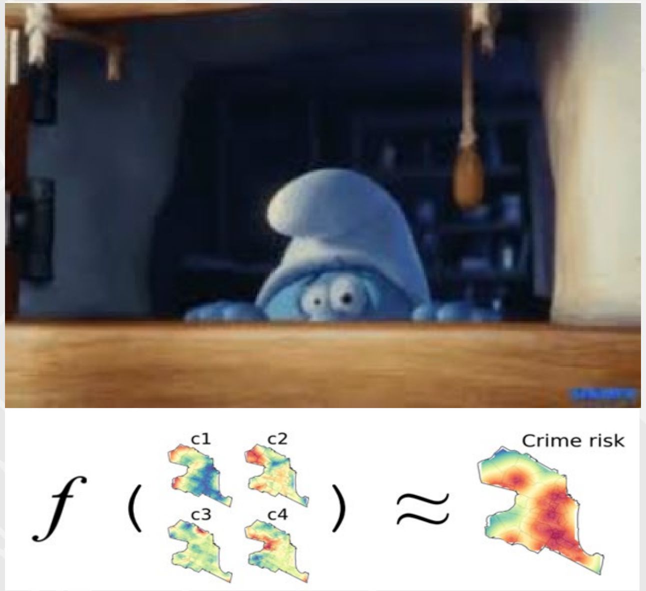
Testing spatially conscious machine learning models to forecast crime.

A case study for the prediction of acquisitive crime in Budapest.

25th Annual Conference of the European Society of Criminology,
3-6 September 2025 Athens, Greece

Ourania Kounadi | Andrea Pődör



ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY
universität wien
Digital Geography
University of Vienna

1. Introduction| Background

What is Spatially conscious machine learning modelling (spatial ML)?

→ Spatial Statistics & Machine Learning

→Three approaches:

1) **Inclusion of spatial features in original algorithms (Feature Engineering)**
2) Hybrid Models with spatial statistics (GW – RF)
3) Spatial Cross Validation strategies (interpolation or extrapolation of geodata)

Deng, He, & Liu, 2023

→ Spatiotemporal dependency into machine learning models to predict robberies in Dallas

→ **Spatiotemporal lag variables** can effectively improve the prediction accuracy of machine learning models.

1. Introduction| Scientific Objectives, motivation

SO 1: Investigate *what spatial features can be used for crime prediction and how they can be incorporated into a ML modelling workflow.*
SO 2: Evaluate the predictive *performance of spatial-conscious machine learning models for crime.*

✓ ***Liu, Kounadi, & Zurita-Milla, 2022***

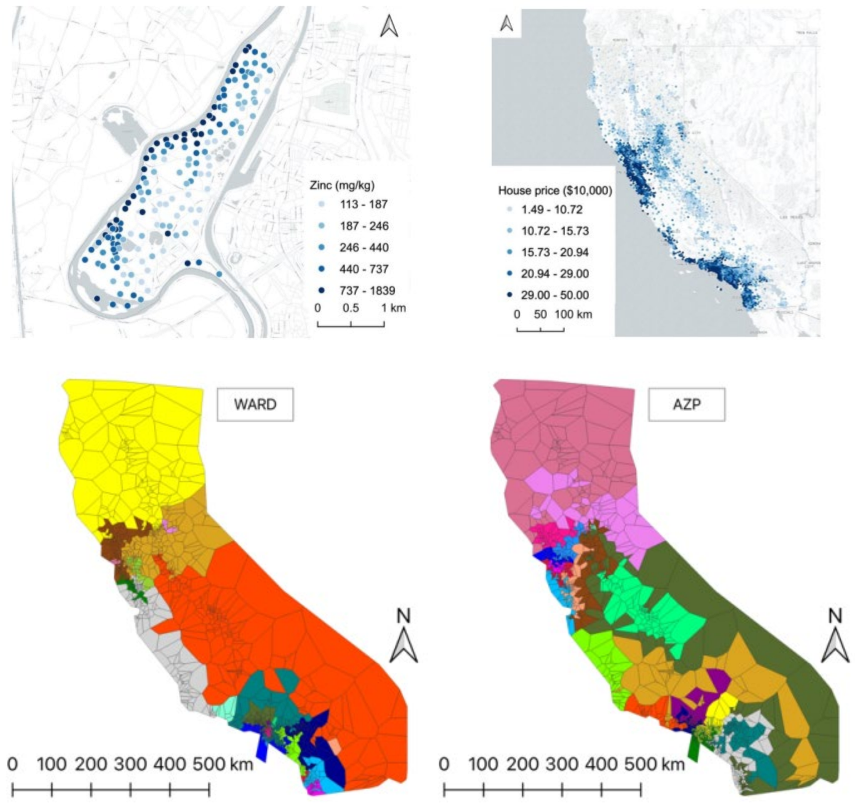
- ***Spatial Lag**: Lower errors and reduce the global spatial autocorrelation of the residuals*

✓ ***Boegl and Kounadi, 2024***

✓ ***Regionalization:** improves R-squared scores, less computational effort than GWR or GW-RF*

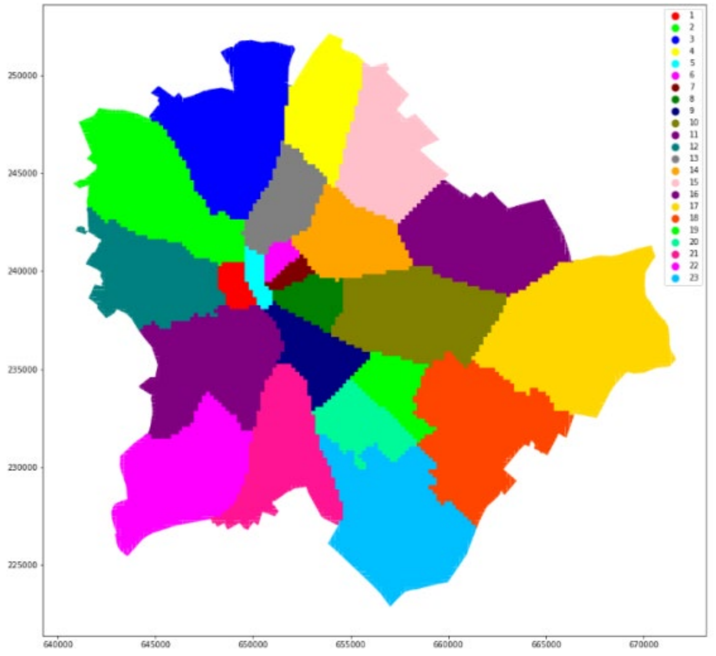
✓ ***Khalfa et al. 2025***

✓ ***Similar methodological approach**: Spatial and temporal unit of analysis, independent features, train-test selection, evaluation metric, fixed number of predicted hotspots = area coverage*

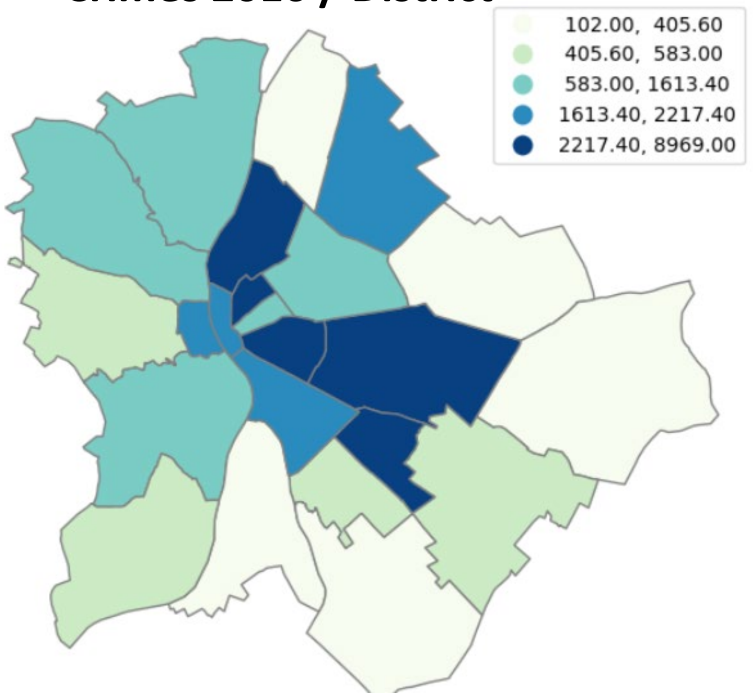


2. Data Description| Study area & crime events

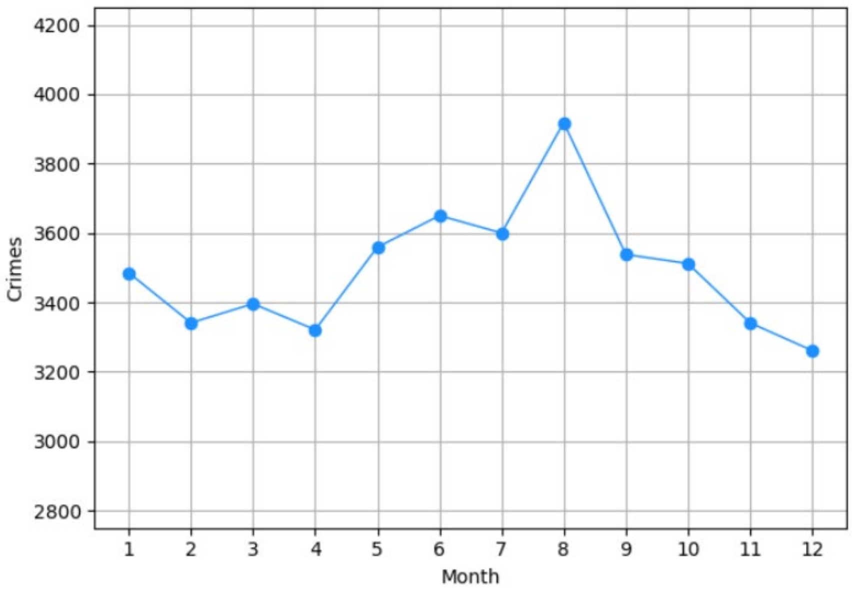
Districts/ Grid cells



Crimes 2016 / District



Crimes 2016 / Month



Spatial unit: 200 * 200 meters grid
Temporal unit: month
Precision: XY point & day

Training 2015 – Testing 2016
Crime data: 2013 until 2016 (≈ 190,000 events)
e.g., burglaries, robberies, car theft, pickpocketing.

2. Data Description| Independent Features (23)

Input Features A
Demographic & Socio-economic features
Total population
Percentage youth
Percentage of non-domestic inhabitants
Percentage of single households
Unemployment rate
Percent of houses occupied by homeowners
Dwelling stock
Environmental Features
Shops
Bars
Cafes
Restaurants
Snack bars/fast food
Green Space
Proximity Features
Train stations
Highways
Tram stops
Bus stops
Subway stops

Input Features A: freely available data from Open Street Map and the Hungarian Census Burau

Input Features B: engineered features from historical crime data

Input Features B
Number of crimes in the previous month
Months since last crime
Number of crimes in the last 12 months
Number of crimes in the same month last year
Number of crimes in the previous month in the neighborhood (district)

3. Modelling Framework| General workflow (1)

Preprocessing

(OSM & Census)

- Spatial Joins
- Joins
- Distance operations
- Topological operations
- Disaggregation

Modelling

(1yr rolling window)

- Random Forest (RF)
- S_lag RF
- Reg_RF
- LISA_RF

Analysis

- Hotspot Transformation
- Hit Rate
- Feature Importance
- Maps

3. Modelling Framework| General workflow (2)

- **Modelling Algorithm**: Random Forest
- **True Hotspots**: grid cells where one or more crimes occurred.
- **Predicted Hotspots**: grid cells with higher probability of a crime to occur.
- **Performance Evaluation:** Recall or Hit Rate (the percentage of the true hotspots that were correctly identified)
- **Interpretation of Predictive Performance**: -Top 5% percentile and a hit rate of 87%.

"***The predicted hotspots cover 5% of the study area and include 87% of the total area where a crime occurred*"*.**

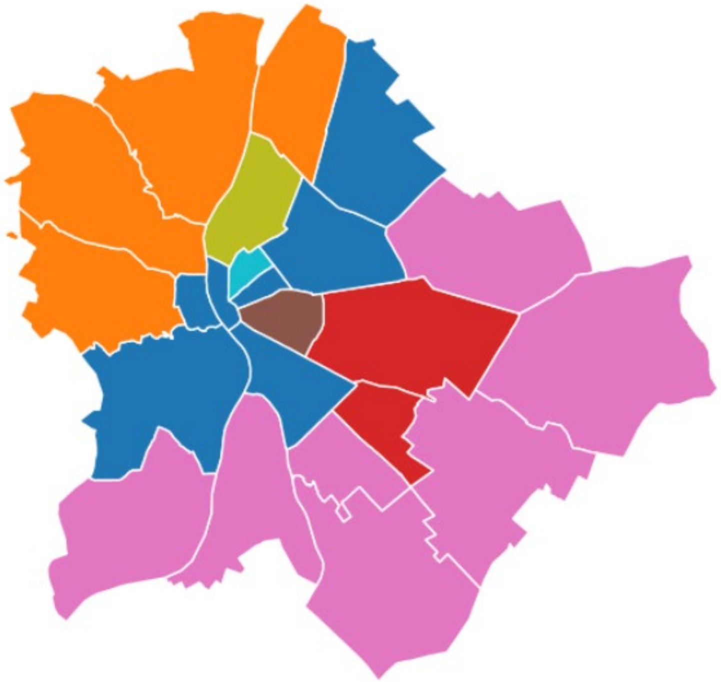
3. Modelling Framework| Spatial ML Models

S_Lag

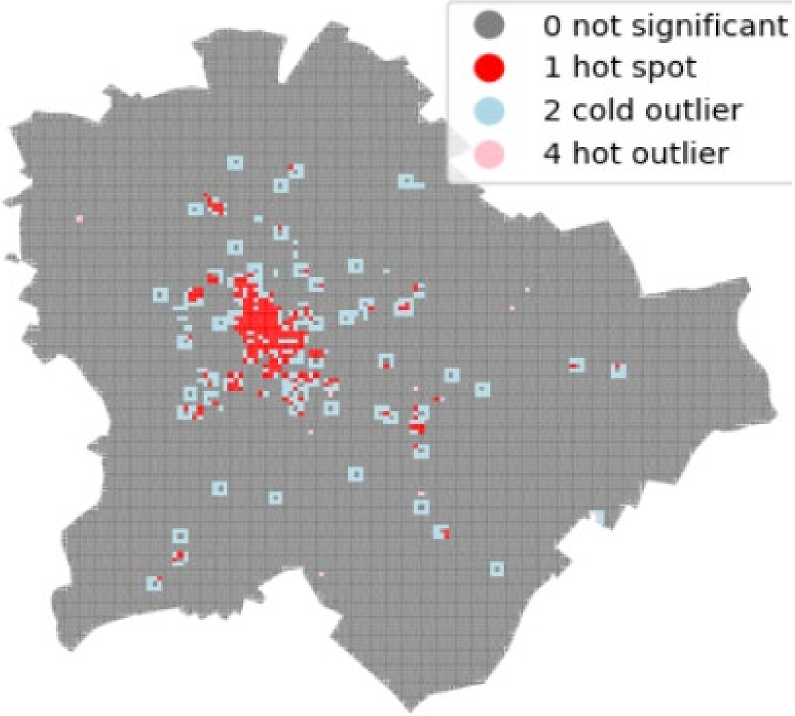
→ **Optimize Spatial Weight Matrix**
  → Moran's I maximize with 1st Order Queen

→ ***Engineer Spatial Lag features for:***
  a) count of crimes in the previous month in the cell and
  b) count of crimes in the last 12 months in the cell

Regionalization with Ward
(7 clusters)



LISA
(Local Moran's I, p-value < 0.05)



4. Results| Feature Importance (December 2016)

RF

1. **Previous month**
2. **Distance to bus**
3. **Previous 12M**
4. **Previous M district**
5. **% non HU**

S_Lag_RF

1. **Previous month**
2. **Distance to bus**
3. **Previous 12M**
4. **Previous 12M_lag**
5. **Same M last year**

....

10. Previous month lag

Reg_RF

1. **Previous month**
2. **Distance to bus**
3. **Previous 12M**
4. **Previous M district**
5. **Distance to train**

....

Middle → Ward clusters

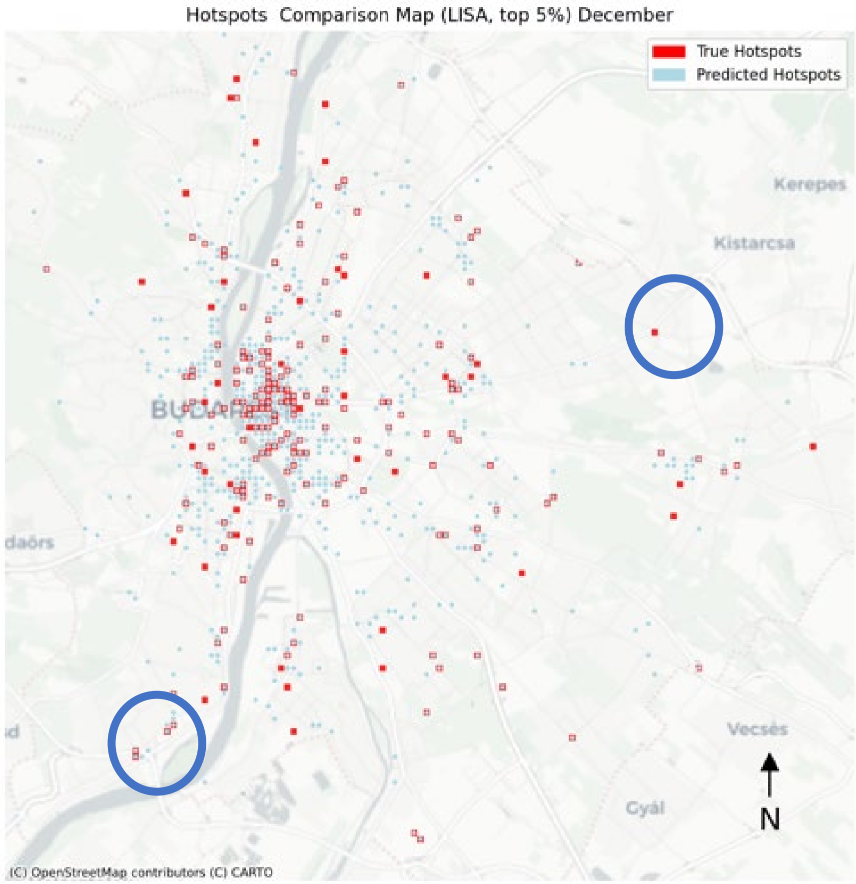
LISA_RF

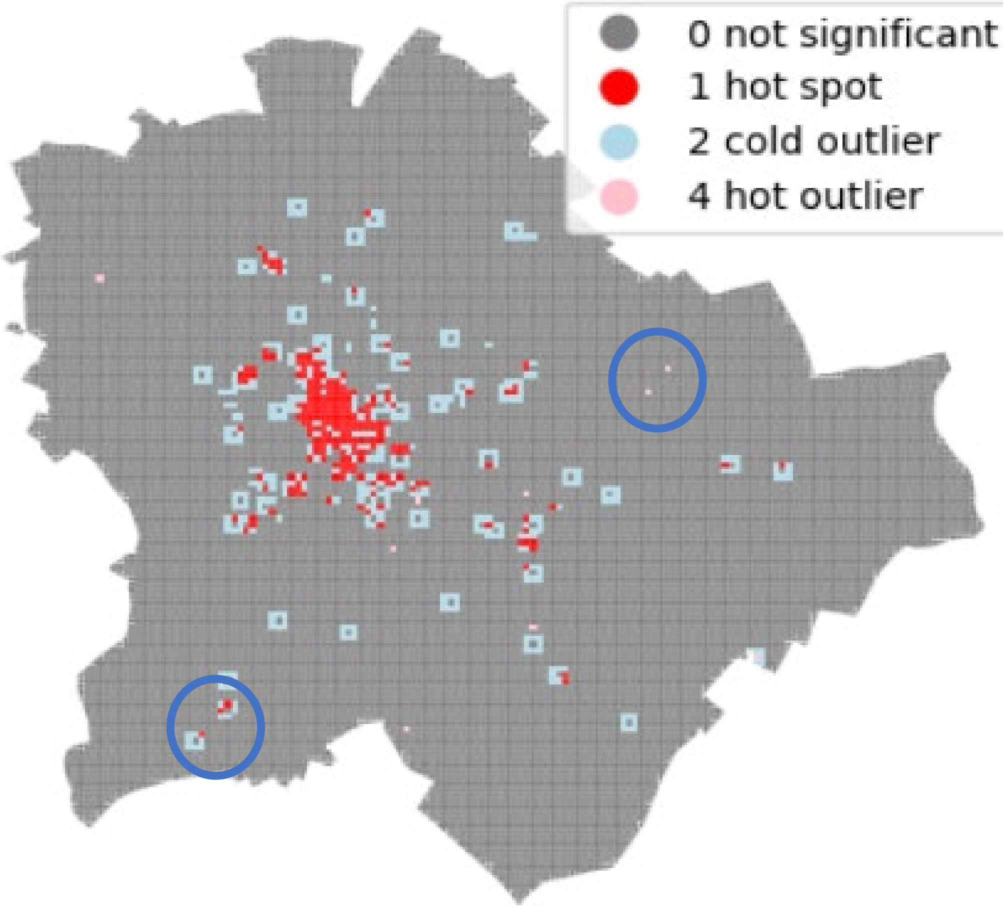
1. **Previous month**
2. **Distance to bus**
3. **Previous 12M**
4. **Previous M district**
5. **Distance to subway**

. . .

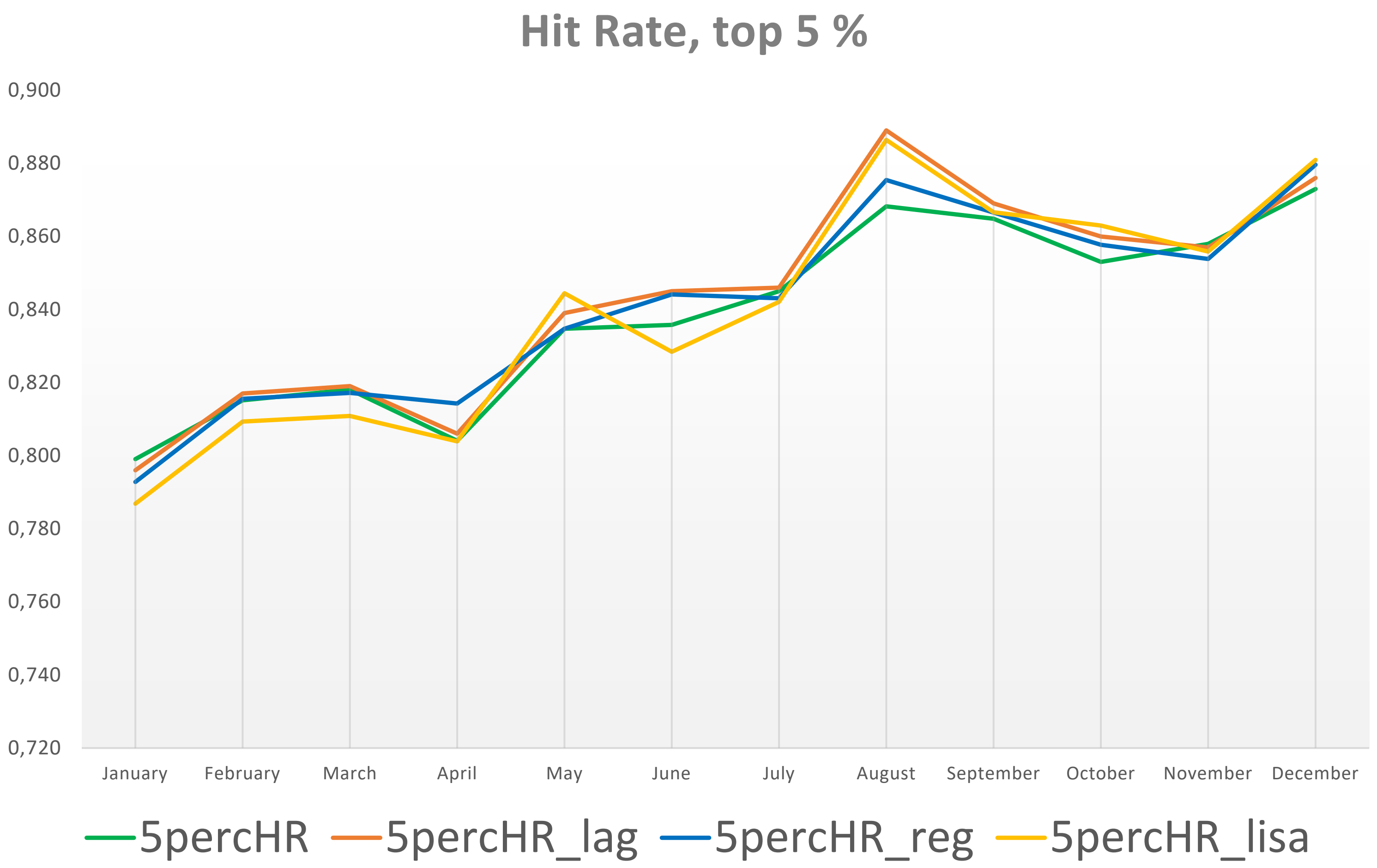
End → LISA features

4. Results| Predictive Maps





4. Results| Hit Rate comparison



AVG over 12 months

RF : 83,9 %

S_lag: 84,3 %

Reg_RF: 84,1 %

LISA_RF: 84%

5. Conclusions

Key Findings

- ✓ Spatially-conscious ML models further improve the predictive performance of traditional ML for forecasting crime.
- ✓ The most significant features are created from fine level spatiotemporal information of historic crime events
- ✓ Spatial lag related features are among the top important features; Spatial lag ML models perform better than Regionalization ML and LISA ML models.

Next Steps

- ✓ **Empirical testing in Budapest & Vienna**: period: 2019 -2023, with various crime types
- ✓ **Spatial features:** regionalization algorithms, number of clusters, LISA method, p-value, processing spatial groupings, additional spatial lag features.
- ✓ **ML algorithm:** tuning the hyperparameters, exploring additional supervised learning algorithms, testing with nested CV could potentially improve further the predictive performance and/or model's generalizability.
- ✓ **Units of analysis (space and time) & crime types**

Testing spatially conscious machine learning models to forecast crime.

A case study for the prediction of Acquisitive crime in Budapest.

25th Annual Conference of the European Society of Criminology

References

- ✓ *Khalfa, R., Snaphaan, T., Ristea, A., Kounadi, O., & Hardyns, W. (2025). Predicting Crime at Micro Places: Comparing Machine Learning Methods Across European Cities. In New Research in Crime Modeling and Mapping Using Geospatial Technologies (pp. 81-111). Cham: Springer Nature Switzerland.*
- ✓ *Boegl, L., & Kounadi, O. (2024). Introducing Spatial Heterogeneity via Regionalization Methods in Machine Learning Models for Geographical Prediction: A Spatially Conscious Paradigm. European Journal of Geography, 15(4), 244-255.*
- ✓ *Liu, X., Kounadi, O., & Zurita-Milla, R. (2022). Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features. ISPRS International Journal of Geo-Information, 11(4), 242.*
- ✓ *Deng, Y., He, R., & Liu, Y. (2023). Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models. Information Sciences, 646, 119414.*

ourania.kounadi@univie.ac.at

podor.andrea@amk.uni-obuda.hu

Thank you for your attention !