URBREATH [101139711]

Systemic Integration of Transformative Technical and Nature-based Solutions to Improve Climate Neutrality of European Cities and Regions and tackle Climate Change: the URBREATH Approach

D3.8 - AI Models for Socioeconomic, Community, Organisational and Citizen Well-being - V2

Project Reference No	URBREATH – 101139711
Deliverable	D3.8 AI models for Socioeconomic, Community, Organisation and Citizen Well-being – Version 2
Work package	WP3: URBREATH data strategy and tools
Type	OTHER
Dissemination Level	PU - Public (fully open)
Date	19/12/2025
Status	Final
Editors	Maria-Nefeli Kousta, Georgios Kopsiaftis, Nikolaos Bakalos, Anastasios Doulamis, Nikolaos Doulamis (ICCS)
Contributors	Task 3.5 participants
Reviewers	Stijn Vranckx (VITO), Emma Gaitán Fernández (FIC)

Document description	This document outlines the models being developed to relate the application of nature-based solutions to socioeconomic and citizen well-being indicators. The collected data will encompass local urban planning, land use, natural environment information, and socioeconomic parameters. This deliverable corresponds to T3.5, with updates scheduled for M36 (December 2026).

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
0.1	20.10.2025	Table of contents	ICCS
0.2	10.11.2025	First draft with model definitions and initial I/O analysis	ICCS
1.0	10.12.2025	Draft ready for internal review	ICCS, LAT40, DEDA
1.1	16.12.2025	Review - Comments and suggestion	VITO, FIC
2.0	17.12.2025	Final version ready for quality check by the coordinator	ICCS
2.0	18.12.2025	Final quality check by the coordinator	LC
2.1	19.12.2025	Minor review completed	ICCS
Final	19.12.2025	Version ready for submission to the portal	LC

Disclaimer

The URBREATH project is co-funded by the European Union under grant agreement ID 101139711. The information and views set out in this document are those of the URBREATH Consortium only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

This deliverable (D3.8 – AI Models for Socioeconomic, Community, Organisational and Citizen Well-being – Version 2) presents the second development cycle of the data-driven and AI-enabled analytical tools developed within WP3. Building on the methodological foundation established in Version 1, this updated deliverable reports substantial progress in refining tools, implementing models, integrating datasets, and preparing for operational use across the URBREATH platform.

The tools developed in this phase address several key aspects of urban well-being, including access to essential services, public transport reachability, crime patterns, property and rental market trends, and operational planning challenges. Each tool has been refined through improved data preprocessing workflows, enhanced statistical and machine learning components, and closer alignment with municipal datasets and requirements.

Collaboration with pilot cities has played a vital role in advancing this work. The dialogue with pilots has facilitated the provision of complete historical crime data and informed the classification of urban sectors, validation of visualisation needs, and alignment of analytical outputs with user expectations. These exchanges ensure that the tools are not only technically robust but also responsive to real urban planning challenges.

Looking ahead to Version 3 and the final deliverable, the modelling framework will be further reinforced by incorporating additional explanatory variables—such as land-use characteristics, demographic indicators, and mobility patterns—to improve interpretability and predictive accuracy. Several tools are now reaching a stage of maturity where they can be integrated into the main URBREATH platform, alongside the development of visual interfaces to support scenario exploration and city-level decision-making.

Overall, this deliverable shows significant progress in developing AI-based analytical tools that support URBREATH's broader goal of promoting evidence-based urban planning and enhancing understanding of socio-economic and community well-being. The work completed in this phase creates a strong technical foundation for the next cycle of development.

Table of Contents

1 INTRODUCTION 8
1.1 PURPOSE AND SCOPE 8
1.2 APPROACH FOR WORK PACKAGE AND RELATION TO OTHER WORK PACKAGES AND DELIVERABLES 9
1.3 METHODOLOGY AND STRUCTURE OF THE DELIVERABLE 10
2 TOOLS, MODELLING FRAMEWORK AND IMPLEMENTATION 12
2.1 REVISIONS TO OBJECTIVES AND SCOPE SINCE VERSION 1 12
2.2 DETAILED TOOLS AND MODEL DESCRIPTION 13
2.2.1 15-Minute Index Tool 13
2.2.2 Public Transport Accessibility Tool 17
2.2.3 Optimal Snow Deposit Analysis Tool 19
2.2.4 Crime Statistics Tool 28
2.2.5 Rent Price Prediction 38
3 TOOL INTEGRATION WITHIN THE URBREATH PLATFORM 41
3.1 ANALYSIS OF INTEGRATION TECHNOLOGIES 41
3.2 CURRENT INTEGRATION STATUS 43
4 CONCLUSIONS 44
5 REFERENCES 45

List of Figures

Figure 1: Comparison of solver outputs for input snow volume V = 1,000,000 m³, number of deposits k = 5. ... 25
Figure 2: Comparison of solver outputs for input snow volume V = 1,000,000 m³, number of deposits k = 25 .. 26
Figure 3: Map presenting the sectors of Leuven municipality (source: Leuven_Sectors). 30
Figure 4: Example output of the residual anomalies method applied to a representative urban crime time series. 33
Figure 5: Example output of the Isolation Forest method applied to a representative urban crime time series. 34
Figure 6: Typical architecture of an LSTM model. 36

Figure 7: Indicative preliminary results of the loss function during model training for vandalism incidents in Section A00 of central Leuven. 37

Figure 8: High-level docker container architecture and communication with Kafka.. 41

Figure 9: Initial implementation of 15-minute index analysis in the URBREATH platform (a)............................. 42

Figure 10: Initial implementation of 15-minute index analysis in the URBREATH platform (b). 43

List of Tables

Table 1: Technical specifications of the 15-minute index tool. 16

List of Terms and Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
CBC	COIN-OR Branch and Cut
CSV	Comma-Separated Values
CP-SAT	Constraint Programming Satisfiability
CUSUM	Cumulative Sum
GPKG	GeoPackage
IDE	Integrated Development Environment
KPI	Key Performance Indicator
KR	Key Result
LSTM	Long Short-Term Memory
LP	Linear Programming
MILP	Mixed Integer Linear Programming
MSE	Mean Squared Error
NbS	Nature-based Solution
OR	Operations Research
ORS	Open Route Services
OSM	Open Street Map
PoI	Point of Interest
QGIS	Quantum Geographic Information System
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
UI	User Interface
UNODC	United Nations Office on Drugs and Crime
WP	Work Package

1 Introduction

This deliverable is the second version of the “AI Models for Socioeconomic, Community, Organisational and Citizen Well-being,' building on the initial D3.7 results. Building on the methodological and conceptual foundations established in the first version, this document reports the progress made in developing, refining, and implementing data-driven and AI-based analytical tools within WP3. The main aim of these tools remains to assist urban planners, local authorities, and project stakeholders in understanding key urban dynamics through quantitative modelling and evidence-based analysis.

While the first version concentrated on establishing the theoretical framework, identifying relevant datasets, and outlining initial modelling concepts, Version 2 shifts towards practical implementation and early validation. Several tools have advanced from conceptual design to functional prototypes, aided by improved data access and stronger collaboration with pilot cities. Notably, new datasets provided by Leuven have facilitated the development of comprehensive temporal models for crime analysis, including time-series decomposition, anomaly detection methods, and an LSTM-based forecasting architecture. Likewise, improvements in geospatial processing, accessibility modelling, and optimisation techniques have broadened the analytical capabilities of the service-accessibility and operational-planning tools.

A key focus in this version is the integration of AI and statistical methods to enable more robust interpretation, prediction, and decision support. The tools now include advanced machine learning approaches alongside traditional statistical techniques, allowing for a deeper understanding of temporal patterns, spatial variability, and behavioural dynamics across various urban indicators. The report also highlights progress in data preprocessing workflows, tool interoperability, and the early development of visualization components that will form part of the final, integrated URBREATH platform. Therefore, Version 2 marks a shift from foundational design towards operational maturity, laying the foundation for the next stage of development. The upcoming version will concentrate on broader feature integration, further methodological refinement, and enhanced interpretability of AI-driven outputs, ultimately supporting the project’s goal of providing actionable analytical tools for urban well-being assessment.

1.1 Purpose and Scope

The purpose of this deliverable is to document the second-stage development of the AI-based analytical tools created under WP3 to evaluate socioeconomic, community, organisational, and citizen well-being. While Version 1 established the conceptual and methodological foundations, Version 2 reports tangible progress in model implementation, data integration, and initial validation activities.

This deliverable broadens the scope to include updated tool prototypes, enriched datasets from partner cities, and advances in statistical and machine learning techniques. These developments improve the analytical capabilities of the modelling framework and strengthen its alignment with municipal needs. More specifically, Version 2:

- Presents implemented and updated AI models, including time-series analysis, anomaly detection, and LSTM-based forecasting.
- Describes improved data workflows, covering preprocessing, harmonisation, and spatial classification.
- Reports early results from crime prediction, accessibility modelling, and operational optimisation tools.
- Outlines methodological refinements and technical steps required for the next development cycle.

The scope of this deliverable is therefore to provide:

- A concise overview of technical progress since Version 1.
- A clear definition of ongoing development needs that will guide the transition toward Version 3 and full tool integration into the URBREATH platform.

This document does not present final or fully validated models, but rather an intermediate milestone that consolidates progress and establishes the direction for subsequent development.

1.2 Approach for Work Package and Relation to Other Work Packages and Deliverables

WP3 employs a data-driven and AI-focused approach to analyse key socioeconomic and community well-being indicators across the URBREATH pilot cities. The overall methodology combines statistical modelling, machine learning techniques, geospatial analysis, and optimisation frameworks to produce quantitative insights that support urban planning and decision-making processes. The development of these analytical tools follows an iterative process, incorporating conceptual design, data gathering, preprocessing, model implementation, and initial validation.

WP3 operates at the intersection of several other work packages and relies on continuous exchanges to ensure coherence and operational relevance. Its approach is structured around the following principles:

- **Dependency on WP2 (Data, Standards, and Integration):** WP3 builds on the data infrastructures and harmonized datasets prepared under WP2. This includes the collection, structuring, and standardization of spatial, temporal, and socioeconomic data used in the modelling tools. Close alignment with WP2 ensures that the analytical methods rely on consistent and high-quality inputs.
- **• Support for WP4 (Digital Platform and Interfaces):** The outputs of WP3—analytical models, indicators, forecasts, and anomaly detection results—form core components of the URBREATH digital platform. WP3 provides the computational engines that WP4 will integrate into user-facing interfaces, dashboards, and decision-support tools.
- **• Feedback loops with Pilot Activities (WP5–WP7):** Collaboration with municipal partners and pilot cities enables the refinement of models and validation of assumptions. Local knowledge informs the selection of relevant indicators, sector classifications, and priority use cases. The crime prediction and accessibility tools developed in this deliverable are directly shaped by such interactions.
- **• Contribution to Project-Level Assessments (WP1 & WP8):** WP3 supplies quantitative evidence and analytical outputs that support project coordination (WP1) and broader impact assessment and evaluation activities (WP8), especially regarding urban well-being indicators and data-driven planning approaches.

The tools in this deliverable complement, modify, and extend those introduced earlier, especially D3.7, which established the conceptual modelling framework. Version 2 advances these tools towards implementation and integration. The AI models and analytical tools presented in D3.8 are designed to operate as modular components that can be integrated into higher-level interpretability and decision-support environments, such as the VIE-AI platform described in D3.10-V1. D3.8 acts as an intermediate milestone before full deployment in future deliverables (including D3.10 – V2). Through this structured workflow, WP3 ensures the development of robust analytical components that are compatible with the project's digital ecosystem and aligned with partner cities' needs.

1.3 Methodology and Structure of the Deliverable

The methodology underlying this deliverable follows an iterative and evidence-based approach aligned with the objectives of WP3. Development of the analytical tools proceeded through four main phases:

1. Data acquisition and preprocessing, including the integration of updated datasets from pilot cities and the harmonisation of spatial, temporal, and socioeconomic information.
2. Model development and refinement, employing statistical methods, machine learning techniques, and optimisation frameworks tailored to each tool.

3. Intermediate testing and validation, where prototype implementations were evaluated using available data to assess feasibility, identify limitations, and guide methodological adjustments.
4. Documentation and alignment with project requirements, ensuring coherence with related deliverables and preparing the tools for subsequent integration into the URBREATH digital platform.

This deliverable summarises the progress achieved in these phases and describes the current development status of each analytical component. It also outlines the next steps required to advance the models towards full operational maturity in the upcoming project period. The structure of the document is as follows:

- **Section 1** introduces the purpose, scope, and methodological approach of the deliverable, and positions WP3’s work in relation to other work packages.
- **Section 2** provides updates on each analytical tool, detailing their objectives, data sources, preprocessing techniques, modelling approaches, technical implementation, and initial results.
- **Section 3** outlines the steps required to integrate the analytical tools into the URBREATH digital platform. It specifies the technical components that need adaptation for platform deployment, including data exchange mechanisms, interoperability requirements, visualisation needs, and API design considerations.
- **Section 4** concludes with reflections on progress so far and presents the roadmap for further development and integration into the URBREATH system.

This structure ensures transparency in the methodological process and gives a clear overview of the technical advancements achieved in Version 2, while laying the groundwork for future deliverables in the next phases of the project.

2 Tools, Modelling Framework and Implementation

2.1 Revisions to Objectives and Scope Since Version 1

Compared to the objectives and tool specifications in Version 1 of this deliverable, several adjustments to the scope have been necessary due to data availability constraints, pilot-specific limitations, and evolving project needs. These deviations do not affect the overall direction of WP3 but refine the focus of tool development to ensure feasibility, methodological soundness, and alignment with the resources provided by pilot cities.

Economic Activity Model – Not Implemented Due to Lack of Data

The economic activity model, initially introduced in Version 1 as part of the analytical suite, cannot currently be implemented. The necessary datasets—covering economic transactions, business activity records, or detailed labour-market indicators at suitable spatial and temporal resolutions—were not available from pilot cities or national sources. Without these data, the development of the model would not meet the required standards of quality or reliability for integration into the URBREATH analytical framework. The model may be reconsidered in later stages of the project if sufficient data becomes available.

Crime Analysis Tool – Implemented Only for the City of Leuven

Version 1 proposed a crime analysis and prediction tool intended for use across multiple pilot cities. However, only the City of Leuven was able to supply complete and consistent multi-year crime datasets (2015–2024). Consequently, the advanced statistical and AI modelling elements, such as time-series analysis, anomaly detection, and LSTM-based forecasting, are implemented solely for Leuven. The tool remains fully adaptable and can be extended to other cities once comparable datasets are available.

Selective Use of AI Methods Across Tools

- While Version 1 outlined the goal of applying AI techniques across all analytical tools, practical implementation has been adapted to the nature and data characteristics of each specific application.
- AI and deep learning methods are used only in areas where they deliver meaningful predictive or analytical value (e.g., crime forecasting).
- For tools where AI is not methodologically appropriate or where datasets are limited, alternative modelling approaches, such as optimization frameworks, geospatial methods, or classical

statistical analysis, have been selected to ensure the scientific robustness and usefulness of results.

All modelling components remain adaptable and may be updated to include AI once additional datasets are provided or new pilot cities show interest.

In summary, these adjustments reflect a data-driven and feasibility-focused refinement of the original objectives. They ensure that the tools developed under WP3 stay methodologically sound, aligned with pilot city capabilities, and ready for further expansion as new data and requirements arise.

2.2 Detailed Tools and Model Description

2.2.1 15-Minute Index Tool

2.2.1.1 Tool Overview and Objectives

The 15-minute index is a tool designed to quantify and visualize urban accessibility. Its main objective is to assess how easily residents can reach essential daily services, such as shops, healthcare, education, food, parks, entertainment, and financial services, within a 15-minute walk or bike ride. To achieve this, the tool applies the 15-Minute City concept by converting urban features into measurable indicators. It calculates walking times from residential areas to Points of Interest (PoIs) and aggregates results on a hexagonal grid. The purpose is to support planners in creating cities that are more sustainable, equitable, and human-centered.

2.2.1.2 Data Sources, Retrieval, and Preprocessing

The tool utilizes two alternative data sources: **OpenStreetMap (OSM)** or **locally supplied geospatial datasets**. When OSM is chosen, two primary datasets are extracted and harmonized:

- **Street network**
  - Only pedestrian-accessible streets are included.
- **Points of Interest (PoIs)**
  - PoIs are organised into eight service categories relevant to socio-economic and well-being analysis: (i) *markets & groceries*, (ii) *restaurants & cafés*, (iii) *education*, (iv) *health*, (v) *banks & post offices*, (vi) *parks*, (vii) *entertainment*, and (viii) *shops*.
  - Extraction is based on a predefined collection of OSM tag sets to ensure cross-city consistency and reproducibility.

The overall data processing pipeline consists of the following steps:

Hexagonal Grid Generation

To enable fine-grained spatial modelling, the tool divides the city into a 250-meter hexagonal grid.

Parks category

All service categories are represented as point features in OpenStreetMap (OSM), except for parks, which are typically mapped as polygons. To ensure accessibility analysis for this category, the tool uses park access points (gates). These gates are derived either directly from OSM or generated algorithmically when missing.

Park Access Point Classification

Each park is assigned access points classified into three types:

- **Type A – Existing Gates (OSM derived)**

Points located within **10 m** of the park boundary, identified using the following OSM tags:

  - `barrier=gate`
  - `barrier=entrance`
  - `entrance=yes`
  - `leisure=park`
  - `leisure=dog_park`

- **Type B – Street-Park Intersections**

If no Type A gates are found, the tool identifies **intersections between the park perimeter and the street network**, using any OSM street tagged as:

  - `highway=*` (all road types)

- **Type C – Virtual Access Points (Generated)**

If no Type A or B points are available, the tool generates **virtual gates every 100 m** along the park perimeter.

2.2.1.3 Modelling Approach and Technical Implementation

The tool requires the following **input parameters** to initialise and run the accessibility analysis:

- **Bounding box:** it defines the area of interest of the analysis, expressed as `[lat_min, lon_min, lat_max, lon_max]`.
- **Category:** service category for which accessibility is assessed (one of the eight predefined categories, or “all” for a combined score).

- **Mode of transportation:** travel mode used in the analysis (foot or bike; default = foot).
- **Measurement criteria:** metric used for accessibility computation, such as time or distance, where distance (and derived time) is measured along the network.
- **Boundary polygon:** optional polygon used to clip or limit the analysis to specific administrative or planning units (e.g., municipal borders, district boundaries).
- **Output folder:** directory where all resulting files will be stored.

The processing pipeline implemented in the tool is structured as follows:

1. Download the street network and relevant PoIs based on the selected service categories or use local PoIs.
2. Generate a hexagonal grid that covers all nodes in the street network.
3. Compute walking times from each street node to the nearest PoI in each category included in the analysis.
4. Aggregate node-level walking times to the hexagon level.
5. Calculate summary metrics for each hexagon:
    a. Average walking time across all categories.
    b. Maximum walking time.

Walking times are calculated assuming a speed of 5 km/h**,** which is slightly above the average adult walking speed, in an attempt to cover most use cases in this first version. As noted in the planned next steps, this parameter will be configurable in future versions to better accommodate different user needs.

The output consists of a **hexagonal vector layer** that includes, for each hexagon, the walking time in minutes for each service category, the average walking time across all categories, and the maximum walking time. The outputs are provided in both **CSV** and **GPKG** formats (EPSG:3857) and are clipped to the specified boundary polygon.

The algorithm is implemented in Python, and all geospatial operations are performed using QGIS in headless mode. This allows processing without the need for a graphical user interface. The tool is fully configured through a .ini parameter file, where all inputs are defined.

The script can be executed from the command line or any Python-compatible integrated development environment (IDE). A standard command-line execution looks like this:

```
.../python3 overallExecutor.py parameters.ini > log.txt 2>&1 &
```

This command runs the process in the background and logs all messages and errors in log.txt. Table 1 contains the technical specifications of the 15-minute index tool, including the programming environment and dependencies.

Table 1: Technical specifications of the 15-minute index tool.

Programming language	Python
Libraries used	Pandana, geopandas, numpy, pandas, osmnet, rtree, pyproj, shapely, geovoronoi, fiona=1.9.5, rasterio, gdal,scipy, beautifulsoup4, from qgis.core import *
Tools needed to implement and run the algorithm	Python, IDE, command line

2.2.1.4 Current Development Status and Next Steps

The tool is currently in an advanced prototype stage and integrates a complete workflow for assessing pedestrian accessibility to key services. It automates data extraction, preprocessing, travel-time computation, spatial aggregation, and indicator generation, producing analysis-ready outputs suitable for mapping, reporting, and further modelling.

The tool currently supports the following capabilities:

- **Extraction and preprocessing** of OSM street networks and PoI datasets, as well as optional local PoI data.
- **Computation of walking times** to the nearest PoI for each service category.
- **Aggregation of results** to a regular hexagonal grid covering the area of interest.
- **Generation of accessibility indicators**, including average and maximum travel-time metrics.
- **Production of outputs**, delivered as:
    - **CSV files** containing walking times and averages, clipped to the optional boundary polygon when applicable.
    - **GPKG layers** representing the hexagonal grid with all the times and averages, also clipped to the provided polygon when applicable.

Building on the current functionality, several enhancements are planned to improve the tool's flexibility, usability, and integration within the broader URBREATH ecosystem. These improvements aim to better support various pilot contexts, enable scenario-based analysis, and offer more detailed accessibility assessments tailored to different user groups.

Planned future developments include:

1. **Refinement of service categories** – Collaborate with pilot cities to identify whether additional or custom categories are needed. Currently, the tool supports predefined categories, with potential future parametrization of categories based on user requirements.
2. **Platform integration** – Incorporate the tool into the URBREATH platform to enable scenario simulations, such as adding new or planned PoIs. For efficiency, recalculation of the 15-minute index could be limited to a **user-defined area** around the new PoIs, using a buffer consistent with walking speed. This could be offered as an optional feature, allowing the user to draw a bounding box before triggering the calculation.
3. **Parametrization of travel speed** – Allow users to adjust walking speed to better reflect the mobility of different populations (e.g., elderly or children), providing more realistic accessibility assessments.

2.2.2 Public Transport Accessibility Tool

2.2.2.1 Tool Overview and Objectives

Building on the foundation of the 15-minute index tool, we are conceptualizing a complementary simulation tool that evaluates how easily residents can access public transportation in urban areas. While the 15-minute index tool measures proximity to key amenities by walking or biking, the proposed Public Transport Accessibility Tool specifically focuses on pedestrian access to transit infrastructure. The goal is to assess how effectively people can reach public transit stops from road-intersection nodes, offering valuable insights into walkability, spatial equity, and the overall functionality of urban mobility networks. The metrics generated will help planners, municipalities, and transit agencies evaluate and enhance the coverage, quality, and inclusivity of public transportation systems.

2.2.2.2 Data Sources, Retrieval, and Preprocessing

The main data source for public transport infrastructure will be OpenStreetMap (OSM). The two datasets to be extracted are:

1. **City street network** and
2. **Public transport nodes** categorized into different types of transport: metro stations, tram stops, bus stops, etc.

Public transport nodes can be obtained using OSM's `public_transport` and mode-specific tagging scheme. In the initial version, we plan to extract metro/subway, tram, bus, and regional or urban rail, based on the pilot city's infrastructure and needs. Preprocessing will include:

- filtering nodes within the administrative boundaries of the respective study area,
- validating coordinate geometry,
- harmonizing transport mode attributes,
- ensuring consistency between stop positions and platforms,
- and integrating the final dataset with the road network graph for subsequent accessibility modelling.

As the tool evolves, additional OSM features, such as route relations (`type=route`, `route=bus/tram/train`) or network-level structures, may be incorporated if required data is available and robust. The city will be divided into a 250-meter hexagonal grid, generated **only where OSM street nodes exist**.

2.2.2.3 Modelling Approach and Technical Implementation

Mapping Transport Modes and Infrastructure

We collect spatial data on key public transport modes, including metro, tram, bus, and regional rail stations, within the city boundary. Each mode is assigned an importance weight reflecting its contribution to efficient and sustainable mobility (e.g., metro > tram > bus). This infrastructure data is integrated with the urban street network, ensuring that accessibility assessments use realistic walking paths rather than straight-line distances.

Scoring Pedestrian Accessibility from Road Nodes

Each road intersection node in the network is evaluated for walking accessibility to nearby public transport stops using OpenRouteService. Walking times to the closest stops are calculated for each node and combined with mode-specific weights. This yields a public transport accessibility score per node, with higher scores indicating better walkability to high-quality transit services.

Aggregating and Visualising Accessibility

Accessibility scores are aggregated into hexagonal grids at various resolutions (e.g., 100 m to 500 m). The score for each hexagon reflects the average accessibility of the nodes it encompasses. These aggregated scores are visualised via colour-coded maps, highlighting well-served areas and revealing gaps in network coverage.

2.2.2.4 *Current Development Status and Next Steps*

The existing functionality exists as an extension of the 15-minute index tool, incorporating additional categories for various public transport modes. A fully independent implementation has not yet been developed. An important next step is to explore whether the objectives of the Public Transport Accessibility Tool can be effectively met by expanding the scope of the current 15-minute index tool. There is a credible argument for integrating both accessibility analyses, amenities, and public transport within a single, unified framework. However, the methodological complexity involved in transport-specific analysis may warrant the development of a dedicated, standalone tool. If a separate implementation is pursued, the initial version is expected to focus on pedestrian access to transit nodes. Future iterations could then expand to include the full public transport network topology, encompassing routes, service frequencies, and transfer points. This would facilitate more comprehensive assessments of urban connectivity, travel times, accessible destinations, and inter-neighbourhood linkages. Further extensions will depend on the availability, granularity, and reliability of the required data sources.

2.2.3 Optimal Snow Deposit Analysis Tool

2.2.3.1 *Tool Overview and Objectives*

The tool aims to identify the optimal snow deposit locations by minimizing the distance traveled from a street to the nearest deposit point. In its initial version of the tool, the user provides as inputs the total amount of snow to be deposited (in cubic meters) and the number of deposit sites to be used. Based on these inputs, the tool provides the coordinates of the selected deposit sites along with the street-to-deposit assignments. The goal is to support both real-time emergency decision-making, where speed and responsiveness are crucial, and long-term planning, which requires stable, high-quality solutions for infrastructure evaluation and resource allocation.

2.2.3.2 *Data Sources, Retrieval, and Preprocessing*

The spatial analysis integrates datasets from two primary repositories: municipal land-parcel data from the City of Tallinn and road network geometries from OpenStreetMap (OSM). This dual-source approach was adopted to establish a robust foundation for identifying optimal snow-deposit locations.

2.2.3.2.1 *Municipal Parcel Data*

A parcel-level land-use dataset was acquired from the City of Tallinn, comprising attributes such as parcel type, total area, ownership tenure, administrative metadata, and specific land-use codes. The data retrieval process prioritized three land-use categories deemed relevant for potential snow storage:

- 016: Public facilities / public land
- 017: Parks and green areas
- 007: Transport and parking areas

2.2.3.2.2 Road Network Data

The road network topology was derived exclusively from OpenStreetMap (OSM). Road geometries were extracted and harmonized to produce a consistent dataset. The processing workflow resulted in a vector layer containing:

- Spatial Geometry: The vector form of the road network layout.
- Segment Length: The geodesic length calculated for each road segment.
- Road Width: The average width estimated from OSM data or assumed where metadata was missing.
- Ploughable Surface Area: Computed surface area indicating the snow-generation capacity of the street network.

This standardized dataset serves as the baseline for estimating snow accumulation volumes and enables the spatial connection between snow-generation sources (road segments) and potential storage sites.

2.2.3.2.3 Semantic Filtering and Classification

To enhance classification precision and refine the broad granularity of official land-use codes, a heuristic keyword-matching algorithm was implemented.

Inclusion Criteria (Contextual Identification)

To retain parcels suitable for snow deposit that might otherwise be obscured by generic administrative coding, an inclusion filter was applied. Parcels were designated as candidate sites if their nomenclature contained terms indicative of:

- Parks/Green Areas: park, haljasala, roheala, mänguväljak
- Parking Areas: parkla, parkimisala, parking, car park

Exclusion Criteria (Institutional Filtering)

A comprehensive exclusion protocol was applied to the parcel name, notes, owner, and administrator fields to eliminate false positives. This step ensured that institutional grounds, such as educational, medical, and cultural facilities, were not erroneously classified as snow-storage sites. Parcels containing the following keywords (or their English equivalents) were systematically removed:

- kool (school), lasteaed (kindergarten), ülikool (university)
- haigla (hospital), muuseum (museum), raamatukogu (library)
- spordikeskus (sports center), staadion (stadium)
- teater (theater), kirik (church), kalmistu (cemetery)

Refinement for Code 016: For parcels categorized under land-use code 016 (Public facilities), which frequently encompass non-park institutional properties, a strict retention rule was enforced: only parcels explicitly identified as parks or green areas via the inclusion filter were retained, provided they remained free of exclusion keywords.

2.2.3.2.4 Site Categorization and Capacity Stratification

Following the filtering process, candidate sites were then stratified into capacity tiers based on total surface area to facilitate storage potential estimation:

- Tier A (High Capacity: Large-scale regional storage sites.): more than 50,000 m²
- Tier B (Medium Capacity: Intermediate storage sites): between 10,000 and 50,000 m²
- Tier C (Low Capacity: Local, distributed storage areas.): between 1,000 and 10,000 m²

Snow-storage capacity estimates were subsequently modeled for each parcel according to its assigned tier.

2.2.3.2.5 Volumetric Capacity Modeling

To translate spatial extent into operational storage potential, a volumetric model was applied to the classified parcels. Storage capacity was estimated using tier-specific parameters defining the average pile height (h) and a usable surface area factor (f). The usable factor accounts for operational constraints, such as maneuvering space, setbacks, and slope stability, which reduce the effective area available for snow piling. Total snow-storage capacity was derived by multiplying the total parcel area by the assigned usable factor and average pile height.

2.2.3.3 Modelling Approach and Technical Implementation

This section outlines the methodological approach used to identify optimal snow deposit locations within the city of Tallinn, based on the total snow volume to be managed and the available candidate sites. The method is designed to support both the initial version of the tool (which relies on simple snow distribution assumptions and static siting) and future enhancements that may include dynamic inputs such as weather forecasts and routing considerations.

Regarding snow demand estimation, the current version of the tool assumes that the total snow volume provided by the user is distributed across streets based on their area. To do this, each street segment is allocated a portion of the total snow volume proportional to its physical size, calculated from the available length and width of each road segment. This way, each street segment is assigned an associated snow volume.

To compute the distances between streets and snow deposit sites, we first determine the centroid for each street segment and each snow deposit location. We then create a road network graph and identify the nearest nodes to the centroids of roads and deposits. These nodes serve as the representatives for distance calculations. To assess the operational suitability of each candidate deposit, the tool calculates the shortest-path travel distance along the actual street network from each street segment to each deposit location, ensuring that distances reflect realistic travel routes rather than straight-line (Euclidean) distances.

The location-selection task is formulated as a Capacitated Facility Location / p-Median problem, a well-established model in operations research for selecting a subset of facility locations under capacity constraints (Daskin, 1997; Laporte et al., 2019; Mirchandani and Francis, 1990). It is part of a broader category of facility location problems that aim to choose from a finite set of facilities to optimally serve demand sites. Each potential facility-demand pair is associated with an assignment cost, often defined as the distance between the respective locations. Additionally, fixed opening costs and capacity constraints (known as capacitated facility location) can be assumed for each facility, and there can also be a limit on the total number of facilities allowed. When the number of facilities is fixed to p, the problem is referred to as a p-median problem.

The capacitated p-Median problem is formulated as follows:

$$\min_{\mathrm{x,y}} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} d_j c_{ij} x_{ij}$$

$$s.t. \sum_{i\in\mathcal{I}} x_{ij} = 1 \; \forall j \in \mathcal{J} (= \text{ each demand point is assigned to one facility}),$$

$$x_{ij} \leq y_i \; \forall i \in \mathcal{I}, y \in \mathcal{J} \; (= \text{assigment allowed only to open facilities}),$$

$$\sum_{j\in\mathcal{J}} d_j x_{ij} \leq C_i y_i \;\; \forall i \in \mathcal{I} \; (= \text{facility capacity constraints are not exceeded}),$$

$$\sum_{i\in\mathcal{I}} y_i = p \;\; (= \text{exactly p facilities open}),$$

where p is the number of facilities to locate, $\mathcal{I}$ is the set of indices of candidate facility locations and $\mathcal{J}$ is the set of demand indices and the decision variables are defined as:

$$x_{ij} = \begin{cases} 1\text{, if demand j is assigned to facility i} \\ 0\text{, otherwise} \end{cases}$$

and

$$y_i = \begin{cases} 1\text{, facility is open} \\ 0\text{, otherwise} \end{cases}$$

In our case, the optimization simultaneously determines which candidate deposit locations to activate, subject to the user-specified number of locations, and which deposit each street segment's snow should be transported to, ensuring that the full snow volume is assigned and no deposit exceeds its storage capacity. The model enforces the following constraints:

- Capacity limits at each candidate deposit.
- Activation of exactly the number of deposit locations specified by the user.
- Unique assignment of each street segment for one selected deposit.

The objective of the optimization is to minimize the total volume-weighted travel distance from each street segment to the snow deposit site assigned to it. This corresponds to minimizing the overall transportation effort required to remove snow from the streets and store it at designated deposit locations. Consequently, the output of the tool is a set of optimal deposit locations and an assignment of snow demand to each selected site.

Based on the available data, the total number of street segments is N = 9718 and the number of possible snow deposit locations is M = 960. The whole model would therefore produce binary variables for each street-deposit pair (NxM), corresponding to approximately 9.3 million binary variables, in addition to 960 variables indicating opened/closed deposits and several million associated constraints. Solving a model of this size directly would be computationally prohibitive. For this reason, we adopted a hybrid approach. An initial clustering step using k-means was performed to aggregate nearby demand points into representative clusters. Demand and distances were also aggregated per cluster. It should be noted that the clustering does not change the optimization model; it only reduces the input size to a computationally feasible level. This is a common approach in large-scale p-median literature (Francis and Lowe, 2019). Possible assignments per cluster were also restricted to the K nearest deposits to further shrink the model.

Given the above, the capacitated p-median was solved on clusters ("demand zones"), and each cluster's assignment was back-mapped to the streets belonging to it. Consequently, the tool outputs the coordinates of the selected snow deposit locations, representing the optimal configuration under the given inputs, as well as an allocation of street segments to the selected deposits, indicating how snow should be distributed spatially into the deposits.

Solver Selection and Computational Considerations

To solve the capacitated p-median model described above, two different optimization back-ends were implemented: the CBC Mixed-Integer Linear Programming (MILP) solver and the OR-Tools CP-SAT solver. Both solvers address the same mathematical formulation but differ substantially in how they process integer decision problems and in the computational guarantees they provide. In the following we outline their characteristics, assumptions, and the reasons for experimenting with both.

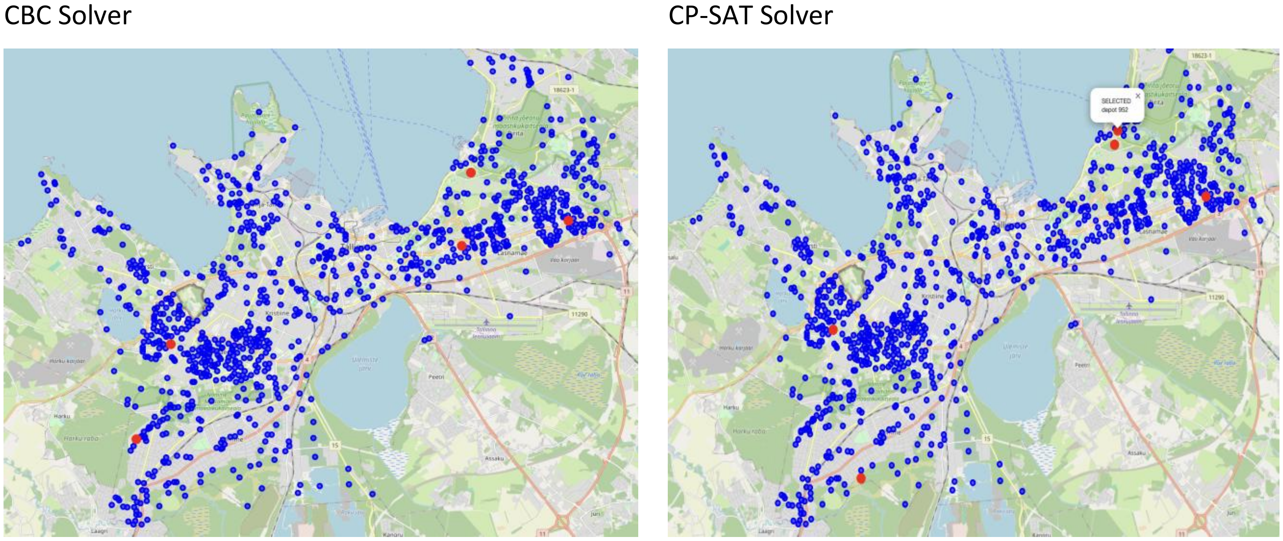
CBC MILP Solver

CBC is an open-source MILP solver implementing classical branch-and-bound with cutting planes. It supports continuous and integer variables and directly accepts the linear objective and constraints of the capacitated p-median formulation in floating-point arithmetic. In our implementation, CBC is used in combination with the sparse representation of assignment variables, which significantly reduces the size of the problem. CBC showcases stable performance for medium-scale models, especially when sparsification is used. However, CBC tends to scale less effectively when the number of binary variables grows into the millions, capacities interact with assignment variables in many-to-one patterns and the LP relaxations are weak (Bertsimas and Tsitsiklis, 1997, 1997; Wolsey and Nemhauser, 1999).

CP-SAT Solver

The CP-SAT solver from OR-Tools is a state-of-the-art constraint programming engine that combines SAT solving, cutting planes, linear relaxation, and local search (Perron et al., 2023). CP-SAT is explicitly designed for large-scale 0–1 problems with combinatorial structure, such as assignment, routing, and location problems. Unlike CBC, CP-SAT requires integer coefficients; therefore, distances are scaled to integer values prior to optimization. Key advantages of CP-SAT include highly effective search heuristics for binary assignment structures, intrinsic ability to exploit sparsity, parallel search across multiple workers, and strong performance on capacitated facility-location-type formulations.

In practice, CP-SAT proved substantially faster than CBC, particularly when the user-defined number of deposits is relatively small. CP-SAT's underlying conflict-driven search allows it to efficiently exploit combinatorial structure, and its anytime behaviour (returning feasible solutions before proving optimality) makes it especially well-suited for interactive decision-support environments. This is critical for applications such as real-time emergency response, where decision-makers benefit from a good solution quickly, even if it has not yet been certified as optimal. By contrast, CBC tends to produce highly robust and stable solutions thanks to its mature branch-and-cut framework and strong linear-programming relaxations. Although CBC may require longer solving times, it often delivers solutions with predictable quality and behaviour across problem instances, which is valuable when decision makers prioritize consistency or when results are used for documentation, auditing, or long-term planning. For these reasons, we chose to offer both solvers to users: CP-SAT for speed and responsiveness, and CBC for reliability and solution robustness. This dual-solver approach ensures that users can select the method best aligned with their operational needs and time constraints.


CBC Solver

CP-SAT Solver

Figure 1: Comparison of solver outputs for input snow volume V = 1,000,000 m³, number of deposits k = 5.

CBC Solver

CP-SAT Solver

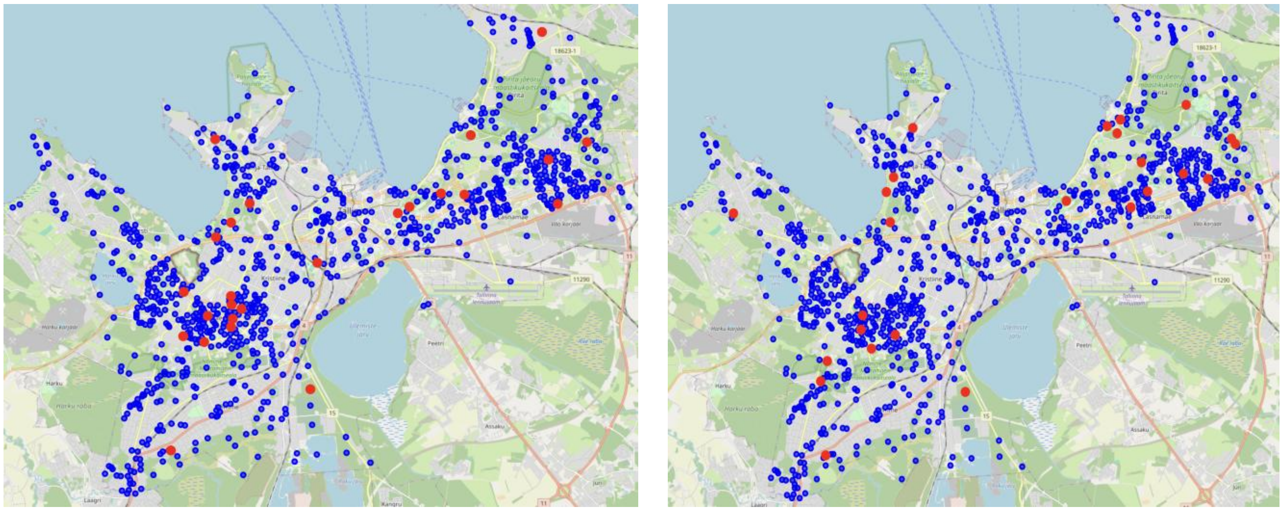


Figure 2: Comparison of solver outputs for input snow volume V = 1,000,000 m³, number of deposits k = 25

To illustrate how these differences manifest in practice, we also visualized the outputs of both solvers for a scenario involving 1,000,000 m³ of snow and a limit of 5 and 25 available deposit locations in Figure 1 and Figure 2, respectively. In these figures, all 960 potential deposit sites are shown in blue, while the sites selected by each solver appear in red. This example helps clarify how CP-SAT and CBC prioritize locations under identical constraints, providing an intuitive understanding of the solvers' behaviour and the trade-offs discussed above.

2.2.3.4 Current Development Status and Next Steps

At the current stage, the tool is implemented as a Python-based command-line application. It provides a complete operational pipeline for generating optimal snow-deposit allocations using capacitated facility location methods. The system is designed as a sequence of modular computational components, each implemented, tested, and integrated into a workflow. This includes computing snow demand for each street segment, clustering demand points, computing shortest-path distances on the road network, and subsequently solving the capacitated p-median optimization model using the solvers described above.

The command-line interface enables users to specify parameters such as the total snow volume, the number of deposit locations to activate, and the specification of the solver (cbc or cp-sat). The tool outputs the coordinates of selected snow deposit sites, along with the assignment of aggregated demand clusters to these sites and their corresponding street assignments. These results can be directly interpreted, further processed in external GIS environments, or visualized with appropriate Python libraries.

Internally, the tool executes a modular pipeline consisting of:

1. Preprocessing and calculating street-level snow demand
2. Constructing the road-network graph and computing realistic travel distances.
3. Performing spatial clustering of street segments.
4. Formulating and solving the capacitated p-median model.
5. Post-processing and back-mapping assignments from clusters to individual street segments.

This pipeline is already automated and reproducible, enabling the tool to be run multiple times with different user-defined inputs. While the optimization core and computational workflow are complete, the tool currently lacks a graphical interface and is intended for use by technically skilled users familiar with command-line environments. Its outputs are available in tabular and geospatial formats but require external visualization tools for mapping and analysis. Integration with the broader project platform and a dedicated UI are not yet implemented.

Overall, the tool is fully functional in its current technical form. The current version of the tool provides a robust foundation for selecting a predefined number of snow deposit sites based on travel distance, snow demand, and site capacity. In future versions, we plan to explore whether it is feasible to use weather forecast data from FIC to dynamically estimate snow amounts, replacing the simple assumption that snow distribution is proportional to road area. This would allow for time-dependent snow distribution instead of a static estimate. Remote sensing sources could also help estimate spatially varying snow distribution. Additionally, information on snow machinery storage from Tallinn's municipality could enable the tool to account for the origin and return of snow removal vehicles, leading to more accurate results that better reflect real-world conditions. The pipeline could also be expanded to include a routing optimization step after selecting an optimal subset of available snow deposit locations, integrating these deposits into a comprehensive location-routing framework. Environmental or regulatory constraints, such as exclusion zones or noise-sensitive areas, could also be added. It’s important to note that any future updates depend on the availability and suitability of the relevant data. Lastly, the goal is to integrate the tool into the URBREATH platform to provide a user-friendly interface. Ultimately, users will be able to view the selected deposits on an interactive map, enabling a more straightforward overview and better management interpretation.

2.2.4 Crime Statistics Tool

2.2.4.1 Tool Overview and Objectives

Socioeconomic conditions significantly influence crime trends and their effects on well-being. Research consistently finds strong links between poverty, unemployment, and increased crime rates. (United Nations Office on Drugs and Crime (UNODC), 2021). For example, neighborhoods with economic disadvantages frequently face higher levels of property crimes, vandalism, and theft. The Crime Statistics Tool aims to help cities understand how crime patterns change over time and how they connect to broader socio-economic and environmental factors. The tool combines geospatial crime mapping with demographic data, urban features, and advanced analysis methods to identify high-risk areas and predict emerging trends. Its main goal within URBREATH is to provide a decision-making framework that assists municipalities in assessing safety conditions, designing targeted interventions, and gaining a better understanding of how crime impacts citizen well-being.

The tool analyzes crime data across both spatial and temporal levels. Spatially, incident locations are mapped and grouped by neighborhood boundaries, helping to identify persistent hotspots and areas where environmental factors, such as poor lighting, limited public spaces, or broken infrastructure, may increase crime risk. Temporally, the tool uses time-series techniques to find seasonal patterns, long-term trends, and irregular variations. These analyses help differentiate between regular cycles and abnormal spikes, which could indicate reporting problems or sudden shifts in local conditions.

To enhance predictive capabilities, the tool includes an LSTM-based modeling component trained on historical crime data from the City of Leuven. This enables the system to capture complex temporal dependencies and produce short-term crime forecasts. Additional diagnostic modules, such as residual-based anomaly detection and Isolation Forest methods, offer insights into unusual patterns and data inconsistencies. Together, these components create a comprehensive analytical workflow that supports evidence-based planning, early detection of emerging safety issues, and a better understanding of the factors influencing crime dynamics in urban areas.

2.2.4.2 Data Sources, Retrieval, and Preprocessing

The crime data used to develop the URBREATH Crime Statistics Tool were provided directly by the City of Leuven. The municipality supplied a set of Excel files, each corresponding to a specific crime category:

- assault and battery (not intra-family)
- bicycle thefts
- burglaries
- vandalism

- noise pollution reports
- thefts from or to vehicles

Each dataset contains essential descriptive information for every recorded incident, including:

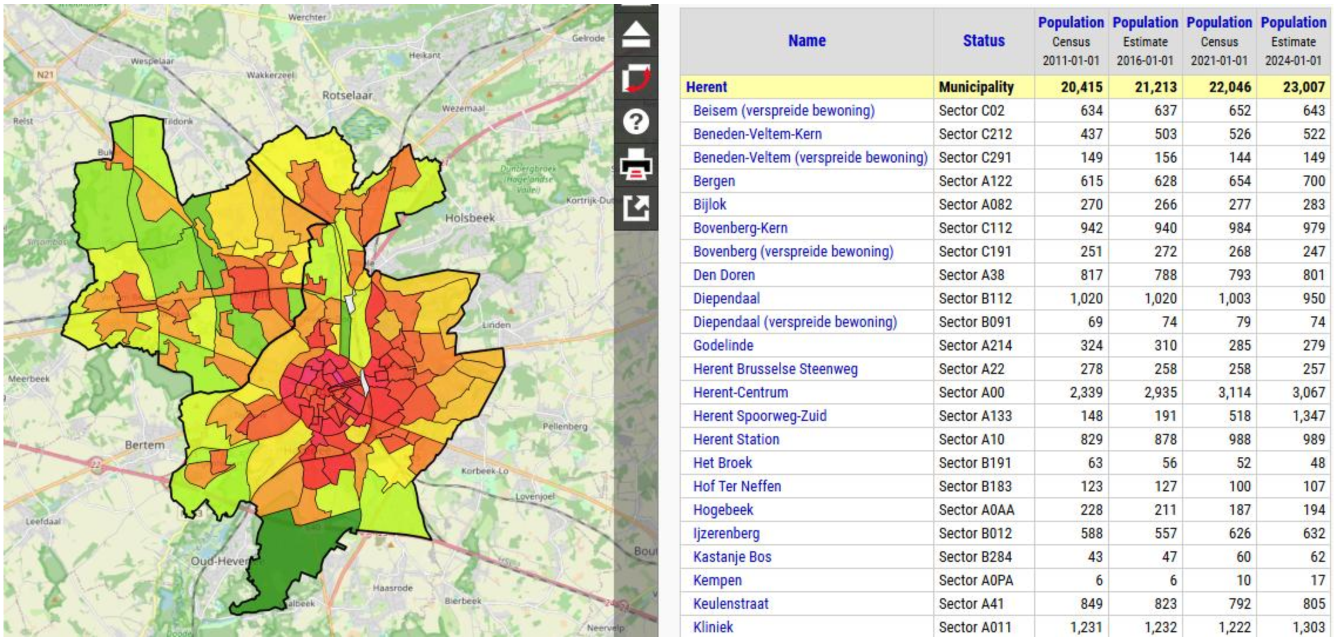
- location data (Leuven statistical sector, postal code, and street name)
- date of occurrence
- time of the incident

The complete set of crime logs spans from 2015 to 2024, offering almost a decade of data suitable for trend analysis, seasonality detection, and predictive modeling.

Upon retrieval, all Excel files were examined for structural consistency and then imported into a unified processing environment. During the initial preprocessing stage, entries with missing or incomplete information, such as absent timestamps or location fields, were removed from the consolidated dataset to ensure data quality and methodological robustness. Additional steps included format harmonization, constructing complete timestamps from separate date and time columns, and validating basic geographic identifiers. Figure 3 is a map of Leuven illustrating the various city sectors for which crime reports are available.

For the initial version of the tool, the individual Leuven sectors were classified based on their location and distance from the central urban area. This simple spatial categorization supports early-stage exploratory analysis and comparison across different parts of the city. In future versions, and in collaboration with the municipality, a more detailed classification system will be developed using additional criteria such as population density, presence of abandoned buildings, socio-economic characteristics, and other indicators of urban vulnerability. These improvements will enhance the analytical value of the tool and make it more applicable to real planning needs.

The resulting cleaned and integrated dataset served as the foundation for subsequent time-series analysis, anomaly detection, and LSTM model training.



Name	Status	Population	Population	Population	Population
		Census	Estimate	Census	Estimate
		2011-01-01	2016-01-01	2021-01-01	2024-01-01
Herent	**Municipality**	**20,415**	**21,213**	**22,046**	**23,007**
Beisem (verspreide bewoning)	Sector C02	634	637	652	643
Beneden-Veltem-Kern	Sector C212	437	503	526	522
Beneden-Veltem (verspreide bewoning)	Sector C291	149	156	144	149
Bergen	Sector A122	615	628	654	700
Bijlok	Sector A082	270	266	277	283
Bovenberg-Kern	Sector C112	942	940	984	979
Bovenberg (verspreide bewoning)	Sector C191	251	272	268	247
Den Doren	Sector A38	817	788	793	801
Diependaal	Sector B112	1,020	1,020	1,003	950
Diependaal (verspreide bewoning)	Sector B091	69	74	79	74
Godelinde	Sector A214	324	310	285	279
Herent Brusselse Steenweg	Sector A22	278	258	258	257
Herent-Centrum	Sector A00	2,339	2,935	3,114	3,067
Herent Spoorweg-Zuid	Sector A133	148	191	518	1,347
Herent Station	Sector A10	829	878	988	989
Het Broek	Sector B191	63	56	52	48
Hof Ter Neffen	Sector B183	123	127	100	107
Hogebeek	Sector A0AA	228	211	187	194
Ijzerenberg	Sector B012	588	557	626	632
Kastanje Bos	Sector B284	43	47	60	62
Kempen	Sector A0PA	6	6	10	17
Keulenstraat	Sector A41	849	823	792	805
Kliniek	Sector A011	1,231	1,232	1,222	1,303

Figure 3: Map presenting the sectors of Leuven municipality (source: Leuven_Sectors).

2.2.4.3 Modelling Approach and Technical Implementation

The modelling approach adopted in the URBREATH Crime Statistics Tool builds on established statistical and mathematical frameworks that have long been used to analyze and predict crime dynamics. Statistical methods provide essential capabilities for understanding temporal and spatial patterns and detecting irregularities. Earlier research has demonstrated how mathematical and computational models can explain the emergence of crime hotspots, capture behavioural feedback loops, and identify conditions under which criminal activity escalates or stabilizes (Short et al., 2008). Comprehensive reviews further highlight how statistical models, ranging from regression-based approaches to more complex dynamical formulations, have become integral tools for analysing crime data, evaluating interventions, and guiding decision-making in urban environments (Sooknanan and Seemungal, 2023).

The reliance on robust statistical modelling is particularly important given the limitations and biases inherent in real-world crime datasets. Studies show that police-recorded crime can be affected by substantial underreporting and that bias tends to increase at finer spatial scales, complicating micro-level analysis and hotspot detection (Buil-Gil et al., 2022). Within this context, URBREATH integrates advanced statistical and machine learning techniques, including time-series decomposition, anomaly detection, and predictive modelling, to mitigate data uncertainties and support reliable interpretation.

In particular, the modelling approach for the URBREATH Crime Statistics Tool aims to capture the temporal and spatial complexity of crime patterns and to turn these insights into actionable insights that can inform planning interventions. Using multi-year crime datasets from the City of Leuven, the tool combines statistical time-series techniques with machine learning methods to analyze historical behaviour, identify anomalies, and forecast future trends. This hybrid approach enables the system to consider both predictable elements, such as seasonality or long-term shifts, and irregular fluctuations that may indicate emerging risks or data inconsistencies. The technical implementation concentrates on developing a robust analytical workflow capable of supporting early pattern detection, hotspot analysis, and predictive modelling across various crime categories and geographical areas.

2.2.4.3.1 Time Series Analysis Tool

Time-series analysis plays a central role in crime prediction and analysis framework developed within URBREATH. Crime data usually shows temporal patterns, such as seasonal fluctuations, weekly cycles, long-term trends, and abrupt changes, making time-series–based analytical tools particularly suitable for detecting irregularities and predicting future behavior. By analyzing historical crime records from Leuven, the tool identifies deviations from expected patterns, highlights anomalous events, and supports proactive safety planning. Outlier detection methods, such as residual-based anomaly analysis and Isolation Forest, are integrated to distinguish normal variability from unusual or potentially concerning shifts in crime activity.

Residual-Based Outlier Detection

Residual-based outlier detection is a classical statistical technique used to identify observations that deviate substantially from expected behaviour in a time series (Hyndman and Athanasopoulos, 2018). It relies on the analysis of residuals, defined as the difference between the observed crime count y_t and the value predicted a fitted model $\hat{y}_t$. This is formally expressed as:

$$e_t = y_t - \hat{y}_t$$

When the forecasting model (e.g., ARIMA, linear regression, or LSTM) accurately captures the underlying trend and seasonality of the series, the residuals should behave like random noise with constant variance and no temporal structure. Outliers correspond to residuals that exceed the range of normal variation or demonstrate unexpected temporal behaviour. A common approach is to compare each residual to a threshold based on the standard deviation of the residual series (Box et al., 2015; Chandola et al., 2009):

$$|e_t| > k \cdot \sigma_e$$

Where σ_e is the is the standard deviation of the residuals and k is a sensitivity parameter (typically between 2 and 3). Residuals falling outside this range are flagged as potential anomalies. Robust variants use the median absolute deviation (MAD) to reduce the influence of extreme values:

$$|e_t - median(e)| > k \cdot MAD$$

Beyond simple thresholding, additional diagnostic tools can be applied. Autocorrelation analysis helps determine whether the residuals exhibit patterns inconsistent with white noise, indicating model misspecification or unmodelled seasonality. Cumulative sum (CUSUM) tests or rolling-window statistics can further detect structural breaks—persistent shifts in crime levels over time.

Outliers detected through residual analysis may reflect:

1. Data inconsistencies or incomplete reporting
2. Short-term external shocks such as festivals, police interventions, or weather events,
3. Long-term shifts in criminal activity due to socio-economic changes
4. Unmodelled behavioural patterns not captured by the forecasting model.

Figure 4 presents an example of the residual-based outlier detection method applied to vandalism cases in Sector A43 of Leuven. The identified anomalies illustrate the method's capability to detect irregular spikes, support the evaluation of model performance, and offer insights into unusual events requiring further investigation.

Residual analysis, therefore, serves as both a diagnostic tool, ensuring that the predictive model is well fitted, and an analytical method for detecting meaningful deviations in crime patterns. It complements more advanced machine learning techniques and forms an essential component of the broader temporal modelling framework used in the URBREATH Crime Statistics Tool.

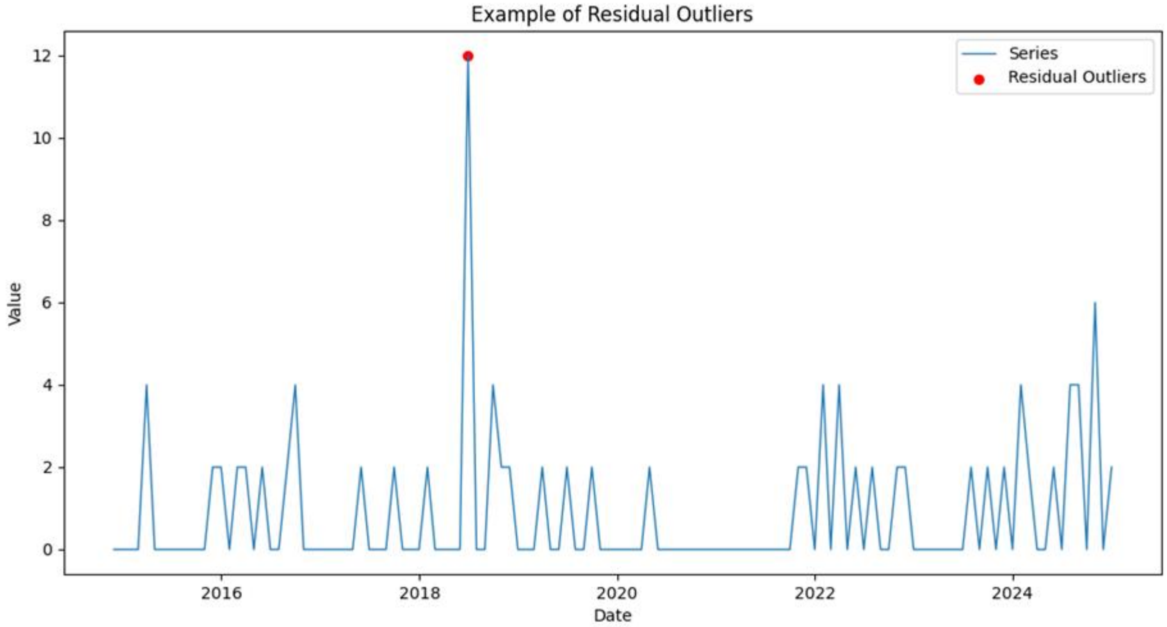


Figure 4: Example output of the residual anomalies method applied to a representative urban crime time series.

Isolation Forest

Isolation Forest is a machine learning technique specifically designed for anomaly detection, especially effective in high-dimensional or irregularly distributed datasets (Liu et al., 2012, 2008). Unlike traditional distance- or density-based methods, Isolation Forest constructs an ensemble of randomly generated binary trees, known as isolation trees, which recursively partition the data space. The fundamental idea is that anomalies are easier to isolate: because they are rare and significantly different from most observations, they generally require fewer random splits to separate from the rest of the data. The path length of an observation, defined as the number of splits needed to isolate it, is the main measure used to assess how anomalous it is.

Formally, the anomaly score for an observation $x$is computed as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where $E(h(x))$ is the expected path length of x across all trees in the forest, n is the sample size, and $c(n)$ is the average path length of unsuccessful searches in a binary search tree. Scores close to 1 indicate strong anomalies, while values near 0 suggest normal behaviour. This formulation allows the algorithm to handle non-Gaussian, skewed, or heterogeneous distributions, conditions commonly found in crime datasets.

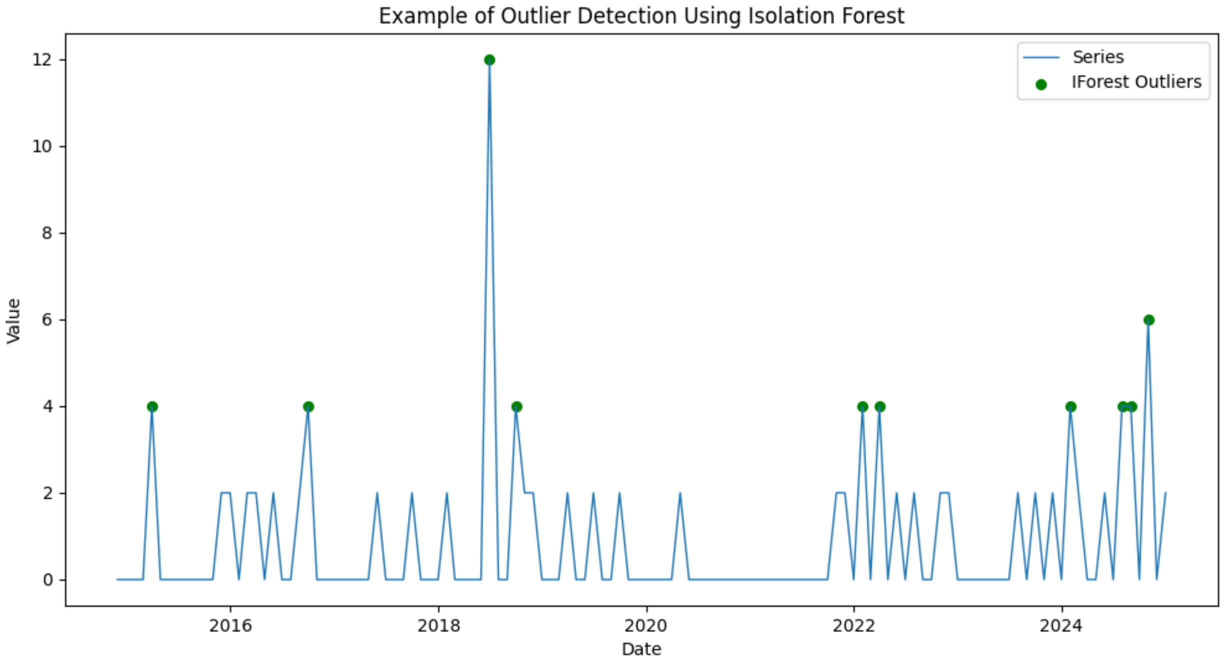


Figure 5: Example output of the Isolation Forest method applied to a representative urban crime time series.

In crime analytics, time-series behaviour often exhibits nonlinear dynamics, abrupt fluctuations, and atypical events that traditional statistical models may not fully capture. Isolation Forest is well suited for detecting such irregularities. It can identify unexpected temporal spikes, sudden drops (e.g., during lockdown periods), changes in variance, or other structural deviations that indicate abnormal crime activity or unusual reporting patterns. Because the method does not assume any specific distributional form, it remains robust across different crime types and neighbourhood profiles.

Figure 5 illustrates an application of the Isolation Forest algorithm to vandalism cases in Sector A15 of Leuven. The detected anomalies indicate periods where criminal activity diverges notably from historical patterns, providing valuable clues for further investigation or targeted policy measures.

2.2.4.3.2 AI crime level prediction tool

Within the URBREATH crime statistics tool, the Long Short-Term Memory (LSTM) architecture is applied to historical crime data from the Leuven municipality to uncover temporal patterns and make predictions about future crime activity. Crime time series often show complex behaviour, including seasonality, gradual structural changes, nonlinear dynamics, and sudden deviations related to exceptional events or reporting anomalies. These features limit the effectiveness of traditional statistical methods, which typically assume linearity or stationarity. LSTM networks, on the other hand, are specifically designed to learn long-term dependencies and adjust to changing patterns within sequential data, making them suitable for modelling crime trends.

By capturing these underlying temporal patterns, the LSTM model offers more precise forecasting than simpler models and improves the system's ability to detect emerging changes in crime levels. This predictive capability supports proactive safety planning by enabling municipalities to anticipate potential rises in specific crime categories, allocate resources more effectively, and evaluate the possible effects of seasonal or long-term behavioural shifts. In the wider URBREATH framework, the LSTM-based prediction tool contributes to a deeper understanding of how urban conditions influence crime dynamics and strengthens the evidence base for enhancing citizen well-being and urban resilience.

Long Short-Term Memory (LSTM) networks

Long Short-Term Memory (LSTM) networks are a specialized class of recurrent neural networks (RNNs) designed to model sequential data by capturing dependencies that evolve over time (Greff et al., 2017; Hochreiter and Schmidhuber, 1997). Unlike traditional RNNs, which struggle to retain information across long sequences due to vanishing gradients, LSTMs incorporate internal memory cells regulated by gating mechanisms. These gates enable the network to selectively store, update, or discard information, making LSTMs particularly effective for time-series regression tasks where recognizing seasonal patterns, long-term trends, and nonlinear temporal relationships is essential.

An LSTM network is generally composed of the following key components:

- **Input Layer:** Processes the multivariate time-series data, such as socioeconomic indicators, environmental variables, or historical crime records.
- **LSTM Units:** Each unit contains a memory cell together with three gating mechanisms that regulate the flow of information:
  - **Input gate:** determines which new information is added to cell state.
  - **Forget gate:** identifies which past information should be discarded.
  - **Output gate:** controls what information is passed to the next layer.

  These gates enable the model to preserve long-term dependencies and mitigate issues such as vanishing gradients.

- **Dense Layer:** Maps the temporal features learned by the LSTM units to a set of predictive representations.
- **Output Layer:** Produces the final regression output, typically representing the predicted value for each future time step.

Together, these mechanisms enable the model to maintain long-range temporal dependencies and learn complex sequential dynamics. After the recurrent layers, a dense layer translates the learned temporal features into a compact predictive representation, and the output layer generates the final regression value for the next time step. Figure 6 presents a typical LSTM architecture, that is also implemented in the Statistical Crime Tool.

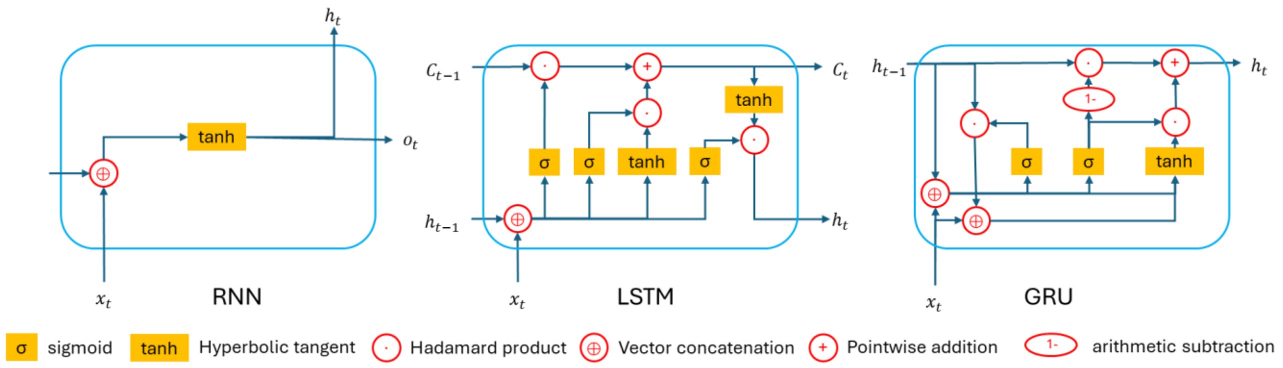


Figure 6: Typical architecture of an LSTM model.

The LSTM architecture implemented in the URBREATH crime prediction tool follows a sequential design tailored for one-step-ahead time-series forecasting. The first LSTM layer has 50 hidden units and is configured to return the full sequence of hidden states, allowing the model to analyse temporal interactions across the entire input window. A dropout layer with a rate of 0.2 is included to mitigate overfitting by randomly deactivating neurons during training. The second LSTM layer, also with 50 units, returns only the final hidden state, effectively summarizing the temporal information into a single context vector. An additional dropout layer provides further regularization.

After the recurrent stage, the model transitions into a fully connected section. A dense layer with 25 neurons and ReLU activation transforms the LSTM output into a richer latent representation capable of modelling additional nonlinear relationships. The final output layer comprises a single neuron that produces the predicted crime value. The network is trained using the Adam optimizer with mean squared error (MSE) as the loss function, which is well suited for continuous time-series forecasting tasks.

LSTM networks are particularly advantageous for crime prediction because crime time series often contain nonlinear behaviours, seasonal fluctuations, and sudden deviations that simpler statistical models cannot effectively capture. By learning these underlying temporal patterns, the LSTM model enhances forecasting accuracy, improves detection of emerging trends, and supports proactive decision-making for urban safety planning. Figure 7 presents preliminary training results, illustrating the loss function for the LSTM model applied to vandalism incidents in Section A00 of central Leuven.

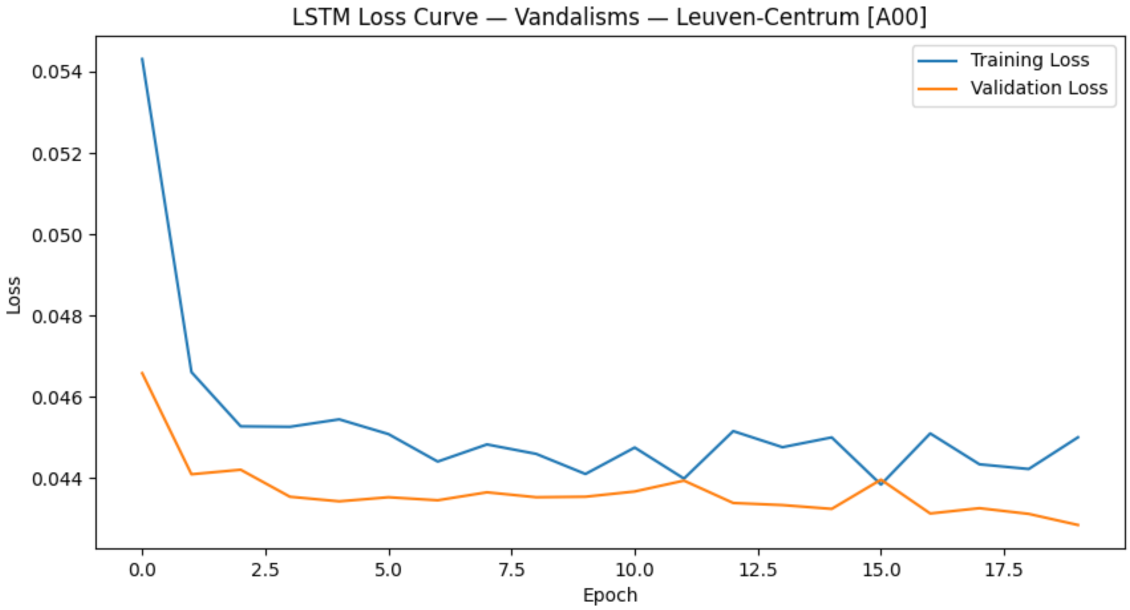


Figure 7: Indicative preliminary results of the loss function during model training for vandalism incidents in Section Aoo of central Leuven.

2.2.4.4 Current Development Status and Next Steps

The AI crime prediction module is currently in an early but functional stage of development. The LSTM architecture has been implemented and tested using the historical crime datasets provided by the City of Leuven, with initial experiments conducted on selected crime types such as vandalism, bicycle thefts, and burglaries. These initial tests show that the model can identify key temporal behaviours—seasonal variations, long-term trends, and sudden deviations—while the integrated outlier detection methods (residual-based analysis and Isolation Forest) provide additional insights into unusual or structurally significant events. Meanwhile, the statistical component of the tool is being prepared to support a set of visualization features that will allow users to explore crime characteristics interactively across multiple dimensions, including:

- crime category,
- time zone or hour of day,
- area or cluster of areas (as defined in consultation with the municipality), and
- temporal evolution of crime activity over the 2015–2024 period.

These visual analytics capabilities will help contextualize spatial and temporal differences, reveal underlying trends, and support more intuitive interpretation by non-technical users.

Preliminary analyses also highlighted several considerations shaping the roadmap for further development. In particular, initial tests found limited correlation between crime levels and the presence of Nature-based Solutions (NbS). This suggests that NbS may not exert a measurable short-term influence on crime or that their potential effects are overshadowed by stronger confounding factors. Notably, the COVID-19 time-frame appears to have had a significant impact on crime patterns, often dominating or masking more subtle effects related to local interventions or small-scale NbS deployments. This reinforces the need for a broader set of explanatory variables to better understand the drivers of crime in Leuven.

Accordingly, the next phase of development will focus on expanding the analytical framework to include additional spatial, demographic, and socioeconomic features. These may encompass land-use patterns, population density, demographic composition, accessibility indicators, and mobility-related variables. Enriching the dataset with such complementary information is expected to enhance predictive performance, support more nuanced interpretation of model outputs, and provide a more comprehensive understanding of urban crime dynamics. A detailed statistical analysis will be carried out across all Leuven sectors, and—where appropriate—areas may be grouped into clusters based on criteria provided by the municipality or local stakeholders. This approach will facilitate comparisons between neighbourhoods with similar profiles and aid in refining the identification of contextual factors influencing crime behaviour.

These enhancements will guide the development of the next version of the tool, aiming to deliver a more robust, informative, and actionable crime analytics module. The improved tool will support evidence-based urban planning, strengthen the municipality’s ability to anticipate emerging risks, and contribute to a better understanding of citizen well-being across different neighbourhoods.

2.2.5 Rent Price Prediction

2.2.5.1 Tool Overview and Objectives

Property prices and rent levels are key indicators of urban economic health, housing accessibility, and overall social stability. Since housing is one of the largest expenses for most households, fluctuations in the real estate market can directly impact residents' quality of life. Rising prices may reflect economic growth or a city's appeal, but they can also hinder access to affordable housing, especially for low- and middle-income individuals (Eurostat, 2021; Goracy et al., 2024). Understanding how property values change over time helps city officials predict affordability issues and plan for more inclusive urban growth.

Housing affordability greatly affects individual and community well-being. When rents or mortgage payments increase faster than household incomes, families often face financial struggles, cut back on essential expenses, and risk displacement. These issues can weaken social ties, increase mental stress, and deepen existing inequalities. Conversely, stable and affordable housing promotes community unity, encourages long-term residence, and allows households to invest in health, education, and daily life (Kadi and Lilius, 2024).

Within URBREATH, the Property Prices and Rent Trends tool seeks to explore how housing market conditions relate to wider socio-economic and environmental factors. The tool combines information on property values and rental prices with demographic indicators, income data, population density, and proximity to services or amenities. This integrated approach helps identify the factors that make some neighbourhoods more affordable than others, and how these patterns influence residents' opportunities and well-being. Insights from this analysis can support local authorities in shaping policies that address affordability concerns and promote balanced urban growth.

The relevance of this tool to URBREATH is its role in understanding how housing affordability interacts with the broader urban environment. By integrating housing data with information on mobility, accessibility, environmental quality, and social conditions, the tool enhances other parts of the URBREATH framework. This integration enables a more comprehensive evaluation of how various elements of the urban system impact citizens' daily lives and overall quality of life.

2.2.5.2 Current Development Status and Next Steps

At the start of the project, the consortium discussed developing this tool with the City of Leuven. The city showed interest in such an analysis, but at that time, it did not yet have access to sufficiently detailed rental price data. As a result, the tool could not be developed during the initial phases of URBREATH. After renewed discussions at the General Assembly held on October 15–17, 2025, Leuven confirmed that it had since gathered the necessary rent price datasets. With this data now available, the WP3 relevant partners can start designing and implementing the Property Prices and Rent Trends tool.

In the upcoming stages of the project, the tool will be developed using the newly provided data along with additional socio-economic variables. Both statistical techniques and AI-based models will be explored to analyze how rental prices relate to demographic trends, land-use patterns, accessibility, and other urban features. The results will help deepen understanding of housing dynamics in Leuven and support evidence-based decisions aimed at improving affordability and well-being across different neighborhoods.

3 Tool Integration within the URBREATH Platform

3.1 Analysis of Integration Technologies

The tools and models outlined in Section 2 are designed to be incorporated into the URBREATH platform to improve their usability and make their outputs accessible to end users. While each tool has its own functional requirements and technical details, the integration method follows a common core structure.

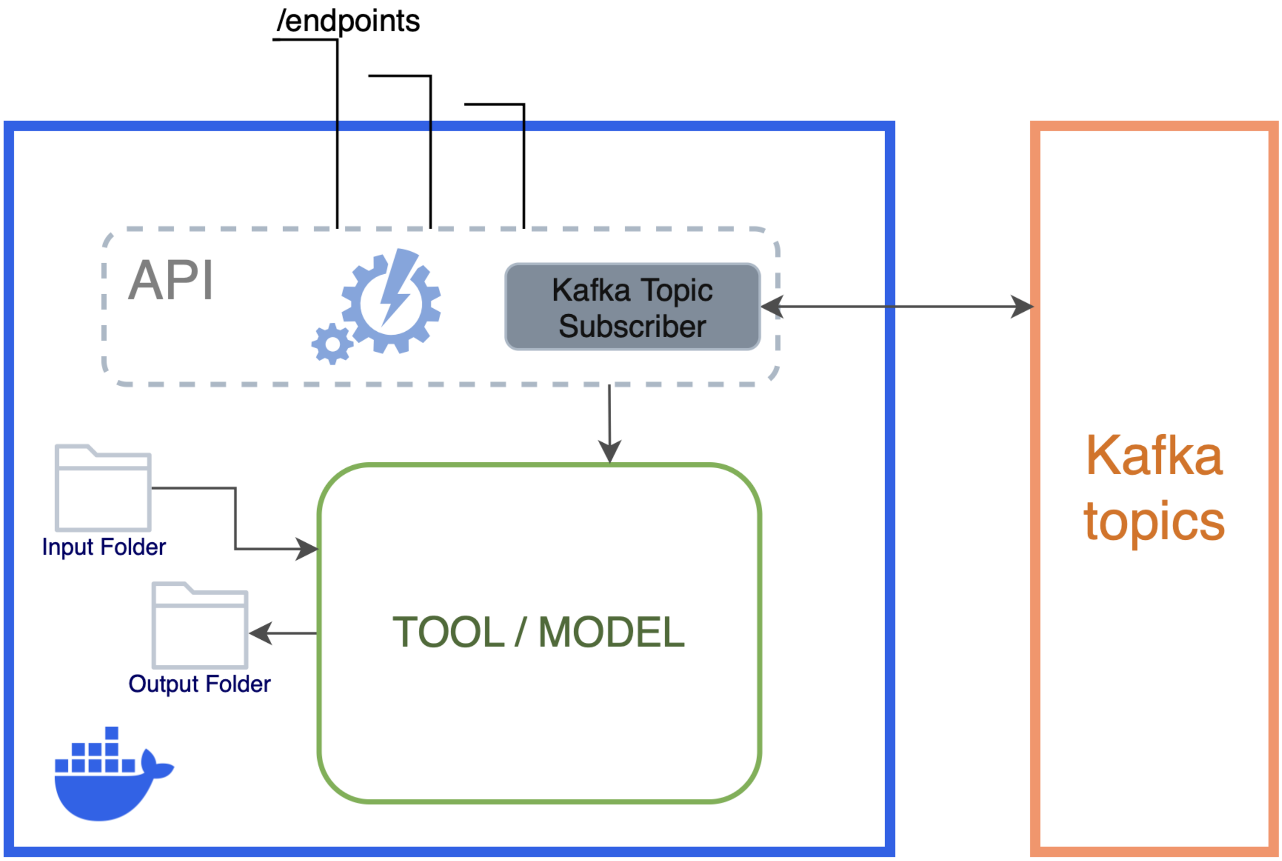


Figure 8: High-level docker container architecture and communication with Kafka.

Most tools—whether analysis tools, scenario simulations, or prediction models—require users to provide specific input parameters, such as preferences, coordinates, or thresholds. Once these inputs are submitted through the platform, the system must reliably communicate with the backend service where the computation occurs. To support this interaction in a scalable and decoupled way, the platform uses a Kafka-based event-driven architecture. Each tool or model is containerized with Docker, ensuring consistent performance, reproducibility, and isolation. The containerized tool exposes its functionality via an API and subscribes to the relevant Kafka topic. When a user submits parameters through the URBREATH interface, a message is published to the appropriate topic. The Dockerized tool then receives this message, performs the requested computation, and sends the results back through the same communication channel. A high-level overview of this architecture, including the Docker container, internal components, and Kafka ecosystem, is shown in Figure 8.

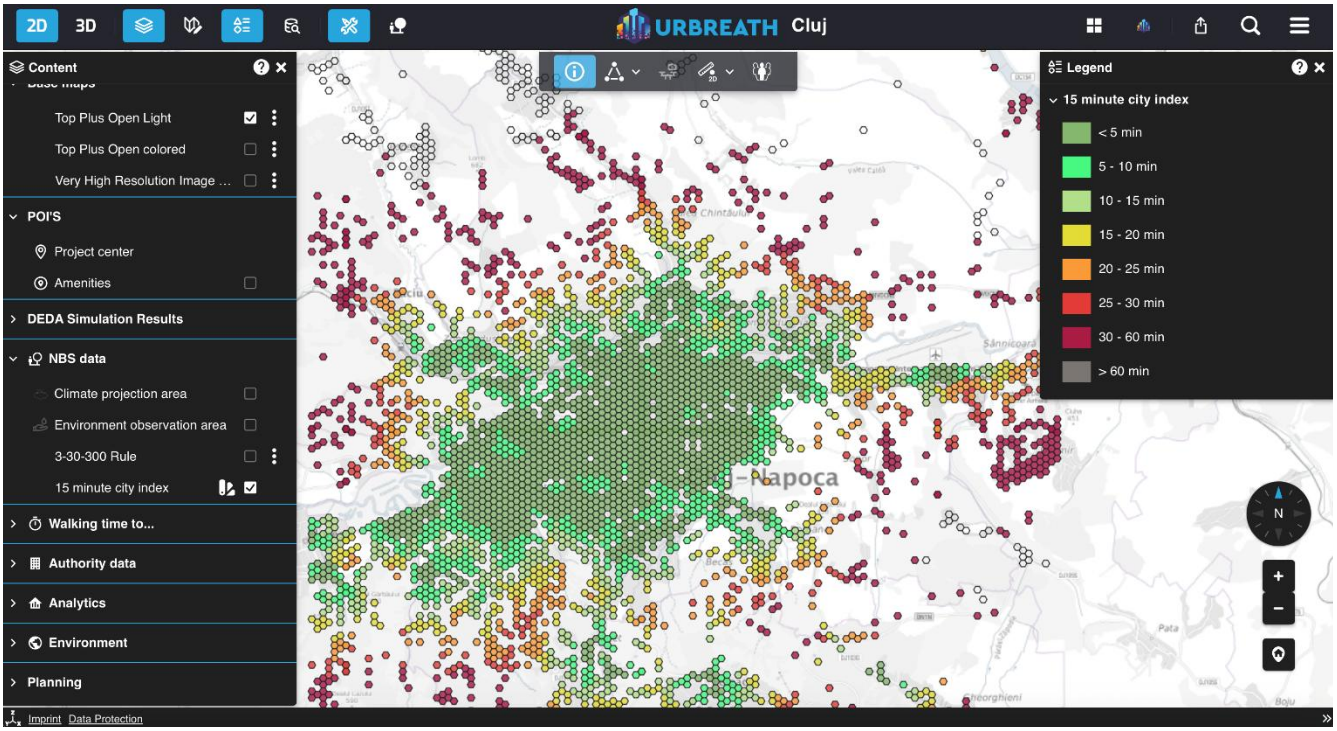


Figure 9: Initial implementation of 15-minute index analysis in the URBREATH platform (a).

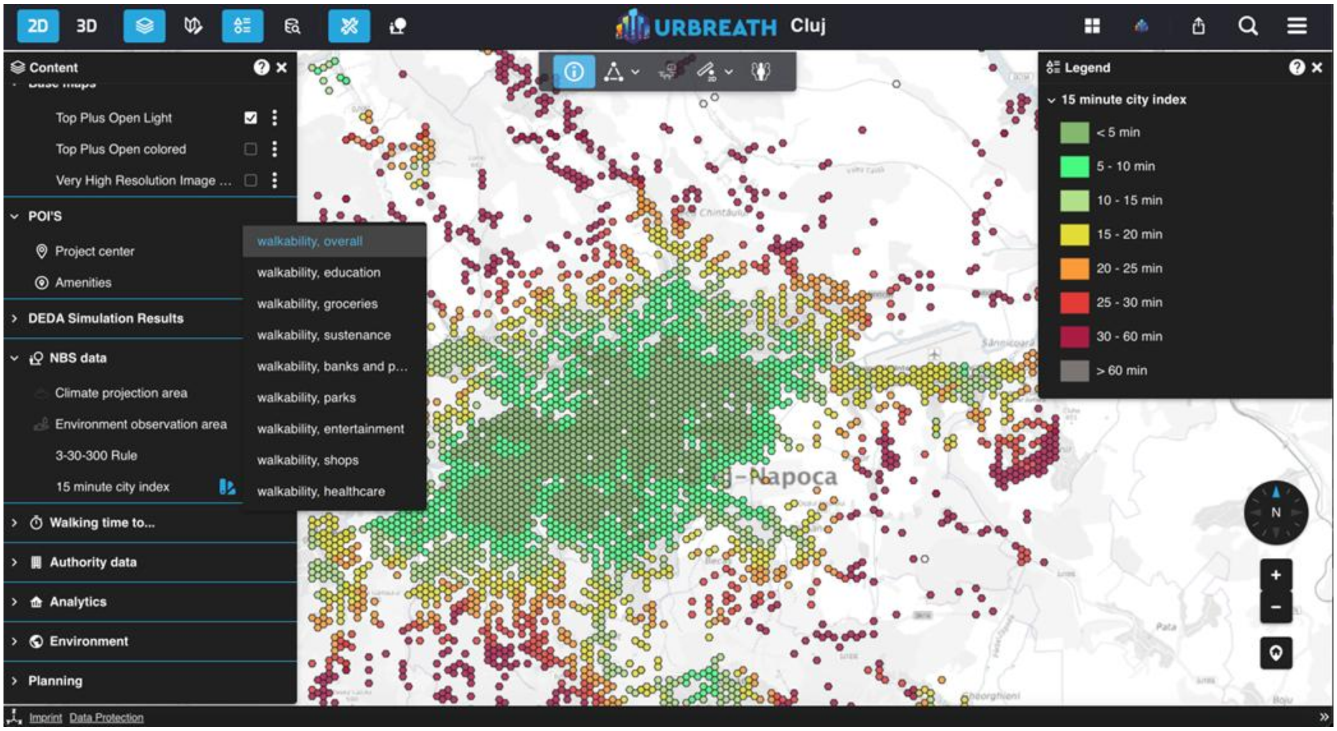


Figure 10: Initial implementation of 15-minute index analysis in the URBREATH platform (b).

3.2 Current Integration Status

The 15-minute index tool is already integrated into URBREATH Virtual City Map for the case study of Cluj-Napoca. The interactive map offers a user-friendly interface for users to explore the analysis results of the 15-minute index, either as an overall score (Figure 9) or focusing on specific categories (Figure 10). A legend on the right side of the interface explains the scoring scale and its corresponding color gradient, making it easy for users to interpret spatial differences throughout the city.

While the current analysis is fully available, the next development step is to create a more dynamic, interactive version of the tool. This improved version will enable users to input a new Point of Interest (PoI), specify its coordinates and category, and activate a localized simulation. The system will then recalculate the 15-minute index within a defined bounding box around the new PoI, demonstrating how such an intervention could affect accessibility conditions in its surrounding area.

4 Conclusions

This second version of Deliverable D3.8 shows significant progress in developing the analytical tools outlined under WP3 to evaluate socioeconomic, community, organisational, and citizen well-being. Building on the conceptual foundations established in Version 1, the tools have advanced into functional prototypes that incorporate sophisticated statistical methods, machine learning techniques, geospatial processing, and optimisation strategies.

Several tools, such as the crime analysis and prediction module, the updated accessibility framework, and the operational optimisation models, have reached a level of maturity that allows for initial validation and sets the stage for integration into the URBREATH platform. The availability of enriched datasets, especially from the City of Leuven, has facilitated comprehensive time-series modelling and anomaly detection workflows, which will serve as reference implementations for future analytical developments.

Meanwhile, the deliverable identifies necessary adjustments to the initial plan, primarily influenced by data availability and pilot-specific constraints. The scope refinement ensures that tool development remains based on reliable evidence and delivers outputs that are both technically sound and relevant for pilot cities. A flexible and adaptive modelling strategy has been adopted, allowing the tools to evolve as additional datasets or new pilot requirements arise.

Looking ahead, the next project phase will focus on enhancing model interpretability, adding more explanatory variables, and improving interoperability with the URBREATH digital platform. Ongoing collaboration with pilots will aid in refining sector classifications, validating model outputs, and expanding the analytical framework to other urban settings. These efforts will support the transition from prototype tools to fully operational elements of the URBREATH ecosystem.

Overall, the work carried out during this reporting period lays a strong methodological and technical foundation for the ongoing development of AI-enabled tools that deepen understanding of urban well-being and support evidence-based planning across partner cities. The tools described in D3.8 support the related KR and KPIs (i.e., create a participatory tool for engaging diverse stakeholders and social innovation strategies) by translating complex urban data into structured indicators, forecasts, and comparative analyses that can be shared with local authorities, practitioners, and community representatives. The fact that at least one pilot city has expressed interest in each tool demonstrates their relevance to real planning contexts and their suitability for engagement with local stakeholders. These analytical outputs can be embedded into participatory platforms, thereby supporting broader involvement of residents and local communities in evidence-based urban planning.

5 References

[1] Bertsimas, D., Tsitsiklis, J.N., 1997. Introduction to Linear Optimization. Athena Scientific.

[2] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

[3] Buil-Gil, D., Moretti, A., Langton, S.H., 2022. The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis. J. Exp. Criminol. 18, 515–541. https://doi.org/10.1007/s11292-021-09457-y

[4] Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput Surv 41, 15:1-15:58. https://doi.org/10.1145/1541880.1541882

[5] Daskin, M., 1997. Network and Discrete Location: Models, Algorithms and Applications. J. Oper. Res. Soc. https://doi.org/10.1057/palgrave.jors.2600828

[6] Eurostat, 2021. Housing Statistics and Reports. European Commission, Eurostat.

[7] Francis, R.L., Lowe, T.J., 2019. Aggregation in Location, in: Laporte, G., Nickel, S., Saldanha da Gama, F. (Eds.), Location Science. Springer International Publishing, Cham, pp. 537–556. https://doi.org/10.1007/978-3-030-32177-2_18

[8] Goracy, D., Maciejewska, A., Maciuk, K., 2024. The Purchasing Potential of EU Residents in the Real Estate Market in the Context of Sustained Development. Sustainability 16, 10373. https://doi.org/10.3390/su162310373

[9] Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: A Search Space Odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28, 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

[10] Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: principles and practice. OTexts.

[12] Kadi, J., Lilius, J., 2024. The remarkable stability of social housing in Vienna and Helsinki: a multi-dimensional analysis. Hous. Stud. 39, 1607–1631. https://doi.org/10.1080/02673037.2022.2135170

[13] Laporte, G., Nickel, S., Saldanha-da-Gama, F., 2019. Introduction to Location Science, in: Laporte, G., Nickel, S., Saldanha da Gama, F. (Eds.), Location Science. Springer International Publishing, Cham, pp. 1–21. https://doi.org/10.1007/978-3-030-32177-2_1

[14] Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-Based Anomaly Detection. ACM Trans Knowl Discov Data 6, 3:1-3:39. https://doi.org/10.1145/2133360.2133363

[15] Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining. Presented at the 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17

[16] Mirchandani, P.B., Francis, R.L., 1990. Discrete location theory.

[17] Perron, L., Didier, F., Gay, S., 2023. The CP-SAT-LP Solver, in: Yap, R.H.C. (Ed.), 29th International Conference on Principles and Practice of Constraint Programming (CP 2023), Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, p. 3:1-3:2. https://doi.org/10.4230/LIPIcs.CP.2023.3

[18] Short, M.B., D'orsogna, M.R., Pasour, V.B., Tita, G.E., Brantingham, P.J., Bertozzi, A.L., Chayes, L.B., 2008. A statistical model of criminal behavior. Math. Models Methods Appl. Sci. 18, 1249–1267. https://doi.org/10.1142/S0218202508003029

[19] Sooknanan, J., Seemungal, T.A.R., 2023. Criminals and their models - a review of epidemiological models describing criminal behaviour. Appl. Math. Comput. 458, 128212. https://doi.org/10.1016/j.amc.2023.128212

[20] United Nations Office on Drugs and Crime (UNODC), 2021. Global Crime Trends Report. United Nations, Vienna, Austria.

[21] Wolsey, L.A., Nemhauser, G.L., 1999. Integer and Combinatorial Optimization. John Wiley & Sons.