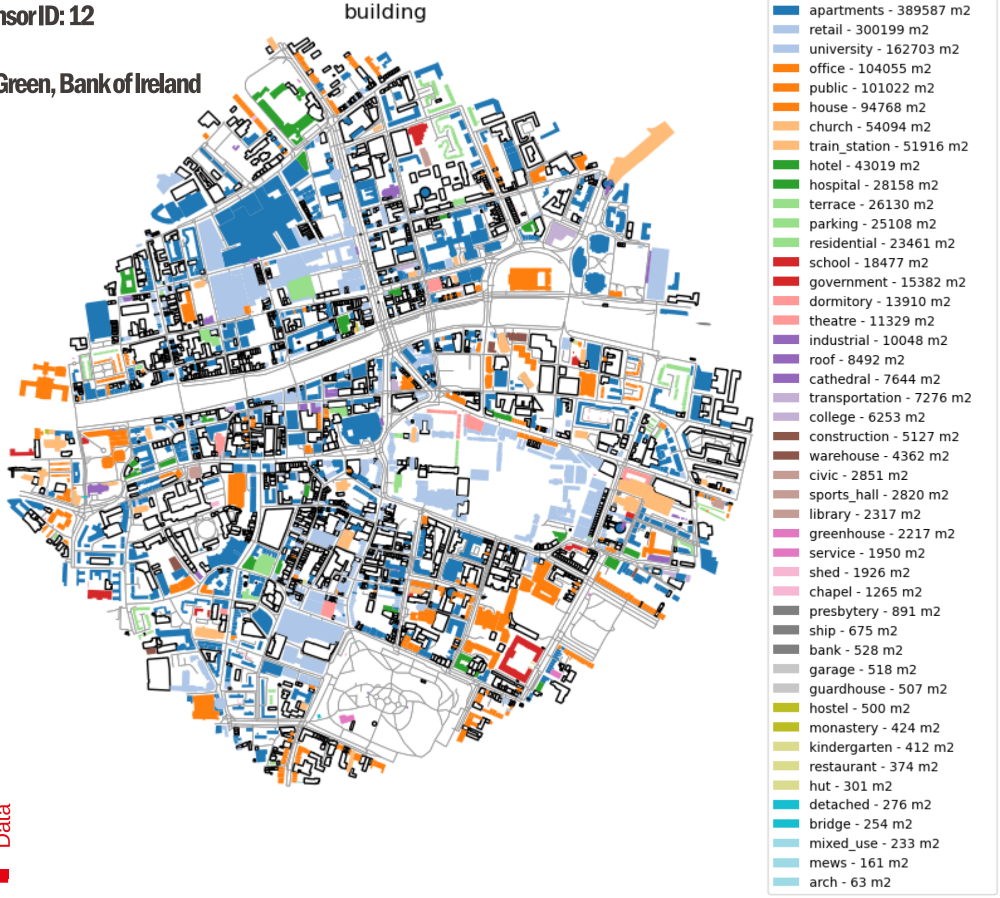
College Green, Bank of Ireland

Determine the closest node → Retrieve 15 mins walk polygon → Inside urban forms

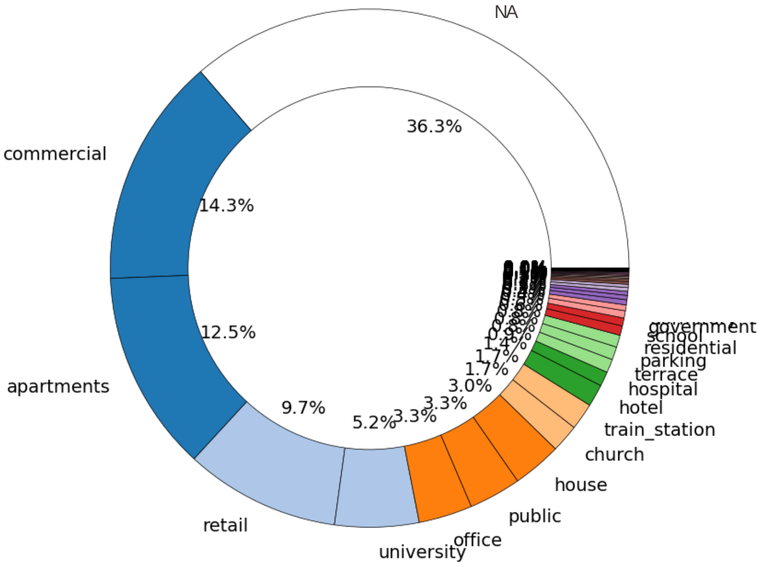
Data - Urban Forms: Polygon Urban Forms

Sensor ID: 12

College Green, Bank of Ireland

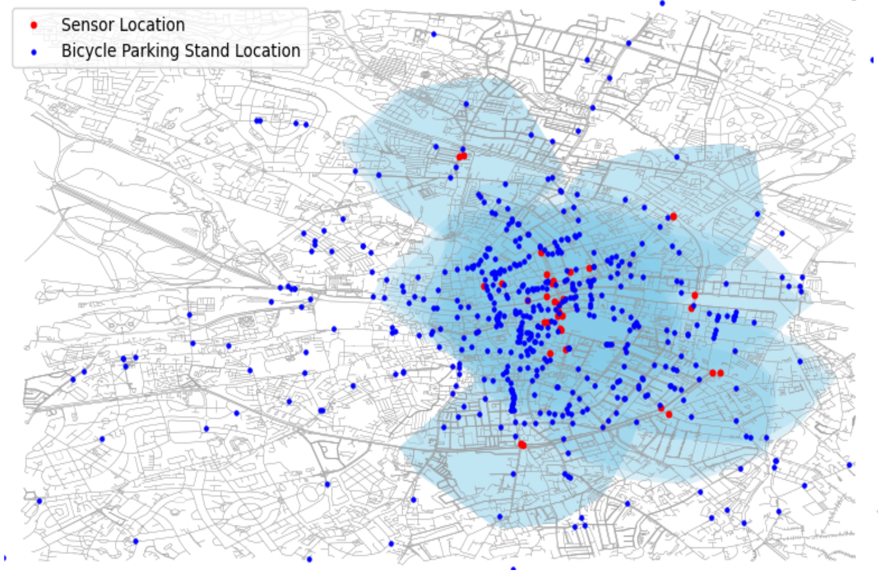


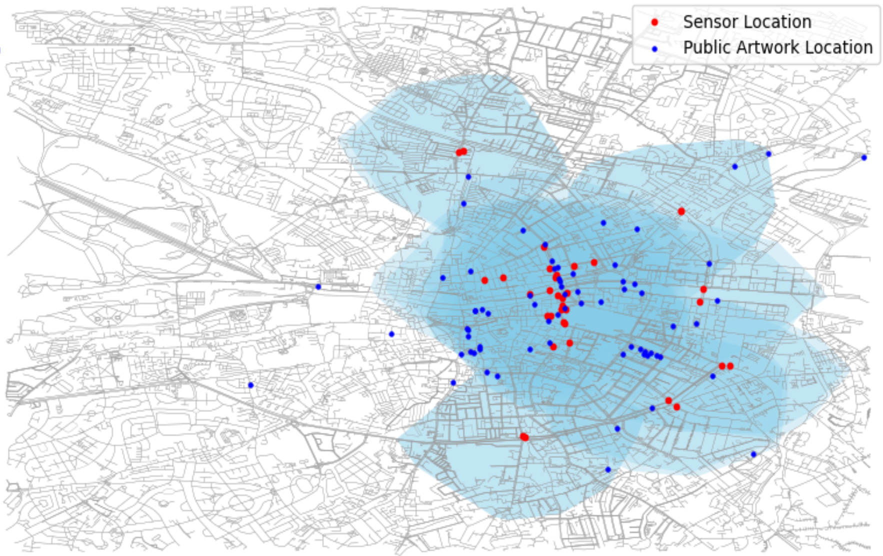
Total Area of Each Building Type (Ordered by Area) within the 15 min walk distance from College Green Counter



Data - Urban Forms: Polygon Urban Forms

Yihan Wang





O U T L I N E

Data Exploration

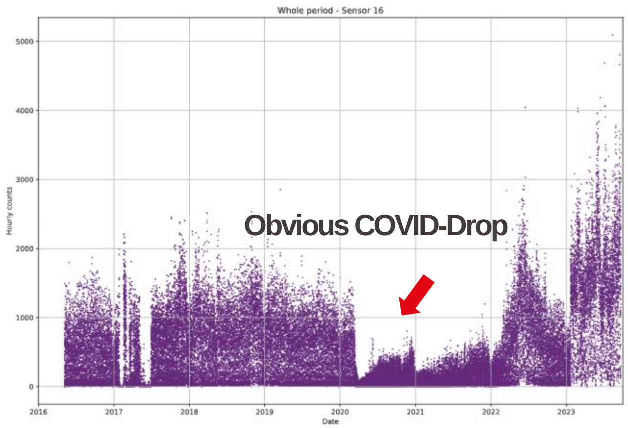
- Pedestrian Counts
- Urban Forms
- Other Features

Model Regression & Interpretation

Model – Data Preprocessing

Melbourne

- Filtering out data with COVID-19 restrictions: (→ 4115348 × 135 dataframe)



Restrictions including:

- School Closing
- Workplace Closing
- Cancel Public Events
- Restrictions on Gatherings
- Close Public Transports
- Protection of elderly people
- Stay at Home Requirements
- Restrictions on Internal Movement
- International travel controls
- Restrictions on Gatherings
- Contact Tracing
- Facial Coverings

- One-hot encoding for categorical features
- Z - score Standardization: $x_{i,j}{}' = \frac{x_{i,j} - min(x_j)}{max(x_j) - min\,(x_j)}$
- Normalization: $x'_{i,j} = \frac{x_{i,j} - mean(x_j)}{std\,(x_j)}$
- Deal with NaN values: fill with feature means

Model – Setup

Interpretable model: linear regression

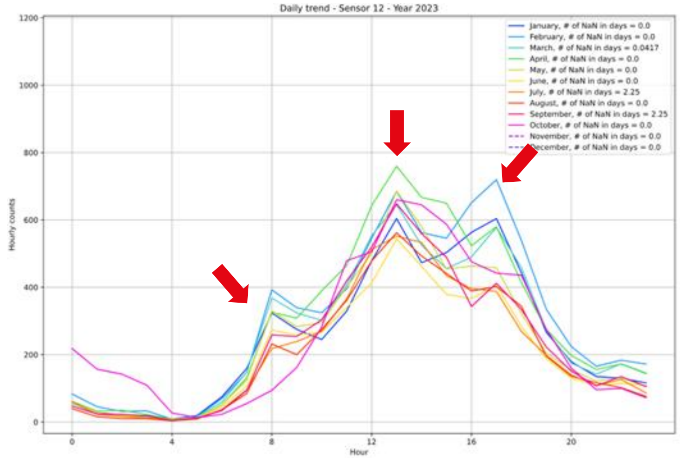
Dependent Variable		Hourly pedestrian counts		Perimeter Mean	Recreational	PopSum	distance	B15_Commercial_per	B5_others_per	S15_street_length_avg
Method		Ordinary Least Squares		Perimeter Stdev	Green View Mean	Men	feature_x	B15_Religious_per	S15_n	S15_circuity_avg
Predictors		Degree	Footprint Proportion	Complexity Mean	Sky View Mean	Women	feature_y	O5_WalkableLength_m	S15_m	S15_self_loop_proportion
cc	ws	Clustering	Footprint Mean	Complexity Stdev	Building View Mean	Elderly	nearest_x	O5_PublicTransportAccessCounts	S15_k_avg	S5_n
speed_u	Node Density	Closeness Centrality	Footprint Stdev	Building Count	Building View Stdev	Youth	nearest_y	O5_Amenity_Shannon_Entropy	S15_edge_length_total	S5_m
speed_v	PageRank	Betweenness Centrality	Commercial	Food	Road View Mean	Children	O15_WalkableLength_m	B5_Accommodation_per	S15_edge_length_avg	S5_k_avg
rh	Street Length	Civic	Social	Healthcare	Road View Stdev	O15_Green Area_m2	O15_PublicTransportAccessCounts	B5_Commercial_per	S15_streets_per_node_avg	S5_edge_length_total
temp_K	day_type	Entertainment	Perimeter Total	Institutional	Visual Complexity	n	O15_Amenity_Shannon_Entropy	B5_Religious_per	S15_street_length_total	S5_edge_length_avg …

Features including: weather, urban form, network property, population…

Model – Exploration

Yihan Wang

- Plain model: $R^2 = 0.287$
- Baseline model: Categorizing hour in day: $R^2 = 0.523$



- Morning Peak: 7, 8, 9
- Lunch Peak: 12, 13, 14
- Evening Peak: 16, 17, 18

Hour categorization helps explaining the variation a lot!

- Adding a predictor of 'year': $R^2 = 0.538$

Other temporal predictor remains unexplored

Model – Exploration

Baseline model: $R^2 = 0.523$

- **Exploring feature interactions: $R^2 = 0.545$**

Intuition: Young people may be influenced by day type more

Intuition: For population groups, the fractions are more meaningful.

Included interactions

Men / PopSum	Youth * Commercial	<day_type * Institutional>	< day_type * B15_Commercial_per >
Youth / PopSum	Youth * Entertainment	<day_type * Recreational>	< day_type * B15_ Religious_per >
Children / PopSum	PopSum * Commercial	<day_type * O15_GreenArea_m2>	< day_type * B5_ Religious_per >
day_type * morning_peak	PopSum * Entertainment	<day_type * O15_WalkableLength_m>	<day_type * Youth>
day_type * lunch_peak	PopSum * Building Count	< day_type*O5_WalkableLength_m >	<day_type * Men>
day_type * evening_peak	<day_type * Commercial>	<day_type * B5_Accommodation_per>	<day_type * Women>
< day_type * O15_PublicTransportAccessCounts >	<day_type * Entertainment>	<day_type * B5_Commercial_per>	<day_type * Elderly>
<day_type * O5_PublicTransportAccessCount>	<day_type * Building Count>	< day_type * B15_Accommodation_per >	<day_type* Children>

Intuition: people may have different public transport use behavior on holidays vs workdays

< >: $p < 0.05$

- **Adding additional features: $R^2 = 0.580$**

(yearly average data: gdp, green gas emission from different sections)

Model – Interpretation (with baseline model)

Decomposition of the residuals

- **Weather features:**

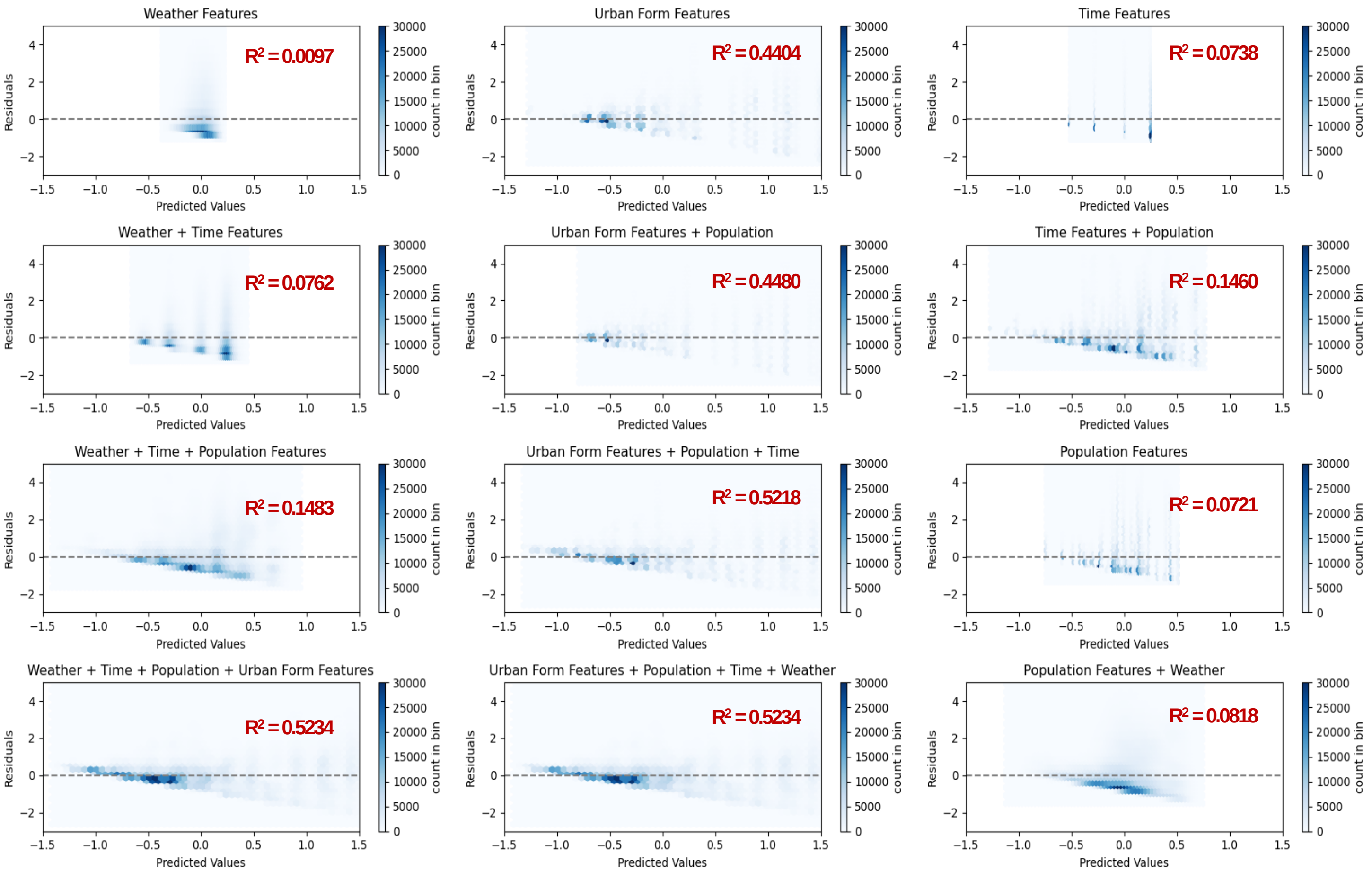
cc, speed_u, speed_v, rh, temp_K, ws

- **Time features:**

day_type, morning_peak, lunch_peak, evening_peak

- **Population features:**

PopSum, Men, Women, Elderly, Youth, Children



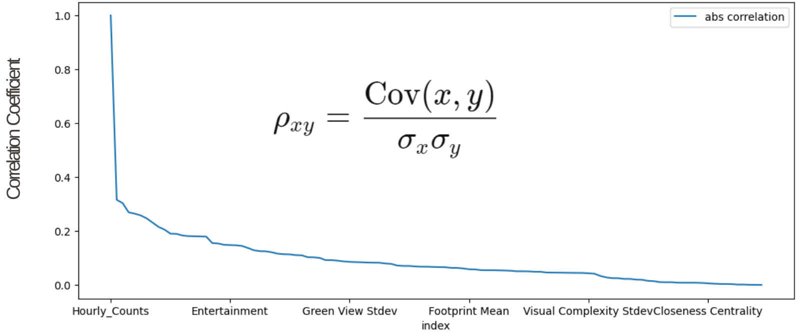
Model – Attempt to lower the feature dimension (baseline model)

Yihan Wang

Baseline model: $R^2 = 0.523$

- ## Correlation analysis

Sorted Correlation Coefficients between Features and Pedestrian Count

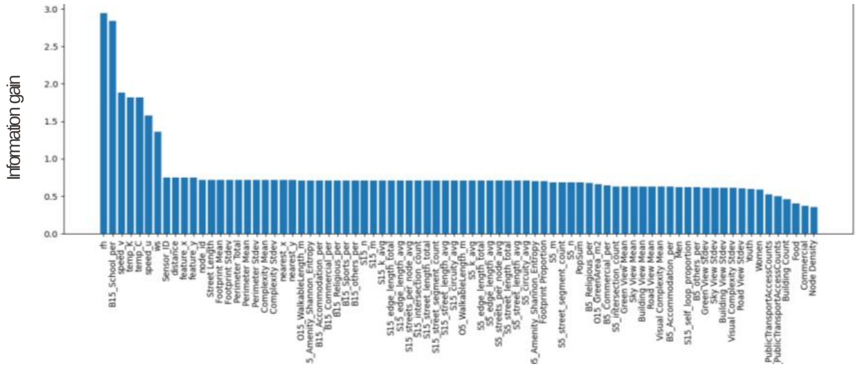


Rule out the features with correlation coefficient < 0.01

➡ $R^2 = 0.509$

cc, S5_edge_length_avg, S5_street_length_avg, S5_circuity_avg, S15_self_loop_proportion, Closeness Centrality, B5_others_per, B5_Accommodation_per, Institutional, speed_u, distance, Building View Stdev, Clustering (Weighted): 13

- ## Information gain



Rule out the features with information gain < 0.04

➡ $R^2 = 0.507$

Helthcare, day_type, lunch_peak, Degree, Clustering(Weighted), Institutional, Civic, evening_peak, Clustering, Eigenvector Centrality, PageRank, Social, n: 14

Summary

- Explored and analyzed urban features from open datasets.
- Developed mathematical models statistically explaining the spatial-temporal relationship between pedestrian flow and different groups of features.
- Implemented traditional machine learning techniques, improving the model R^2 up to **0.580**, with Luis F. Miranda-Moreno et al.[1] highest $R^2 = \mathbf{0.60}$.

[1]. Luis F. Miranda-Moreno and David Fernandes. Modeling of Pedestrian Activity at Signalized Intersections Land Use, Urban Form, Weather, and Spatiotemporal Patterns