

From Static Data to Dynamic Mobility: A Spatio-Temporal Method for Student Mobility Estimation

Quinten van de Korput

Student number: 02011126

Supervisors: Prof. dr. Ivana Semanjski, Prof. dr. ir. Sidharta Gautama

Counsellor: Elham Fakhraian

Master's dissertation submitted in order to obtain the degree of Master of Science in Computer Science Engineering

Academic year 2024-2025

Preface

I am proud to present my master's thesis on the compelling topic of student mobility. Its completion marks the culmination of five years of dedicated study, spanning both my bachelor's and master's degrees in Computer Science Engineering. Beginning this academic journey amid the COVID-19 pandemic posed significant academic and social challenges, some of which continued well into the later stages of my education. I am deeply grateful to those who supported and guided me throughout this time.

This thesis offered a unique opportunity to bring together my interests in urban planning and the technological focus of my studies. The process has been both intellectually enriching and personally rewarding, offering an ideal environment for growth and learning.

I would like to extend my sincere thanks to my supervisor, Prof. Dr. Ir. Sidharta Gautama, whose guidance and expertise were essential throughout this project. I am also thankful to my friends, who not only shared this journey with me but also offered valuable perspectives that helped shape my work.

Permission of Use on Loan

The author gives permission to make this master dissertation available for consultation and to copy parts of this master dissertation for personal use. In all cases of other use, the copyright terms have to be respected, in particular with regard to the obligation to state explicitly the source when quoting results from this master dissertation.

Quinten van de Korput, May 2025

Explanatory Note

This master's dissertation is part of an exam. Any comments formulated by the assessment committee during the oral presentation of the master's dissertation are not included in this text.

Note on the Use of Generative AI

In the preparation of this thesis, Generative Artificial Intelligence (AI) was utilized as a supportive tool to enhance the quality and efficiency of the writing and research process. It was applied in several key areas:

- **As a spellchecker and grammar checker:** AI was employed to identify and correct typographical errors in individual words and to ensure grammatical accuracy within sentences.
- **For evaluating writing style and identifying improvement points:** The AI was used to analyze the writing style, offering suggestions for improvement in areas such as conciseness, flow, and tone.
- **In improving code quality:** Where applicable, AI tools were consulted to review and suggest enhancements for code segments, aiming for better efficiency.
- **To aid in searching academic sources:** Generative AI proved helpful in the initial stages of literature review by assisting in the generation of relevant keywords and search terms. This facilitated more targeted and effective searches for academic sources and scholarly articles.

It is important to note that while AI provided valuable assistance, it served as an auxiliary tool, and all critical decisions, interpretations, and the ultimate articulation of ideas were undertaken by the author.

From Static Data to Dynamic Mobility: A Spatio-Temporal Method for Student Mobility Estimation

by Quinten van de Korput

Supervisors: Prof. dr. Ivana Semanjski, Prof. dr. ir. Sidharta Gautama

Counsellor: Elham Fakhraian

Master of Science in Computer Science Engineering

Ghent University

Academic year 2024-2025

Abstract

Student population numbers have been growing in recent years, and students can significantly impact their host cities. However, student-focused mobility studies remain limited, partially due to data scarcity. Existing data sources for student housing information, registration systems and surveys, are often infeasible or too abstract. This creates a need for methodologies that can infer spatial distributions from aggregated data. To address this, a student housing location estimation method is proposed, using road networks and student housing numbers influenced by proximity-based influence factors calibrated by survey data. Mobility studies frequently use population-level activity-based simulations to evaluate scenarios, but some subpopulations, such as students, follow rigid schedules that are often underutilized. A framework that can utilize these rigid schedules in simulation is needed for more detailed analyses. To capture this, a student mobility model guided by utility choice models and class schedules is developed and applied within an agent-based simulation framework. Assuming proximity to campus, train stations, and the city center as influence factors, the housing location estimation achieved a 9.4% error on the student numbers for survey districts. The framework generalizes across urban contexts, apart from recalibration if the urban contexts are too different. The modular simulation framework, centered on class activities, captures key mobility patterns such as peak hours, mode choice and attendance variability. While the full choice models are still under evaluation, simplified variants show the promise of the methodology to capture real-world behavior. Both of the proposed methodologies enhance the feasibility of student-related studies, enabling better assessments of safety, accessibility, economic opportunities, student housing quality and sustainable transport options.

Keywords: Student Mobility, Agent-based Simulation, Housing Location Estimation, Spatio-Temporal Modeling

From Static Data to Dynamic Mobility: A Spatio-Temporal Method for Student Mobility Estimation

Quinten van de Korput

Supervisor(s): Prof. dr. ir. Sidharta Gautama, Prof. dr. Ivana Semanjski

Abstract—Urban areas face increasing pressure from growing populations, including students, creating challenges for sustainable mobility and housing. However, a lack of detailed data on student movements hinders effective planning and analysis.

This research introduces a spatio-temporal method to overcome data limitations. It first proposes a technique to estimate the spatial distribution of student housing by combining aggregated survey data with influence factors such as proximity to campuses, the city center, and train stations. Second, it develops a student mobility model that uses these estimated housing locations and university class schedules to model individual student mobility patterns and choices over time.

Applied to a case study in Ghent, the housing estimation method successfully approximated the known distribution of students across city districts, while providing an opportunity for aggregation on different levels. The student mobility model simulation visualized dynamic student concentrations around campus areas corresponding to class times and demonstrated the potential to analyze activity peaks and evaluate planning scenarios. The results highlight the framework's utility for gaining insights into student mobility with limited data, though further calibration of choice models is necessary for enhanced accuracy.

Keywords—Student Mobility, Agent-based Simulation, Housing Location Estimation, Spatio-Temporal Modeling

I. INTRODUCTION

The amount of people living in metropolitan areas is projected to increase, which arises challenges around sustainable urban development, mobility, and housing. [7] In the meantime, student numbers have been climbing steadily putting pressure on so-called student cities. [16] They face challenges due to significant student numbers that impact the city economically, socially, culturally and physically. [14] However, analyzing students and their impact is difficult because of a lack of data.

Shifting towards sustainable mobility has been found crucial to avoid environmental collapse. [1] To achieve sustainable mobility, insight is needed into mobility patterns. Technological solutions have been proposed for sustainable mobility, but are at a risk of reinforcing mobility injustices. [13] As cities grow, understanding and addressing the mobility patterns of specific population groups, such as students, becomes increasingly more important. By analyzing a specific subgroup, such as students, we can better understand their unique mobility needs and behaviors. [10]

This paper aims to address the lack of detailed data on student mobility by proposing a spatio-temporal method for estimating student housing locations and simulating their mobility patterns. By leveraging aggregated data and agent-based modeling, insights can be provided into student mobility that can inform urban planning and policymaking. The research is guided by two main research questions:

- RQ1: How can the spatial distribution of student housing be estimated using incomplete or aggregated data?
- RQ2: How can student mobility be realistically modelled and simulated using estimated housing locations and class schedules?

An overview of the relevant literature is provided in section II. Section III introduces two methodologies: one for estimating the spatial distribution of student housing based on survey data and influencing factors, and another for simulating student mobility using fixed class schedules and utility functions. A case study on the city of Ghent is presented in section IV to validate the proposed methods. Finally, section V discusses the findings, answers the research questions, and outlines future research directions.

II. LITERATURE REVIEW

A. Student Housing

Student housing data can be obtained through registration systems, collaborations with housing initiatives, and surveys. While registration systems offer detailed data, they might not be feasible because of technical or privacy limitations. [8] Surveys can capture characteristics but are limited by aggregation level. [15] Key influences on student housing choice include proximity to campus, sustainable transport, train stations, and student neighborhoods. [14, 18]

B. Origin-Destination Matrix Estimation

Origin destination matrices (ODM) have become a key method for representing human mobility. They indicate the number of individuals traveling from origin to destination over a given period of time. Traditional methods of constructing an ODM include travel surveys, individual- and population level models and traffic counts. [2] Stopher and Greaves [17] highlight that travel surveys are resource-intensive and time-consuming.

Traffic count data has been the most used way of determining ODMs, but the reliability heavily depends on the counting locations and the amount of counting locations. [2] More recently, a shift is made to mobile approaches using GPS or cellular data that do not rely on set measurement points. While they are good at capturing temporal and spatial repetitiveness, they do not contain data concerning the population demographics. [5] This makes them less suitable for evaluating subpopulation mobility patterns.

C. Mobility Simulation

ODMs can also be generated from simulations. Simulations allow for the modeling of individual or group mobility patterns based on predefined rules. One of the original models in human mobility is the four-step model, which is a trip-based model described by: trip generation, trip distribution, mode choice and route assignment. [11] Krizek [9] points out that trip-based models are limited because people often combine trips into chains of trips. These tour-based models can also be seen as a subset of activity-based models since the tours are described by the activities within the tour. Activity-based models are the most versatile and are grounded in the idea that travel is derived from the need to participate in activities. [12] Activity data still needs to be collected through surveys.

Agent-based modelling allows granular simulation because each individual is represented separately and follows a set of rules, including decision processes, movement patterns and interactions with other agents. [4] The accuracy of the model depends heavily on the input data, and they can become computationally expensive. [6]

Mazzulla [10] simulate student mobility, but are limited to the campus. An extension to the complete city is desired.

D. Utility Model

Utility models have become widely used to simulate choices related to activities, travel modes and destinations. The utility of an option i is described by:

$$U_i = V_i + \epsilon_i$$

with V_i the deterministic component, modeled as a linear combination of attributes and ϵ_i a stochastic component capturing unobserved factors. The main downside is the need to calibrate the parameters using real data. [3]

III. METHODOLOGY

There are 2 different methodologies proposed in this paper. The first one focuses on the student housing location estimation, while the second one uses this estimation and class schedules to model and simulate student intra-city mobility.

A. Student Housing Location Estimation

The student housing location estimation assumes several influence factors that are approximately independent. All of these influence factors are proximity based. Intensity functions are defined that indicate the intensity of student housing in

function of the distance to an influence factor. The student housing distribution is influenced by the proximity to campuses, the city center and train stations.

$$I_{campus}(d_{campus}), I_{center}(d_{center}), I_{station}(d_{station})$$

More campuses in an area attract more student interest. To capture this, a campus density factor is computed:

$$\rho_{campus,i} = \sum_{j=0}^N (d(\phi_j, \lambda_j, \phi_i, \lambda_i) < R) \quad \forall \text{ nodes } i$$

with R a defined radius, d the great circle distance, ϕ the longitude and λ the latitude. These intensity functions and $(\rho_{campus,i} + 1)$ are combined through multiplication and computed for each node in a graph representing the road network of the city. The additional factor 1 for the campus density guarantees everywhere within the city can be an intensity of student housing. Utilizing the road network of the city helps in capturing the housing density. Areas with closely spaced streets tend to have denser housing.

After assigning each node a student housing intensity, these factors are rescaled to match the correct number of students in student housing. The remaining students are assumed to be commuter students and are assigned to the train station nodes taking into account the relative size of the train station through number of travelers and connections.

The intensity functions are determined excluding large student housing projects since they can locally change the spatial distribution drastically. To determine the intensity distributions from survey data, they are discretized into a number of values. An isotonic regression is used to interpolate taking into account a non-increasing constraint. A genetic heuristic is used to determine the intensity arrays, the crossover operation is defined as:

$$C_1 = [P_1[:k] || P_2[k:n]], C_2 = [P_2[:k] || P_1[k:n]]$$

with k the crossover point, P the parents and C the children. It assumes a better solution can be formed by combining the first part of one good solution with the last part of another. The fitness is evaluated by calculating the MAE between the student housing assignment and survey data.

B. Agent-based Student Mobility Simulation

The methodology leverages spatial data (student housing locations) and temporal data (class schedules) to model student mobility. Figure 1 shows the simulation flow for an agent, centered around the class schedule. The resulting ODM shows when students begin moving, aggregated in 15-minute intervals. Travel time is calculated from the current to the destination node based on movement speed, helping determine when the student should leave to arrive on time.

Each agent is assigned a class schedule taking into account study program student numbers. Additionally, each student is assigned to a node in the road network, with a higher probability for nodes closer to the student's campuses:

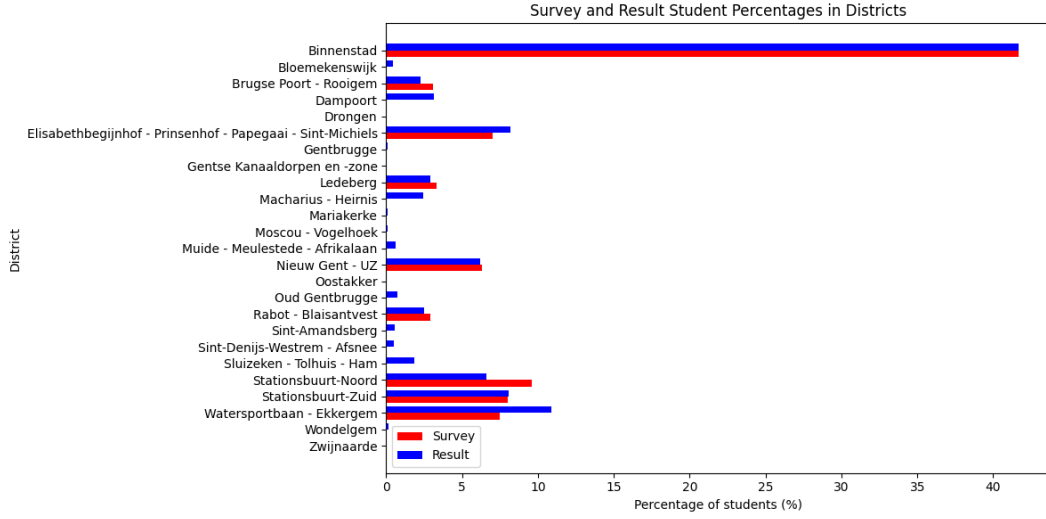


Fig. 2: Percentage of students in districts. Comparison of survey data and methodology result.

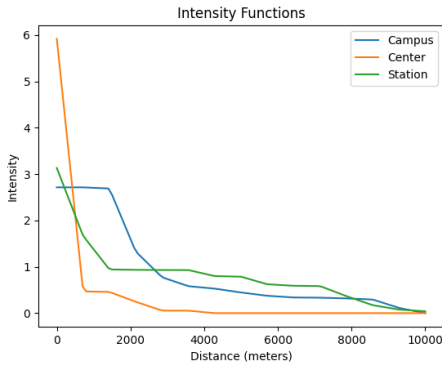


Fig. 3: Intensity functions of results for the Ghent case study.

find a near-optimal solution. The genetic heuristic parameters were chosen to make sure the execution had a high chance to converge. Figure 3 shows the resulting intensity functions for the city of Ghent.

The spatial distribution results are good. Figure 2 shows the percentage of student within each district for the methodology and according to the survey data. The best MAE achieved is 303.25. For the 29,000 students needed to be assigned to the city, this results in an error of 9.4% on the survey district's student numbers.

The main benefit of the spatial distribution is the ability to aggregate on different levels than the survey. Some care has to be taken to guarantee the error on the different aggregation level does not become too large.

The methodology generalizes to different urban contexts. The results of the literature have been shown through a case study on Kortrijk, that they are not immediately transferable. This means recalibration of the intensity functions might be needed in a different context.

B. Agent-based Student Mobility Simulation

The simulation was performed on the bachelor students of the Faculty of Engineering and Architecture at Ghent University.

There was no data available to calibrate the choice models. Thus, a simplified choice model was used for evaluating the simulation model. Each student received a probability of choosing a certain choice. Additionally, the class attendance probability decays each week by a decay factor as well. They are drawn from distributions chosen by practical considerations:

Class Attendance	$N(0.95, 0.1)$
Decay Rate	$N(0.02, 0.02)$
Home Break	$N(0.1, 0.3)$
Weekend Home	$N(0.85, 0.2)$

The travel mode probabilities were chosen constant:

$$P\{Car\} = 1/24, P\{Public transport\} = 3/24$$

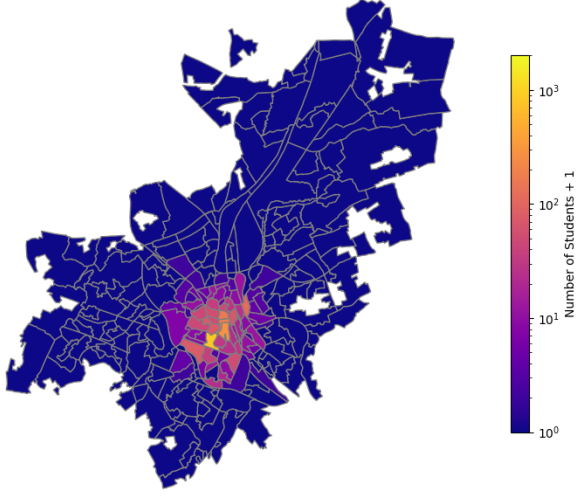
$$P\{Bike\} = 2/3, P\{Walk\} = 4/24$$

Figure 4 shows the student spatial distribution at different times on the first day of the first semester. At 8:00, figure 4a, the students are still in their student home or assigned to the train station in case of commuter students. By 12:00, figure 4b, most courses have started and a clear gathering around the campus statistical sectors can be noticed. The simulation shows potential to evaluate the spatial distributions of students through time.

The origin-destination flows can be visually observed at flowmap.city¹. Through the simulation it is also possible to follow one particular agent. This can provide even more granular insights. Information can also be gathered of the type of location the student is at. The student might be in the statistical sector of the campus but might be taking a break there.

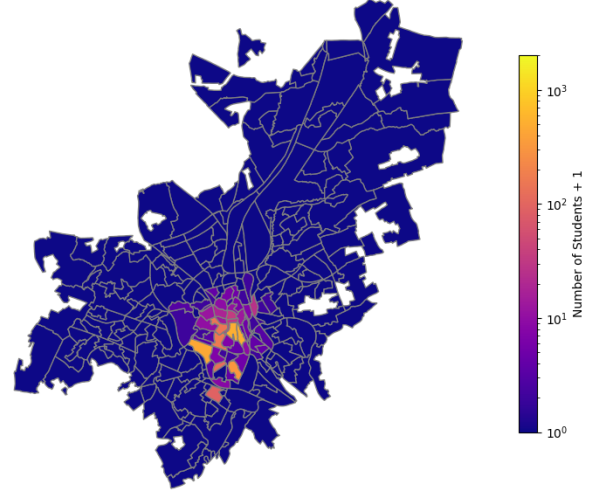
¹<https://app.flowmap.city/public/c66108de-e3c1-4a63-a3ae-65ee011feb29>

Student Distribution in Gent, Belgium at 2024-09-23 08:00:00



(a) 8:00

Student Distribution in Gent, Belgium at 2024-09-23 12:00:00



(b) 12:00

Fig. 4: FEA bachelor students spatial distribution over time

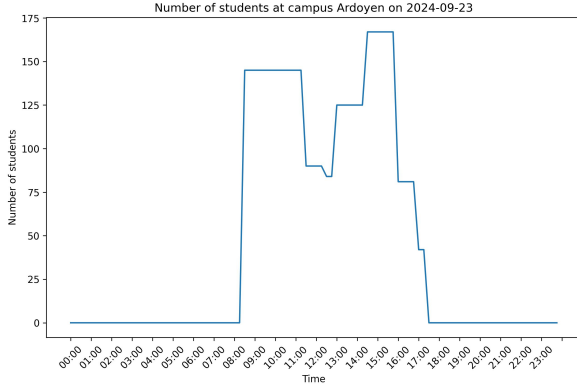


Fig. 5: Number of students at campus Ardoyen on the first day of the semester

The peak hours at specific nodes can also be analyzed. For instance, figure 5 illustrates the number of students present at campus Ardoyen throughout the first day of the semester. This type of analysis can help identify periods of high activity, which may be useful for planning purposes such as optimizing public transport schedules or campus facilities.

The mobility model allows evaluation of different scenarios. This way, policy decisions can be evaluated before implementation. For instance, the impact of increasing public transport frequency or adding new bike lanes can be simulated to assess their effect on student mobility patterns. Similarly, changes in housing policies, such as the development of new student housing projects, can be analyzed to understand their influence on spatial distributions and travel behaviors.

V. DISCUSSION AND CONCLUSION

This paper proposed two methods to address the gap within urban studies related to the student population group. The proposed solutions excel through their requirement for barely any data, while still achieving good results.

The proposed student housing location estimation framework provides a robust methodology for estimating the spatial distribution of student housing. By leveraging key influence factors such as proximity to campuses, city centers, and train stations, the framework achieves good results, as demonstrated in the case study on Ghent. While it is not a substitute for detailed registration systems, it offers a viable alternative when such systems are not feasible due to privacy or technical constraints. Furthermore, the methodology is applicable to different urban contexts when the correct student housing location influence factors are determined. The model is not able to capture all influence factors for student housing. Future work should evaluate a good set of influence factors to achieve good results. The output intensity distributions can not be transferred to completely different urban contexts.

The student mobility model complements the student housing location estimation by modelling student mobility patterns based on estimated housing locations and class schedules. The model is employed in a simulation framework. The simulation results show the ability to capture key mobility patterns while following the rigid class schedule structure, providing valuable insights into student mobility. However, validation of the choice models is necessary to enhance the accuracy and reliability of the simulation results. The stochastic choices are able to capture variability and heterogeneity in mobility behavior. The simulation framework has been able to simulate different scenarios, which can help decision makers.

The main contribution of this thesis is the ability for city councils to be able to reason more objectively about student influenced decisions. This can guide decisions impacting safety, sustainable transport modes or accessibility. Additionally, the methodology reduces the privacy risk associated with other heavily data dependent approaches.

Both methodologies are generalizable to different urban contexts. Even better, they can be made generalizable to different population groups. By adjusting the influence factors in the housing location estimation and tailoring the utility models in the mobility simulation, the frameworks can be adapted to analyze other population groups that are characterized by rigid schedules and a set spatial distribution within a city. This flexibility enhances the applicability of the proposed methodologies, making them valuable tools for urban studies and planning across diverse scenarios.

Future research into the student housing location estimation should properly test the generalizability of the methodology as well as under which conditions the intensity functions are generalizable to different urban contexts as well. This could speed up the location estimation in new cities. Extending the simulation framework to include additional student activities, such as leisure or part-time work, could provide a more comprehensive understanding of student mobility. This would require identifying the locations of these activities and calibrating the models using real movement data, while maintaining the low data requirements of the current methodologies.

Human decisions within the real-world are influenced by their surroundings through friends, family and neighbors. Incorporating social networks will increase the accuracy of decision models for students. For example, a break on campus becomes more attractive if friends also take a break on campus. To finalize, machine learning techniques can be explored based on the output of simulations. If a machine learning technique can be trained through simulation and then fine-tuned on a real case, this will greatly help the generalizability of the methodologies and might remove the need for detailed simulations within each urban context.

Ultimately, the dual-framework approach introduced in this paper highlights the potential of flexible, low-data urban modeling for a range of practical applications. As urban environments grow more complex and data privacy concerns intensify, tools that balance accuracy with minimal data dependency will become increasingly valuable. By bridging the gap between theoretical modeling and actionable planning, this research contributes a scalable foundation for data-informed decision-making in cities, especially in domains where population-specific behaviors—like those of students—play a pivotal role. With further validation and development, these methodologies can become integral to inclusive, adaptive, and efficient urban planning strategies.

REFERENCES

- [1] David Banister. “The sustainable mobility paradigm”. In: *Transport policy* 15.2 (2008), pp. 73–80.
- [2] Hugo Barbosa et al. “Human mobility: Models and applications”. In: *Physics Reports* 734 (2018), pp. 1–74.
- [3] Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press, 1985.
- [4] Eric Bonabeau. “Agent-based modeling: Methods and techniques for simulating human systems”. In: *Proceedings of the national academy of sciences* 99.suppl_3 (2002), pp. 7280–7287.
- [5] Francesco Calabrese et al. “Understanding individual mobility patterns from urban sensing data: A mobile phone trace example”. In: *Transportation research part C: emerging technologies* 26 (2013), pp. 301–313.
- [6] Andrew Crooks, Christian Castle, and Michael Batty. “Key challenges in agent-based modelling for geo-spatial simulation”. In: *Computers, Environment and Urban Systems* 32.6 (2008), pp. 417–430.
- [7] Rares Halbac-Cotoara-Zamfir et al. “Uncovering demographic trends and recent urban expansion in metropolitan regions: A paradigmatic case study”. In: *Sustainability* 12.9 (2020), p. 3937.
- [8] Evelien Vanderlinden Johan Desseyn and Sarah Huysmans. *STUDENTENHUISVESTING IN DE GENTSE WOONSTAD meet- en monitoringsystemen: “Op welke adressen wonen studenten?”*. 2021.
- [9] Kevin J Krizek. “Neighborhood services, trip purpose, and tour-based travel”. In: *Transportation* 30 (2003), pp. 387–410.
- [10] Gabriella Mazzulla. “An activity-based system of models for student mobility simulation”. In: *European Transport Research Review* 1 (2009), pp. 163–174.
- [11] Michael G McNally. “The four-step model”. In: *Handbook of transport modelling*. Vol. 1. Emerald Group Publishing Limited, 2007, pp. 35–53.
- [12] Rolf Moeckel et al. “Agent-Based Simulation to Improve Policy Sensitivity of Trip-Based Models”. In: *Journal of Advanced Transportation* 2020.1 (2020), p. 1902162.
- [13] Anna Nikolaeva. “The politics of non-knowing, smart technology and just mobility transitions: A case study and research agenda”. In: *Environment and Planning F* (2024), p. 26349825241228591.
- [14] Darren P Smith. “‘Studentification ication’: the gentrification factory?” In: *Gentrification in a global context*. Routledge, 2004, pp. 73–90.
- [15] Stad Gent. *Gentse kotstudenten: resultaten van de bevraging studentenhuysvesting bij 7500 studenten in 2024*. Nov. 2024. URL: https://stad.gent/sites/default/files/media/documents/20250121_rapport%20bevraging%20studentenhuysvesting_2024.pdf.
- [16] Stad Gent. *Tendensen*. 2024. URL: <https://hoeveelin.stad.gent/tendensen/> (visited on 04/28/2024).
- [17] Peter R Stopher and Stephen P Greaves. “Household travel surveys: Where are we going?” In: *Transportation Research Part A: Policy and Practice* 41.5 (2007), pp. 367–381.
- [18] Judith Thomsen and Terje Andreas Eikemo. “Aspects of student housing satisfaction: a quantitative study”. In: *Journal of Housing and the Built Environment* 25 (2010), pp. 273–293.

Contents

Preface	i
Permission of Use on Loan	ii
Explanatory Note	iii
Note on the Use of Generative AI	iv
Abstract	v
1 Introduction	1
2 Literature Review	3
2.1 Studentification	3
2.2 Accessibility Studies	4
2.3 Origin-Destination Matrix Estimation	6
2.4 Human Mobility Simulation	9
2.4.1 Agent-based modelling	12
2.4.2 Utility model	13
3 Student Housing Location Estimation	15
3.1 Current Difficulties	15
3.2 Estimation Quality	16
3.3 Estimation Method	17
3.3.1 Data Collection	17
3.3.2 Student Assignment	18
3.3.3 Intensity Distribution Determination	22
3.3.4 Methodology Parameters	29
3.4 Ghent Case Study	30
3.4.1 Data Collection	31
3.4.2 Intensity Distribution Determination	32
3.4.3 Student Assignment	41
3.5 Kortrijk Case Study	47
3.5.1 Data Collection	47
3.5.2 Student Assignment	48
3.6 Conclusion	48

3.7	Validity	51
4	Agent-Based Student Mobility Simulation	53
4.1	Current Landscape	54
4.2	Methodology	54
4.2.1	Simulation Structure	56
4.2.2	Agent Logic	58
4.2.3	Decision Models	60
4.3	Case study	64
4.3.1	Baseline	65
4.3.2	Full Attendance	69
4.3.3	Randomized Node Choice	71
4.4	Possible Improvements	71
4.5	Conclusion	73
4.6	Validity	73
5	Societal Impact	75
6	Conclusion and Future Research	77
	Bibliography	85

List of Figures

3.1	Road network graph of Ghent, Belgium collected from OpenStreetMap. {Red node: campus of Ghent University, Blue node: train station}	19
3.2	Complete student assignment method: From independent intensity functions to a combined function assigned to all nodes in the road network.	20
3.3	Intensity array options within an iteration of the alternating pass algorithm for a currently best array $[1, 1]$ with $\eta = 0.1$. {blue: non-increasing, red:increasing}	25
3.4	First two values in a length-10 non-increasing intensity array for the genetic algorithm over different generations. {Red: 10% fittest individuals}	27
3.5	Crossover operation of the genetic heuristic on 2 parents of length 10 with crossover index 4.	28
3.6	Mutation operation of the genetic heuristic on an array of length 10.	29
3.7	Districts of Ghent: colored areas indicate the percentage of students residing in each district, based on survey data.	32
3.8	Evolution of metric score over generations in the genetic heuristic for intensity determination.	33
3.9	Comparison MAE and MSE for the percentage of students assigned to districts.	35
3.10	Comparison of intensity arrays with the non-increasing parameter set to true or false.	36
3.11	Node intensity for intensity functions with and without campus density factor.	38
3.12	Difference in intensity functions between polynomial fitting and isotonic regression on approximately the same intensity array.	38
3.13	Variation in determined intensity functions using identical parameters, caused by randomization effects.	40
3.14	Evolution of MAE and intensity functions determined using the alternating pass approach.	40
3.15	The percentage of students within districts at the output of the alternating pass approach.	41
3.16	MAE evolution over generations in the genetic algorithm.	42
3.17	Percentage of students within districts at the output of the genetic approach	43
3.18	Intensity arrays and corresponding functions from the genetic heuristic.	43
3.19	Student housing node intensity at the output of the genetic heuristic.	44
3.20	Number of students in each statistical sector after aggregating student intensities from the genetic heuristic.	45
3.21	Student density for each statistical sector after aggregation intensities from the genetic heuristic.	46
3.22	Histogram of distances to the nearest campus for each assigned student.	47
3.23	Road network graph of Kortrijk, Belgium collected from OpenStreetMap. Red node: campus of higher education institution, Blue node: train station	49
3.24	Student node intensity in Kortrijk using intensity functions derived from Ghent.	50

4.1	Programming class structure of the simulation framework.	56
4.2	Student mobility model: agent simulation logic.	59
4.3	Example of attendance probability evolution in the simplified attendance choice model, with a starting probability of 0.9 and a decay rate of 5%.	62
4.4	Probability distribution used in the different choice models.	65
4.5	Student spatial distribution at various times on the first day of the semester.	67
4.6	Changes in student location types across different hours on the first day of the semester: break indicates break on campus	68
4.7	Number of students on the Ardoyen campus over time during the first day of the semester.	68
4.8	Class attendance rate throughout the semester.	69
4.9	Path of a single agent during the first day of the semester: warmer colors indicate trips later in the day.	70
4.10	Results of the full attendance scenario on the first day of the semester.	71
4.11	Spatial student distribution using random assignment to nodes.	72

List of Tables

3.1	Parameters student housing location assignment: options indicate the values evaluated. . .	29
3.2	Parameters intensity functions determination: options indicate the values evaluated	30
3.3	Hyperparameters alternating pass algorithm for intensity functions determination	30
3.4	Hyperparameters genetic algorithm for intensity functions determination	30
3.5	Metric scores for different sets of parameters randomly selected from the evaluated parameters. {k = polynomial degree, n = set size distances}	34

Chapter 1

Introduction

Societies are an ever-evolving mechanism influenced by numerous factors. Driven through technological innovations, they have become more mobile. In parallel, the number of people living within metropolitan areas keeps growing. [26, 54] This significantly increases the importance of urban mobility studies. These studies model and evaluate the specific mobility patterns within or between metropolitan areas. The results of these mobility-related studies are used in decision-making, influencing various aspects of urban planning and transportation systems. Recently, the student population has been growing at a faster rate than the overall population. Especially within the so-called student cities, student numbers are on the rise. These cities are heavily impacted by higher education institutions and their students. [67] An increasing number of students prefer to live in student housing. Student housing is purposely dedicated for students, providing them with a unique living environment. Furthermore, students prefer to live close to each other, creating student neighborhoods. [72] These neighborhoods are subject to various social and economic dynamics, e.g. studentification. [62]

While human mobility has become a well-studied field, it lacks opportunities to study specific population groups. The mobility patterns of population groups are hidden within the overall population's mobility. [43] However, evaluating specific mobility patterns of a population subset can provide crucial insights into the population's needs. It can also increase accuracy of existing mobility evaluations. Combining this with the fact that students are often not included within population level studies, a gap within mobility studies becomes clear. Student mobility has been understudied even though their movement patterns are subject to class schedules and thus partially predictable.

One of the challenges with student studies is the absence of student-specific data. An important data point is the location of student housing. Student housing location data is often absent. When the student housing information is collected on a house specific level, this is usually achieved through a registration system. Where a registration system is not an option because of legal or technological constraints, alternative methods need to be explored. [33] These alternative methods, the most prominent of which are surveys, suffer from their aggregation level. A result of the data absence is a limited number of student related studies.

The goal of this thesis is to propose a methodology for inferring detailed student housing locations from aggregated data and to develop a student mobility model, based on these locations and class schedules,

for use in a simulation framework to study student-related mobility. To achieve this, this thesis answers following research questions:

- RQ1: How can the spatial distribution of student housing be estimated using incomplete or aggregated data?
 - RQ1.1: What approximately independent spatial factors can be used in a student housing location estimation?
 - RQ1.2: How can the spatial dependency functions be inferred from aggregated data?
- RQ2: How can student mobility be realistically modelled and simulated using estimated housing locations and class schedules?
 - RQ2.1: What are the choice models needed for a student mobility simulation?
 - RQ2.2: How can class schedules help guide student mobility simulation?

First, chapter 2 provides the research context for this study. It covers student impact on their host cities, origin-destination matrix determination, human mobility estimation techniques and accessibility studies. The context is then utilized in chapter 3, which proposes a methodology to determine a student housing spatial distribution from aggregated survey data by determining influence factors and approximating them as independent. Chapter 4 utilizes these student housing locations and class schedules in a student mobility model that is supported by choice models. Next, chapter 5 explains the societal contributions of this thesis through a societal reflection. Finally, in chapter 6, the answers to the research questions are formulated, and potential future research directions are discussed.

Chapter 2

Literature Review

This chapter introduces essential concepts and theoretical frameworks from the literature that underpin this research. An additional goal is to give insight into the impact of students on a city. The concepts discussed here are fundamental for understanding and contextually placing the practical implementations presented in chapter 3 and chapter 4.

The chapter begins with a comprehensive overview of studentification to describe the perceived urban impact of students. Together with accessibility studies, it can gain insight into student points of interest (POI) and student housing. These POI can be related to origin-destination matrices. Both accessibility studies and origin-destination matrices provide an analytical framework for understanding movement patterns within urban environments. The literature review then explores relevant student mobility and traffic flow studies, examining their methodologies and findings. Additionally, it addresses the evolution and application of agent-based approaches in human mobility simulation, highlighting their advantages for modeling complex student movement behaviors. By synthesizing these concepts and theoretical frameworks, this literature review positions the current research within the broader field of mobility studies.

2.1 Studentification

Students can have an enormous impact on their surroundings and host city. A recurring corresponding phenomenon is studentification. Studentification, short for student gentrification, describes the economic, social, cultural and physical transformations within university cities that are related to the in-migration of students of higher educational institutions. This in-migration differs from other migration due to its seasonal and cyclic nature. [62] Student populations are present during lesson periods and leave during holidays and weekends. The main driver for studentification is the growth in student populations over the last decades. The ratio of students in student cities has increased significantly. [59] Analyzing the transformations associated with studentification gives insight into the impact of students onto their host cities.

The main economic impact is associated with the recommodification of housing into multiple student housing units that are rented out separately. A consequence of this recommodification is an inflation in property prices. [62] Housing is mostly remodified within student neighborhoods. In reaction to this

trend, cities are trying to redistribute and reinvent student housing through new developments. [35] This redistribution is sometimes in contrast with student preferences. Students prefer student housing close to campus and student services. [72] These new developments might not be build close to campus due to several environmental constraints. The main constraint on development sites is a lack of development space close to the higher education institutions. Social constraints also play a role. Studentification has an additional economic impact through student services offered. Student services organize themselves around student housing changing the economic landscape of the neighborhood. [62]

The social impact of studentification is defined by the replacement or displacement of existing population groups by a transient, generally young population group. [62] The transient nature of the student population stems from their limited study duration and their absence from the city for parts of the year. This results in new patterns of social concentration and segregation. The gathering of young people with shared cultures, lifestyles and consumption patterns is defined as the cultural transformation of studentification. The gathering of students within a neighborhood also has a physical impact which cannot be neglected. The physical transformation is often an initial upgrade of infrastructure because of the conversion to student accommodation, but in a later stage studentification can lead to a downgrade depending on the local context. [49, 62] Significant efforts need to be made to maintain the improvement of the student neighborhoods.

While studentification is closely related to gentrification, there are differences between both concepts. Studentification is partially driven by higher education institutional actors, but the students drive a demand for the type of accommodation. Gentrification is driven by higher-income individuals looking for different types of housing, amenities or neighborhood character without a direct involvement of an institution. [62] The transformations attributed to studentification show the potential impact of students on the urban landscape in their host cities. Students also impact systems within the city. For example, they impact the transportation system not only in numbers but also through their unusual mode selection and mobility patterns. [50]

The relationship between higher educational institutions (HEIs) and their host city is complex. The mutual impact between universities and a city has been found to be both positive and negative. However, this positive influence is not automatically achieved and requires work to create favorable conditions. Besides the impacts of studentification, student cities can also benefit from increased job opportunities and an increased pool of intellect. [10]

2.2 Accessibility Studies

Urban accessibility studies have existed for several decades. These studies have a focus on the urban service delivery. The goal of these studies is to evaluate and make access to services equitable. Accessibility measures can be divided into place-based measures and people-based measures. Place-based measures take into account the proximity to a service, where people-based measures include space-time information from individual-level travel data. [51] Spatial accessibility thus incorporates the ease of access to an infrastructure of interest. This has numerous policy implications, e.g. proposing additional resources in areas with poor accessibility. These accessibility studies become more accurate with more geospatial information available. [52] A more recent concept in assessing accessibility within a city is the 15-minute city (FMC). This concept has gained traction in the years following Covid-19 and most of the current research of accessibility studies focusses on this concept.

Moreno et al. [48] introduce the concept of the FMC as a way to evaluate and reduce inequality within

a city. The goal is to construct cities where all essential services are reachable within 15 minutes by foot or by bike. It fits within the broader movement of chrono-urbanism, which let's time (co)-define urban planning. Moreno et al. [48] define 6 categories of essential services: housing, work, commerce, healthcare, education and entertainment. The FMC is not rigidly defined and flexible to other cities by expanding the concept to an x-minute city. The essential services categories are described for the total population. When assessing the accessibility of a population subset, some categories might not apply. For the higher education student subset of the population, the categories work and education can be combined. While the FMC provides an easy to use framework for evaluating accessibility within cities, it also receives criticism. At first, the FMC concept introduces a good classification of required services, but does not address differences within these categories. Fulfilling some categories completely within a city is not always feasible. For example, elementary education might be possible to achieve within the 15-minute concept, but a university not because of space and class sizes limitations.

Allam et al. [4] show the potential of the 15-minute city to shape a net-zero future while also warning to make changes supported by a human approach that incorporates inhabitants. They indicate the ability to reduce green house gasses by reducing the amount of car trips within a city. Abbiasov et al. [1] evaluate the FMC using mobile GPS data. They caution for segregation through the 15-minute city because of the potential exorbitant increase in local trips. Khavarian-Garmsir, Sharifi, and Sadeghi [34] go further in warning for a lack of incorporation of different needs within social groups, energy-efficiency and culture and heritage. They also indicate implementation limitations that are not addressed within the concept.

Willberg, Fink, and Toivonen [80] compare differences within the movement speed of the population based on age, weather conditions and road conditions. This has an influence on the proximity measure of x minutes where the movement speed has to be taken into account. After the introduction of the FMC, research into an evaluation framework has been done to evaluate a city's compliance with the FMC. Gaglione et al. [23] provide an evaluation of the city of Naples, Italy taking into account different age groups. Ferrer-Ortiz et al. [21] evaluate the FMC for the city of Barcelona, Spain. The studies emphasize the importance of granular data to accurately assess accessibility and compliance with the FMC framework. Without such data, evaluations risk oversimplifying the complexities of urban accessibility and may overlook critical disparities among different population groups or service categories. Future research should aim to address these limitations by incorporating more detailed datasets and exploring innovative methodologies to provide a comprehensive evaluation of the FMC's implementation and impact.

Teixeira et al. [71] summarize implementation practices of the FMC around the world. They report an increase in FMC initiatives. It can be noticed, that the concept is mostly implemented in cities in North-America and Europe, while slowly also occurring on other continents. There is a large variation in the degree of radicalism between different implementations. This summary also indicates the limitations through data. The FMC can be a good framework provided some additional research, adaptations to the corresponding city and taking into account potential limitations of the concept.

Urban studies often incorporate multiple dimensions determined from the city's challenges. Moreno et al. [48] mention density, proximity, diversity and digitalization as dimensions to incorporate with their concept of the FMC. The dimensions can be expanded with flexibility, connectivity and human scale urban design. [34] The dimensions evaluated differ between studies. One downside of this approach is the incomparability of the different dimensions used within different studies. However, a benefit is the choice to select the dimensions that contain challenges for that city, without the need to cover all dimensions.

In studies evaluating the 15-minute city, accessibility is mostly defined from an origin point towards the services. Determining this origin point clearly allows researchers to contextualize their findings. These origin points are usually the residence of the consumer. However, data in most cities or municipalities is not available on the granularity of the household. To circumvent this, data is aggregated over statistical sectors or neighborhoods. The choice of the origin point is important as it can have a big impact on the results and validity of the study. The smaller the geographical areas used, the more representative the results will be to the actual situation. This is one of the reasons that makes every accessibility study different from each other, making it more difficult to compare results.

In a 2022 study evaluating the concept of the x-minute city, the origin points were defined as the centroid of a US census block and statistical areas in New Zealand. [40] Barcelona was evaluated using the cadastral parcels of the city. [21] They used dissemination blocks for Vancouver and census tracts for Turin to evaluate the city. [30, 68] Even though each of these studies uses different names for the statistical sectors, they all refer to the same principle. These statistical sectors can vary significantly in size, even within the same country or city. [40] This variability makes a standardized evaluation of accessibility across different cities difficult and promotes city specific evaluations.

The student population’s points of interest (POIs) overlap with those of the general population in some categories but differ significantly in others. For instance, while categories like healthcare are relevant and similar for both groups, the categories of education and work can be combined for students, as their primary ”work” is their education. For students, the combined category of education and work primarily revolves around higher education institutions (HEIs). These include university campuses, libraries, research facilities, and other academic buildings. Additionally, housing for students is often characterized by smaller units, such as dormitories, shared apartments, or purpose-built student accommodations, which are typically located near campuses or in areas with high student density. [77] Students frequently engage with commerce and entertainment POIs, but their preferences and behaviors may differ from the general population. Student-oriented entertainment often differs from that of the general population, as it is frequently organized by students themselves, though they also participate in broader entertainment offerings. Additionally, students often favor more affordable options.

2.3 Origin-Destination Matrix Estimation

The concepts of human mobility and origin-destination matrices (ODMs) are closely related. Human mobility studies focus on understanding movement patterns of individuals or groups within a specific area. This includes spatial movement, modality, individual mobility and population mobility. Origin-destination matrices are crucial metrics within the broader concept of human mobility. They indicate a number of individuals traveling between locations over a given period of time. [6] Barbosa et al. [6] provide an extensive overview of the field of human mobility, referring to ODMs and models used within human mobility. Human mobility studies can serve as input in determining ODMs, but ODMs can also be used in deducting patterns within mobility and to evaluate policy decisions. To understand the essence of origin-destination matrices further, methods to create ODMs are discussed.

Travel surveys are one of the traditional methods for constructing an origin-destination matrix. These surveys collect data directly from individuals about their travel behavior, including trip origins, destinations, purposes, modes of transport, and travel times. The data is typically gathered through questionnaires, interviews, or travel diaries. [6] Stopher and Greaves [69] highlight that while travel surveys provide detailed and accurate information, they are resource-intensive, time-consuming, and often suffer from

low response rates. Additionally, the accuracy of the data depends heavily on the willingness and ability of respondents to recall and report their travel behavior accurately. The responsiveness of travel surveys has dropped in the last couple of years. This also leads to an imbalance in the participated households. Households that travel more or are larger have lower response rates. [69] Stopher and Greaves [69] propose new avenues for travel surveys including paid panels, using GPS devices and even the bypassing of surveys all together through simulation. [60]

Despite these challenges, travel surveys remain a valuable tool for understanding travel patterns and constructing ODMs. They are particularly useful in contexts where other data sources, such as traffic counts or mobile data, are unavailable or insufficient. However, the high cost and effort associated with conducting large-scale travel surveys have led researchers to increasingly explore alternative methods for ODM estimation.

Several models exist to infer human mobility that are the result of decades of research. They can be classified into individual-level models and population-level models. The best known population model is the gravity model. The main assumption in this model is that the amount of travels between two regions is a decreasing function of the distance. The model is a gross simplification of travel flows and often falls short of capturing actual observed travel flows. [6] It still remains quite popular in transportation problems. Simini et al. [61] expand the concept of a gravity model into deep learning. The attractiveness of a region in the gravity model is based on the population size. Zhang and Li [83] adapt this concept by proposing a different attractiveness measure using the activity space.

In addition to the gravity model, several other classical population-level models have been proposed over the years, such as the radiation model, intervening opportunities model, and the random utility model. These models vary in complexity and assumptions, often focusing on factors like opportunity distribution or individual decision-making behavior. While they have provided foundational insights into human mobility, many struggle to capture real-world variability and temporal dynamics, especially when compared to recent data-driven approaches. [6]

An individual-level model is Lévy flight. It is a class of random walks that is composed of many small displacements with an occasional large displacement. The large displacements in Lévy flights can be considerable but occur at the same time as a smaller step which is unrealistic. They are a rough approximation to human trajectories. [6] Rhee et al. [57] validate the relation to Lévy flight by an analysis using GPS data. They indicate this is valid within a range of 10 kilometers and the network is likely to be disconnected. The Lévy walks model is randomized and does not capture temporal and spatial regularity. This is explored by Gonzalez, Hidalgo, and Barabasi [25] through mobile data.

Beyond Lévy flights, researchers have developed several other individual-level models that aim to capture human mobility patterns more accurately. These models focus on understanding and predicting individual movement behaviors rather than aggregate flows between regions. While population-level models like gravity models work well for large-scale predictions, individual-level models offer insights into personal movement choices, routines, and deviations. [6]

Individual-level models include time-geographic approaches, agent-based models, continuous-time random walks, preferential return models, and exploration and preferential return (EPR) models. These frameworks incorporate various behavioral aspects such as spatial memory, frequent visitation patterns, and the tendency to explore new locations. [6] Unlike pure random processes, these models attempt to capture the underlying decision mechanisms that govern human movement choices.

Bera and Rao [9] present a comprehensive state-of-the-art overview of origin-destination matrix estimation methods using traffic count data. A distinction can be made between static and dynamic estimations. In static methodologies, the traffic flows are considered as time-independent and there is a focus on long term planning and design goals. Dynamic estimations evaluate smaller network sections to determine short-term strategies. The review explores different various estimation techniques for static estimations, including gravity models, information minimization (IM) and entropy maximization (EM), statistical approaches, path flow estimation models and neural networks for both congested and uncongested networks. Dynamic approaches are more complex and use techniques like Kalman filtering and state-space models. Bera and Rao [9] consider limitations of determining ODMs from traffic data. The reliability depends highly on the counting locations and the amount of counting locations. The amount of counting locations is frequently less than the amount of O-D pairs requiring additional constraints that are not always easy to define. Rostami Nasab and Shafahi [58] indicate that even after decades of research determining ODMs from traffic data is still challenging. They infer issues with classical approaches that boil down to sensitivity on inputs and ignorance to weights within the network.

Within ODM estimation using traffic counts, the amount of traffic sensors needed relative to the network size is sometimes not discussed. However, the amount of traffic sensors and location plays a crucial role in the accuracy of the method. Yang and Zhou [82] address this by defining a minimum number of counting points for a percentage of the total flow using an integer linear programming approach. This is only if the counting locations are optimal. However, this might not be possible in a real setting through numerous real world limitations. The methodologies in ODM estimation often include an already existing partial or complete origin-destination matrix. [9, 58] In some contexts this matrix is not available, which makes most of these methodologies unusable within that context.

More recently a shift is made into ODM estimation using mobile location data from GPS, Wifi or cellular connections. [13, 16, 17, 32] Iqbal et al. [32] propose a methodology using Call Detail Records (CDR) in combination with a limited number of traffic counts to predict origin-destination matrices. They use the CDRs, which contain tower locations and caller IDs, to create a transient OD matrix from tower-to-tower. The traffic data is then used to reiterate this into a node-to-node ODM. With rescaling they arrive to a proper ODM which is validated using a real world scenario. Calabrese et al. [17] use mobile data to evaluate individual mobility. They make an important distinction between individual mobility and vehicular mobility. Traffic counts are limited to the vehicular mobility and don't gain insight into the entire mobility patterns within a city. Calabrese et al. [17] warn that mobile data does not contain information about the demographic of the users since this is private data. Other workarounds have to be found to address this and link mobile location data with demographic data. A common approach is using census data in combination with a home location. This home location can be deducted from the nights CDRs and as one of the places humans spend a significant amount of time. Users within a mobile context are often determined to be random, which cannot be guaranteed. [17]

Gonzalez, Hidalgo, and Barabasi [25] compare mobile methods to Lévy flights, an older approach for estimating human mobility. They found that mobile methods perform better because they can capture the temporal and spatial repetitiveness in human movement. Older methods are more randomized based and do not capture this repetition properly. Lévy flights were found to give a good estimation for short distances, but were outperformed by a mobile approach. Kraft et al. [37] compare a GPS based method to travel diaries. Travel diaries are dependent on participation, whereas GPS not necessarily. The mobile method based on GPS outperformed the travel diaries for understanding human travel. Even social media

has become a source of travel information. Cheng et al. [19] propose a method that uses social media inferred data to determine origin-destination matrices. They integrate it with existing travel surveys to increase accuracy.

Mobile approaches have proven to be a good solution in analyzing human mobility. Their limitations in demographic information make them less suitable for evaluating a subset of the entire population. Additionally, cooperation is required between cities and cellular providers. These cellular providers do not offer their data for free, increasing the cost of using mobile data.

In conclusion, the study of human mobility and origin-destination matrices has evolved significantly over time, with methodologies progressing from traditional travel surveys and mathematical models to sophisticated data-driven approaches. While each method offers unique advantages, surveys providing detailed demographic information, traffic counts offering objective measurement, and mobile data capturing real-time movements at scale, they also present distinct limitations that researchers must navigate. Population-level models like the gravity model offer simplicity but often lack precision, while individual-level models provide granular insights but may struggle with generalizability. The emergence of mobile data approaches represents a promising direction for ODM estimation, offering unprecedented temporal and spatial resolution, though challenges remain regarding demographic representation, privacy concerns, and data access. Moving forward, hybrid approaches that combine the strengths of multiple methodologies may prove most effective in capturing the complex dynamics of human mobility. As cities continue to evolve and transportation needs change, developing accurate and efficient methods for ODM estimation remains a critical research frontier with significant implications for urban planning, transportation management, and public policy.

2.4 Human Mobility Simulation

Simulation methods have proven to be an important tool in estimating and modelling human mobility. Without the need to conduct surveys or collect traffic data, mobility can be modelled on a more granular scale. Several different methods for modelling human mobility exist: the four step model (FSM), tour-based models, activity-based models and recently also machine learning models. [14, 38, 41, 44] These models and simulations can be used to construct more detailed origin-destination matrices. To understand the different approaches and their benefits and drawbacks, an overview is given of each and linked to simulation practices.

One of the older models is the FSM. The model describes what is needed for transport system analysis. The FSM is an example of trip based models. Each trip is described as an independent unit of analysis, focusing on the movement from an origin to a destination. The FSM consists of four sequential steps: trip generation, trip distribution, mode choice, and route assignment. [44]

1. Trip Generation: In this step, measures of trip frequency are developed. Trips are represented by their end points and attractions.
2. Trip Distribution: Next the trips are distributed to match the attractions of locations and the real travel impedance.
3. Mode Choice: Trip tables are adapted to represent real mode distributions.
4. Route Assignment: Modal trips are assigned to the real modal networks.

While the FSM has been widely used in transportation planning, it has limitations. It assumes trips

are independent and does not account for trip chaining or the interdependence of trips. Additionally, it focuses on aggregate data, because granular data is not available, which may overlook individual travel behaviors. Despite these drawbacks, the FSM remains a foundational model in transportation analysis and serves as a basis for more advanced models like activity-based approaches. [44] Algiers et al. [3] describe a trip-based system implemented for the city of Stockholm, Sweden. They model behavioral choices using the logit frameworks augmented with additional constraints.

Rasouli and Timmermans [56] combined four commonly cited limitations of a trip-based model. These limitations include insufficient spatial and temporal resolution, lack of behavioral foundation, lack of integration in model estimations and interdependency of trip generation, destination choice, mode choice and route choice. These limitations led to the development of new models and FSM extensions to resolve all or some of these limitations.

According to Krizek [38], trip models are limited in capturing real human behavior because people often combine different trips into a chain of trips. This corroborates the interdependency between trips. These tours start and end at the same location with one or more intermediate stops where a person can spend a limited amount of time. This start and end point is usually the residence of the person making the trip. The average length of these tours is limited to a few stops. This is confirmed through surveys to correspond with the correct length of real tours. These tours are described by the activities a person does within the tour. [38, 45]

Krizek [38] found that the duration and length of tours are influenced by the accessibility within a neighborhood. Their study revealed that tours in more accessible neighborhoods tend to have fewer stops but occur more frequently. This suggests that improved accessibility can lead to shorter, more frequent trips, potentially reducing overall travel time. They also indicate that this does not imply tours are only within their neighborhood. They found lengthier tours are also still made.

Research has been done into tour recommendations. These recommendations can serve as an input to simulation or as a recommendation for users. Lim et al. [39] propose a tour-based recommender for tourists taking into account user preferences and average visiting times constrained by the duration, the start location and the end location. They rely on the assumption that people spend more time at locations they are interested in. This assumption can be used within simulations as well for certain point of interest categories. However, when not on vacation, people also spend a significant amount of time at work and stores, not of their own volition but because of necessity.

Tour-based models are a subset of activity based models, since they are completely described by the activities within a tour. Activity based models offer a more comprehensive framework by considering the full range of activities individuals engage in throughout the day, rather than focusing solely on trips or tours. These models aim to capture the interdependencies between activities, their sequencing, and the constraints imposed by time and space. [7, 14, 43, 53]

Activity-based models (ABMs) are grounded in the idea that travel is derived from the need to participate in activities. Unlike trip-based models, ABMs consider the entire activity schedule of an individual, including the timing, location, and duration of activities. This allows for a more realistic representation of travel behavior and its relationship with land use and urban form. [14, 46]

Bowman [14] describes an operational activity-based model system that integrates activity scheduling with travel demand forecasting. They highlight the advantages of ABMs in capturing the complexity of human

behavior, including the influence of sociodemographic factors, land use patterns, and transportation policies. Moeckel et al. [46] warn that activity models have not yet been able to overcome all limitations of trip-based models. They propose a system based on some mandatory trips, usually work or education, and some discretionary trips that a user can choose within a day to improve the accuracy of the determined activity schedules. This better reflects the real world, where people have mandatory commitments every day and plan the rest of their day around these.

Activity data is collected through surveys and panels. Participants keep track of a travel diary which indicates where and when they perform an activity. This travel diary information can be used as an estimation of the demand of schedules and locations. Schedules comprise the entire activity chain of a person within a day, locations are the locations where activities are performed. [7] A set of possible schedules is used to determine types, after which each individual in the simulation gets assigned a sample schedule with locations so that the resulting traffic patterns correspond to real patterns. [7, 29, 43, 55]

Adding social networks allows the modelling of interaction between individuals in the simulation. Shared trips can be modelled more precisely. [7, 24] Galland et al. [24] use social networks to predict carpooling. They combine the social network with a trip based model. When two individuals close in the social network have a similar path, they get a probability to carpool to their respective destinations.

While ABMs provide a good methodology for simulations, they are limited by the amount of data collected. Bellemans et al. [7] indicate they used data from GPS systems, questionnaires for social networks, questionnaires for preferences and a panel for activity and event data. Great care has to be taken to avoid collection bias when using all these different data sources. ABMs are also limited by estimation of activities and using zonal attractiveness instead of complete activity schedules for each individual. Collecting these activity schedules for each individual is unfeasible but within certain groups these activity schedules can be constructed more detailed. For example, students have a rigid schedule for most of their day, where the location is also known. This allows hybrid approaches where part of the schedule is fixed and the rest can be approximated using existing methods. Moeckel et al. [46] argue that the activity-based model must perform at least as well as a trip-based model in validation to justify the added computational complexity.

Machine learning approaches to human mobility have been on the rise in recent years. It is seen as the next step following simulations to predict mobility more accurately. There are a number of models that have the opportunity to capture spatiotemporal data. These models range from time-series models like ARIMA to neural networks like recurrent neural networks (RNNs). [81] Luca et al. [41] provide an overview of deep learning approaches, Xie et al. [81] go broader and describe machine learning models.

Asad et al. [5] utilize machine learning in determining traffic flows between train stations. They collect data from RFID sensors at the entrances and exits of stations. Similar methods can be used on the urban scale, but with different input data. Xie et al. [81] list some open datasets for urban flow prediction. They also denote that data sparsity is a big problem. For machine learning specifically, significant amounts of data is required. This makes it less suitable within a large amount of cities where this data is not yet collected. [41]

The main benefit from machine learning arises when urban mobility flows can be determined from descriptive data. This removes the need for traffic flow data when predicting human mobility. This data would still be needed during training. [41] There are still some unanswered questions around machine learning techniques for urban mobility. It is unknown how much impact the aggregation level has on the

performance of the model. Tunability to other urban areas would be beneficial to gain the most benefits from machine learning. It has not yet been proven that this is possible and that one technique would work well within a different context. [41]

2.4.1 Agent-based modelling

Human mobility estimation and simulation can happen on different scales. Part of the mobility techniques focus on determining mobility flow from frequencies, attractiveness and average time spend. [39, 79] These techniques are also used in origin-destination matrices creation methodologies like the population-level models described in section 2.3. Machine learning techniques are also implemented on a less granular level. Other models try to understand mobility on a more granular level by representing individuals separately. This same distinction can be made for simulations. When simulating individuals, this is often referred to as Agent-based modeling. [12]

Agent-based modeling is a powerful approach for simulating human mobility at a granular level. It focuses on the behavior of individual agents, which can represent people, vehicles, or other entities, and their interactions within an environment. Each agent operates based on a set of rules, which can include decision-making processes, movement patterns, and interactions with other agents or the environment. This bottom-up approach allows for the emergence of complex system-level behaviors from simple individual-level rules. [12]

Agent-based models are particularly well-suited for modeling human mobility because they can incorporate heterogeneity in agent characteristics, such as age, income, or preferences, as well as spatial and temporal dynamics. For example, agents can be assigned specific schedules, modes of transportation, and destinations, enabling the simulation of realistic travel patterns. [42] One of the key advantages of agent-based models is their flexibility. They can be used to model a wide range of scenarios, from daily commuting patterns to the spread of infectious diseases through human movement. For this, agent-based models integrate various data sources, such as census data, GPS traces, and social network information, to enhance the realism and accuracy of the simulations. [27]

However, agent-based models also have limitations. They are computationally intensive, especially when simulating large populations or complex environments. The accuracy of the model depends heavily on the quality and granularity of the input data, as well as the assumptions made in defining agent behaviors. Calibration and validation of agent-based models can be challenging, requiring extensive data and iterative testing. [20] Despite these challenges, agent-based models remain a valuable tool for understanding and predicting human mobility. They provide insights into the interactions between individual behaviors and system-level outcomes, making them a critical component of modern urban planning, transportation management, and public policy development.

Several implementations of agent-based models already exist. Within simulation, they are often combined with activity-based models to get the most accurate representation of the real world. [7, 14, 24, 29, 43, 46, 53, 55] Agent-based models are not limited to simulation models but can also be used when collecting data. Tornero, Martínez, and Castelló [73] model agents to collect origin-destination matrices from intelligent road networks. These road networks contain infrastructure to vehicle (I2V) and vehicle to infrastructure (V2I) communication to track the agents within the system.

2.4.2 Utility model

Agents within simulations are modelled to have independent choices about the activities, travel modes, travel paths supported by their preferences. Utility models are mathematical frameworks used to represent the decision-making processes of agents in simulations. These models assume that agents make choices to maximize their utility, which is a measure of satisfaction or benefit derived from a particular option. Utility models are widely used in transportation and mobility studies to simulate choices related to activities, travel modes, destinations, and routes. [8]

The most common utility model is the discrete choice model, which evaluates the utility of each available option and assigns probabilities to each choice based on their relative utilities. The utility U_i of an option i is typically expressed as:

$$U_i = V_i + \epsilon_i$$

where V_i is the deterministic component of the utility, often modeled as a linear combination of attributes (e.g., travel time, cost, comfort), and ϵ_i is the stochastic component, capturing unobserved factors and randomness in decision-making. [74]

The multinomial logit (MNL) model is a widely used discrete choice model. It assumes that the stochastic components ϵ_i are independently and identically distributed following a Gumbel distribution. The probability P_i of choosing option i is given by the softmax:

$$P_i = \frac{\exp(V_i)}{\sum_{j \in C} \exp(V_j)}$$

where C is the set of all available choices. [8]

Utility models can be extended to account for more complex decision-making processes. For example, nested logit models allow for hierarchical choice structures, where decisions are made in stages (e.g., first choosing a travel mode, then a route). Mixed logit models relax the independence assumption of the MNL model, allowing for correlations between choices and heterogeneity in agent preferences. [74]

In agent-based simulations, utility models are used to assign preferences and decision-making rules to individual agents. For instance, an agent might choose a travel mode based on the utility derived from attributes like travel time, cost, and environmental impact. By incorporating utility models, simulations can capture realistic decision-making behaviors and provide insights into how changes in policies or infrastructure might influence mobility patterns. [27]

Mazzulla [43] uses logit structures with utility functions to model student choices for the type of tour made, the activities and the transport mode choice. Their research is limited because it only takes place on one campus and does not incorporate mandatory activities. Raney and Nagel [55] also use utility functions where choices are made according to activity plans. Bowman [14] shows a tree based choice model using utility functions representing real decision tree structures. The utility model has become standard in choice models. [45, 46]

Utility functions contain parameters that have to be determined using real data. This is the main downside of utility functions. Miller, Roorda, and Carrasco [45] explain a log-likelihood maximization technique for utility function parameter estimation. Another downside is the possible simplification of factors contributing to the decision. Only a limited number of factors is used within a utility function. They have to be carefully selected to give a good approximation of the real world without overly complicating

the decision process.

Chapter 3

Student Housing Location Estimation

When analyzing student movement, student housing locations function as the origin and destination of most of the students' travels. This highlights the importance of understanding where students reside in their student city. Student housing is influenced by various factors, including university location, rental prices, access to transportation, access to student amenities and more. This complexity makes it difficult to generate completely accurate predictions. As a result, universities and city councils spend significant efforts in determining the locations of student housing through registration systems or surveys. However, acquiring and using this data can present legal and technical challenges. Determining the location of student housing within a city is thus both an intriguing and challenging problem. The difficulties provide an opportunity for a method estimating student housing location from aggregated data. If a sufficiently accurate estimation could be determined for student housing locations from survey data, the estimation could serve as valuable input for student-related studies. Especially if the estimation method allows studies on different aggregation levels.

This chapter will first address the challenges associated to current methods determining student housing locations. It will then define what constitutes as a reasonable estimation. Next, a detailed description of the proposed estimation method will be provided. Finally, the method will be evaluated and applied to the city of Ghent.

3.1 Current Difficulties

To fully comprehend the difficulties associated with collecting student housing information, a basic overview of the landscape of student accommodation is required. Student housing varies significantly between countries. Even within a single country, there are additional differences from city to city, increasing complexity. Student accommodation is influenced by the local housing market, policy decisions, university decisions, the city environment and other factors. These factors make it challenging to analyze student housing on a broader scale than the city level.

Student accommodation can be classified into distinct categories. The primary differentiation is between on-campus and off-campus housing. On-campus student housing is most often organized by the university, on campus and in proximity to the academic buildings. Off-campus housing entails all the other types of student accommodation within the city. This type of housing is not necessarily situated near the

university campus. Off-campus housing is present in multiple different forms. The first category comprises large-scale off-campus housing. This type of housing is typically developed and organized by private cooperations that see market opportunities. A single large-scale student housing project can host tens to hundreds of students. A second category are students that rent a home or apartment on the private market as student accommodation. A third category includes traditional student homes. They are characterized by students renting one room within a student building. The complexity and size of the private student housing sector makes it difficult to collect information about the location of student housing. [63]

Thorough information collection about all student housing within a city requires registration systems. Each student is responsible for registering their student housing within this system. This approach has been successfully applied in some cities, but is not suitable within every context. Students are not always compelled to register their student housing. When they are, enforcement measures are needed to guarantee students will enter the information. Introducing these measures comes with legal difficulties. The results are information systems that are often incomplete. Additionally, privacy requirements in the registration system may sometimes restrict the use of the collected data. [33]

To circumvent the registration of each student, cities cooperate with universities and larger off-campus private housing initiatives. This allows information collection of larger-scale student housing. Even though limited, this data contains information about the location, number of rooms and sometimes the available amenities. The downside of this approach is the lack of information on the other types of off-campus housing. Another alternative method to a registration system, is collecting data through a survey. [63] With enough student participation, it is possible to collect characteristics of the student housing situation within a city. The advantage is that this collection can be done without a risk of violating privacy. A disadvantage is the requirement for enough participation making it almost impossible to capture all required student housing information this way. Since not every student fills in this survey, extrapolation is needed to the entire student population. Additionally, the data is fixed to the aggregation level of the survey. It is not straightforward to use this data within studies on different aggregation levels.

3.2 Estimation Quality

A quality student housing estimation can be used as input for student-related studies in the absence of other detailed data sources. To analyze student mobility, one of the goals of this thesis, the type of student housing is mostly irrelevant while the location is essential. However, the information available about a certain type of student accommodation can improve the quality of the location estimation. Quantifying what constitutes a quality student housing location estimation is an important step in assuring the validity of the methodology. This quantification will then be used to evaluate the proposed method.

As indicated before, student housing location depends on several factors. Despite this, student housing is mostly concentrated around campuses. This is a consequence of students defining proximity to campus as one of the most important factors in deciding on where to live during their studies. [72, 84] However, literature also indicates that students are sometimes willing to commute longer to campus if this comes with a reduction in rent prices. [18] Combining this with information that there are students living in their parental homes within the city, which are more distributed around the city, it becomes clear that a quality location estimation requires a large concentration of student housing around the campuses, but also incorporates students living further away. The methodology should also capture commuter students living outside the city.

Research also indicates studentification, which results in clustering of students within certain neighbor-

hoods. Studentification is a form of gentrification, where students move towards a new neighborhood. Student amenities will then be organized in the neighborhood, which increases the attraction of the neighborhood to students even more. Studentification creates student neighborhoods that become characterized by a high concentration of student amenities. [31] The location estimation should reflect this clustering and indicate student neighborhoods where expected. These student neighborhoods often lie close to both the campus and the city center.

In certain countries, a large part of the student population living in student housing return home during weekends, exam periods and holidays. In Belgium, students prefer sustainable transport options, resulting in mostly train travel for longer distances. As effect, this group chooses their student housing location in function of the distance to a train station with a connection to their hometown. [66]

To quantify the quality of a student housing location estimation, several metrics can be considered. A first metric considered is the distance to the nearest campus. This should indicate a significant number of students living within a short distance of the campus and also indicate a more even distribution for the remaining students. Another metric to consider is the density of student housing, indicated by the number of students per square meter. Calculating this for neighborhoods makes it easy to check against known student neighborhoods.

Finally, the location estimation should be validated against known data points from survey data where possible. This will make sure that the results from the metrics have merit in a real world scenario. This can also indicate what values of the metrics are considered to be good. A real world validation is important in assessing the accuracy and reliability of the estimation, ensuring the location estimation reflects the student housing situation within a city.

3.3 Estimation Method

The literature indicates that student housing locations are related to the city center, train stations and campus locations. These relations are exploited when estimating the student housing locations. The methodology aims to generate a spatial distribution of student housing that can be aggregated on different levels. The proposed methodology for student housing location estimation consists of three main steps. The first step is the collection of relevant data. Next, an intensity distribution in function of the distance to campus locations, the city center and train stations is determined from survey data. Step three applies this intensity function to generate a spatial distribution of student housing.

3.3.1 Data Collection

The data collection process starts by obtaining a map of the city, more specifically a graph that represents the city's road network. This graph can be downloaded from OpenStreetMap through a Python package called osmnx. [11] Figure 3.1 visualizes the road network of Ghent, Belgium. While the location estimation could be done on another representation of the city, the road network serves a crucial role in the proposed method. Students will be assigned to nodes within the road network. A benefit from this approach is that a denser road network creates a more dense assignment of students. Denser road networks have more connected roads, which often correlate with denser housing. While the student assignment is applied on the road network, this spatial representation has no inherent meaning. This allows a floating point number of students to be assigned to a node. The meaning only emerges when nodes are aggregated according to districts, after which the number of students can be rounded to the nearest integer. It is important to refrain from making conclusions on the node specific student locations.

The next piece of information needed are the campus coordinates. More often than not these locations are not immediately available, and they need to be collected manually. This is not an issue, because within a city there are a limited number of campuses. These campuses are assigned to the closest node within the road network. Figure 3.1 shows the campus nodes of Ghent University that lie within the city boundaries. Likewise, the coordinates of the city center have to be determined. A first approach is to take the centroid of the graph. The centroid will not always coincide with the real city center because bounds of cities can have irregular shapes where the center is not located close to the centroid, e.g. the shape of Ghent makes the centroid lie geographically more north than the real city center. To increase accuracy it is recommended to manually provide the coordinates of the real city center.

A significant percentage of students does not have the ability to rent a room during their studies and do not live within the student city itself. These students commute to classes from their home in a different city or town. In Belgium, most of these students commute by train. There are other modes of transport that can be used as well. Commuter students contribute to the student movements within a city once they arrived there. To account for commuter students, and since most of them come by train, they are assigned to the nodes corresponding to the train stations. The major train stations in Ghent can be seen as the blue nodes in figure 3.1. Collecting information about the train stations is thus crucial for the evaluation of student mobility within a city. To make the student assignment to the train stations as realistic as possible, additional information is required about the percentage of commuter students that use each train station. Given this information, a number of students can be accurately assigned to the station nodes. Additionally, the station information influences the student housing locations, enhancing the importance of this data.

Large student housing projects play a significant role in the overall student housing landscape. Incorporating these projects in the estimation method is crucial for improving the accuracy of the student housing location estimation. These projects have a high density of students, which can significantly impact the overall estimation and spatial distribution of students within a city. To include large student housing projects, the locations of these projects need to be identified, together with the amount of students hosted by each of these locations. Incorporating this information makes the method better reflect the actual student housing situation within the city.

3.3.2 Student Assignment

The collected data can be leveraged in the student housing location assignment. The methodology uses three one-dimensional functions that describe the student intensity of a node in function of the distance to respectively the city center, the campus and the train station. While these student intensity functions are not completely independent, they can be approximated as independent providing the opportunity for a simplified combined intensity function. These intensity functions incorporate the conclusions made based on the literature that the student housing locations are dependent on the proximity to the city center, campuses and stations. The three intensity functions are combined into one three-dimensional intensity function. Next, a number of students for each node is calculated using this intensity function. Figure 3.2 shows the entire process for the student assignment. Subsequently, a more detailed description of each step is given.

The three one-dimensional student intensity functions are able to capture dependencies on the distances to the city center, the campuses and the train stations. They are an approximation of a more complex problem into smaller, less complex, subproblems. As this is an approximation, there is a difference between



Figure 3.1: Road network graph of Ghent, Belgium collected from OpenStreetMap. {Red node: campus of Ghent University, Blue node: train station}

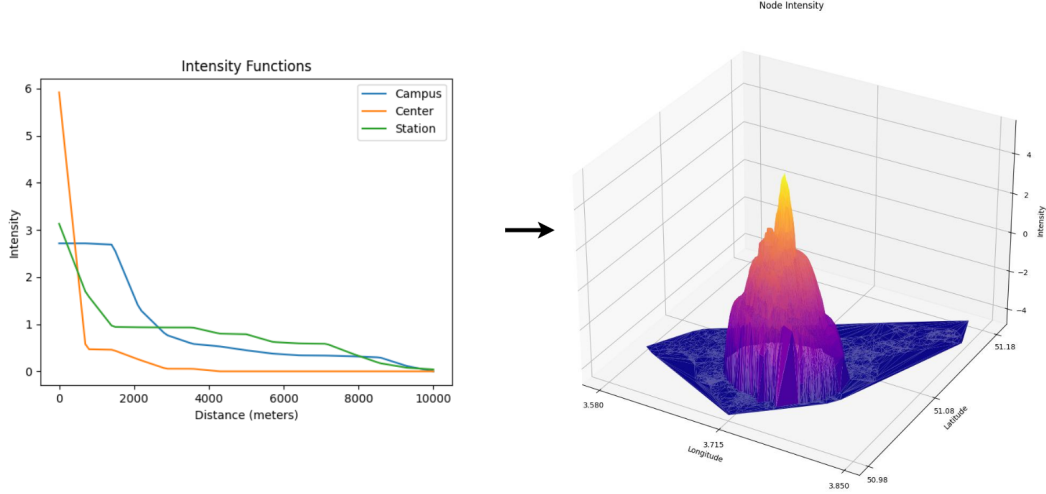


Figure 3.2: Complete student assignment method: From independent intensity functions to a combined function assigned to all nodes in the road network.

the combined intensity function using the independent functions and the real intensity distribution. A variety of intensity functions can be defined with different constraints. One interesting constraint is non-increasing functions, which capture that students prefer to live rather closer than further. The independent intensity functions can be manually selected, determined from real student housing data or selected using an optimization approach based on survey data to find a reasonable solution. A manual choice requires extensive domain knowledge to make an educated decision. Where this knowledge is missing, the data-driven approach and optimization approach are good alternatives. The data-driven approach requires enough detailed student housing location data where the optimization approach does not have this constraint because it only requires student housing survey data as input. The optimization method is further described in section 3.3.3.

In both the data-driven and optimization approaches, a flexible representation of the set of intensity functions proves beneficial. A first flexible set are polynomials of a varying degree. A higher degree polynomial will be able to capture more complex problems, but has a risk to overfit to the provided data points. Polynomials, by default, can not capture the non-increasing constraint. It is possible, using specific methods, to achieve a non-increasing polynomial within a domain. Another method is using regression on the data points to determine the functions. The use of an isotonic regression allows easy incorporation of the non-increasing constraint. Both of these sets of functions can be determined from a list of known function points. This property is exploited in the optimization approach to determine the intensity functions.

The intensity functions are defined in function of a distance. A choice can be made about which distance is used, where the most straightforward distance is the minimum distance. However, for student housing it might be interesting to allow a random selection from a set of minimum distances to campuses. This might have as benefit a more even distribution of students that can take into account students not always going to the closest campus. Both options for the distances will be evaluated. Additionally, a high density of clustered campuses increases the amount of students around the cluster even more. More campuses mean more students have class within that region, which in turn results in more students wanting to live close to this cluster of campuses. This is also a factor in the occurrence of student neighborhoods. To accommodate this clustering the intensity function is multiplied by a campus density number. This

campus density is the amount of campuses within a certain radius of the node. A value of one is added to the campus density to ensure that nodes without a campus in the defined radius can still be assigned students. A comparison will be made against the method without this additional factor. A more detailed choice can be made for this factor if the amount of students with classes on a campus is taken into account. This goes out of scope for this thesis.

Dependency intensity function	Symbol intensity function
Campus	$I_{campus}(d_{campus})$
Center	$I_{center}(d_{center})$
Station	$I_{station}(d_{station})$

The next step is to combine the independent intensity functions into one three-dimensional intensity function. This is achieved by multiplying the separate intensity functions and the campus density:

$$I(d_{campus}, d_{center}, d_{station}) = I_{campus}(d_{campus}) \cdot I_{center}(d_{center}) \cdot I_{station}(d_{station}) \cdot (\rho_{campus} + 1) \quad (3.1)$$

where ρ_{campus} , the campus density, is defined as in equation (3.2); N is the number of campuses, λ the longitude, ϕ the latitude, R a defined radius and d the great circle distance.

$$\rho_{campus,i} = \sum_{j=0}^N (d(\phi_j, \lambda_j, \phi_i, \lambda_i) < R) \quad \forall \text{ nodes } i \quad (3.2)$$

Finally, a number of students is assigned to each node by calculating the intensity for each node in the network. The intensity function is scale independent, which means rescaling is needed to assign the correct amount of students. This rescaling is achieved by normalizing the assigned intensities and then multiplying by the total number of students in student housing. This total number of students in student housing can be defined as a percentage of the total student population. A choice can be made about large student housing projects. They can be treated as all other student housing and thus captured within the intensity functions, or they can be treated separately and not include them in the intensity functions. The second approach is more logical since the large student accommodations are outliers on certain nodes compared to the rest of the city. Treating the large student housing projects separately increases accuracy of the intensity functions. First the amount of students within large student housing is subtracted from the total number of students in student housing. The remaining students are assigned using the described method, after which the students of the large student housing projects are again added to their corresponding nodes. The intensity values are rescaled for each node:

$$S_i = \frac{I_i}{\sum_{j=1}^N I_j} \times N_{students} \times p_{housing} \quad (3.3)$$

where N is the total number of nodes, S_i the amount of students at node i , $N_{students}$ the total number of students and $p_{housing}$ the percentage of students living in student housing.

This rescaling ensures that the sum of the assigned students equals the total number of students in student housing:

$$\sum_{i=1}^N S_i = N_{students} \times p_{housing} \quad (3.4)$$

The intensity, as defined, for each node can be a float number of students. Even after rescaling the students assigned to a node can be floating point numbers. Of course in the real world, part of a student does not exist. This is solved when aggregating nodes according to a higher level of abstraction e.g. districts. In this aggregation, the values are rounded to the nearest integer. Analysis on student mobility will rarely be done on each node in a road network, because that is often too specific and would require information about where each student exactly lives. The information might also be part of another system that works on a higher abstraction level, not requiring the most detailed information. Cities are also interested in abstracted movements. This method can thus provide an estimation of the student housing locations. It is important to note again, that only after aggregation, meaning is given to the amount of assigned students.

Lastly, the commuter students are assigned to the train stations:

$$S_i = S_i + N_{students} \times (1 - p_{housing}) \times p_{station,i} \quad \forall \text{ nodes } i \in X \quad (3.5)$$

where $p_{station,i}$ is the percentage of commuter students assigned to that station and X is the set of station nodes.

3.3.3 Intensity Distribution Determination

The accuracy of the proposed methodology is completely dependent on the choice of intensity functions. An automated approach is needed to explore possible functions because of the large amount of options when data is lacking and not enough knowledge is available. There are an unlimited number of continuous intensity functions that can be defined. This search space is too large to exhaustively evaluate, therefore the functions are represented by a limited number of intensity values for defined function points. On these intensity values, a continuous intensity function is fitted taking into account the constraints of the function. Two types of function fitting are evaluated, monotonically decreasing regression fitting and polynomial fitting.

The optimization works iteratively to improve the determined functions in each step towards a better solution. A better solution to the problem lies closer to the real survey data collected about student housing. Several metrics can be chosen that define the distance to the survey data. Common metrics include the mean absolute error (MAE) and the mean squared error (MSE). The MSE accounts for a preference toward smaller errors by penalizing larger deviations more heavily. The MAE treats all errors equally and is less sensitive to outliers. In the equations for MAE and MSE, $\mathbf{P}_i$ indicates the percentage of students living within a district according to the survey data and M is the number of districts.

$$MAE = \sum_{i=0}^M |\mathbf{S}_i - \mathbf{P}_i \cdot N_{students} \cdot p_{housing}| \quad (3.6)$$

$$MSE = \sum_{i=0}^M (\mathbf{S}_i - \mathbf{P}_i \cdot N_{students} \cdot p_{housing})^2 \quad (3.7)$$

Survey data is not available on node level. So to evaluate, an aggregation is needed onto the abstraction level of the survey data. This summarized value can be used in the defined metric. The higher abstraction level can be defined as a set of polygons, districts, over the city. To determine the students within a district a summation is performed on the students of all nodes within that district:

$$\mathbf{S}_i = \sum_{j \in \text{district } i} S_j \quad (3.8)$$

Two optimization approaches are defined, an alternating pass algorithm and a genetic algorithm. Both approaches can be adapted to accommodate varying numbers of intensity functions. For the student housing, the three defined intensity functions are determined from survey data by applying the optimization methodologies. The discrete representation of the intensity functions is paramount in both optimization approaches. This representation is essentially an array for each intensity function, containing values of the function at certain distances. For the student housing location this comes down to three independent arrays.

$$\mathbf{I}_{campus} = [i_{campus,d_1}, i_{campus,d_2}, \dots, i_{campus,d_{max}}] \quad (3.9)$$

$$\mathbf{I}_{center} = [i_{center,d_1}, i_{center,d_2}, \dots, i_{center,d_{max}}] \quad (3.10)$$

$$\mathbf{I}_{station} = [i_{station,d_1}, i_{station,d_2}, \dots, i_{station,d_{max}}] \quad (3.11)$$

These intensity arrays need to be initialized. The initialization value will have an impact on the resulting intensity functions and the computational speed. Starting from a solution closer to the optimum, will significantly speed up finding the optimum intensity array. Furthermore, a different optimum might be found because different intensity arrays are explored during the approach. The rescaling of the intensities to students, makes that the intensity functions can be independent of scale. As an effect, multiple solutions can exist that are fundamentally equivalent. This increases the chance in finding the same optimal solution regardless of the initialization choice. In each iteration of both approaches, several intensity arrays are evaluated after which a choice is made that moves towards better intensity functions and thus a better solution.

The student housing location assignment requires a continuous function because distances can occur that are not part of the array. Two options of conversion to a continuous function are supported: isotonic regression and polynomial fitting. A polynomial of a certain degree, k , is fitted to the values of $\mathbf{I}_{campus}$, $\mathbf{I}_{center}$ and $\mathbf{I}_{station}$ to compute I_{campus} , I_{center} and $I_{station}$ respectively. The degree of the polynomial is limited because a higher degree increases the risk of overfitting; moreover, an excessively high degree can lead to a poorly conditioned polynomial fit. A good selection has to be made for the polynomial degree. The coefficients, a_k , b_k and c_k in the following equations are determined:

$$I_{campus}(d_{campus}) = \sum_{j=0}^k a_j d_{campus}^j \quad (3.12)$$

$$I_{center}(d_{center}) = \sum_{j=0}^k b_j d_{center}^j \quad (3.13)$$

$$I_{center}(d_{station}) = \sum_{j=0}^k b_j d_{station}^j \quad (3.14)$$

Isotonic regression fits a free-form line to the array, with a condition that it is non-increasing using a weighted least squares fit. Since this isotonic regression is non-increasing this is a good choice to enforce a non-increasing constraint on the intensity functions. Polynomials do not enforce this constraint, but can give an overall decreasing intensity function when the intensity array is decreasing that might only locally increase. Polynomials can also be used when there is no non-increasing condition on the intensity function.

Alternating pass algorithm

The first optimization approach is an alternating pass algorithm. In this method, each iteration consists of subsequently optimizing one intensity array while keeping the other arrays constant until all arrays are optimized for this iteration. The algorithm, specified for the student location, consists of an initialization step and an iterative part. In each iteration, each intensity array is optimized consecutively by evaluating possibilities where the elements can change by a step size η or remain unchanged. Iterations run until no improvement is made anymore within one iteration.

1. Initialize $\mathbf{I}_{campus}$, $\mathbf{I}_{center}$, and $\mathbf{I}_{station}$.
2. Repeat until no further improvement in the metric:
 - (a) For each of $\mathbf{I}_{campus}$, $\mathbf{I}_{center}$, and $\mathbf{I}_{station}$:
 - i. Generate all variations with elements $\pm\eta$ or unchanged.
 - ii. Evaluate each by fitting a continuous function and using student assignment and the chosen metric.
 - iii. Select the best-scoring array.

For each array, the first step is to determine the new possibilities for this iteration. These are determined starting from the current best array. The new possibilities are all arrays where each element of the current best array independently of the others can increase by η , decrease by η or remain unchanged. A first constraint that is always present is that the intensity array is always positive or zero. No negative students exist. As mentioned before, a non-increasing constraint can be added to the intensity functions. This can be taken into account in determining which arrays to evaluate. The non-negative condition can be captured by:

$$i_{d_k} > 0 \quad \forall k \quad (3.15)$$

A non-increasing array is defined by:

$$i_{d_k} \geq i_{d_{k+1}} \quad \forall k \quad (3.16)$$

where i_{d_k} represents the intensity at distance d_k .

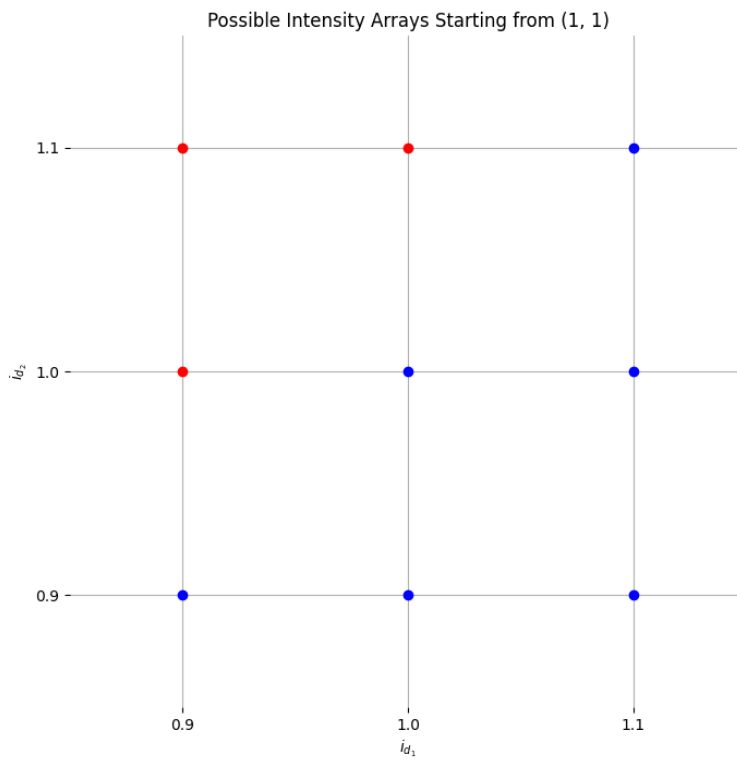


Figure 3.3: Intensity array options within an iteration of the alternating pass algorithm for a currently best array $[1, 1]$ with $\eta = 0.1$. {blue: non-increasing, red:increasing}

An intensity array of length 2 can be represented within a 2D space. Figure 3.3 shows all possible arrays starting from an initial one that was $[1, 1]$. This visualization gives more insight into all possible intensity array options. It can be extended to longer intensity arrays. The amount of options of intensity arrays grows exponentially with the length of the array and is $O(3^n)$. From this it can be deduced that this approach is only viable for arrays of limited length. Longer arrays, will result in computational difficulties evaluating all possibilities. The non-increasing constraint partially limits the amount of possibilities that need to be evaluated, but even in that case increasing length will eventually result in computational issues.

The intensity functions are now determined for all possible arrays. With these intensity functions, the student housing assignment can be applied as explained in section 3.3.2. After which aggregation according to districts and evaluation using the MSE or MAE can occur. The best scoring intensity array is selected and becomes the new optimal intensity array.

The alternating pass algorithm will always find the same result ensuring consistency, but has as major downside high computational requirements. While the solution moves towards a better solution, there is no guarantee this solution is the same when the intensity arrays are optimized in a different order due to possible dependencies between arrays when they are not completely independent.

Genetic algorithm

The second proposed optimization technique is a genetic heuristic. A genetic algorithm is often used in optimization problems because it offers a good balance between performance and finding optimal or near-optimal solutions. It is especially interesting when a better solution might contain combined characteristics of two other solutions. A genetic algorithm is based on principles of natural selection where solutions evolve over time with fitter solutions surviving generations. It starts from an initial population of possible solutions, each represented by intensity arrays, the discrete representation of the intensity functions. Each individual is evaluated by determining the intensity functions, using student housing location assignment and a chosen metric to determine a fitness score based on the survey data. The algorithm iterates through a process of evolutionary operators including selection, crossover and mutation to evolve the population towards better/fitter solutions during each generation.

Figure 3.4 shows the evolution of a non-increasing intensity array in the genetic approach over different generations. The solution moves towards fitter individuals, which results in the movement to the higher values for these first and second array intensities. A spread out is observed that shows the variety of individuals in the population. This spread out also indicates the scale independence of the intensity functions due to the spread out of the fittest individuals. The 10% fittest individuals are indicated, to more properly see the evaluation.

The steps in the genetic heuristic are:

1. Initialize the population with individuals represented by $\mathbf{I}_{campus}$, $\mathbf{I}_{center}$ and $\mathbf{I}_{station}$.
2. Repeat for each generation:
 - (a) Evaluate the fitness of each individual in the population by fitting a continuous function and using student assignment and the chosen metric.
 - (b) Select the fittest individuals to carry over to the next population.

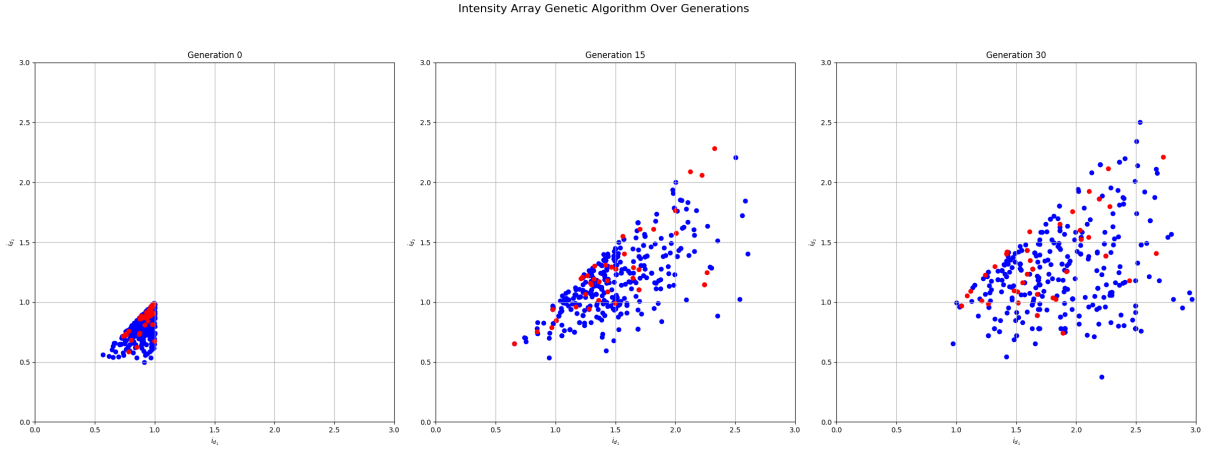


Figure 3.4: First two values in a length-10 non-increasing intensity array for the genetic algorithm over different generations. {Red: 10% fittest individuals}

(c) Repeat until the new population has reached the population size:

- i. Select two parents based on fitness.
- ii. Perform crossover (with a given rate) to produce offspring.
- iii. Mutate offspring (with a given rate).
- iv. Add offspring to the new population.

(d) Use the new population in the next generation.

Initializing the population is accomplished by generating random intensity arrays, $\mathbf{I}_{campus}$, $\mathbf{I}_{center}$ and $\mathbf{I}_{station}$, for each individual. A random array with values between 0 and 1 is created of the defined length. Due to the scale independence of the intensity functions, other ranges initialization values, instead of 0 to 1, are also possible. A greater range of intensity values is achieved through mutation when needed. When non-increasing intensity is a requirement, the randomized array is sorted in decreasing order. The size of the population, is a parameter that has to be chosen. It plays a role in how fast better solutions are found. A larger population will evaluate more arrays in a generation, but comes with a higher computational load. The randomization aspect of the initialization phase, creates variance in how fast a near-optimal or optimal solution can be found. If by chance, randomly the optimal solution is chosen, the algorithm could end quickly. If only solutions far away from the optimal are initialized, the algorithm will have to run for more generations to achieve the near optimal solution. An early cut off point can be determined to improve execution time. This was not employed in this thesis.

The fitness score is the chosen metric, MAE or MSE, which can be calculated by fitting a continuous function to the intensity arrays, use this function to assign student housing to nodes in the network, aggregate assigned students according districts and finally calculating the metric score between the result and the survey data. Both MAE and MSE have to be minimized, this results in individuals with lower metric scores to be the fittest individuals of the population. A percentage of the fittest individuals is selected to survive the generation. A percentage of solutions that survive that is too high, will reduce the capacity to evaluate alternative solutions. If the percentage is too small, it will reduce the capacity to keep the fittest solutions. Therefore, a balanced percentage of the population should be selected to carry over to the next population.

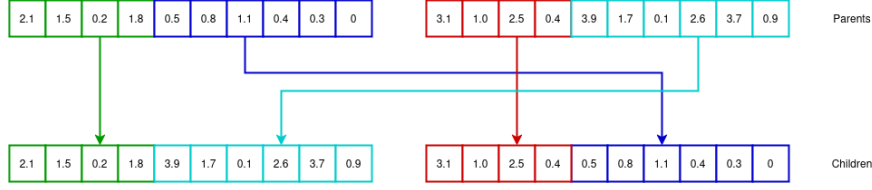


Figure 3.5: Crossover operation of the genetic heuristic on 2 parents of length 10 with crossover index 4.

The individuals selected for crossover are often referred to as the mating pool. To make sure fitter individuals are selected more, their probability of selection is higher. This probability is calculated by normalizing all fitness scores, computing the negative exponential distribution of these normalized fitness scores and finally normalizing the results of the exponential distribution again. Randomness is achieved through the probabilities, which allows less fit individuals to mate as well to keep variance in the set of solutions. However, this randomness is skewed towards fitter individuals, since fitter individuals are more likely to create even fitter offsprings.

When two parents are selected, and thus part of the mating pool, crossover can occur between them. The crossover rate indicates the probability of 2 parents to engage in crossover. If no crossover is applied to the parents, they are copied outright and become their own children. They have a chance to mutate, just like their children, in the next step. The crossover is based on the principle that the offspring inherits characteristics from both parents. For the intensity arrays, this is implemented as follows: for each intensity array (e.g., $\mathbf{I}_{campus}$, $\mathbf{I}_{center}$, $\mathbf{I}_{station}$), a random crossover point is selected. The first part of the array up to the crossover point is taken from one parent, and the remaining part is taken from the other parent. This creates two offspring arrays, one for each parent.

For example, let $\mathbf{I}_{campus}^{(p1)}$ and $\mathbf{I}_{campus}^{(p2)}$ represent the campus intensity arrays of parent 1 and parent 2, respectively. If the crossover point is at index k , the offspring arrays are defined as:

$$\mathbf{I}_{campus}^{(child1)} = [\mathbf{I}_{campus}^{(p1)}[0 : k], \mathbf{I}_{campus}^{(p2)}[k : n]]$$

$$\mathbf{I}_{campus}^{(child2)} = [\mathbf{I}_{campus}^{(p2)}[0 : k], \mathbf{I}_{campus}^{(p1)}[k : n]]$$

This process is repeated for the other intensity arrays ($\mathbf{I}_{center}$ and $\mathbf{I}_{station}$) to generate complete offspring solutions. Each crossover point is chosen independently for each array, to ensure more variation. The crossover ensures that the offspring inherit traits from both parents, promoting diversity in the population while preserving good characteristics. Figure 3.5 visualizes the crossover operation between 2 parents of length 10, with crossover point at index 4.

Each individual coming out of the crossover has a probability to mutate. This probability is called the mutation rate. The mutation operator introduces random changes to the offspring arrays to maintain diversity in the population and explore new solutions. If the individual mutates, a random value between -0.5 and 0.5 is added to each value in each of the arrays. The range of this random value should be in proportion to the initialization range. Figure 3.6 visually shows the mutation operation for an intensity array of length 10. If no mutation occurs, a copy is returned of the original array. The output arrays of the mutation step, mutated and not-mutated, are added to the new population.

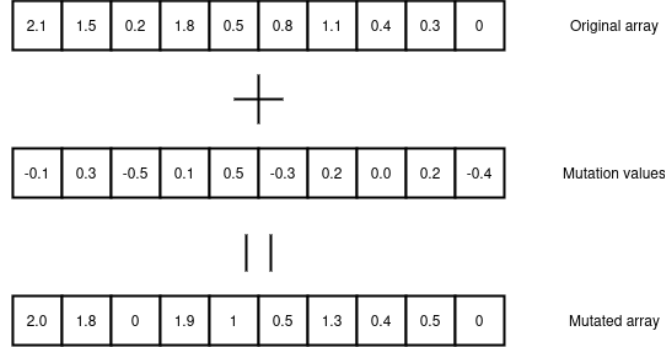


Figure 3.6: Mutation operation of the genetic heuristic on an array of length 10.

Parameter	options
Radius (m)	[250, 500, 750]
Large housing	[True, False]
Set size of minimum distances to campus	[1, 2, 3]

Table 3.1: Parameters student housing location assignment: options indicate the values evaluated.

When the intensity functions need to be non-increasing, the arrays are sorted in decreasing order after crossover and mutation. The crossover and mutation are repeated until a full new population is generated. This new population will be the population at the start of the next generation, when the old population dies off. The genetic algorithm provides a robust and flexible approach to finding the best intensity functions for the student housing location estimation, especially when the search space is large and complex.

3.3.4 Methodology Parameters

The methodologies of the student housing location assignment and the optimization approaches for selecting intensity functions, are openly described without making definitive choices. The goal was to keep the methodology flexible to be adapted to different scenarios and localization problems. This results in a set of parameters that can be chosen, and they will have an influence on the outcome or speed with which the solution is achieved. All parameters were discussed within the methodology, but are now briefly summarized again. The options in the tables are the ones evaluated to compare the different parameter choices, however several other options exist. The table gives an indication of the type of the parameter, from which all possibilities for the parameter can be determined. Table 3.1 shows the parameters that can be chosen within the student assignment itself. The radius indicates the radius in meters of the circle in which campuses are counted for the campus density. Large housing indicates if large housing is separately added or incorporated through the intensity function. The set size denotes from the list of distances to campuses, how many of the minimal distances can be chosen from to use as distance in the campus intensity function.

Table 3.2 contains all parameters within the intensity determination methodology. The fitting approach indicates if a polynomial fitting is used or an isotonic-regression fitting. The regression can only be used when the non-increasing parameter is True, because it requires a monotonically decreasing array. When a polynomial is chosen, its degree must also be determined. The condition was made to set the polynomial degree to be smaller than the length of the array, because otherwise the resulting polynomial has a high chance to be poorly conditioned and can not be trusted. The metric indicates the use of the mean absolute error or the mean squared error. The MSE will penalize large mistakes, sacrificing smaller

Parameter	Options
Fitting approach	[Polynomial, Regression]
Polynomial degree	[4, 8, 12]
Non-increasing	[True, False]
Metric	[MAE, MSE]
Array length	[10, 15, 20]
Campus density	[True, False]

Table 3.2: Parameters intensity functions determination: options indicate the values evaluated

Parameter	Options
Step size	[0.1]

Table 3.3: Hyperparameters alternating pass algorithm for intensity functions determination

ones. The MAE will explain how many students are wrongly assigned in the city, while treating all errors equally. The array length determines the length of the intensity arrays, which also has an impact on the fitted function and how much it can be controlled. Campus density indicates if the additional campus density factor is used within the intensity function or is ignored.

Both optimization approaches have additional hyperparameters that mostly do not have an impact on the found solution, but the speed with which a solution can be found. Table 3.3 shows the additional step size, η , parameter for the alternating pass algorithm. Table 3.4 indicates the intrinsic genetic algorithm parameters. These are quite general for the genetic heuristic and with their explanation in the previous section, no further explanation is needed here.

The search space of methodology parameters is extensive. To provide some insight in the possible effects of each, they are evaluated for the case study of Ghent, Belgium. The evaluated parameters are listed within the parameter tables. Even with these seemingly limited options, the combined total of options is still quite large. Due to limited computational power, efficiency constraints of Python packets and time limitations, a structured approach is used in the next section that avoids evaluating all possibilities, but will give insight on the impact of each parameter.

3.4 Ghent Case Study

The proposed student housing location estimation method is applied to the city of Ghent, Belgium. Ghent is a vibrant student city with a significant number of students attending Ghent University and other higher education institutions. While Ghent has a number of higher education institutions, the focus will lie on the university students for this case study. Ghent University has about 50,000 students in the 2024-2025 academic year. [76] In the city of Ghent, an estimate of 58 percent of university students live in student housing. [63] The method will thus assign 29,000 students, see equation (3.17), in student housing and will assume all other students are commuter students.

Parameter	Options
Population size	[400]
Generations	[100]
Mutation rate	[30%]
Crossover rate	[50%]
Selection fittest size	[10%]

Table 3.4: Hyperparameters genetic algorithm for intensity functions determination

$$N_{students} = 50000 \cdot 0.58 = 29000 \quad (3.17)$$

3.4.1 Data Collection

As mentioned in the methodology, the first step is to collect a graph of the city of Ghent. The resulting graph can be found in figure 3.1. Ghent University has campuses in the region of Ghent, but also in Kortrijk, Bruges and Songdo, South-Korea. For this analysis only the campuses in Ghent itself will be selected, since the focus lies on student mobility in Ghent. To detail this even more, only campuses that fall within the graph of the city will be used in the student housing assignment. The corresponding nodes are colored red in figure 3.1. For Ghent these campuses are:

Id	Campus	Latitude	Longitude
1	Aula	51.051232127160986	3.7246330730671504
2	Boekentoren	51.0464138949851	3.7247732669376443
3	Coupure	51.05285881796822	3.7088778441920063
4	Heymans	51.025602461940174	3.730260686763223
5	Dunant	51.04906879806054	3.703747457149687
6	Ledeganck	51.03658416355151	3.724341223538734
7	Mercator	51.04390640143318	3.712262746292744
8	Proeftuin	51.02450133891323	3.7401662115784156
9	Rommelaere	51.046316522642655	3.7197647923255235
10	Schoonmeersen	51.03318718337439	3.7034998087368387
11	Sterre	51.02678707542553	3.7162196994767625
12	Tweakerken	51.04304481213696	3.728521507140288
13	Ufo	51.04651669349979	3.7272516534456392
14	UZ	51.0244896685205	3.726113769566922
15	Ardoyen	51.01270127263122	3.7085425375719128

The next step is to determine the train stations within the city to which commuter students will be assigned. In total, there are 6 train stations in the greater Ghent region. However, 2 of these are large stations, with significant passenger numbers and a lot of connections, while the others are smaller stations with less frequent service. [64] Since commuter students come from a lot of different places and prefer a good connection to their student city, only the 2 larger stations are taken into consideration:

Id	Station	Latitude	Longitude	Estimated % of students
1	Sint-Pieters	51.03633517061647	3.710910841805177	85 %
2	Dampoort	51.05553982643968	3.7414885986806086	15 %

A choice is made to set Sint-Pieters as the station 85 percent of the commuter students use daily. No data could be found on the actual station commuter students use, thus a choice was made to test the methodology that is proportionate to the size and number of connections of each train station. The distance towards the campuses also influenced the percentage of students at a train station. Sint-Pieters lies closer to more campuses and has better public transport connections than Dampoort, which results in

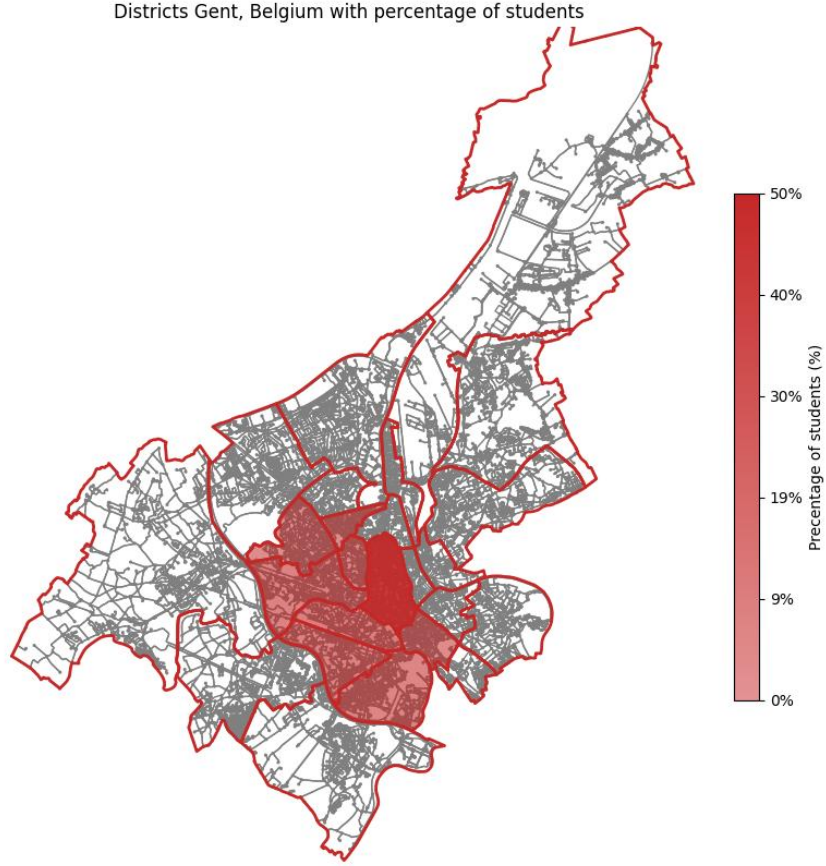


Figure 3.7: Districts of Ghent: colored areas indicate the percentage of students residing in each district, based on survey data.

even more students traveling through Sint-Pieters. Large student housing information was also provided by the city of Ghent. [65] Additionally the real city center is added, since it originally was too north.

3.4.2 Intensity Distribution Determination

Before assigning students, an intensity function needs to be determined. A lack of knowledge and data about the complex student housing situation forces the use of the optimization approach using survey data to determine a quality intensity function that describes the student housing locations in Ghent. This survey data is available on the abstraction level of districts. The data is only available for most of the center districts, so these will be used in determining the metric score. The districts of Ghent, and those used for evaluation can be seen in figure 3.7. In each evaluation, the students are assigned using the proposed method and the metric is computed over these aggregated results and the survey data.

As indicated in tables 3.1 to 3.4, the parameter options within the methodology are quite extensive. Even limiting the parameters to the options in the tables, still results in hundreds of parameter sets that need to be evaluated. To work around this, first a random selection of parameters is evaluated. The best

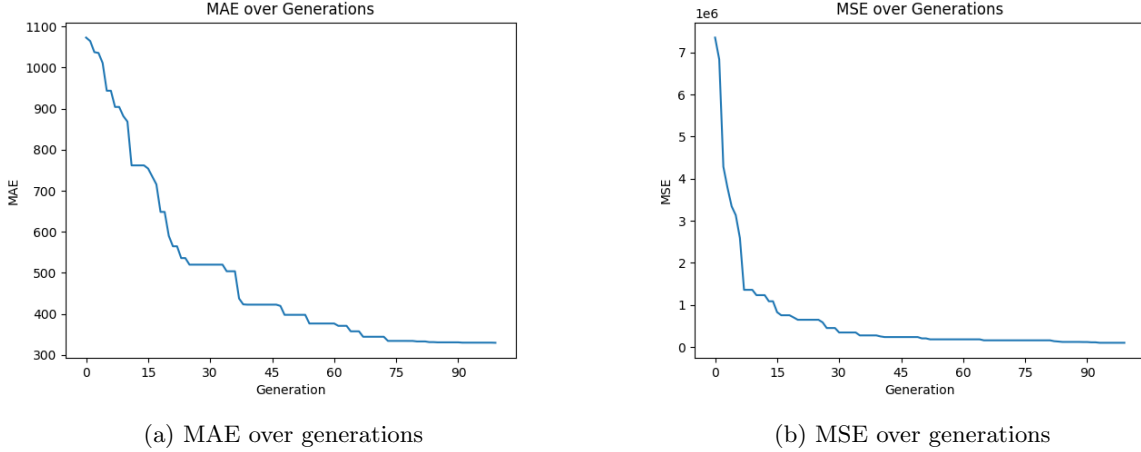


Figure 3.8: Evolution of metric score over generations in the genetic heuristic for intensity determination.

resulting parameters are selected and used to do a parameter by parameter evaluation. This means that one parameter of the best solution can change, while the others remain as in the current best solution. In the end, this will result in an evaluation of the impact of each parameter on the performance, and in a selection of one of the best parameter combinations. This combination is more extensively evaluated afterwards.

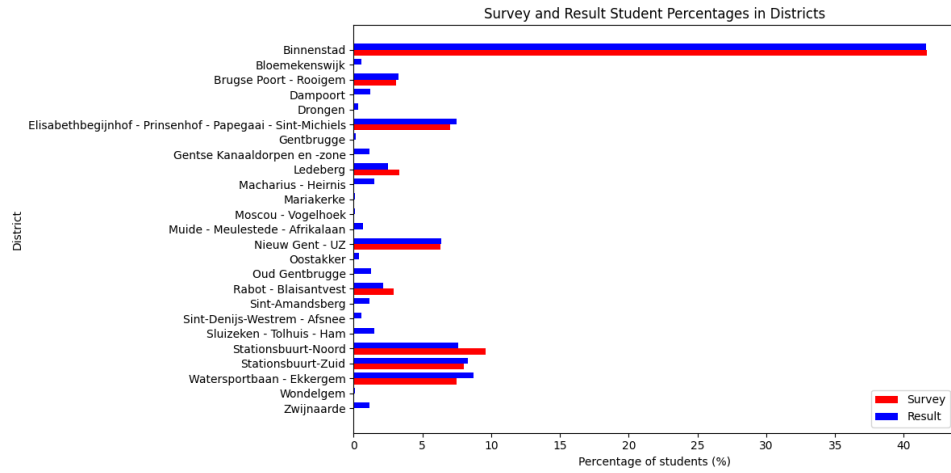
Genetic Heuristic

Because of the high number of options, the genetic algorithm is chosen for the parameter evaluation since it is the fastest performing one. While the algorithm has random factors, making a near-optimal solution not guaranteed after a set number of generations, a number of generations is selected where the improvement of the metric has mostly stagnated. Since the algorithm hyperparameters do not have much impact on the solutions, but mostly influence the execution time, they are kept constant during the entire evaluation. This means that each set of parameters is run for 100 generations with a population of 400 individuals, a mutation rate of 30%, a crossover rate of 50% and in each generation the 10% best individuals are guaranteed survival to the next generation. Figure 3.8 shows the stagnation of the metric scores and explains the choice for the algorithm parameters. This stagnation is detected on all runs performed, but no explicit guarantee can be given that improvements would not be made after 100 generations.

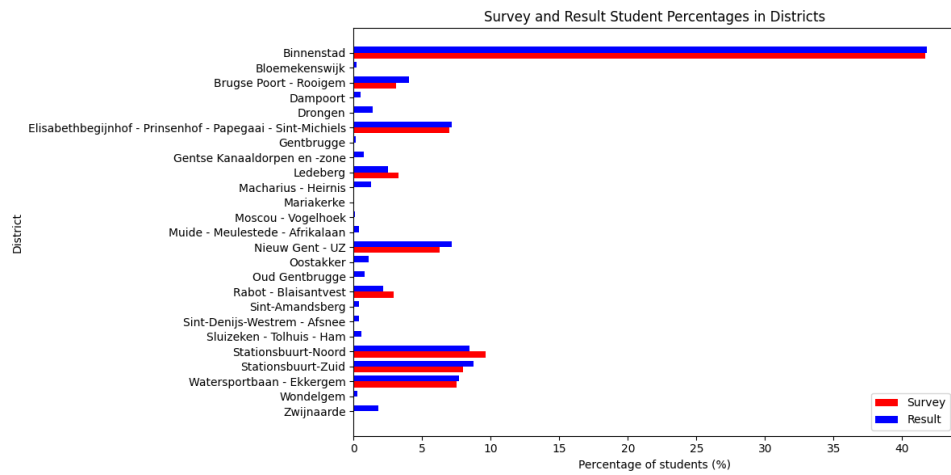
42 random sets of parameters were initially evaluated to select a starting point for the in-depth analysis of each parameter. Table 3.5 shows the corresponding parameters and metric score for all evaluated sets. A first selection is made by choosing the best performing set of parameters for the mean absolute error metric. Starting from this, each parameter is evaluated separately in the order of occurrence in table 3.5. For each comparison, the metric scores are given, together with a more detailed explanation explaining the cause and meaning of the results when necessary. The first parameter is the used metric itself, thus a comparison between the mean absolute error (MAE) and the mean squared error (MSE).

metric	non-increasing	n	campus density	large housing	radius	regression	k	length	best score
MAE	False	1	False	True	750	False	4	20	498.68
MAE	False	1	False	True	750	False	12	15	188.98
MAE	False	2	False	True	250	False	8	20	409.52
MAE	False	3	False	True	500	False	8	10	389.37
MAE	False	3	True	False	500	False	4	15	677.52
MAE	False	3	True	False	750	False	12	20	256.56
MAE	True	1	False	False	250	False	4	20	791.37
MAE	True	1	False	False	750	True	8	20	555.48
MAE	True	1	True	False	500	False	4	20	780.01
MAE	True	1	True	True	250	True	12	20	429.06
MAE	True	2	True	False	750	False	12	20	560.25
MAE	True	2	True	False	750	True	12	15	494.05
MAE	True	2	True	True	500	True	8	10	419.72
MAE	True	2	True	True	500	True	8	15	329.51
MAE	True	3	False	True	750	True	12	15	484.29
MSE	False	1	False	False	500	False	4	15	693419.79
MSE	False	2	False	True	500	False	4	15	388450.85
MSE	False	2	False	True	500	False	4	20	404429.10
MSE	False	2	False	True	750	False	4	20	388690.75
MSE	False	2	True	False	750	False	4	15	546573.73
MSE	False	3	False	False	750	False	12	20	102047.47
MSE	False	3	True	True	250	False	8	15	214525.73
MSE	False	3	True	True	750	False	4	20	337150.15
MSE	True	1	False	False	250	True	8	20	507102.45
MSE	True	1	False	False	500	True	4	20	506842.43
MSE	True	1	False	False	750	False	4	15	1309531.27
MSE	True	1	False	True	500	False	8	10	199524.68
MSE	True	1	False	True	750	False	8	15	228827.31
MSE	True	1	True	False	500	True	8	20	535017.28
MSE	True	1	True	True	250	True	8	20	362171.76
MSE	True	2	False	False	500	False	8	10	370314.96
MSE	True	2	False	True	500	False	4	15	470256.13
MSE	True	2	True	False	250	True	8	20	574131.79
MSE	True	2	True	False	750	False	8	20	349773.88
MSE	True	2	True	True	500	False	8	15	267834.95
MSE	True	2	True	True	750	True	4	20	288216.25
MSE	True	3	False	False	750	False	4	15	1219121.35
MSE	True	3	False	True	250	False	4	20	514232.74
MSE	True	3	False	True	750	False	4	15	476458.01
MSE	True	3	True	False	250	False	8	10	327329.34
MSE	True	3	True	True	250	True	4	20	341656.87
MSE	True	3	True	True	750	True	4	15	197647.77

Table 3.5: Metric scores for different sets of parameters randomly selected from the evaluated parameters. {k = polynomial degree, n = set size distances}



(a) MAE



(b) MSE

Figure 3.9: Comparison MAE and MSE for the percentage of students assigned to districts.

metric	non-increasing	n	campus density	large housing	radius	revaluation	regression	k	length	best score
MAE	False	1	False	True	/	False		12	15	188.98
MSE	False	1	False	True	/	False		12	15	44387.91

The two metrics can not be directly compared, since they have a different meaning. No immediate decision can be made on the better result from the metric scores themselves. Figure 3.9 shows the resulting student percentages and the survey student percentages within the districts for both options. The MSE error decreases the largest errors by compensating elsewhere. This is clear on the neighborhood, Stationsbuurt-Noord, where the error reduces, while other errors in different neighborhoods increase. The mean absolute error gives an average number of students over each neighborhood that are assigned to the wrong neighborhood. The meaning of this metric is simple to deduce. More importantly, an easy connection can be made between the MAE and the amount of students wrongly assigned. The mean squared error will assign students more wrongly, just to make sure the error in one neighborhood is not too high. Because of the interpretability, the MAE, seems the right choice for this problem and will be used in continued evaluations.

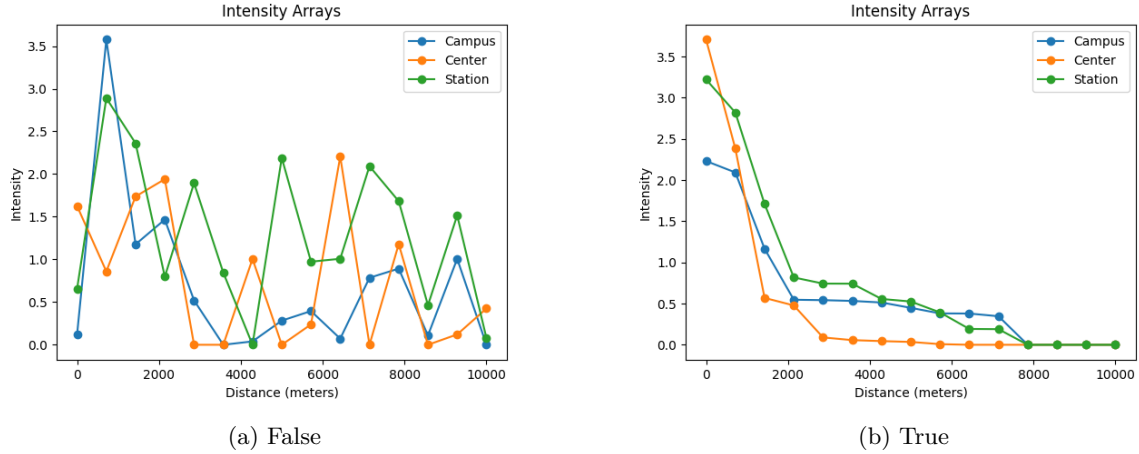


Figure 3.10: Comparison of intensity arrays with the non-increasing parameter set to true or false.

The next parameter is the decision to constraint the intensity array to a non-increasing array. This constraint also translates to the determined intensity function after fitting, but a polynomial fit will still allow areas of non-increasing intensity values. Figure 3.10 shows the intensity arrays for the case with and without a non-increasing constraint. The metric score of the non-increasing intensity arrays, is higher, worse than the score of the unconstrained intensity arrays. The unconstrained intensity arrays have more opportunities to capture the extremely complex situation of student housing, but comes with significantly reduced interpretability. The non-increasing arrays can be explained from literature indicating the decrease of student interest in neighborhoods with increasing distance to the campus, center and stations. A MAE of 403.23 means about $403.23 \cdot 9 \approx 3630$ students are wrongly located within the 9 districts data is available from. On the grand scheme of the 29000 students assigned into student housing this is an error of at least 12 % since the error is not known on the districts without survey data. This is acceptable for a very rough estimation, but a smaller error, is desired to improve accuracy. The explainability is extremely important in this real world application of student housing locations, thus there is a preference for the non-increasing intensity arrays. A different selection of the other parameters, will improve the score to a more acceptable value.

metric	non-increasing	n	campus density	large housing	radius	regression	k	length	best score
MAE	False	1	False	True	/	False	12	15	188.98
MAE	True	1	False	True	/	False	12	15	403.23

The size of the set of minimum distances to choose from for the campus intensity, has no significant effect on the resulting score. These differences can be attributed to the randomized elements of the genetic algorithm. In the overall intensity and the intensity within one district, there are no significant differences. Consequently, the choice is made to keep this set size to one distance, which results in the minimum distance to a campus being used in the campus student intensity function.

metric	non- increasing	n	campus density	large housing	radius	regression	k	length	best score
MAE	True	1	False	True	/	False	12	15	403.23
MAE	True	2	False	True	/	False	12	15	404.63
MAE	True	3	False	True	/	False	12	15	405.02

Treating the large student housing projects separately is beneficial for the overall result. Large student housing projects, have a significant impact on student numbers within certain neighborhoods, increasing the student density drastically. This local peak in student density is hard to capture within the student intensity functions. The intensity functions will consequently perform worse on the entire student housing location assignment. This can be attributed to the fixed location of these projects and students having different decision patterns when it comes to large student housing projects. Not everywhere within a city, large developments can occur, so they are fixed in location. Students will select these projects due to convenience and possibly reduced prices. This means, students do not primarily choose based on the distances towards the campuses, city center and stations for larger developments. It makes sense to treat them separately and let the intensity functions focus on the student housing location distribution independent of the larger student housing developments.

metric	non- increasing	n	campus density	large housing	radius	regression	k	length	best score
MAE	True	1	False	True	/	False	12	15	403.23
MAE	True	1	False	False	/	False	12	15	596.65

The next two parameters, campus density and radius, are related to each other. The radius indicates the radius of the circle in which the campuses are counted for the campus density factor. If the campus density is false, the radius has no effect on the intensity. Figure 3.11 shows the difference in intensity functions with and without the application of the campus density factor. It is clear that no campus density factor spreads out the intensity function more, while the intensity function with a campus density factor has higher peaks around campuses relatively reducing other areas of the city. When the radius of the campus density increases, the intensity function has even higher peaks but also spreads out more over the city. Making a balanced choice based on these results is difficult because the small improvement due to campus density with radius 500 could be a side effect of the randomized aspects in the genetic heuristic. However, the results give insight in the effect of the campus density on the geographic interpretation of the intensity function. For the rest of this section, the best combination will be used, with the campus density turned on, and the radius set to 500 meters.

metric	non- increasing	n	campus density	large housing	radius	regression	k	length	best score
MAE	True	1	False	True	/	False	12	15	403.23
MAE	True	1	True	True	250	False	12	15	426.54
MAE	True	1	True	True	500	False	12	15	396.11
MAE	True	1	True	True	750	False	12	15	422.96

The selected fitting function and the degree of the polynomial are also related. When the regression is used, which can only be fitted for non-increasing arrays, the polynomial degree has no effect on the outcome. If a polynomial function is used, k indicates the degree of the polynomial. The degree of the

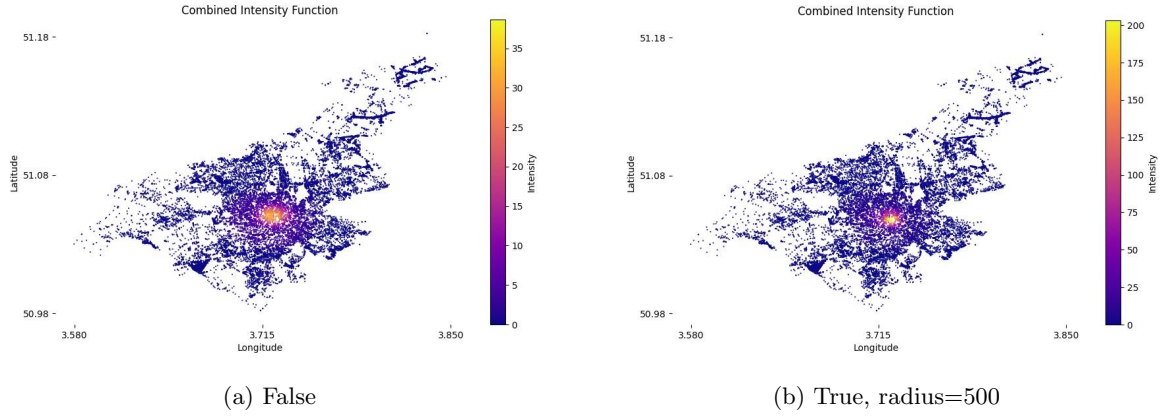


Figure 3.11: Node intensity for intensity functions with and without campus density factor.

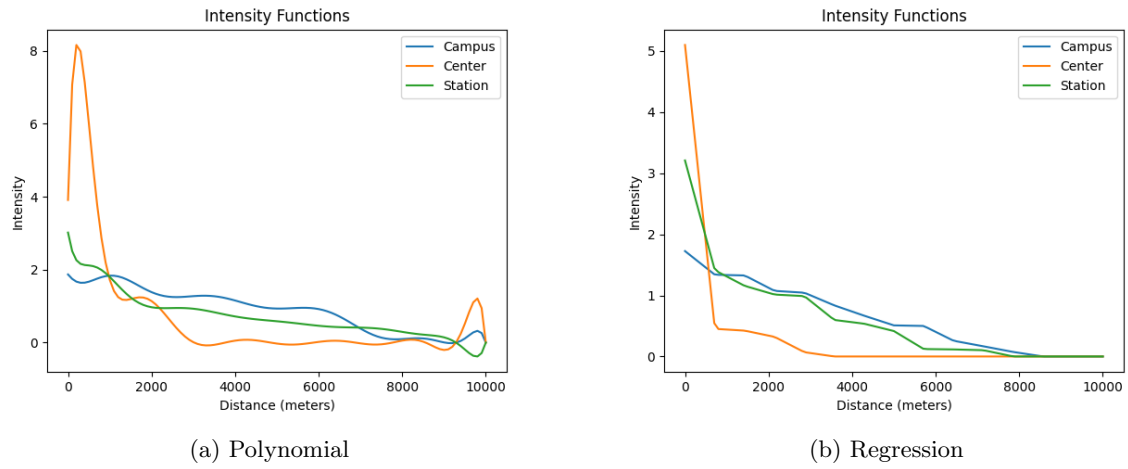


Figure 3.12: Difference in intensity functions between polynomial fitting and isotonic regression on approximately the same intensity array.

polynomial can not be equal or higher than the length of the array, since that would result in a possible badly fitted polynomial. Figure 3.12 shows the intensity functions fitted for an almost equal intensity array. There are some big differences in the intensity functions. The most visible one, is the variation in the polynomial and that the polynomial is increasing in some intervals. The higher polynomial degree will be able to capture more complex patterns, but might overfit. The selection of a decreasing intensity array, makes that even the polynomial can be related to the student housing location choices students make. However, this relation is clearer for the regression function.

The length of an array determines the control over the intensity functions. A longer increases the intensity resolution. This comes at a cost of execution time. It will take longer to find the optimal or near-optimal solution and there is a reduction in variance for each element at the same population size. In principle longer arrays, should improve accuracy with a good selection of algorithm parameters. The selected algorithm parameters, make finding optimal or near-optimal solutions for a length of 20 difficult. From the choices between regression, length and polynomial degree the best performing one is chosen. This also corresponds to the most explainable one, because of the use of a regression fitting function. The length of 15 performs well, while not suffering yet from a relatively too small population size.

metric	non- increasing	n	campus density	large housing	radius	regression	k	length	best score
MAE	True	1	True	True	500	True	/	10	413.03
MAE	True	1	True	True	500	True	/	15	341.90
MAE	True	1	True	True	500	True	/	20	425.75
MAE	True	1	True	True	500	False	4	10	539.76
MAE	True	1	True	True	500	False	4	15	544.59
MAE	True	1	True	True	500	False	4	20	546.80
MAE	True	1	True	True	500	False	8	10	346.23
MAE	True	1	True	True	500	False	8	15	449.12
MAE	True	1	True	True	500	False	8	20	446.40
MAE	True	1	True	True	500	False	12	15	418.80
MAE	True	1	True	True	500	False	12	20	384.70

The manner of exploring the parameters is not exhaustive. There might be interaction between different parameters, which means some combinations might be even better than the parameters selected through the parameter by parameter comparison. However, the selected parameters provide an adequate solution for this use case and show the capabilities of the proposed methodology. The parameter comparison gives guidance in the effects and meaning of each parameter in the methodology. In an ideal world, more computational time is available to determine the best intensity functions. The selected parameter combination will be used to evaluate and compare the genetic and alternating pass optimization approaches in the next section.

The genetic approach sometimes suffers from its randomized aspects. While these aspects ensure variance in the population, they make each run through different and as a result it can take a while until an equivalent optimal or near-optimal solution is reached. Figure 3.13 shows two intensity arrays of two different runs with the same best parameters selected in this section. These intensity arrays show different results. More importantly, the scores are different for each run after 100 generations. To increase the chances to get the same result each time, the population size and number of generations could be increased. This increases the runtime of the algorithm, but ensures the best solution can be found more easily.

The parameter comparison is based on the meaning of most parameters. Only the score is used for a decision about the set size of minimum distances and the large housing. The difference is high enough in the large housing to draw conclusions. No difference was detected in the case of the minimum distance set size, so choosing the minimum distance does not invalidate the conclusions.

run	metric	non- increasing	n	campus density	large housing	radius	regression	k	length	best score
1	MAE	True	1	True	True	500	True	/	15	333.46
2	MAE	True	1	True	True	500	True	/	15	363.82

Alternating Pass Algorithm

The selected parameters will be used to briefly evaluate the alternating pass algorithm. The length of the array has to be limited to at most 10. Because of the high computational requirements an evaluation of $3^{15} = 14348907$ options is not feasible. The alternating pass is run with a step size, η , equal to 0.1 and starts with an array of all ones. The MAE is 477.87 after the alternating pass approach. Figure 3.14a

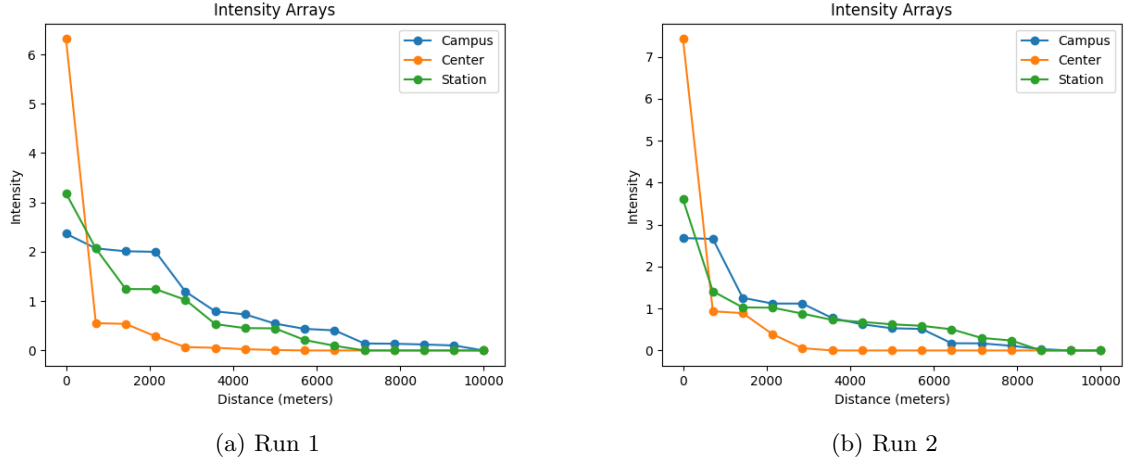


Figure 3.13: Variation in determined intensity functions using identical parameters, caused by randomization effects.

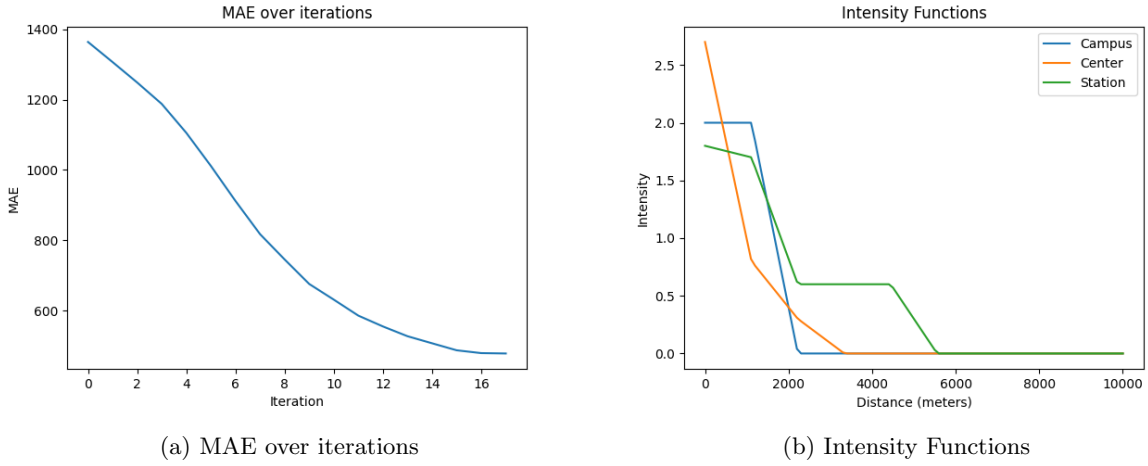


Figure 3.14: Evolution of MAE and intensity functions determined using the alternating pass approach.

shows the evolution of the metric score over different iterations. It indicates a minimum is achieved quickly, however as seen from the genetic algorithm, this is not the global minimum. The algorithm is limited by the step size, the length and the initialization values in determining the best solution. The step size limits the evaluated values to the original values plus or minus a multiple of the step size. While there is some scale independence, not every equivalent solution is evaluated because of the step size.

The isotonic regression performs really well when the length of the intensity array increases, so there is more control over the intensity function. The alternating pass approach is limited in intensity array length, because of the high set of options needing to be evaluated. This also reduces control over the intensity functions due to the lower resolution of intensity values. The initialization values play a big role in which solutions are evaluated in each iteration and thus limit the alternating pass algorithm further. Figure 3.14b shows the resulting intensity functions. These functions are more rigid than when a longer intensity array is used. The results, figure 3.15, have some similarities with the results of the genetic approach. However, the errors made are significantly higher, as can be deduced from the metric score.

The alternating pass approach has as main benefit control over which intensity arrays are evaluated, but this is also its main limitation. The control increases the repeatability of the method. The alternating

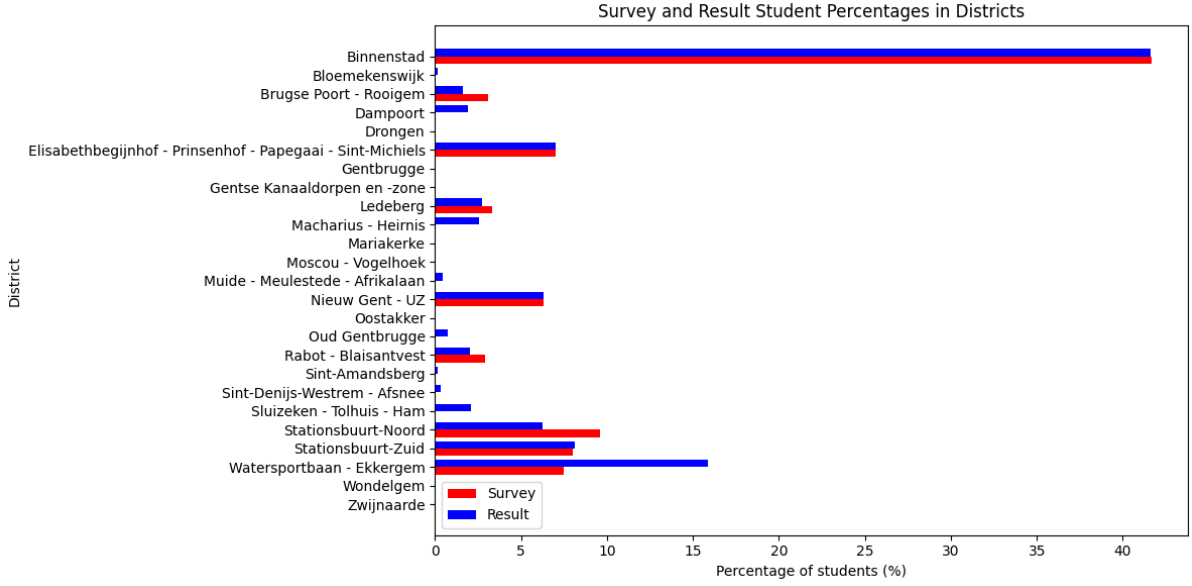


Figure 3.15: The percentage of students within districts at the output of the alternating pass approach.

pass algorithm is particularly useful when the search space is relatively small, such as when the intensity arrays are short or when the constraints significantly limit the number of valid variations. However, as the length of the intensity arrays increases, the number of possible variations grows exponentially, making the approach less practical for larger problems. The step size is too limiting in evaluating all good intensity arrays within reasonable time. Despite its limitations, the alternating pass approach offers a deterministic and systematic way to explore the search space. The genetic heuristic is better in finding optimal solutions.

3.4.3 Student Assignment

In this section, the complete student assignment method is evaluated. The intensity function and student assignment parameters used are the best parameters from the previous section. Firstly, the genetic approach is run, with 800 individuals in the population and 300 generations. This is a significantly longer and larger population size than in the previous section to make sure the near-optimal solution is reached before diving into the analysis of the student assignment.

metric	non- increasing	n	campus density	large housing	radius	regression	k	length	score
MAE	True	1	True	True	500	True	/	15	303.25

Evaluating if the genetic approach with the adapted parameters has reached a near-optimal solution can be achieved by analyzing the variation of the MAE score over different generations. Figure 3.16 shows the evolution of the mean absolute error over the different generations. A first conclusion that can be deduced from these graphs is that a near optimal solution has been reached. The score improvement stagnated after about 150 iterations. Smaller improvements are still made in the generations afterwards, but these are minimal. The differences between subsequent iterations reduce to negligible values as well as the MAE stays almost constant in the last generations. The improvements are random because of the randomized elements in the genetic heuristic. This results in a substantial improvement in some iterations, while in other generations the best found individual stays the same. The overall population might still

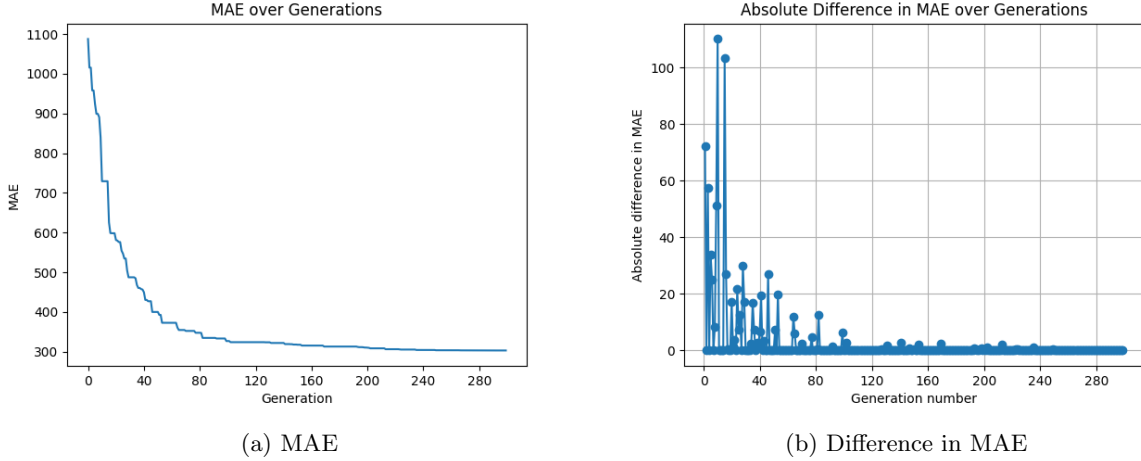


Figure 3.16: MAE evolution over generations in the genetic algorithm.

be improving even when the best individual remains constant, which allows even better solutions to be possibly found in subsequent iterations. The genetic approach can be stopped early when no significant improvement is detected over several generations. This stopping condition is hard to define because of the randomized aspects that might do nothing for a few generations but then could start improving again.

The score of 303.25, means $303.25 \cdot 9 = 2729.25$ students are assigned to the wrong neighborhood. This is an error percentage of at least 9.4%, because additional mistakes can be made in the districts without data. This is a good result for an estimation of student housing using this simplified methodology for a complex problem. It also indicates that if there are additional dependencies between student housing location and city elements, additional improvements can be made. Figure 3.17 indicates the estimated percentage of students in student housing in each district next to the real values of the known districts. There are only some neighborhoods where the error is significant, Stationsbuurt-Noord has too few students assigned, while Watersportbaan-Ekkergem has too many students assigned. The amount of students in Watersportbaan-Ekkergem is too high because the neighborhood contains four campuses, but three of these campuses are small campuses compared to the center or more southern campuses. This increases the number of students in this neighborhood beyond what is observed in the real scenario. Stationsbuurt-Noord has only one campus and is a bit further from the city center. Thus, fewer students are assigned to this neighborhood. However, this is an interesting location for students that have classes on multiple campuses, because it lies between the center campuses and the more southern campuses of Ghent University while also being close to the main train station. Students will choose a location that balances these distances, instead of one where they are only close to one of their campuses.

The resulting intensity arrays are given in figure 3.18a, the intensity functions determined using regression are given in figure 3.18b. The use of the isotonic regression makes the intensity functions completely non-increasing over the entire distance. Some intensity functions have interesting characteristics. The center student intensity function weighs a lot towards smaller distances, with a longer almost zero tail. This means the distance to the city center is extremely important in determining the student intensity at that location. This characteristic can also be seen in the amount of students in each district. The most students in student housing live within the district Binnenstad, which is the district that contains the city center. The longer tail allows other districts to have students as well.

A plateau can be detected in the first 1.5 kilometers for the campus intensity function after which a

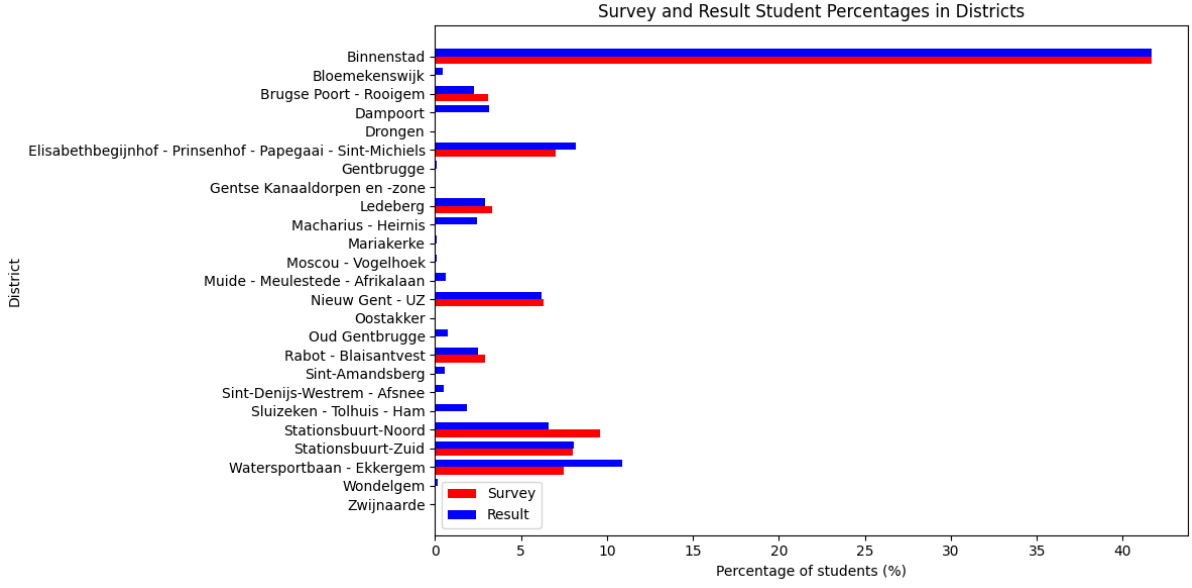


Figure 3.17: Percentage of students within districts at the output of the genetic approach

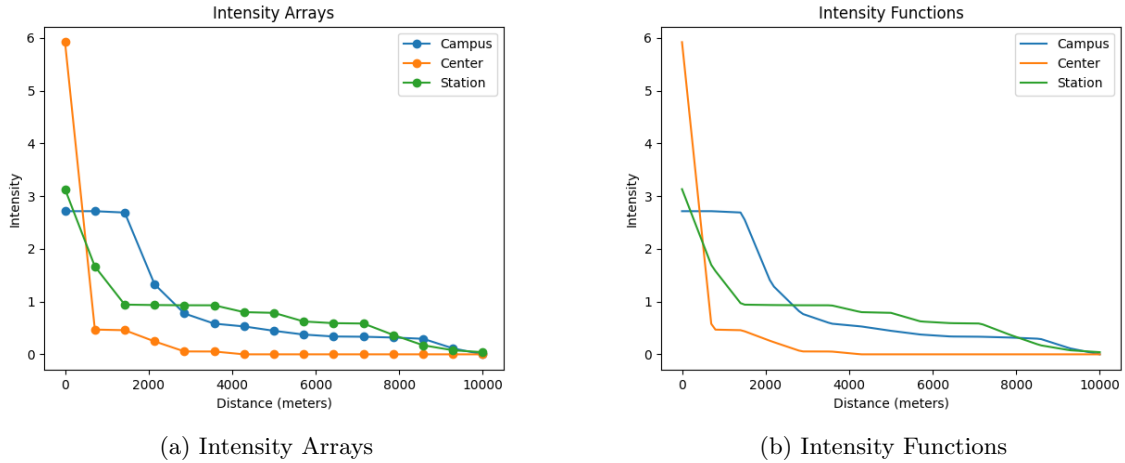


Figure 3.18: Intensity arrays and corresponding functions from the genetic heuristic.

significant drop can be detected. This indicates that students are fine living within a certain distance, but beyond this distance the attraction of neighborhoods decreases rapidly. The station intensity function also shows interest in student housing close to train stations, because of the easy connections to home. The interest decreases in the first few kilometers, afterwards the distance to the station becomes less important. An explanation for this is the high cyclability of the city of Ghent and students using bikes to move towards the train stations. Furthermore, Ghent is a relatively small city. Starting from the city center, in about five kilometers you have left the main city region. This means the value of the combined intensity function around this distance should move towards zero. This is achieved by the intensity function of the city center, that becomes almost zero or zero after several kilometers. Since all intensity functions are multiplied together, this also makes the combined intensity function near-zero or zero when moving towards the city bounds.

The student housing assignment according to the determined intensity function can be seen in figure 3.19, without the assigned commuter students. Figure 3.19a shows the student assignment with the large

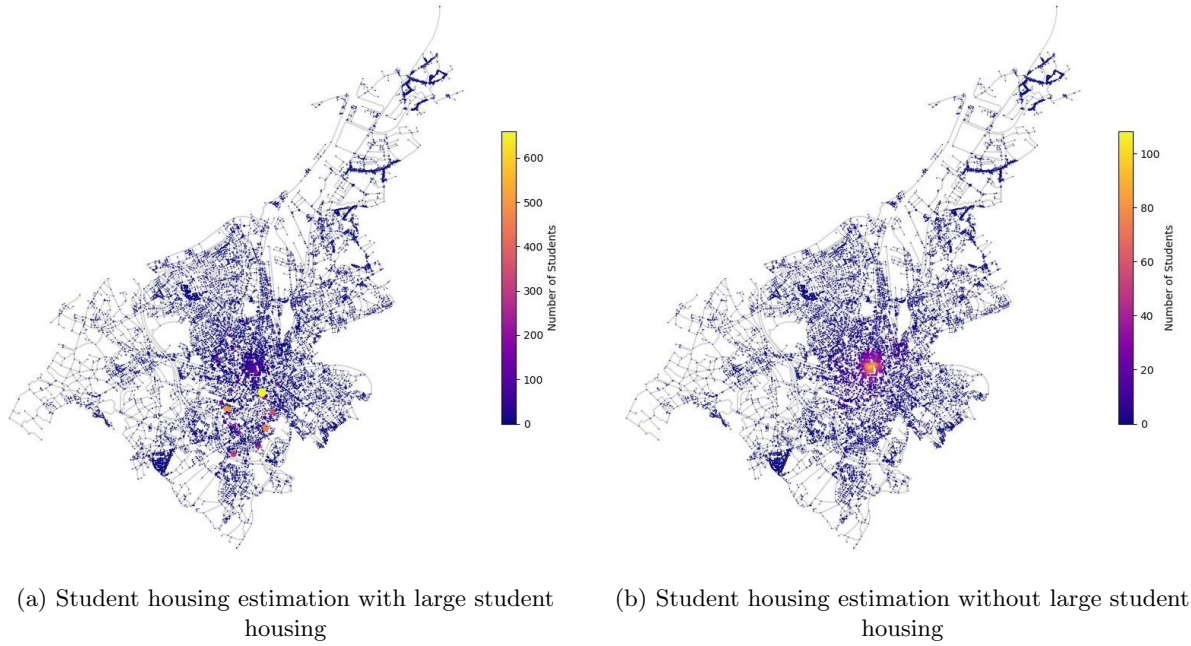


Figure 3.19: Student housing node intensity at the output of the genetic heuristic.

housing project present. It shows clearly that some nodes peak really high, where large student housing projects are present. It can be interesting to look at the student assignment using this intensity function if the large student housing projects are not present. Figure 3.19b shows the student assignment without the large student housing developments. A clear high intensity can be seen around the city center, that spreads out in the entire Binnenstad district. There is a spread outward from the city center, accompanied by several smaller clusters near specific campuses.

The student assignment using intensity functions has no inherent meaning until aggregation happens on a higher abstraction level. To evaluate the student housing assignment and intensity functions, this aggregation can also be done on different abstraction levels than the districts. This allows to see if the intensity functions generalize well to the entire city and are not overfitting to the districts. Another interesting abstraction level are the statistical sectors, which allows other data to be added in future work. The statistical sectors give a more fine-grained aggregation, allowing more detail in the student mobility study. Figure 3.20 shows the aggregated values for the statistical sectors of Ghent. Evaluation on the abstraction level of statistical sectors is more difficult, because no data is available on this abstraction level. However, a less formal evaluation based on known student neighborhoods is possible.

The highest number of students are in the statistical sector known to be the student neighborhood in the city. This is extra enhanced by some large student housing projects in that sector. Surrounding this sector are some other high intensity statistical sectors, that are known to contain a lot of students. The number of students is not the correct metric for student neighborhood detection, because the statistical sectors differ in size. The student density is better fit in determining which neighborhoods can be classified as student neighborhoods. Figure 3.21 shows the student density for the different statistical sectors. Only a small difference can be detected in the top student neighborhoods between the student density and number of students. The student density indicates the same top neighborhoods but in a different order. Some smaller student neighborhoods got a boost now. The top neighborhoods remain mostly the same with only some small differences. The aggregation and neighborhood analysis shows that the

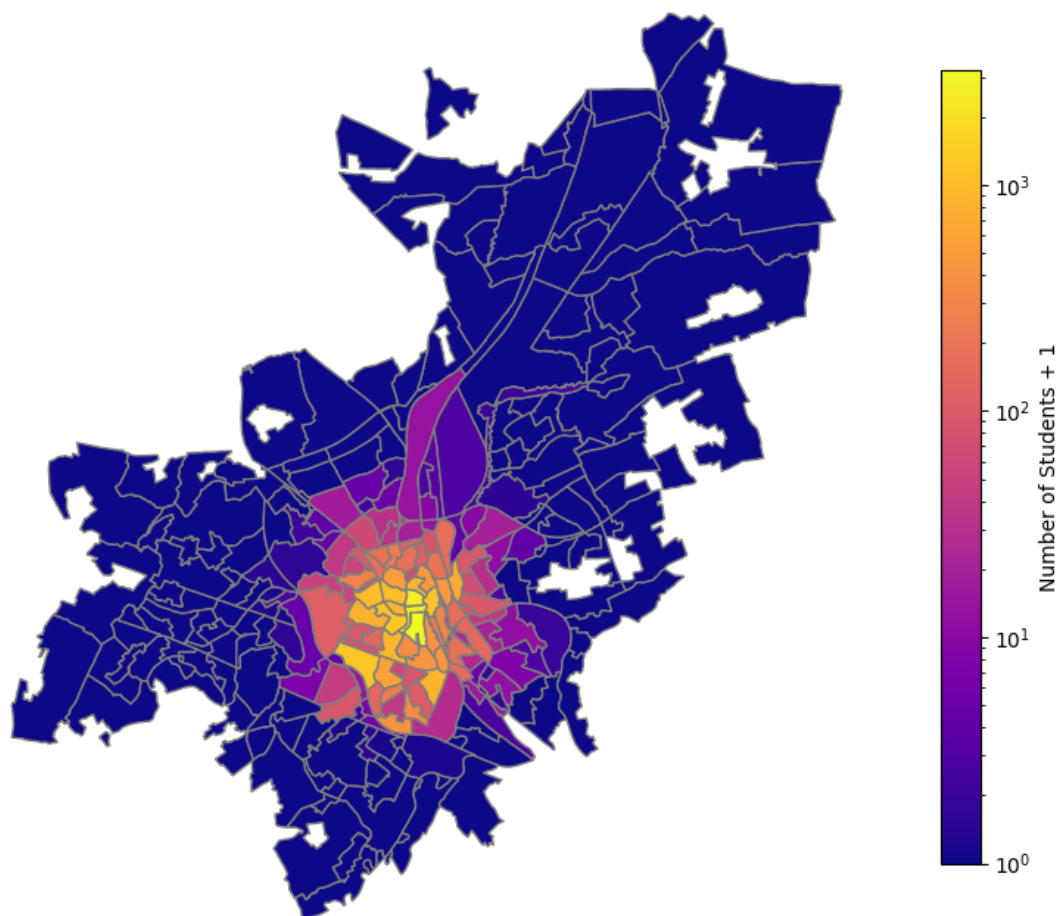


Figure 3.20: Number of students in each statistical sector after aggregating student intensities from the genetic heuristic.

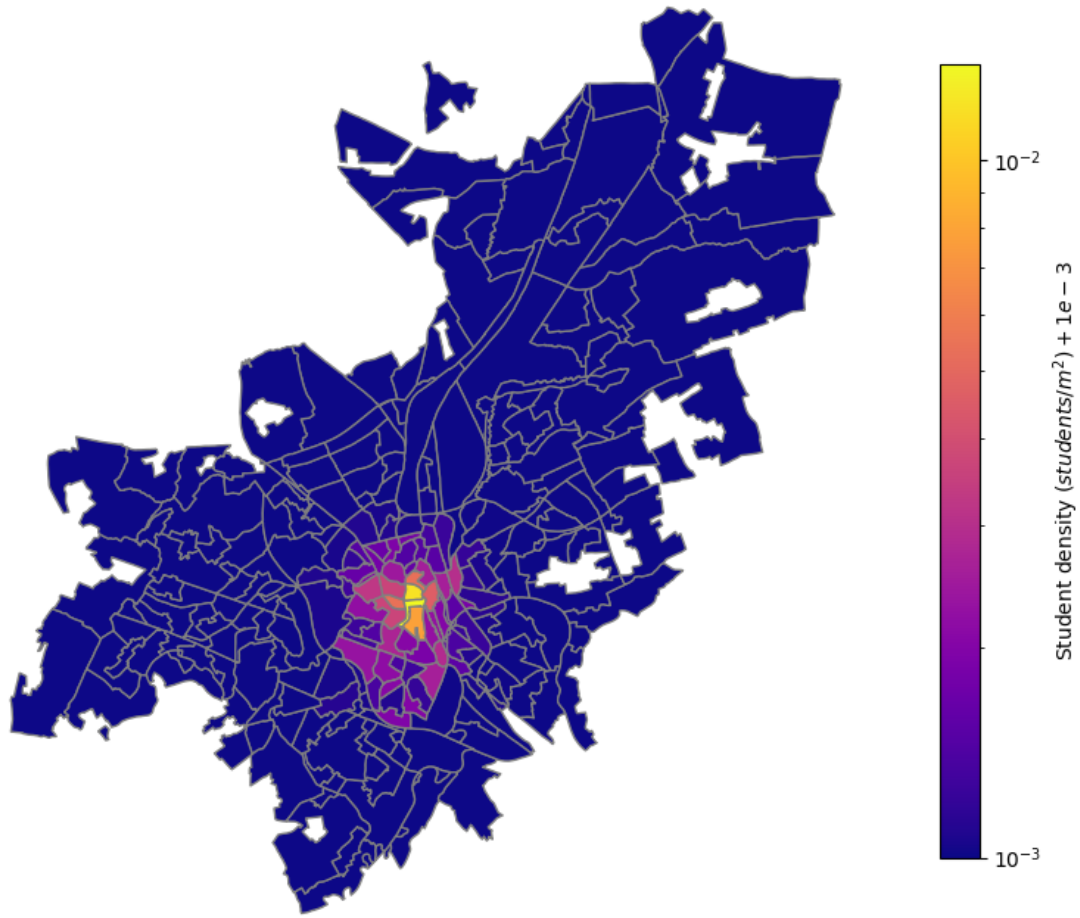


Figure 3.21: Student density for each statistical sector after aggregation intensities from the genetic heuristic.

methodology also generates good results on a different abstraction level. There are definitely additional mistakes because the difference in abstraction level, but the results seem to indicate these mistakes are minor. If data is available on different abstraction levels, an even more detailed student intensity function can be determined. Errors introduced by aggregation increase with finer levels of aggregation. Care must be taken when aggregating student intensities to ensure the error remains within acceptable limits.

Next to the student density for neighborhood detection, the campus distance was found as another interesting metric for the student assignment. The mean minimum distance to a campus is 597.06 meters indicating once more that students prefer to live close to their campuses. There is a standard deviation of 368.22 meters on this minimum distance, which is logical because of the big variation in distances. Figure 3.22 shows a minimum distance histogram of the entire student assignment. There are 2 peaks that stand out, these are the train stations in Ghent, since commuter students are assigned here. The rest of the distances indicate the same characteristic that students are housed close to their campuses.

The genetic heuristic proves to be a reliable way in determining explainable and good results for the student housing location estimation. Additionally, the proposed student housing location estimation

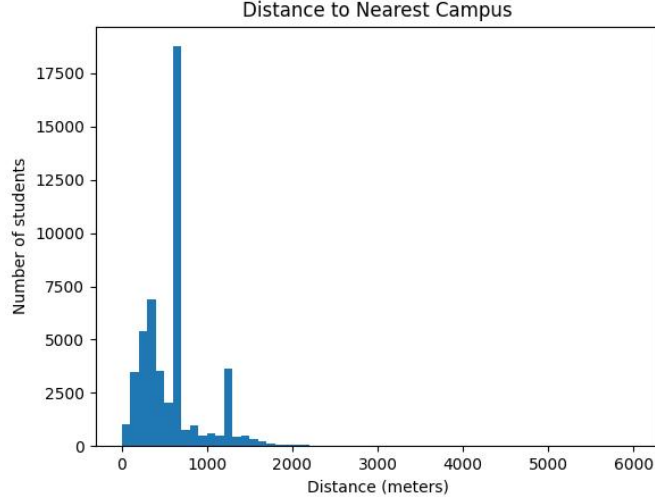


Figure 3.22: Histogram of distances to the nearest campus for each assigned student.

methodology, supported by either the genetic heuristic or the alternating pass algorithm, provides a robust framework for addressing the challenges of estimating student housing locations in urban environments. It is especially useful in situations with missing data and where different levels of abstraction than those in the survey data are desired. While there are a lot of possible parameters in the student assignment, a selection can be made based on results and explainability. The parameters are still part of the methodology, because of possible interest in other problems. The methodology is able to capture both citywide distribution as distributions on a smaller abstraction level for student location estimation. It is not possible to capture the local distribution of students on an abstraction level that is too small. Aggregation partially masks mistakes, but this masking effect diminishes as the aggregation level becomes too small because the node intensities are not always complete students. The outcome of the student housing location estimation can be used in student related studies.

3.5 Kortrijk Case Study

To evaluate the generalizability of the determined intensity functions, the intensity functions determined for Ghent will be applied on Kortrijk without any adaptation. No data is available to evaluate using metrics, however some websites indirectly indicate where student housing is located within the city.

3.5.1 Data Collection

The first step is to once more collect the relevant data. Figure 3.23 shows the road network, the train station and the campuses in Kortrijk. There is only one train station in Kortrijk, so all commuting students are assigned to this train station. There is no clear list to be found of all the higher educational institutions within the city, since they are from a variety of providers. A manual search and selection found these campuses:

Id	Campus	Latitude	Longitude
0	Howest-weide	50.82418440116774	3.249609935584367
1	Howest-level	50.82745845931422	3.2544683714826674
2	Howest-square	50.82527952551063	3.3060241143428124
3	KULeuven	50.806190472980255	3.292689276186894
4	UGent	50.82541267371603	3.2499976615685706
5	Vives	50.80721175823506	3.2844407261426323

No large student housing information was found publicly available for Kortrijk, thus this is omitted for the evaluation.

3.5.2 Student Assignment

Applying the intensity function determined for the city of Ghent, to the city of Kortrijk results in the student assignment of figure 3.24. A website summarizing student accommodation in Kortrijk for students searching housing indicates a distribution of student housing around the city center and the neighborhood south of the train station. [36] The student assignment does have a large amount of student housing around the city center, but almost none south of the train station. The determined intensity functions for Ghent are not directly transferable to the city of Kortrijk. One reason is the decreased importance of the train station on the student housing distribution because the station is located really close to the city center. This results in an incorrect boost of the neighborhood between the city center and the train station. Furthermore, Kortrijk is a less dense city than Ghent, the resulting intensity functions for Kortrijk would need to be more spread out over distances than for Ghent.

Kortrijk is a significantly smaller city in population than Ghent, which is also not considered in the intensity functions. This combination of large differences between Ghent and Kortrijk make the intensity functions not directly applicable. However, with small and substantiated changes to the intensity arrays a good student distribution might be found. An even more accurate approach would be to determine the student intensity functions for Kortrijk directly using survey data and the optimization method.

3.6 Conclusion

The proposed student housing location estimation methodology demonstrates a robust framework for estimating student housing locations in urban environments. By leveraging intensity functions based on distances to campuses, city centers, and train stations, the method captures key factors influencing student housing distribution. The methodology indicates that a complex urban distribution depending on several factors can be divided into smaller approximated independent relationships to each of the factors separately. The genetic heuristic proves to be an effective optimization approach for determining these intensity functions from survey data, offering a balance between performance and explainability.

The case study of Ghent validates the methodology, achieving a reasonable error margin in estimating student housing locations. The results align with known student neighborhoods and provide insights into student density and proximity to campuses. Relations between students choices and distances can be deduced from the intensity functions. While the alternating pass algorithm offers a systematic alternative, its computational limitations make it less suitable for larger problems compared to the genetic heuristic.

The application to Kortrijk highlights the importance of tailoring intensity functions to the specific



Figure 3.23: Road network graph of Kortrijk, Belgium collected from OpenStreetMap. Red node: campus of higher education institution, Blue node: train station

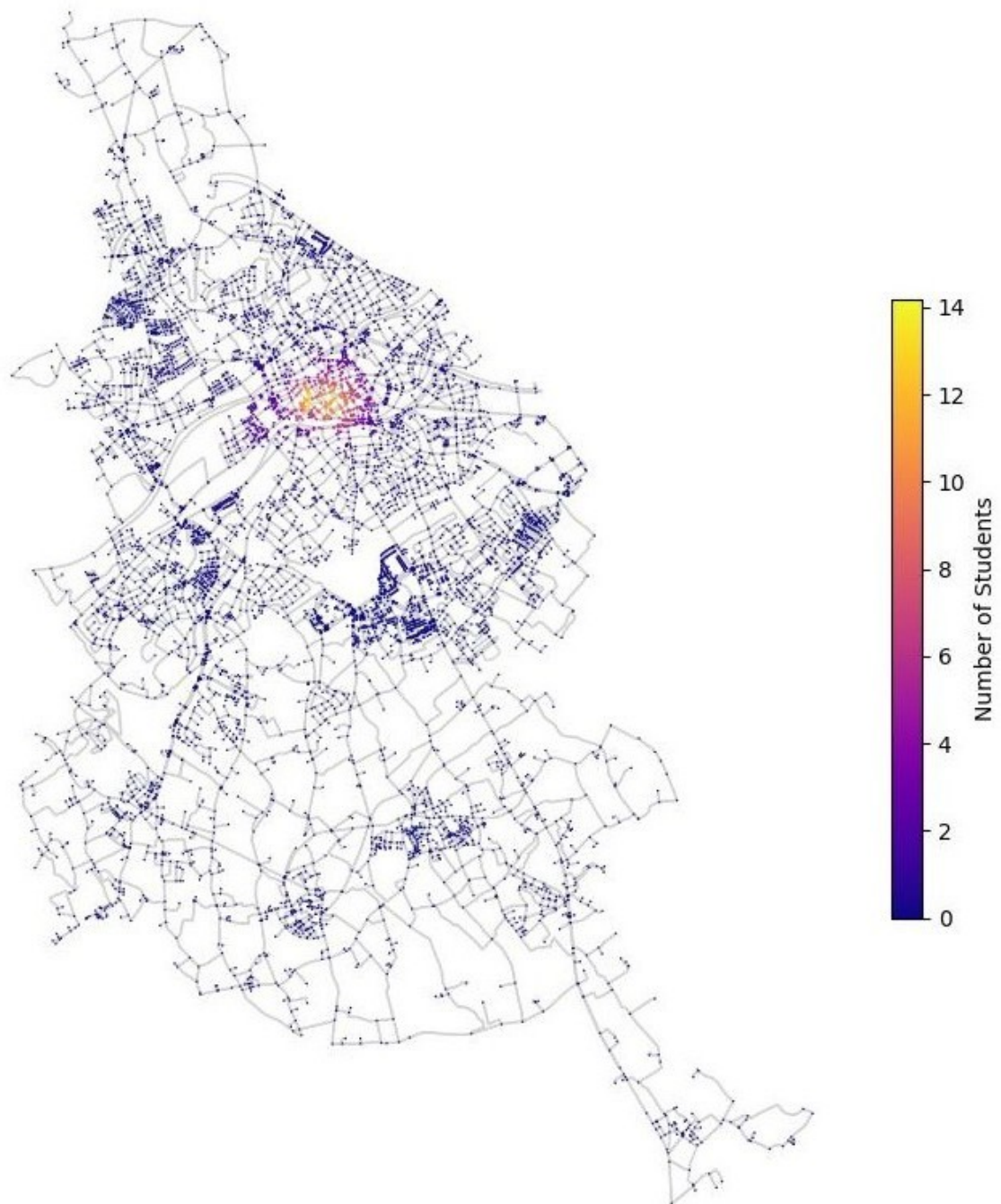


Figure 3.24: Student node intensity in Kortrijk using intensity functions derived from Ghent.

characteristics of each city. Found intensity functions are not always immediately transferable to different urban environments. Differences in urban density, campus distribution, and transportation infrastructure sometimes necessitates localized adjustments to the intensity functions. The best solution is to preform the intensity function determination for the city if it vastly differs from the cities of which the intensity function is already determined.

Overall, the methodology offers a flexible and scalable solution for estimating student housing locations, with potential applications in other urban studies. While not a substitute for detailed registration systems, it provides a valuable alternative for cities with limited data availability, enabling informed decision-making in student-related studies on different aggregation levels.

3.7 Validity

The methodology has as goal to understand the complex situation of student housing locations properly through an instantiation, a concrete realization of the model. This qualifies for the research principles of design science research (DSR). [28] Evaluating the method against threats to validity is essential within DSR to achieve a quality result. Therefore, the methodology is checked against different threats to validity: internal validity, external validity, construct validity and conclusion validity.

External validity is the validity of applying the conclusions in a different context outside this study. The resulting intensity functions have been proven to be not directly applicable within a different urban context. In similar cities, the resulting intensity function might be applicable because of similar urban context. The methodology of determining the intensity functions can be applied in different contexts and is independent of the use case. The data-driven approach allows the method to be tailored to specific characteristics within a city. The methodology is validated against a real world situation proving the validity in a real application. The use of optimization techniques makes the methodology flexible to apply in a different context. The flexibility allows the selection of different types of correlations in different cities, but they have to be independently determined, and the conclusions made on correlations in this study might not be directly applicable to a city of different size. The assumption that student housing location is dependent on factors like campuses, train stations and the city center is supported through literature, reducing the risk of inapplicability in other contexts. A different culture plays a role in the modes of transportation and thus assigning the commuter students to a train station might not be applicable in some other cities.

Construct validity concerns the metrics used within the evaluation. Firstly, the evaluation through validation of a real world scenario shows that the intensity functions can capture the complexity of a student housing location distribution. Distance to campus and student density has shown to give indications on the quality of the result and the known student neighborhoods. However, these metrics might be too limited if the student housing locations are oversimplified. Additionally, no quality ranges are defined for these metrics since they are really context dependent. This makes the metrics suitable for analysis, but not as quality metrics without additional case studies and extensive evaluation. How much the discrete representation of student intensity functions indicate the real intensity functions also falls under construct validity. Of course, a discrete representation is limited in its accuracy. However, when the length of the intensity arrays is long enough, the intensity arrays have proven to indicate corresponding intensity functions quite well. The fitting choice plays a role in the explainability of the results. Regression is easy to relate the intensity functions and intensity arrays to a real world scenario, this is more difficult for the polynomial fit.

The internal validity is the extent to which a piece of evidence supports a claim about cause and effect within the context of the study. The method simplifies the student housing location market. Other factors like affordability, availability and safety are not considered. The intensity functions are however representing this complex situation well for the use case. Aggregation in the optimization and in the eventual student assignment are done on a different abstraction level. There is a risk that a different abstraction level will not capture the correct characteristic accurately. This difference in aggregation achieves good results for the Ghent use case, but there is no guarantee this will work within a different context. Additional research will have to point this out. The causal relationship between student housing locations and campuses, city centers and stations is not formally proven. However, the correlation is deducted from extensive research into preferences of students, supporting this conclusion. Parameters have an effect on the outcome. Other combinations of parameters might result in better conclusions and results. The case by case basis, does not evaluate all combinations, but each parameter independently to conclude meaning. There is a possibility, better results can be achieved using different approaches to the parameter decision. The initialization choices are a threat to the sensitivity of the methodology. It is believed this threat is partially mitigated through the scale independence, but it still remains, especially for the alternating pass algorithm.

The conclusion validity entails the validity of the methods to draw reasonable conclusions. A proper quantitative evaluation was not possible through the significant amount of manual tasks that have to be completed for each city. A qualitative approach was adopted, but was also limited to the survey data. The lack of data disallowed a more advanced qualitative approach. However, by the application to the real world, the conclusion that the method is applicable to real world scenarios is supported. Drawing conclusions from this real world scenario is valid as long as the generalizability is taken into account.

Chapter 4

Agent-Based Student Mobility Simulation

The second goal of this thesis is to propose a method for estimating student mobility. One of the key elements for determining this accurately are the student housing locations. All mobility tours can be defined starting and ending at the housing location. Within chapter 3, a methodology is covered for estimating the student housing locations in the absence of data. The next step is to convert these locations and class schedules to the spatio-temporal domain of human mobility. Students can have a significant impact on their host city. Examples of this impact include studentification [62] and an increase in transportation demand [50]. Student intracity mobility information can help by providing guidance for city council decisions regarding students or the broader city population. The objective of this chapter is to put forward a student mobility model that can be used within agent-based mobility simulations.

During this chapter, a discrete-event simulation (DES) is described for student mobility simulation based on spatial housing information and temporal class schedules. To achieve a granular system this will be an agent-based simulation using the utility framework for choice modelling. Each agent follows a set of defined rules influenced by class schedules that guide their behavior. Not every step of the four step model is incorporated. The focus lies on trip generation and trip distribution, but the simulation also implements mode choice. An extension can be added to implement route choice to complete the four step model. The focus lies on student mobility and their patterns as part of a broader simulation framework.

First, the current situation of simulations and student related studies is summarized in section 4.1 to provide the required context. In section 4.2, the methodology and simulation framework is described. It includes an explanation of the simulation structure, the utility functions and the methods used within the simulation. A simplified variant of the entire simulation structure is executed in section 4.3 through application on the city of Ghent and Ghent University. The aim is to prove that the proposed schedule-based model is flexible and can model real world scenarios. Possible improvement avenues are explored in section 4.4. Finally, the methodology is critically analyzed for its validity in section 4.6.

4.1 Current Landscape

Students can have a significant impact on the systems within a city. Their presence influences housing markets, transportation systems and local businesses. This impact is the main driver for student related studies. These student related studies try to evaluate the characteristics of the student population and their impact. Research has been done into the transformations related to the influx of student populations in the urban landscape, often referred to as studentification. [62] Student mobility mode choices are also studied extensively. While these studies focus on students, a missing puzzle piece is that of student mobility itself.

Human mobility in general has been studied for several years. Either mobility is inferred from available data or simulated. The models used within these studies evolved from more general models focusing on entire populations to agent-based models simulating individual movements and recently also towards machine learning approaches. [6, 7, 9, 32, 38, 44, 81]. A clear direction is chosen for more automatic approaches. Despite the evolution towards individual level simulations, most of these only work accurately when focusing on the entire population to make sure the average flows validate with the real world scenario.

The main challenge for group specific models and simulation techniques is the lack of required data. Both the input data and validation data is often missing for the specific population group. An example of this is demographic data that is missing from traffic counts and mobile data making it impossible to determine to which population group to attribute the data, resulting in most of the human mobility techniques to be irrelevant for a population group specific study. This also makes machine learning techniques difficult to implement. Tunability of these techniques has not yet been tested enough to make this a viable option in the absence of required data.

Agent-based models using logit functions have become the standard in modeling human behavior and mobility because of their flexibility and relatively low data requirements. However, their application to student populations remains significantly limited. Furthermore, current implementations are restricted to campus environments, neglecting the broader urban context. [43] This is because trips within a campus can be attributed to only the student population, making it easier to develop a student related simulation model.

Most activity models go broad in the possible activity schedule decision of the agents. Students have a different activity schedule than most of the population due to its rigidity during class hours. Utilizing this realization and student housing locations, an easier to use student simulation can be constructed where only limited additional data is required to model student preferences. Simulation models need to be calibrated using counting data. This data is missing for students. A possibility is to set up these student counters. Another possibility is to add the student mobility model to a synthesis simulation model that can be calibrated by normal traffic counts. The student mobility model is proposed in the following section.

4.2 Methodology

The goal of this section is to provide an overview of the student mobility model through the agent logic. Additionally, the entire simulation framework used for evaluation is covered. Every part of the simulation is described in detail. The output of the simulation is an origin-destination matrix on a specific aggregation level. This aggregation level could be districts, statistical sectors or something else.

The aggregation level has an impact on the granularity of the results. The more granular the aggregation, the more impact each individual student agent will have in the simulation. A less granular aggregation level is however easier to validate using real data. A balanced choice needs to be made. The origin-destination matrix will contain the amount of students within a certain time period traveling between two aggregation zones. Let n be the number of zones. The OD matrix $T^{(\tau)}$ is defined as:

$$T^{(\tau)} = \begin{bmatrix} t_{11}^{(\tau)} & t_{12}^{(\tau)} & \cdots & t_{1n}^{(\tau)} \\ t_{21}^{(\tau)} & t_{22}^{(\tau)} & \cdots & t_{2n}^{(\tau)} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1}^{(\tau)} & t_{n2}^{(\tau)} & \cdots & t_{nn}^{(\tau)} \end{bmatrix} \quad (4.1)$$

Where $t_{ij}^{(\tau)}$ is the number of trips started from zone i (origin) to zone j (destination) during the time interval τ .

The origin-destination matrix is a well-established form for the collection of mobility information. It provides a structured way to represent the flow of agents between different zones over a specific time interval. This matrix is particularly useful for urban planning, transportation modeling, and understanding mobility patterns. In the context of this simulation, the OD matrix serves as the primary output, summarizing the mobility behavior of the student population. By aggregating individual agent movements into this matrix, it becomes possible to analyze the overall impact of student mobility on the urban environment.

An additional desire is the ability to keep track of the individual students which allows a more detailed evaluation of the simulation and the resulting student mobility. To achieve this, the individual agent movements are also tracked during the simulation. The simulation framework described in this chapter aims to generate such an OD matrix and agent tracking capabilities by modeling the daily activities of students. These daily activities are constrained by class schedules that drastically influence student behavior.

Before describing the inputs of the simulation, the impact of the τ parameter has to be discussed. The parameter τ indicates the time interval for each origin-destination matrix. It was chosen to make the duration of this interval constant for each origin-destination matrix. The length of such an interval defines the level of detail in the output. A shorter interval duration results in more OD matrices at the output for the same simulation duration, providing more information within the day. The main downside of shortening the interval is an increase in execution time and storage requirements. These storage requirements will be touched upon later in this chapter.

The methodology leverages spatial data (student housing locations) and temporal data (class schedules) to simulate the individual movements. The student housing locations are determined using the methodology in chapter 3. Since the graph from that chapter is essentially the road network of the city, the simulation starts from the same graph representation complemented with the student housing information, campus information and station information as in the previous chapter.

To achieve a flexible schedule input, an intermediate representation of the class schedule is used. This representation holds the amount of students attached to this schedule and then a list of the classes. Each entry within this list is represented by the node corresponding to the class location, the start time of the class and the end time of the class. The campus information in the graph can be used to determine the class location node before aggregation. The start and end time of the class is represented as an integer

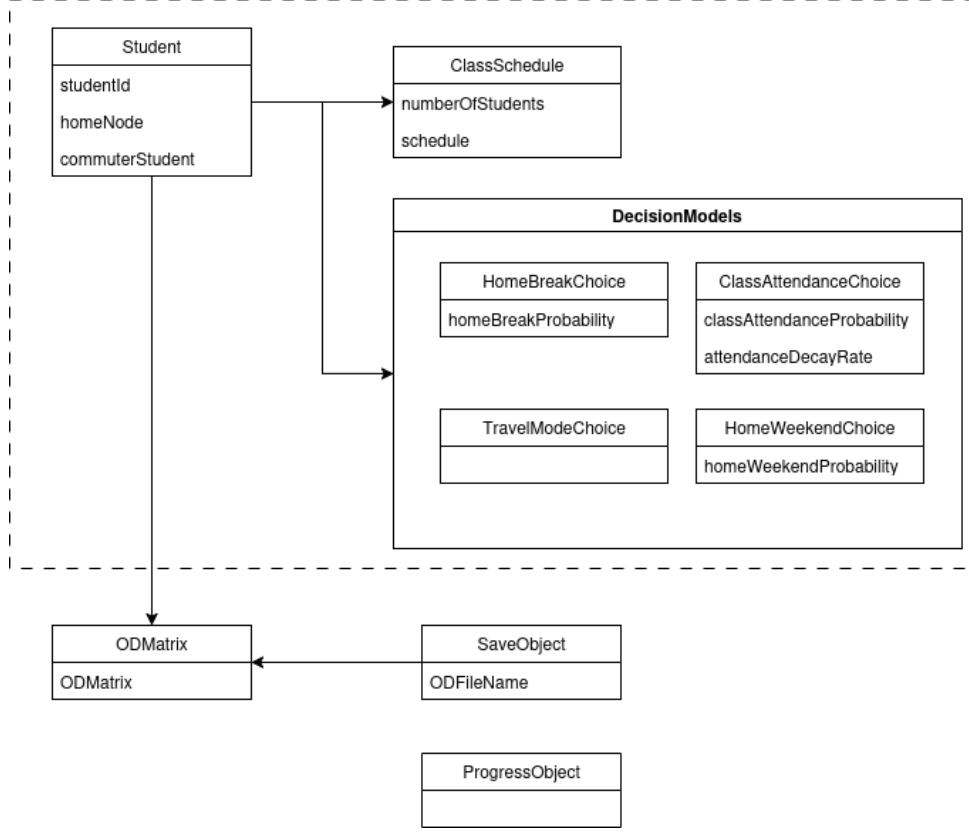


Figure 4.1: Programming class structure of the simulation framework.

representing the time interval in which the class occurs. It is beneficial to make the intervals match with the start and end times of classes. This is an additional constrained on the possible values for the OD matrix intervals.

Let C be the number of classes, v the graph node corresponding with the class location, θ_s the start time and θ_e the end time of the class as integers indicating the interval. Then the schedule can be defined as:

$$S = [(v_0, \theta_{s,0}, \theta_{e,0}), (v_1, \theta_{s,1}, \theta_{e,1}), \dots, (v_C, \theta_{s,C}, \theta_{e,C})] \quad (4.2)$$

Where S is the entire schedule. There is a non-overlapping assumption on the class schedules. It is not possible for a student to be needed in more than one place at the same time. In a real world scenario, students can opt for overlapping classes, especially when elective courses are involved. It is assumed this is a minimal number of students which can be ignored for now, and it could be further developed later.

4.2.1 Simulation Structure

With the input and the output defined, a closer look into the simulation structure itself is now taken. The choice was made for a discrete event simulation, however other frameworks for the extensive version of the agent-based simulation might be more applicable for certain contexts. An example of such a framework is MATsim, which is a framework for agent-based transport simulations. They take into account several mobility factors to improve simulation accuracy. [78] For the proof of concept, where no agent interaction is modelled and travel paths are simplified, the DES approach is feasible. A full implementation of the model will require more in depth scenarios that are more difficult to model using

a discrete event simulation.

Within the DES simulation framework there are different instances that process events either created by the instance itself or another instance. Since there is no interaction yet, each instance will process its own generated events at the proper time. Figure 4.1 shows the entire simulation structure. There are three classes that participate in the DES framework. These are Student, SaveObject and ProgressObject, where student is the main class containing all the logic. The other objects are there to resolve some limitations. The objects within the dotted line are student dependent. The remaining objects are shared among the entire simulation.

Each student is defined by its student ID, which allows tracking of individual students. Each student is assigned a home node and a variable indicating whether the student commutes to class from their hometown and thus does not live in student housing. The student IDs are sequentially increasing for each assigned student. The assignment of the other variables is a bit more complex. To improve accuracy, students in student housing are assigned to nodes closer to their campuses as determined within the student housing location estimation.

Assume an array of remaining students $U = [u_0, u_1, \dots, u_m]$ with m the number of nodes in the graph. Assume the remaining number of commuter students is stored in the variable K_c . Before assigning a student, it is checked if there are still students left to be assigned. Firstly, it is decided if a student is a commuter student. This is done by taking into account the remaining students to be assigned.

$$Pr\{commuter = True\} = \frac{K_c}{\sum_{u=0}^m u_m}, Pr\{commuter = False\} = 1 - Pr\{commuter = True\} \quad (4.3)$$

If the student is a commuter student, the home node is selected from the stations using the station probabilities from the input data in the student housing location estimation. There is of course a guarantee that the students within one node can not go below zero, this is checked at each assignment. When the student is not a commuter student an assignment is made with probabilities depending on the distances towards campus nodes within the schedule. To make sure the probabilities sum to one, they are normalized at the end. Node probabilities are calculated as follows:

$$Pr\{node\ i\} = \frac{u_i}{\sum_{j=0}^C d(v_j, node\ i) + \epsilon} \quad (4.4)$$

$$Pr\{node\ i\} = \frac{Pr\{node\ i\}}{\sum_{j=0}^m Pr\{node\ j\}} \quad (4.5)$$

where ϵ is a small value to avoid division by 0 and d is the great circle distance between two nodes. The array of possible nodes is updated by reducing the selected node by one. Furthermore, if the user is a commuter student, the value K_c is reduced by one as well. The node probabilities are dependent on the distances to all the student's campuses, this is the summation term in equation (4.4). Nodes that lie closer to the student's campuses get a higher probability. The amount of students at each node is also taken into account to keep the spatial distribution of students mostly intact during this assignment step.

The assignment of commuter students can be improved taking into account percentages of commuter

students within a study program. Certain study programs have fewer commuting students than others. This would require data on this level, which can be hard to collect. So for now all students are found to be equal in respect to commuting.

Each student is connected to a class schedule. This schedule is also used in the node assignment. The schedule works as the main activity entity to control the simulation. Right now the simulation is limited to class based activities. Each student has a set of decision models that can be controlled by their preferences. Section 4.2.3 dives deeper in these decision models and how they are defined. Furthermore, the student instance has access to a shared OD matrix to register movements.

The OD matrices tend to be very sparse. Trips often occur to and from the same locations, in this case, the campuses. However, storing this matrix for each time interval during program execution runs into memory limitations. Assume n zones within the city, the OD matrix of one interval then contains n^2 elements. Assuming several steps within an hour, quickly several OD matrices have to be kept track of during execution. To resolve the memory requirements, a save object is added to the simulation that also has access to the origin-destination matrices.

DES simulation happens chronologically based on scheduled events. The save object is scheduled periodically to store the past OD matrices in a different format, after which the object resets the OD matrices for the upcoming intervals. The results are stored in a comma separated file in the format: origin, destination, count, time. The origin and destination represent zones, the count indicates the number of students traveling, and time marks the start of the time interval. While this representation has the potential to need more storage, when a sparse matrix is involved there is an overall gain compared to the OD matrix. Furthermore, it can be written back to main storage, not relying on memory to store the entire OD matrices for an entire semester or year. A final progress object is present within the simulation to indicate when a day is finished, so simulation progress can be tracked. This is achieved by outputting the execution time after each simulation day.

4.2.2 Agent Logic

Each agent within the system follows the same logic that can be influenced by the agent's preferences. The logic dictates which decisions can be made on an agent by agent basis while keeping the rigidity provided by the class schedule. This is essentially the student mobility model. Figure 4.2 shows the entire simulation flow for the students. The simulation structure is quite complex even though only a few options are implemented. An entire implemented system needs to rely on abstraction levels where this complexity is hidden from a potential user. To explain the structure properly, it is explained step by step in detail. A separation is made between the loop section and the finalization steps.

When the simulation starts, on the first Monday of the semester, the student is assumed to be present in their parental home in case of a commuter student or their student accommodation for the other students. The main controlling events for the simulation are the students' classes. This means that most of the events handled by the DES simulation are the start of the class and the end of a class. So first it is checked if the end of the schedule is reached. If this is the case, the agent returns to their home if they are not home yet, and they are removed from the simulation.

When there are still classes left, the student waits until the start of the next class. The student can then decide if they attend the class or decide to skip it. If they skip it, they just wait until the next one where the same decision is asked. However, if the end of the schedule is reached they finalize their simulation.

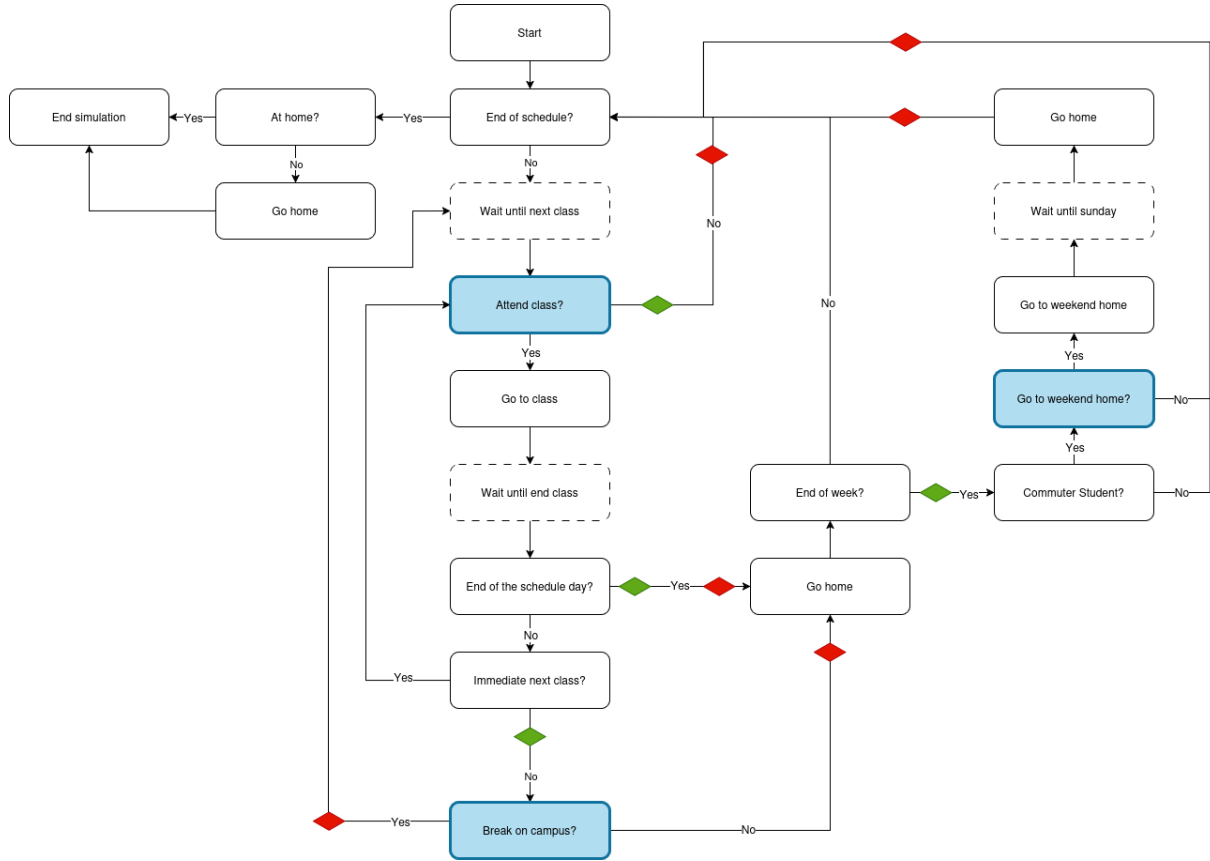


Figure 4.2: Student mobility model: agent simulation logic.

When they decide to attend the class, they move towards the corresponding campus and are assumed to stay there until at least the end of the lecture. In a real world scenario, students might leave the lecture during the breaks. This could be captured by dividing each lecture into their sections while taking into account the dependency between them.

When the class ends and the end of a day is reached within the schedule, the student returns to their home. If however, the student immediately has another lecture, the question is posed if they will attend this lecture. If there is a break scheduled before the next class, the student has to decide on where to spend this break. They can take the break back at home, or they can remain on campus until their next class starts.

When going back home and the end of the week is reached, the students in student housing have to decide if they will return to their parental home (weekend home) for the duration of the weekend. When they go home they start their trip to a train station within a few hours after their last class. The exact departure time is part of the decision process. It is assumed they remain at their weekend home until Sunday evening. Their arrival time in the student city is within a few hours before 10pm. This arrival time is also part of the decision process. Once they have returned to their student city, they resume the normal operation for the coming week. Restricting travel times to after students' last class and Sunday evening is a significant limitation of the approach. However, it can give a first indication of the impact of students returning to their parental home for the weekend.

The current simulation logic has a limited number of possible activities. An agent can go to class, take a break or return home for the weekend. From this list only the classes are specific, it is not specified

what a student does during their break or the weekend. A more comprehensive system should capture additional activities, which could be achieved by extending an existing agent-based mobility framework with the student mobility model. Within figure 4.2, the possible activity blocks are indicated through diamonds. When a green diamond is encountered, the activity block starts and a red diamond indicates the end of the activity block. Since the class activities are rigid for students, they are part of the main simulation structure and are not included in the activity blocks. Within an activity block a student could perform a set of other activities that need to be specified.

Whenever a student makes a movement, the travel time is determined to decide the origin-destination matrix interval in which it needs to be recorded. The trip start time is calculated starting from the activity start time and then deducting the travel time. The trip start time denotes the interval used to record the trip in the ODM. The travel time is mode specific. The student can decide based on environmental factors and preferences which travel mode to use.

In the current simulation flow, four decision models can be determined: class attendance, break at home, weekend at parental home and the travel mode. In figure 4.2, the choice models are indicated by a blue background. The travel mode model is not present in the figure but is applied for each movement. At each of these steps in the simulation, the student makes a decision based on environmental influences and preferences. The decision models are further described in the next section.

4.2.3 Decision Models

The agent within the simulation has several decisions to make. In the simulation logic, these primarily involve choosing whether to attend a class, where to take a break, whether to visit the parental home on the weekend, and which mode of transport to use. A well evaluated approach to modelling these decisions is the use of utility functions. Utility functions are mathematical representations used to model decision-making processes. They quantify the perceived benefit or utility of a particular choice based on various influencing factors. The factors can be characteristics of the user or environmental values. Each decision is associated with a specific utility function, which incorporates relevant factors and their corresponding weights. [8]

The general form of a utility function is:

$$U = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n + \epsilon \quad (4.6)$$

where:

- U is the utility of the choice,
- $x_1, x_2, \dots, x_n$ are the influencing factors,
- $\beta_1, \beta_2, \dots, \beta_n$ are the weights or coefficients representing the importance of each factor,
- ϵ is a random error term capturing unobserved influences.

For each decision model, the utility function is defined based on the relevant factors. The coefficients β can be estimated using data or expert knowledge. Approaches to determine the coefficients include regression, calibration using aggregate data and stated preference surveys. [8] The influencing factors can change over time in the simulation as could happen in a real-world setting. Personal preferences are

introduced through variations in the determined β coefficients. In the default case, each agent has the same decision process and the same preferences. By varying the coefficients, their preferences will differ while keeping the average coefficients correct. The decision-making process then involves comparing the utilities of all available options and selecting the one with the highest utility.

Since data scarcity also remains an issue within this section of the thesis, the utility functions are defined but can not be used to test the simulation principles. Nevertheless, they offer a framework whose parameters can be refined through future research or real-world application. To evaluate the simulation system itself, an alternative simplification of the choice-model is employed. The following subsections describe the utility functions and the simplified choice model for each decision model in detail, including the factors considered and their significance.

For the agents' decisions to reflect the real distribution, a significant number of students is required in the simulation. The more agents present, the more accurately the decisions represent the real scenario. Typical agent counts are in the thousands.

Class Attendance

The first choice dimension is related to the class attendance. For each class, a student can decide whether to attend this class or skip it. This boils down to a binary decision. One way to define a binary utility function is by specifying general utility functions for both options and selecting between them based on their probabilities. However, a simpler approach defines only one utility function for one option. The option is then taken if the utility is greater than 0 and the alternative is taken when the utility is lower or equal to zero. The utility function for the class attendance is defined as:

$$U_{Attend} = \beta_{StCourse} \cdot StCourse + \beta_{IntCourse} \cdot IntCourse + \beta_{Weather} \cdot Weather + \beta_{TTime} \cdot TTime + \epsilon_{Attend} \quad (4.7)$$

$$y_{Attend} = \begin{cases} 1 & U_{Attend} > 0 \\ 0 & U_{Attend} \leq 0 \end{cases} \quad (4.8)$$

in which *StCourse* is a cardinal variable indicating the start time of the course. *IntCourse* represents a cardinal variable indicating how interesting a course is. Determining this variable would require significant effort in the form of surveys. It is less realistic in a real setting, so could be removed as observed variable and could instead fit into the unobserved variables through ϵ_{Attend} . *Weather* is a value indicating the quality of the weather at the travel time. A deterministic method for assigning a value to weather quality is required. *TTime* indicates the travel time from the current location to the correct campus. These influencing factors were selected based on their observability and their potential impact on students' decisions regarding class attendance.

The simplification used for simulation evaluation corresponds to each student having a probability to attend class which decays every week. The longer a semester continues the lower the attendance rate would be. The probabilities and decay rates are selected from a normal distribution with a given mean while guaranteeing the probabilities are between 0 and 1, and the decay rate cannot increase the probability. Figure 4.3 shows an example of the attendance decay starting from a probability of 0.9 with a decay rate

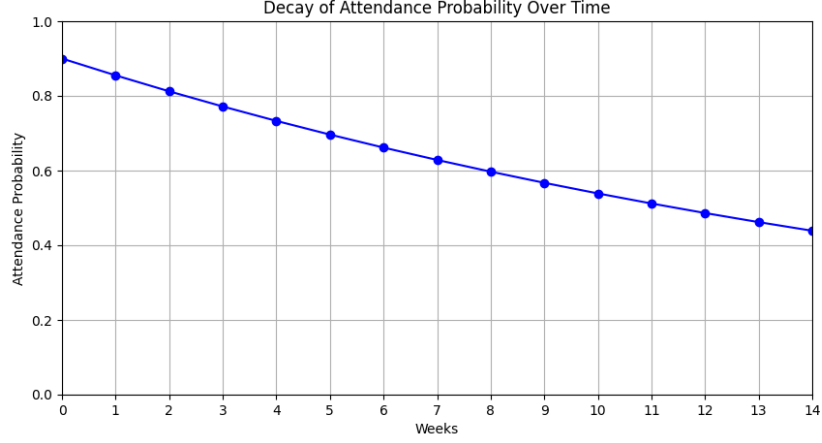


Figure 4.3: Example of attendance probability evolution in the simplified attendance choice model, with a starting probability of 0.9 and a decay rate of 5%.

of 5%. Applicability in a real world setting requires the determination of the correct distribution from real class attendance data.

Home Break

The next choice happens in between classes where the student can decide to go home for the break duration or remain on campus. Again this is a binary decision, so the same interpretation of the utility model is used. The utility function incorporates influences from the duration, travel time and need for social interaction. The function is defined as:

$$U_{HomeBreak} = \beta_{BreakDuration} \cdot BreakDuration + \beta_{TTime} \cdot TTime + \beta_{Social} \cdot Social + \epsilon_{HomeBreak} \quad (4.9)$$

$$y_{HomeBreak} = \begin{cases} 1 & U_{HomeBreak} > 0 \\ 0 & U_{HomeBreak} \leq 0 \end{cases} \quad (4.10)$$

where $TTime$ is the trip time. $BreakDuration$ is a cardinal value expressing the time between classes. $Social$ indicates the need of social interaction for the student. The social interaction is more difficult to model. The evolution of the average social desire and interaction could be determined by conducting surveys. When it is infeasible to add this to the simulation, it could be left out and added to the unobserved variables in $\epsilon_{HomeBreak}$.

The simplified choice model assigns each student a probability of taking a break on campus, randomly drawn from a normal distribution with a specified mean and standard deviation. For a real world application, the correct distribution has to be determined. This distribution can be related to the probability distribution at the output of the utility model. Essentially they should be the same. However, the utility model allows explainability for certain choices and allows correlation between different types of choices.

Weekend at Parental Home

The next decision about the weekend location is similar to the one for the break. Only *Social* is replaced by *Family* in this equation. The *Family* variable indicates the need or desire for social time with family members. Again it can be noted that this is harder to model within the simulation. This could be left out when this is not achievable. The corresponding utility function is:

$$U_{WeekendHome} = \beta_{WeekendDuration} \cdot WeekendDuration + \beta_{TTime} \cdot TTime + \beta_{Family} \cdot Family + \epsilon_{WeekendHome} \quad (4.11)$$

$$y_{WeekendHome} = \begin{cases} 1 & U_{WeekendHome} > 0 \\ 0 & U_{WeekendHome} \leq 0 \end{cases} \quad (4.12)$$

A potential extension of this utility function is a variable indicating if they have social events in their student city during the weekend. For this, a more general activity selection should be present within the simulation. The simplified model is defined in the same way as for the home break decision model.

Travel Mode

The final choice model concerns the travel mode. The travel mode decision involves selecting the mode of transportation for each trip. The problem is not binary and is defined as a multivariate logit model. The utility functions for the available travel modes (Car, Public Transport, Bike, Walk) are defined as follows:

$$U_{Car} = \beta_{TTime} \cdot TTime + \beta_{TCost} \cdot TCost + \beta_{ParkingSpace} \cdot ParkingSpace + \epsilon_{Car} \quad (4.13)$$

$$U_{PublicTransport} = \beta_{TTime} \cdot TTime + \beta_{TCost} \cdot TCost + \beta_{NTransfers} \cdot NTransfers + \beta_{Freq} \cdot Freq + \epsilon_{PublicTransport} \quad (4.14)$$

$$U_{Bike} = \beta_{TTime} \cdot TTime + \beta_{Weather} \cdot Weather + \epsilon_{Bike} \quad (4.15)$$

$$U_{Walk} = \beta_{TTime} \cdot TTime + \beta_{Weather} \cdot Weather + \epsilon_{Walk} \quad (4.16)$$

where *Weather* and *TTime* (travel time) are defined as in previous utility functions. For the car mode, an additional variable is present for the travel cost, *TCost*, and the availability of parking spaces through *ParkingSpace*. The public transport utility function contain some additional variables related to the service delivery. *NTransfers* indicates the amount of transfers needed to complete the entire trip. *Freq* represents the frequency of the public transport for the entire trip.

To calculate the probability between each alternative, the softmax function is used. It is defined to generate values between 0 and 1 where a higher utility score results in a higher probability. The softmax is defined by:

$$P_{mode\ i} = \frac{exp(U_i)}{\sum_{mode\ j} U_j} \quad (4.17)$$

These probabilities are then used to make a decision about the travel mode. A simplified model assigns probabilities for each of the transport modes with the guarantee that they sum to 1. Again a distribution would need to be determined from data or surveys.

$$P_{Car} + P_{PublicTransport} + P_{Bike} + P_{Walk} = 1 \quad (4.18)$$

4.3 Case study

The simulation methodology is tested on a set of students from Ghent University. Firstly a baseline is computed using the simplified choice models from section 4.2.3. Afterwards this baseline is compared against two different scenarios. The goal of this comparison is to showcase the method's capability in evaluating different scenarios for decision-making. The metrics used for comparison and evaluation are origin-destination flows, class attendance rates, and travel mode distributions.

The inputs of the methodology are the student housing locations, the class schedules and distributions used within the choice models. The student housing accommodation estimates from chapter 3 are used to determine the nodes to which students can be assigned. The non-aggregated nodes are used for more granularity in the agent simulation, while making sure the amount of students assigned for simulation within each statistical sector does not cross the meaningful amount of students assigned to the sector by the student housing location estimation. The origin-destination matrix at the output is defined by using the same statistical sectors as zones to limit the amount of different aggregation levels.

The class schedules and their number of enrolled students was collected from university websites. The data from the university groups the amount of students by the number of years they are enrolled within the same study program. There are students that take longer to complete their degree than the standard duration. A decision must be made regarding which year to assign these students, ensuring that the student counts for each schedule accurately reflect the correct number of students. For one study program of three years the students are assigned as follows:

$$N_{S,1} = N_{R,1} + \frac{\sum_{i=4}^{\inf} N_{R,i}}{3} \quad (4.19)$$

$$N_{S,2} = N_{R,2} + \frac{\sum_{i=4}^{\inf} N_{R,i}}{3} \quad (4.20)$$

$$N_{S,3} = N_{R,3} + \frac{\sum_{i=4}^{\inf} N_{R,i}}{3} \quad (4.21)$$

where $N_{S,i}$ represents the amount of students assigned to the schedule year i and $N_{R,j}$ represents the amount of students enrolled for the j^{th} year in the study program. The amount of students will become zero after several enrollment years, making the summation in practice not really infinite. An equivalent expression can be determined for study programs with a different length.

In the evaluation, the included study programs were limited to the bachelor degrees of the Faculty of Engineering and Architecture (FEA). The method is easily adaptable to other faculties or the entire

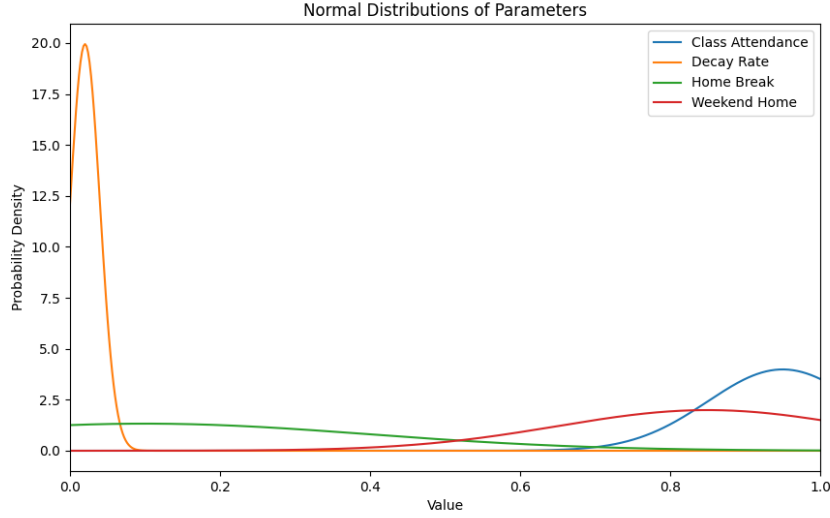


Figure 4.4: Probability distribution used in the different choice models.

university. It only requires additional collection of student numbers which was a manual task within this thesis.

For the baseline, the simplified choice models are used, where the probabilities of the different travel modes are constant for each agent and the remaining probabilities are selected using a normal distribution.

Class Attendance	$N(0.95, 0.1)$
Decay Rate	$N(0.02, 0.02)$
Home Break	$N(0.1, 0.3)$
Weekend Home	$N(0.85, 0.2)$

All values are selected from the range 0 to 1 using the normal distributions to denote the chance a value is selected. Figure 4.4 visualizes the different normal distributions.

The probabilities used for the transport modes are:

$$P\{Car\} = 1/24, P\{Public\ transport\} = 3/24, P\{Bike\} = 2/3, P\{Walk\} = 4/24 \quad (4.22)$$

The distribution and probabilities were selected based on contextual understanding and domain knowledge. It is advised to use the real choice models whenever possible. Additionally, it is important to note that the travel modes are not connected to the travel distance in the simplified model. This makes it possible to reason on averages in travel modes but not on the travel mode of one student in particular.

4.3.1 Baseline

First, the baseline option is explored in detail. The assumptions as described above are evaluated through several figures and metrics. The baseline simulation results are visualized in figure 4.5. The figures show the distribution of students across the city at different times on the first day of the semester. At 8:00 (figure 4.5a), most students are either at their student homes or commuting by train towards the city for their first class. By 12:00 (figure 4.5b), a significant portion of students are concentrated around the campuses, reflecting class attendance during midday. Some students still remain at their homes, since

they decided not to attend their classes or do not have classes in the morning of the first day of the semester. By 16:00 (figure 4.5c), the distribution begins to spread out as students finish their classes. There are still a significant number of students in class. Finally, by 20:00 (figure 4.5d), all classes have finished and all students have returned to their homes.

Another way of visualizing movements of students is by looking at what type of location the student is. Figure 4.6 shows in an interval of one hour the movements students make between different types of locations. Between 8 am and 9 am (figure 4.6a) it is clear that most of the students move from home towards school for classes. Figure 4.6b showcases that during breaks, some students stay on campus for a break, some students return home and others are still in class between the hours of 11 am and 12 pm. This alternative view on the student mobility can be used to infer patterns and to better understand how different spaces are used throughout the day.

The origin-destination (OD) flows for the first day can be visually observed at [flowmap.city](https://app.flowmap.city/public/c66108de-e3c1-4a63-a3ae-65ee011feb29)¹. Since the origin-destination is a sparse matrix, a table or figure showing the OD matrix contains barely any information. The visualization gives an opportunity to evaluate the different flows by selecting zones of interest or selecting periods of interest. Within this thesis, it becomes more logical to analyze the number of students at a campus over time to see movements to and from this node. Figure 4.7 shows the amount of students at campus Ardoyen. This campus is used quite often within the faculty. It becomes clear that the amount of students increases when the classes start, which is equivalent to students moving from their student houses towards the campus. There is a smaller reduction around the lunch hours, this can be explained by students going home in between classes, fewer classes around lunch and students moving toward other campuses for their next classes. When the amount of classes picks up again in the afternoon, an increase in students can be seen again. By the end of the day, when more and more classes end, the number of students on campus reduces until all students are home again.

The attendance rates over the semester are shown in figure 4.8. The decay introduced in the decision model can be noticed over the entire semester. The attendance rates decrease slowly over the passing weeks. There is fluctuation in subsequent days even within one week due to the randomization elements in the decision process. The more agents within the simulation, the more the average attendance will correspond to the decay function with the average attendance probability and the average decay rate.

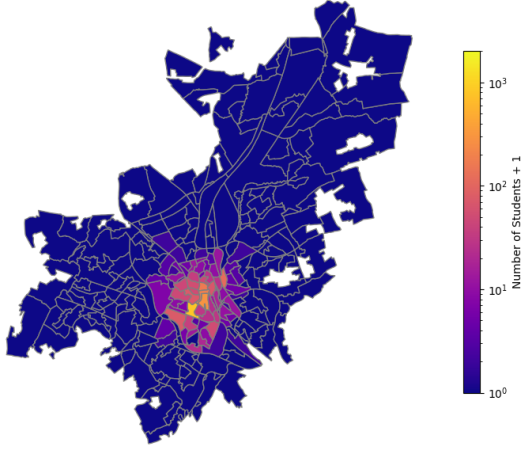
To better understand the simulation's dynamics, it is insightful to follow the path of an individual agent throughout a single day. This allows for a detailed examination of the decisions made by the agent and the resulting movements. Figure 4.9 visualizes the path of a randomly selected agent during the first day of the semester. A warmer travel path indicates travel later in the day. The agent starts their day at their student accommodation and travels to campus for their first class. After their first class, they move towards a different campus for the next class. When the classes end on this campus, the agent moves back home for the evening.

By analyzing individual agent paths, it becomes possible to identify patterns and anomalies in the simulation. This can help refine the choice models and improve the overall accuracy of the simulation. Furthermore, it gives more insight about the exact routes traveled and used for travel time calculations.

The results clearly show that the simulation methodology itself is a valid approach for simulation student mobility, but calibration of the choice models is required. It is limited by the choice models. An accurate choice model will result in an accurate simulation. Overall, the baseline simulation provides a reasonable

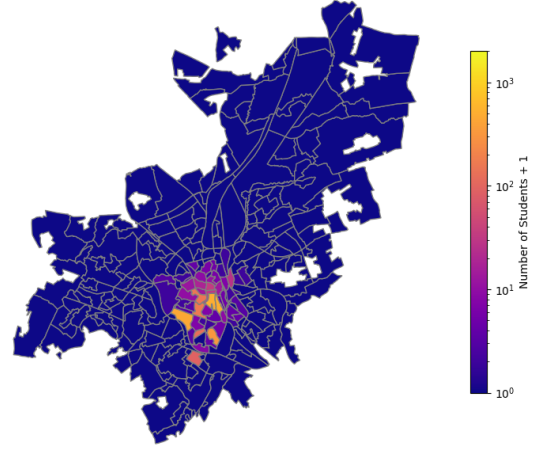
¹<https://app.flowmap.city/public/c66108de-e3c1-4a63-a3ae-65ee011feb29>

Student Distribution in Gent, Belgium at 2024-09-23 08:00:00



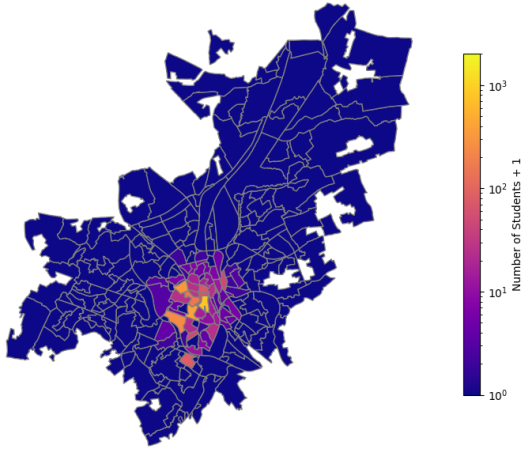
(a) 8:00

Student Distribution in Gent, Belgium at 2024-09-23 12:00:00



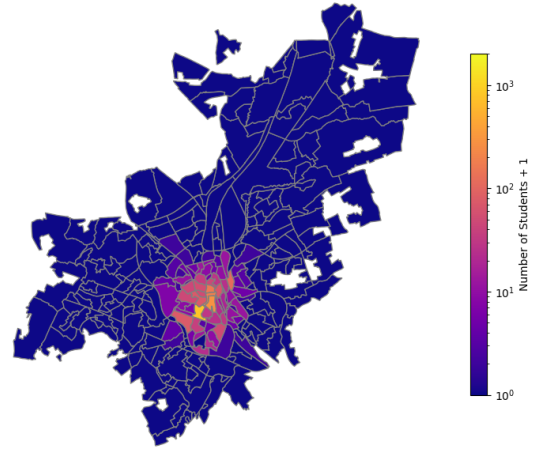
(b) 12:00

Student Distribution in Gent, Belgium at 2024-09-23 16:00:00



(c) 16:00

Student Distribution in Gent, Belgium at 2024-09-23 20:00:00



(d) 20:00

Figure 4.5: Student spatial distribution at various times on the first day of the semester.

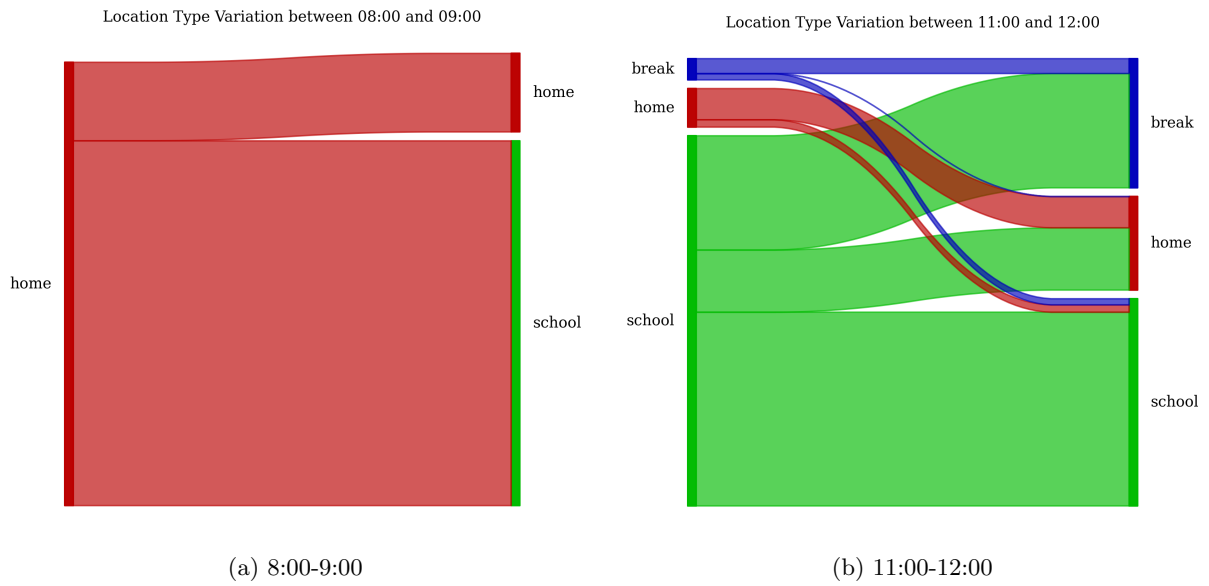


Figure 4.6: Changes in student location types across different hours on the first day of the semester: break indicates break on campus

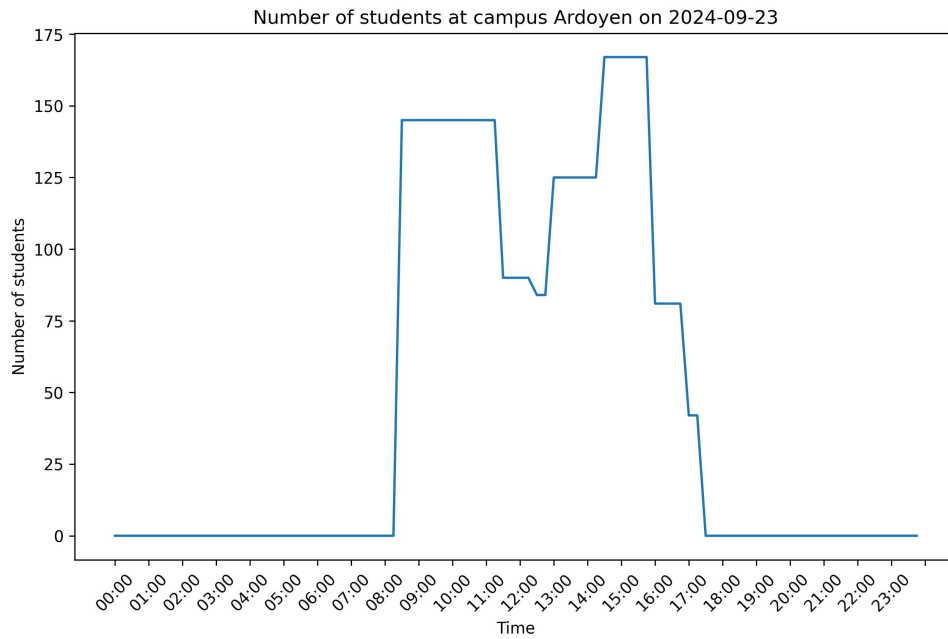


Figure 4.7: Number of students on the Ardoyen campus over time during the first day of the semester.

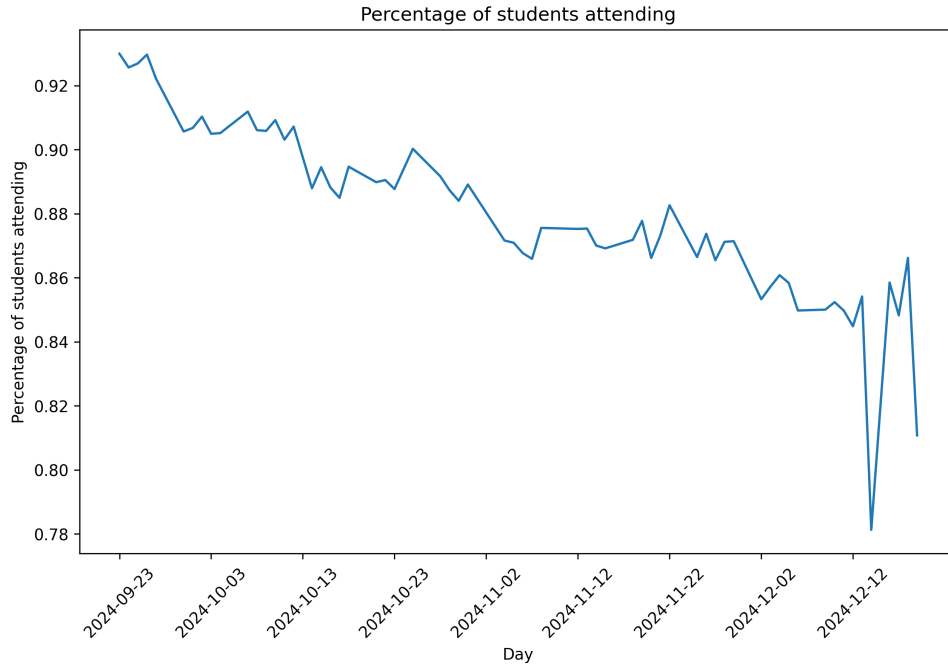


Figure 4.8: Class attendance rate throughout the semester.

approximation of student mobility patterns, given the simplified assumptions. Some work is still needed to verify the distributions within the simplified choice models. The results highlight the potential of the proposed methodology to capture key mobility trends while leaving room for further refinement and validation. To showcase the ability to simulate different scenarios, two scenarios are simulated and compared against the baseline. The first is the scenario where all students come to class. Secondly, a comparison is made with a random node selection for student housing, thus ignoring the dependence on campuses within the simulation steps.

Another point to consider is that each different simulation with the same parameters will result in slightly different results. This can be attributed to the stochastic elements within the simulation model. These stochastic elements provide more realism. To make sure the elements perform the same in subsequent simulations, a seed can be given at the start.

4.3.2 Full Attendance

In this scenario, it is assumed that all students attend every class without exception. This eliminates the variability introduced by the class attendance decision model. The goal is to evaluate the impact of full attendance on the overall mobility patterns and compare it to the baseline scenario.

The results of the full attendance scenario are visualized in figure 4.10a. The distribution of students across the city shows a higher concentration around campuses during class hours compared to the baseline. This is expected, as all students are now attending their classes. However, there are still some students at home because they do not have class at this time. The OD flows also indicate increased movement between student accommodations and campuses. The concentration of students at campus Ardoyen also increased and is now only influenced by the amount of students that have class there and whether they stay on campus during their break. This can be seen in figure 4.10b. The attendance rates over the semester are constant at 100%.

Agent 1 path on 2024-09-23



Figure 4.9: Path of a single agent during the first day of the semester: warmer colors indicate trips later in the day.

Student Distribution in Gent, Belgium at 2024-09-23 12:00:00

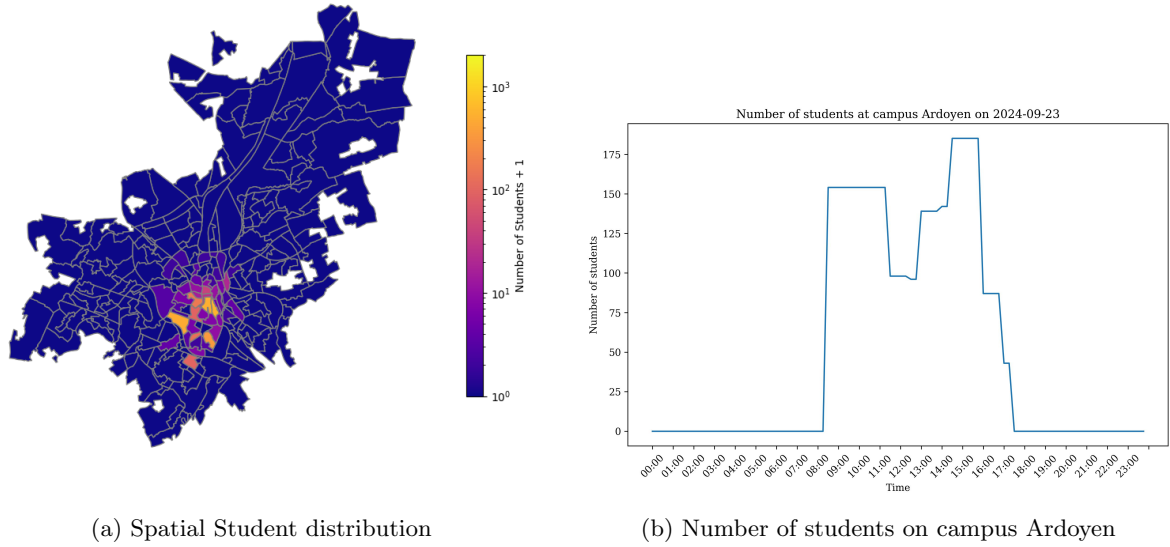


Figure 4.10: Results of the full attendance scenario on the first day of the semester.

4.3.3 Randomized Node Choice

In this scenario, the assignment of students to nodes is randomized, disregarding the proximity to campuses. This tests the sensitivity of the simulation to the spatial distribution of students. The student housing locations remain the same on the input. But afterwards when deciding in which student housing a student lives, the distances to their campuses and the number of students left at each node are not used to determine the node probabilities.

Figure 4.11 shows the student housing distribution for the randomized node selection. This distribution is more evenly dispersed compared to the baseline. However, the differences are small. This can be attributed to the campus distance factor not playing a large role in node selection for the simulation. The attendance rates remain similar to the baseline as these are not directly influenced by the node assignment.

4.4 Possible Improvements

During this chapter some possible improvements were already mentioned. During this section, the goal is to summarize and extend these improvements to give a complete picture of the possibilities with student simulation.

A first possible improvement is to implement the simulation using a more advanced and senior library, such as MATSim (Multi-Agent Transport Simulation). These libraries are still event based to discretize the simulation and gain good performance while providing an additional abstraction layer that improves usability. The libraries also offer additional functionalities and robust infrastructure, while keeping it flexible for the integration of custom models or scenarios. Some of these tools also provide visualization support and integrations with GIS data. Integration with existing code bases is often possible. Using existing libraries makes the solution more standardized but gives away some control over the simulation details.

Student Distribution in Gent, Belgium at 2024-09-23 00:00:00

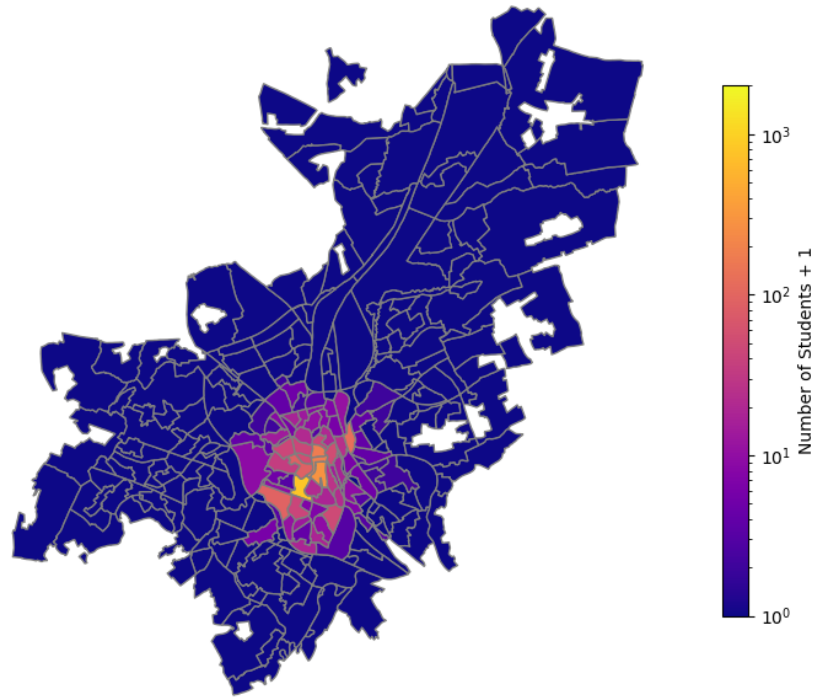


Figure 4.11: Spatial student distribution using random assignment to nodes.

Another improvement would be the addition of social networks within the simulation. Students and in a more general case people make decisions based on their social interactions with friends, family, neighbors and strangers. Also, mobility times and mode choices are directly dependent on the social interactions. A simulation where these social interactions are modelled, could achieve more accuracy. There is however a downside to the social modelling. It is difficult to capture this social information. This would require surveys to gather as much data as possible. After that, this data has to be structured to be able to deduct complete social networks for the simulation.

Not every course is equally interesting and not every student is equally interested in the same course. Including measures of interest could possibly improve the accuracy of the simulations. These are also difficult to measure and would require surveys of students. Furthermore, when courses in a curriculum change, a recollection of these interest measures need to be done. This makes it less suitable in real-world scenarios except when measuring systems are continuously in place.

Weather influences many decisions from transport modes to attendance. It is beneficial to add a weather model to the simulation. When simulating the future, weather predictions need to be added based on previous data increasing complexity. Weather predictions are difficult to get accurately. Recently techniques like machine learning are used to predict weather. A solution would be to simulate different weather scenarios to avoid inaccurate simulations through weather input.

Some improvements are part of the complete utility model proposed in section 4.2.3. As mentioned before, they could be left out of the utility functions to employ a more basic model. The use of the utility model is already a large improvement over the simplified decision models. To make full use of the utility model, the correct parameters need to be determined. These parameters should be transportable to

similar urban environments. However, when the metropolitan situation differs significantly, recalculation of the coefficients is needed.

Agents within a simulation travel at different speeds instead of the average speed of the travel mode. To account for this, speed preferences could be added to the agent's information to more accurately predict the travel times. Heterogeneous agent profiles improve the realism in the simulation. This could be achieved by grouping students into types.

Lastly, the activity model can be worked out for students. This would acquire student movements outside the class hours and improve the extent to which the simulation could provide insights on mobility patterns and can guide policy decisions. It could be achieved within a broader simulation framework that integrates general activity models, with the constrained student mobility model added specifically for student simulation.

4.5 Conclusion

This chapter introduced a student mobility model that leverages spatial housing data and temporal class schedules. The model is applied in a simulation framework for evaluation. The proposed framework employs a discrete-event simulation (DES) approach, with agents representing individual students. The simulation outputs include origin-destination matrices and individual agent tracking, providing insights into mobility patterns and their impact on urban environments.

The methodology was tested on a case study involving students from Ghent University. The baseline simulation demonstrated the validity of the approach, capturing key mobility trends such as class attendance, travel mode distributions, and spatial-temporal student distributions. Two additional scenarios, full attendance and randomized node assignment, highlighted the flexibility of the framework in evaluating different mobility scenarios.

While the results are promising, the simulation's accuracy is heavily dependent on the choice models and input data. Simplified decision models were used for evaluation due to data scarcity, but the utility-based decision framework offers a robust foundation for future refinements. Improvements such as integrating advanced simulation libraries, incorporating social networks, weather models, and heterogeneous agent profiles, as well as extending the activity model, can further enhance the simulation's realism and applicability.

In conclusion, the proposed student mobility model provides a valuable tool for understanding and analyzing student mobility. It offers a flexible framework that can be adapted to various urban contexts and scenarios, supporting data-driven decision-making for urban planning and transportation policies. However, further validation and refinement of the choice models are essential to ensure the simulation's reliability and accuracy in real-world applications.

4.6 Validity

The simulation's validity is checked using the same approach applied in chapter 3 acquired from design science research (DSR). [28] The simulation is evaluated against threats to validity: internal validity, external validity, construct validity and conclusion validity.

External validity is the validity of applying the conclusions in a different context outside this study. The simulation framework itself can be implemented for several cities and contexts quite easily. It does

require the determination of the correct utility functions. These utility functions are context-dependent and need to be calibrated for each specific urban environment. The calibration process involves collecting data on student mobility patterns, preferences, and environmental factors, which can be challenging in data-scarce regions. However, the flexibility of the framework allows it to adapt to different urban settings, provided that sufficient data is available for calibration. The generalizability of the simulation framework is supported by its modular design, which separates the core simulation logic from the input data and decision models. This modularity ensures that the framework can be extended or modified to accommodate different urban contexts, transportation systems, and student populations. The framework's external validity is limited by the assumptions made in the choice models and the availability of input data. The simulation framework can also easily be adapted for scenarios with different population groups and different types of schedules.

Construct validity concerns the metrics used within the evaluation. The metrics used in this simulation include origin-destination flows, class attendance rates, and travel mode distributions. These metrics are well-established in transportation and mobility studies, providing a robust foundation for evaluating the simulation's outcomes. However, the construct validity is limited by the assumptions made in the simplified choice models and the lack of real-world data for validation. The metrics are suitable for analyzing general trends and patterns but may not capture the full complexity of individual student behaviors. Future work should focus on refining these metrics for the specific use case and incorporating additional data sources to enhance the construct validity of the student mobility model.

Internal validity is the extent to which a piece of evidence supports a claim about cause and effect within the context of the study. The simulation framework assumes that the input data and choice models accurately represent the real-world behavior of students. However, the simplified choice models used in this study introduce potential biases and limitations. For instance, the decay function for class attendance is an approximation and may not fully capture the complexities of student behavior over a semester. Additionally, the lack of real-world validation data limits the ability to confirm the causal relationships modeled in the simulation. Future work should focus on collecting empirical data to validate the assumptions and refine the choice models, thereby enhancing the internal validity of the simulation framework.

Conclusion validity entails the validity of the methods to draw reasonable conclusions. The simulation framework provides a structured approach to modeling student mobility, leveraging well-established methodologies such as discrete-event simulation and utility-based decision models. The conclusions drawn from the simulation are reasonable within the context of the assumptions and input data. However, the lack of real-world validation data limits the ability to confirm the accuracy of the simulation's predictions. Despite these limitations, the simulation offers valuable insights into student mobility patterns and serves as a robust tool for exploring various scenarios and their potential impacts on urban environments.

Chapter 5

Societal Impact

To recapitulate, the amount of students within a student city can be highly significant. For example, within the city of Ghent, there are 87.638 students for a population of 270.473 people. [67] Not all students are counted in the population, but about 25% of the people within the city on a weekday are students. These numbers already indicate the immense size of the student population. The student city faces numerous challenges because of this. Urban mobility is a known problem that remains one of the main challenges facing cities. [70] Past research efforts and solutions to the mobility problem focused on the broader population, but rarely on a specific subgroup. This thesis main societal contribution is the ability for city councils to be able to reason more objectively about student influenced decisions. Additionally, existing simulation frameworks can be improved by incorporating the student mobility model. However, the broader societal impact extends further. The goal of this chapter is to contextualize the research into society, discussing the research impact through a societal reflection.

Overall health is influenced by numerous factors. One factor that can promote overall health is the choice of sustainable transport modes. Healthy transport modes have been shown to reduce depression, anxiety, stress and a variety of other health conditions. They also help reduce air pollution which is closely linked to health outcomes. Improved air quality can lower the risk of health issues. [15] The proposed simulation methodology can contribute by providing insight into mobility flows and travel modes for the specific student population. This information could support decision-makers in implementing more targeted and evidence-based measures promoting sustainable and active transport. Student mobility often falls through the cracks of urban systems, despite being a significant population group. Collecting information about this subgroup also improves current existing decision processes regarding active mobility.

Education itself may also be impacted by the proposed methodologies. One of the goals in education is to achieve inclusive and equitable quality education. [75] Analyzing student attendance through mobility patterns can provide insights into the quality of education. Lower attendance rates are often linked to course quality or student interest. [47] Student mobility information, more specifically attendance, can guide quality improvement efforts within higher education institutions by highlighting less-attended courses, which may indicate underlying quality issues.

Achieving equitable education remains a challenge despite consistent improvements. [2] One factor contributing to equity is travel time to and from educational institutions. Travel time can affect student

attendance, which can be linked to academic performance. [47] A student living closer to campus may have more time to spend studying or doing different activities. Insights into inequalities in student travel patterns can inform measures aimed at addressing them. Student housing information can support studies focused on urban accessibility. Broader inequalities can also be studied through human mobility. Gender inequalities can be linked to experiences of sexual harassment. [22] Information about infrequently traveled routes can help law enforcement identify potential risk areas for sexual harassment and other crimes.

As indicated by the 15-minute city concept, accessibility can improve inclusivity, safety and sustainability. [48] For accessibility measurements, information about movements and student housing locations is essential. Additionally, travel times and mode choices can be derived from the simulation output. Travel times can also be used to evaluate accessibility to public transport. In a full activity-based system, accessibility to other services could also be incorporated and evaluated. Accessibility in urban areas is a metric that helps policymakers decide how public spaces are utilized. For example, decisions about which streets are converted into bicycle streets are influenced by mobility patterns.

Student housing locations can be related to the quality of accommodation within a neighborhood. This can provide insights into the quality and safety of student housing. Higher educational institutions play a significant role in innovation as they serve as knowledge hubs for the city. [62] Student housing infrastructure projects can be guided by the student housing location estimates as well as the mobility simulation. Understanding the relationships students prefer with regard to points of interest (POIs) can help identify areas of interest for student housing development that accommodate to the students preferences, thus also highlighting investment opportunities. Mobility patterns can further guide efforts to improve transportation infrastructure. The simulation tool gives a model to evaluate these mobility patterns.

Data is one of the main bottlenecks in student-related studies. Often, at the start of such studies, there appears to be sufficient available data. However, as the study progresses, data limitations become more apparent. This thesis addresses these challenges by providing methodologies that rely on limited data, reducing the need for comprehensive data collection systems. This lowers the data collection burden for cities and enables earlier, lower-cost decision-making. When data is available, it is often related to mobility on the scale of the entire city population. As mentioned at the beginning of this chapter, students are not always explicitly included in this population. The proposed methodology offers a way to evaluate a specific subset of the population. This goes beyond just the student population and includes other subgroups within broader mobility data. Insights into these subpopulations can provide more targeted information for decision-makers.

The use of very limited and aggregated data also reduces privacy risks. Reducing the requirement for detailed, individual-level data is a positive development, as it also improves ease of use with respect to privacy regulations. Finally, it should be noted that the methodology is broad in the sense that similar case studies can easily be conducted in other cities. This contributes to impact on a scale larger than the city level. Additionally, the methodologies can be applied in other fields where similar problems of data scarcity arise. The student housing location estimation model can also be used to estimate different types of spatial distributions using appropriate influence factors. The simulation, guided by both a rigid, schedule-based component and a more flexible component, can be applied in any context where a schedule defines mobility, provided that the choice models are selected and adapted to the specific use case.

Chapter 6

Conclusion and Future Research

Throughout this work, two methodologies were proposed that have the ability to contribute to the field of student intra-city mobility. The first, a student housing location estimation, addresses the data scarcity around student housing by providing a spatial distribution that is calibrated using aggregated data. Secondly, these student housing locations are combined with class schedules in a student mobility model that can be employed in agent-based simulations to study student mobility patterns. By focusing on the student population within this study, it contributes to research into this less-studied and interesting population group. Furthermore, the methods can be applied in different contexts for different subsets of the population providing alternatives when heavily data-dependent studies are not an option. The case study on the student city of Ghent has shown practical usability.

By incorporating proximity to campuses, train stations and city centers as independent spatial influence factors, the student housing location estimation effectively captured key determinants of student preferences. This indicates that spatial distributions can be deducted by determining some approximately independent influence factors. Furthermore, the dependence on these influence factors can be easily determined from heavily aggregated survey data using discretization and a genetic heuristic. It allows for incorporating constraints on the distributions to achieve a more explainable result. Error metrics were used to compare the generated distributions to survey-based ground truth data. The use of the genetic algorithm introduces a data-driven mechanism for tuning the model without requiring exhaustive ground truth data sets. By using the road network as the assignment level, housing density is approximated even better. The use of the road network also allows for aggregations on different spatial abstraction levels like districts or statistical sectors.

The application on the city of Ghent and Kortrijk shows the framework's flexibility and validity, while also emphasizing the need for localized influence factor determination and calibration for cities with a different urban structure. The estimation approach is especially viable for cities and academic institutions that try to understand student housing locations. Furthermore, it is valuable for studies requiring student housing location data as the input.

While the model has shown to approximate student housing locations with an error of less than 10%, some limitations persist. The methodology does not capture all influence factors for student housing, which reduces the correlation with real-world scenarios. Additionally, when applying the model to a completely

different context apart from students, the dependence factors need to be re-evaluated and calibrated. Also, finer aggregations are not necessarily equally accurate as the original calibration aggregation. These limitations highlight the need for further refinement and contextual adaptation of the methodology to improve its generalizability and accuracy.

An answer is formulated for RQ1, that corresponds to the student housing location estimation. *How can the spatial distribution of student housing be estimated using incomplete or aggregated data?* By dividing the problem into approximately independent intensity distributions for different influence factors and by using the road network to assign the number of students. *RQ1.1: What approximately independent spatial factors can be used in a student housing location estimation?* Student housing is influenced by proximity to campuses, train stations and the city center. *RQ1.2: How can the spatial dependency functions be inferred from aggregated data?* By using a discretization approach combined with some distribution constraints like non-increasing functions. A heuristic approach can be used to determine the discrete intensity values from survey data.

The second contribution is a student mobility model that integrates class schedules, choice models and estimated housing data to produce mobility insights. This mobility model is employed in an agent-based simulation framework. The simulation framework demonstrates how rigid class schedules can be integrated into a broader simulation. It is clearly indicated in the modular design where it can be extended by a broader activity-based simulation. Core patterns within student mobility, such as peak hours, travel mode choices and attendance variability are captured through the simulation. Several opportunities exist to represent these patterns, for example agent-paths for individual movement patterns, student counts over time for attendance and peak hour variability, and origin-destination matrices for student flows. The simulation thus captures both temporal rhythms and spatial movements of students.

Agents within the simulation follow stochastic rules informed by the choice models. These choice models incorporate factors such as travel time, travel cost, and personal preferences to determine the likelihood of selecting a particular mode of transportation or route. By integrating these probabilistic rules, the simulation captures the variability and heterogeneity in student mobility behavior. The utility functions presented offer a reliable way of representing the choice model. The simplified choice models, where agents sample from choice distributions, serve as an alternative but come with reduced explainability.

The simulation framework can be used to evaluate different scenarios. These evaluations can support decisions regarding urban planning, such as optimizing public transportation routes, improving infrastructure for active modes of transport, or assessing the impact of policy changes on student mobility. Again there is a low dependency on high-resolution input data.

There are also some limitations to this approach. First of all the choice models need to be calibrated using survey data, which might not be available. The current simulation only addresses classes as potential activities and is limited in differentiation between different agents. This could be solved by employing the student mobility model into a population-broad simulation framework that supports more activities.

An answer is formulated for RQ2, which corresponds to the agent-based student simulation framework. *How can student mobility be realistically modelled and simulated using estimated housing locations and class schedules?* By integrating estimated housing locations with class schedules into a student mobility model employed in an agent-based simulation framework that uses choice models to guide student behavior. *RQ2.1: What are the choice models needed for a student mobility simulation?* Choice models that incorporate factors such as travel time, travel cost, and personal preferences for class attendance, break

location, travel mode and weekend location choices are needed to realistically simulate student mobility. *RQ2.2: How can class schedules help guide student mobility simulation?* Class schedules provide temporal constraints and activity locations that serve as the primary drivers for student mobility patterns and behavior within the simulation.

To finalize this chapter, future research directions are suggested. The student housing location estimation could be improved by adding an unobserved dependency to the observed influence factors similarly to utility functions. This unobserved intensity factor can capture broader characteristics of the spatial distribution while also ensuring explainability through the observed factors. It has to be guaranteed that the unobserved factor will not become dominant over the observed factors, as this would reduce the interpretability of the model. Additional mathematical work needs to be done to understand if there is an optimal solution for the intensity functions that can be calculated in reasonable time. Moreover, machine learning techniques could be explored for the student housing location estimation when enough data is collected.

Future research should investigate the impact of campus sizes and station sizes on the accuracy of the student housing location estimation. This includes determining whether larger or smaller campuses and stations influence the spatial distribution differently and how these variations affect the model's performance. Additionally, research should focus on identifying the exact set of influence factors that contribute to the estimation, as this could refine the methodology and improve its accuracy.

Another important avenue for future research is exploring the generalizability of the methodology across different urban contexts. Specifically, it should be examined whether the output of the methodology for one city can be applied to other cities with similar urban characteristics. This would involve identifying key urban context similarities and testing the transferability of the model to ensure its broader applicability.

The student mobility model needs to be validated in future research using known travel patterns and using survey data to calibrate the choice models. Empirical data regarding class attendance or student registration would improve model accuracy and robustness. This would require deeper collaboration between higher educational institutions and municipalities. Future research could evolve the currently limited simulation into a full activity-based model where the agents perform more daily tasks than attending class. This would require developing realistic activity schedules and incorporating additional data sources. Another approach is to incorporate the student mobility model into a broader simulation framework for both extended activities and choice calibration.

To improve the behavioral modelling, future work could explore reinforcement learning and social influence on decisions. These would respectively allow agents to adapt over time and make the decision more realistic. Reinforcement learning would allow agents to adapt their behavior over time based on experiences, leading to more dynamic and realistic decision-making processes. Social influence modeling could incorporate peer effects and group dynamics, making the decision-making process more reflective of real-world scenarios where individuals are influenced by the behavior and preferences of others.

While this thesis is focussed on students, the core methodologies are adaptable to other population subgroups whose mobility patterns are shaped by routine activities and spatial constraints. Each subgroup would require context-specific influence factors and behavioral rules, necessitating custom calibration of both the estimation and simulation components. Extending the methodology in this direction could support comparative mobility studies and inclusive urban policy planning.

Bibliography

- [1] Timur Abbasov et al. “The 15-minute city quantified using human mobility data”. In: *Nature Human Behaviour* 8.3 (2024), pp. 445–455.
- [2] Mel Ainscow et al. *Developing equitable education systems*. Routledge, 2013.
- [3] STAFFAN Algers et al. “Stockholm model system (sims): Application”. In: *7th World Conference of Transportation Research*. Sydney Australia. 1995, pp. 16–21.
- [4] Zaheer Allam et al. “The ‘15-Minute City’ concept can shape a net-zero urban future”. In: *Humanities and Social Sciences Communications* 9.1 (2022), pp. 1–5.
- [5] Syed Muhammad Asad et al. “Mobility prediction-based optimisation and encryption of passenger traffic-flows using machine learning”. In: *Sensors* 20.9 (2020), p. 2629.
- [6] Hugo Barbosa et al. “Human mobility: Models and applications”. In: *Physics Reports* 734 (2018), pp. 1–74. URL: <https://www.sciencedirect.com/science/article/pii/S037015731830022X>.
- [7] Tom Bellemans et al. “Implementation Framework and Development Trajectory of FEATHERS Activity-Based Simulation Platform”. In: *Transportation Research Record* 2175.1 (2010), pp. 111–119. URL: <https://doi.org/10.3141/2175-13>.
- [8] Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press, 1985.
- [9] Sharmintha Bera and K. V. Krishna Rao. *Estimation of origin-destination matrix from traffic counts: the state of the art*. 2011. URL: <http://hdl.handle.net/10077/6182>.
- [10] Leo Van den Berg and Antonio Paolo Russo. “The Student City. Strategic Planning for Student Communities in EU Cities”. eng. In: European Regional Science Association (ERSA), 2003. URL: <https://hdl.handle.net/10419/116215>.
- [11] Geoff Boeing. *Modeling and Analyzing Urban Networks and Amenities with OSMnx*. 2024. URL: <https://geoffboeing.com/publications/osmnx-paper/>.
- [12] Eric Bonabeau. “Agent-based modeling: Methods and techniques for simulating human systems”. In: *Proceedings of the National Academy of Sciences* 99.suppl.3 (2002), pp. 7280–7287. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.082080899>.
- [13] Patrick Bonnel, Mariem Fekih, and Zbigniew Smoreda. “Origin-Destination estimation using mobile network probe data”. In: *Transportation Research Procedia* 32 (2018), pp. 69–81. URL: <https://www.sciencedirect.com/science/article/pii/S2352146518301649>.
- [14] John Lawrence Bowman. “Activity based travel demand model system with daily activity schedules”. PhD thesis. Massachusetts Institute of Technology, 1995.

- [15] C40 Cities. *Why shifting to green and healthy transport modes delivers vast rewards for cities*. 2024. URL: https://www.c40knowledgehub.org/s/article/Why-shifting-to-green-and-healthy-transport-modes-delivers-vast-rewards-for-cities?language=en_US (visited on 04/28/2025).
- [16] Francesco Calabrese et al. “Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area”. In: (2011).
- [17] Francesco Calabrese et al. “Understanding individual mobility patterns from urban sensing data: A mobile phone trace example”. In: *Transportation Research Part C: Emerging Technologies* 26 (2013), pp. 301–313. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X12001192>.
- [18] Pamela Charbonneau, Laura C. Johnson, and Jean Andrey. “Characteristics of University Student Housing and Implications for Urban Development in Mid-sized Cities”. In: *Canadian Journal of Urban Research* 15.2 (2006), pp. 278–300. URL: <http://www.jstor.org/stable/26192464> (visited on 04/07/2025).
- [19] Zesheng Cheng et al. “Integrating Household Travel Survey and Social Media Data to Improve the Quality of OD Matrix: A Comparative Case Study”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.6 (2020), pp. 2628–2636.
- [20] Andrew Crooks, Christian Castle, and Michael Batty. “Key challenges in agent-based modelling for geo-spatial simulation”. In: *Computers, Environment and Urban Systems* 32.6 (2008), pp. 417–430. URL: <https://www.sciencedirect.com/science/article/pii/S0198971508000628>.
- [21] Carles Ferrer-Ortiz et al. “Barcelona under the 15-Minute City Lens: Mapping the Accessibility and Proximity Potential Based on Pedestrian Travel Times”. In: *Smart Cities* 5.1 (2022), pp. 146–161. URL: <https://www.mdpi.com/2624-6511/5/1/10>.
- [22] Olle Folke and Johanna Rickne. “Sexual Harassment and Gender Inequality in the Labor Market*”. In: *The Quarterly Journal of Economics* 137.4 (May 2022), pp. 2163–2212. URL: <https://doi.org/10.1093/qje/qjac018>.
- [23] Federica Gaglione et al. “Urban accessibility in a 15-minute city: a measure in the city of Naples, Italy”. In: *Transportation Research Procedia* 60 (2022), pp. 378–385. URL: <https://www.sciencedirect.com/science/article/pii/S2352146521009509>.
- [24] Stéphane Galland et al. “Multi-agent simulation of individual mobility behavior in carpooling”. In: *Transportation Research Part C: Emerging Technologies* 45 (2014), pp. 83–98. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X14000035>.
- [25] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. “Understanding individual human mobility patterns”. In: *nature* 453.7196 (2008), pp. 779–782.
- [26] Rares Halbac-Cotoara-Zamfir et al. “Uncovering Demographic Trends and Recent Urban Expansion in Metropolitan Regions: A Paradigmatic Case Study”. In: *Sustainability* 12.9 (2020). URL: <https://www.mdpi.com/2071-1050/12/9/3937>.
- [27] Alison J Heppenstall et al. *Agent-based models of geographical systems*. Springer Science & Business Media, 2011.
- [28] Alan R. Hevner et al. “Design Science in Information Systems Research”. In: *MIS Quarterly* 28.1 (2004), pp. 75–105. URL: <http://www.jstor.org/stable/25148625> (visited on 04/01/2025).

- [29] Christian Hofer, Georg Jäger, and Manfred Füllsack. “Generating realistic road usage information and origin-destination data for traffic simulations: augmenting agent-based models with network techniques”. In: *International Conference on Complex Networks and their Applications*. Springer. 2017, pp. 1223–1233.
- [30] Kate Hosford, Jeneva Beairsto, and Meghan Winters. “Is the 15-minute city within reach? Evaluating walking and cycling accessibility to grocery stores in Vancouver”. en. In: *Transportation Research Interdisciplinary Perspectives* 14 (June 2022), p. 100602.
- [31] Phil Hubbard. “Geographies of Studentification and Purpose-Built Student Accommodation: Leading Separate Lives?” en. In: *Environment and Planning A: Economy and Space* 41.8 (Aug. 2009), pp. 1903–1923.
- [32] Md. Shahadat Iqbal et al. “Development of origin–destination matrices using mobile phone call data”. In: *Transportation Research Part C: Emerging Technologies* 40 (2014), pp. 63–74. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X14000059>.
- [33] Evelien Vanderlinden Johan Desseyn and Sarah Huysmans. *STUDENTENHUISVESTING IN DE GENTSE WOONSTAD meet- en monitoringsystemen: “Op welke adressen wonen studenten?”*. 2021.
- [34] Amir Reza Khavarian-Garmsir, Ayyoob Sharifi, and Ali Sadeghi. “The 15-minute city: Urban planning and design efforts toward creating sustainable neighborhoods”. In: *Cities* 132 (2023), p. 104101. URL: <https://www.sciencedirect.com/science/article/pii/S0264275122005406>.
- [35] Chloe Kinton et al. “New frontiers of studentification: The commodification of student housing as a driver of urban change”. In: *The Geographical Journal* 184.3 (2018), pp. 242–254. URL: <https://rgs-ibg.onlinelibrary.wiley.com/doi/abs/10.1111/geoj.12263>.
- [36] Kotwest. *Koten in Kortrijk*. 2025. URL: <https://www.kotwest.be/nl/koten/kortrijk> (visited on 04/07/2025).
- [37] Stanislav Kraft et al. “Travel diaries, GPS loggers and Smartphone applications in mapping the daily mobility patterns of students in an urban environment”. In: *Moravian Geographical Reports* 28.4 (2020), pp. 259–268.
- [38] Kevin J Krizek. “Neighborhood services, trip purpose, and tour-based travel”. In: *Transportation* 30 (2003), pp. 387–410.
- [39] Kwan Hui Lim et al. “Personalized tour recommendation based on user interests and points of interest visit durations”. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 1778–1784.
- [40] T.M. Logan et al. “The x-minute city: Measuring the 10, 15, 20-minute city and an evaluation of its use for sustainable urban design”. en. In: *Cities* 131 (Dec. 2022), p. 103924.
- [41] Massimiliano Luca et al. “A Survey on Deep Learning for Human Mobility”. In: *ACM Comput. Surv.* 55.1 (Nov. 2021). URL: <https://doi.org/10.1145/3485125>.
- [42] C.M. Macal and M.J. North. “Tutorial on agent-based modeling and simulation”. In: *Proceedings of the Winter Simulation Conference, 2005*. 2005, 14 pp.-.
- [43] Gabriella Mazzulla. “An activity-based system of models for student mobility simulation”. In: *European Transport Research Review* 1 (2009), pp. 163–174.
- [44] Michael G McNally. “The four-step model”. In: *Handbook of transport modelling*. Vol. 1. Emerald Group Publishing Limited, 2007, pp. 35–53.

- [45] Eric J Miller, Matthew J Roorda, and Juan Antonio Carrasco. “A tour-based model of travel mode choice”. In: *Transportation* 32 (2005), pp. 399–422.
- [46] Rolf Moeckel et al. “Agent-Based Simulation to Improve Policy Sensitivity of Trip-Based Models”. In: *Journal of Advanced Transportation* 2020.1 (2020), p. 1902162. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/1902162>.
- [47] Elisabeth Moores, Gurkiran K. Birdi, and Helen E. Higson and. “Determinants of university students’ attendance”. In: *Educational Research* 61.4 (2019), pp. 371–387. URL: <https://doi.org/10.1080/00131881.2019.1660587>.
- [48] Carlos Moreno et al. “Introducing the “15-Minute City”: Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities”. In: *Smart Cities* 4.1 (2021), pp. 93–111. URL: <https://www.mdpi.com/2624-6511/4/1/6>.
- [49] Takashi Nakazawa. “Expanding the scope of studentification studies”. In: *Geography Compass* 11.1 (2017), e12300. URL: <https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/gec3.12300>.
- [50] Sean Nash and Raktim Mitra. “University students’ transportation patterns, and the role of neighbourhood types and attitudes”. In: *Journal of Transport Geography* 76 (2019), pp. 200–211. URL: <https://www.sciencedirect.com/science/article/pii/S0966692318304691>.
- [51] Tijs Neutens et al. “Equity of Urban Service Delivery: A Comparison of Different Accessibility Measures”. In: *Environment and Planning A* 42.7 (2010), pp. 1613–1635. URL: <https://doi.org/10.1068/a4230>.
- [52] Jinwoo Park and Daniel W. Goldberg. “A Review of Recent Spatial Accessibility Studies That Benefitted from Advanced Geospatial Information: Multimodal Transportation and Spatiotemporal Disaggregation”. In: *ISPRS International Journal of Geo-Information* 10.8 (2021). URL: <https://www.mdpi.com/2220-9964/10/8/532>.
- [53] Ram M. Pendyala et al. “Florida Activity Mobility Simulator: Overview and Preliminary Validation Results”. In: *Transportation Research Record* 1921.1 (2005), pp. 123–130. URL: <https://doi.org/10.1177/0361198105192100114>.
- [54] Curtale R et al. *Outlook and demographic perspectives for EU’s rural regions. A modelling-based exercise*. 2025.
- [55] Bryan Raney and Kai Nagel. “An improved framework for large-scale multi-agent simulations of travel behaviour”. In: *Towards Better Performing Transport Networks*. Routledge, 2006, pp. 305–347.
- [56] Soora Rasouli and Harry Timmermans. “Effects of travel time delay on multi-faceted activity scheduling under space-time constraints: A simulation study”. In: *Travel Behaviour and Society* 1.1 (2014), pp. 31–35. URL: <https://www.sciencedirect.com/science/article/pii/S2214367X13000033>.
- [57] Injong Rhee et al. “On the Levy-Walk Nature of Human Mobility”. In: *IEEE/ACM Transactions on Networking* 19.3 (2011), pp. 630–643.
- [58] Mojtaba Rostami Nasab and Yousef Shafahi. “Estimation of origin–destination matrices using link counts and partial path data”. In: *Transportation* 47.6 (2020), pp. 2923–2950.

- [59] Joanna Sage, Darren Smith, and Phil Hubbard. “The Rapidity of Studentification and Population Change: There Goes the (Student)hood”. In: *Population, Space and Place* 18.5 (2012), pp. 597–613. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/psp.690>.
- [60] Li Shen and Peter R. Stopher and. “Review of GPS Travel Survey and GPS Data-Processing Methods”. In: *Transport Reviews* 34.3 (2014), pp. 316–334. URL: <https://doi.org/10.1080/01441647.2014.903530>.
- [61] Filippo Simini et al. “A deep gravity model for mobility flows generation”. In: *Nature communications* 12.1 (2021), p. 6576.
- [62] Darren P Smith. “‘Studentification ication’: the gentrification factory?” In: *Gentrification in a global context*. Routledge, 2004, pp. 73–90.
- [63] Stad Gent. *Gentse kotstudenten: resultaten van de bevraging studentenhuysvesting bij 7500 studenten in 2024*. Nov. 2024. URL: https://stad.gent/sites/default/files/media/documents/20250121_rapport%20bevraging%20studentenhuysvesting_2024.pdf.
- [64] Stad Gent. *How to get to Ghent?* 2025. URL: <https://stad.gent/en/international-students/living/transport-travel/how-get-ghent> (visited on 03/18/2025).
- [65] Stad Gent. *List of Large Student Housing in Gent*. 2025.
- [66] Stad Gent. *Omgevingsanalyse*. 2018. URL: https://stad.gent/sites/default/files/page/documents/Omgevingsanalyse2018-NLweb3_1_0.pdf.
- [67] Stad Gent. *Tendensen*. <https://hoeveelin.stad.gent/tendensen/>. 2024. URL: <https://hoeveelin.stad.gent/tendensen/> (visited on 04/28/2025).
- [68] Luca Staricco. “15-, 10- or 5-minute city? A focus on accessibility to services in Turin, Italy”. en. In: *Journal of Urban Mobility* 2 (Dec. 2022), p. 100030.
- [69] Peter R. Stopher and Stephen P. Greaves. “Household travel surveys: Where are we going?” In: *Transportation Research Part A: Policy and Practice* 41.5 (2007), pp. 367–381. URL: <https://www.sciencedirect.com/science/article/pii/S0965856406001182>.
- [70] *Sustainable Urban Mobility in the EU: No substantial improvement is possible without Member States’ commitment*. Special Report 06/2020. European Court of Auditors, 2020. URL: <https://op.europa.eu/webpub/eca/special-reports/urban-mobility-6-2020/en/>.
- [71] João Filipe Teixeira et al. “Classifying 15-minute Cities: A review of worldwide practices”. In: *Transportation Research Part A: Policy and Practice* 189 (2024), p. 104234. URL: <https://www.sciencedirect.com/science/article/pii/S0965856424002829>.
- [72] Judith Thomsen and Terje Andreas Eikemo. “Aspects of student housing satisfaction: a quantitative study”. In: *Journal of Housing and the Built Environment* 25 (2010), pp. 273–293.
- [73] Rafael Tornero, Javier Martínez, and Joaquín Castelló. “A multi-agent system for obtaining dynamic origin/destination matrices on intelligent road networks”. In: *Proceedings of the 6th Euro American Conference on Telematics and Information Systems*. Association for Computing Machinery, 2012, pp. 157–164. URL: <https://doi.org/10.1145/2261605.2261629>.
- [74] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.
- [75] United Nations. *The 17 Sustainable Development Goals*. <https://sdgs.un.org/goals>. 2015. URL: <https://sdgs.un.org/goals> (visited on 04/28/2025).

- [76] Universiteit Gent. *Over de Universiteit Gent*. 2025. URL: <https://www.ugent.be/nl/univgent> (visited on 03/18/2025).
- [77] Ann Verhetsel et al. “Housing preferences among students: collective housing versus individual accommodations? A stated preference study in Antwerp (Belgium)”. In: *Journal of Housing and the Built Environment* 32 (2017), pp. 449–470.
- [78] Kay W Axhausen, Andreas Horni, and Kai Nagel. *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.
- [79] Siying Wang et al. “Incorporation of intra-city human mobility into urban growth simulation: A case study in Beijing”. In: *Journal of Geographical Sciences* 32.5 (2022), pp. 892–912.
- [80] Elias Willberg, Christoph Fink, and Tuuli Toivonen. “The 15-minute city for all? – Measuring individual and temporal variations in walking accessibility”. In: *Journal of Transport Geography* 106 (2023), p. 103521. URL: <https://www.sciencedirect.com/science/article/pii/S0966692322002447>.
- [81] Peng Xie et al. “Urban flow prediction from spatiotemporal data using machine learning: A survey”. In: *Information Fusion* 59 (2020), pp. 1–12. URL: <https://www.sciencedirect.com/science/article/pii/S1566253519303094>.
- [82] Hai Yang and Jing Zhou. “Optimal traffic counting locations for origin–destination matrix estimation”. In: *Transportation Research Part B: Methodological* 32.2 (1998), pp. 109–126. URL: <https://www.sciencedirect.com/science/article/pii/S0191261597000167>.
- [83] Xinyuan Zhang and Nan Li. “An activity space-based gravity model for intracity human mobility flows”. In: *Sustainable Cities and Society* 101 (2024), p. 105073. URL: <https://www.sciencedirect.com/science/article/pii/S2210670723006832>.
- [84] Jiangping Zhou. “From better understandings to proactive actions: Housing location and commuting mode choices among university students”. en. In: *Transport Policy* 33 (May 2014), pp. 166–175.

