Synthetic Agents

Data mined activity patterns for an activity-based travel demand model

Amanda Barbosa Jardim

1st Supervisor / Prof. Dr. Axel Häusler
2nd Supervisor / Andrea Kondziela, MA

MID / Specialization Computational Design
Technische Hochschule Ostwestfalen-Lippe / 2020

Synthetic Agents
Data mined activity patterns for an activity-based travel demand model

Master Thesis

Amanda Barbosa Jardim
Matriculation number / 15416088

1st Supervisor / Prof. Dr. Axel Häusler
2nd Supervisor / Andrea Kondziela, MA

Collaborator / Maximilian Müh

In partial fulfilment of the requirements for the degree of
Master of Engineering

Master of Integrated Design / Computational Design
Detmolder Schule für Architektur und Innenarchitektur
Technische Hochschule Ostwestfalen-Lippe

I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism. Any use of the works of any other author, in any forms, is properly acknowledged at their point of use.

[redacted]

Amanda Barbosa Jardim
Detmold, 16/09/2020

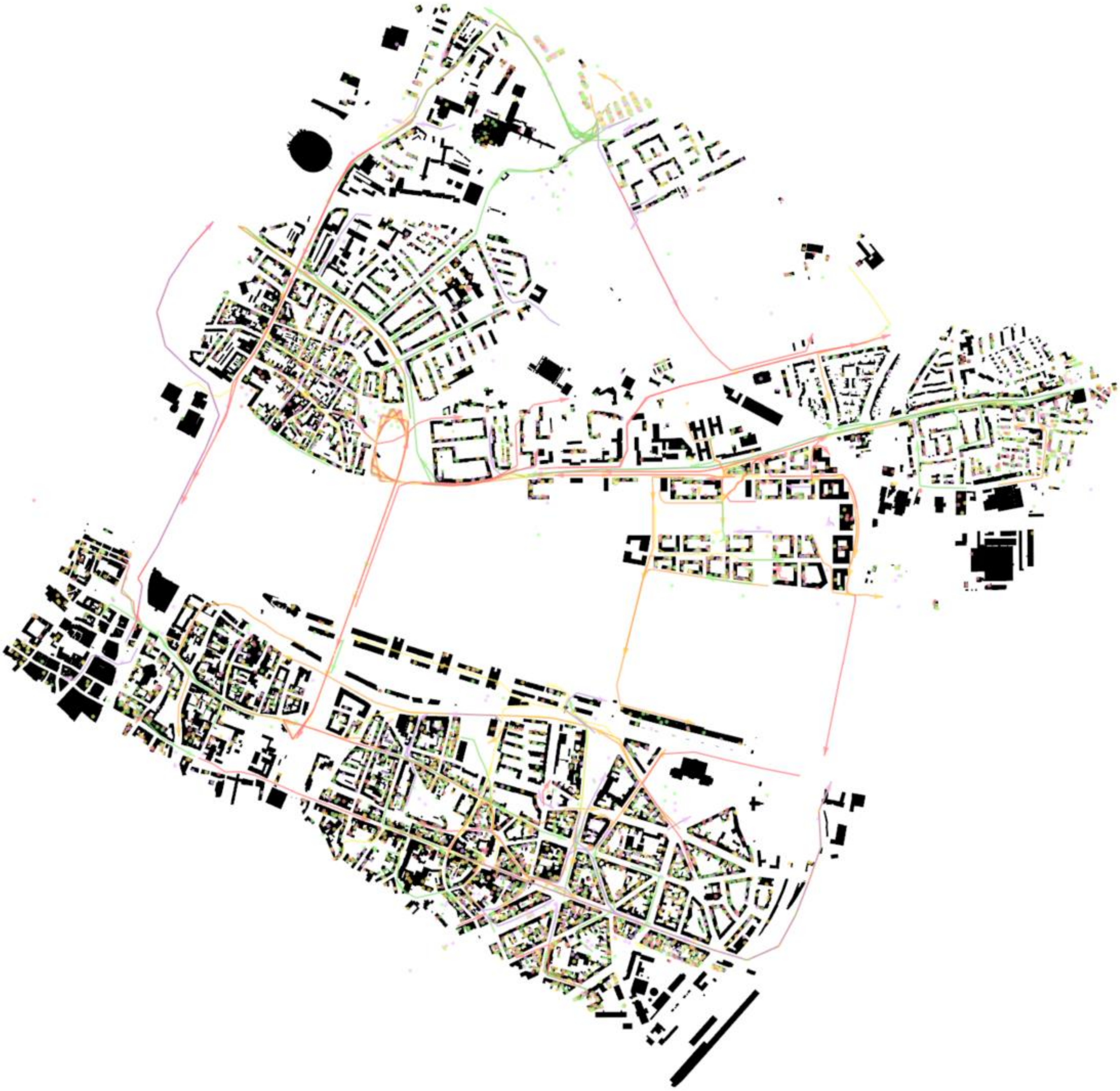


Figure 1 / Trajectory map of mobility simulation in Deutzer Hafen.

Abstract

When designing a city, one of the most important aspects to consider is the mobility patterns of its citizens. Who they are, where they go and when they go will determine how a city should be planned to attend those needs. Agent-Based Models (ABMs) have been long used by planners to simulate these patterns using data from surveys, which usually covers a small sample of the population and are too time and money consuming to update on a regular basis. Social media with its constant stream of data may offer a great opportunity for urban planners to have constantly updated data about citizens. Location-Based Social Networks (LBSN) like Twitter and Instagram allow its users to share their location when posting. By following a user's posts, it is then possible to track their activity patterns and even determine socio-economic aspects. The aim of this master thesis is to use social media data to create a synthetic population for an activity-based travel demand model for the future Deutzer Hafen district in Cologne, Germany. The ABM will be a joint work with Maximilian Müh, who will also build a Tangible User Interface (TUI) table for visualization of the model and interaction with stakeholders.

Keywords: synthetic population, social media, data mining, activity patterns, agent-based model, activity-based travel demand model, urban planning, urban design

Chapters marked with * were written in collaboration with Maximilian Müh

Table of Contents

List of Figures

List of Tables

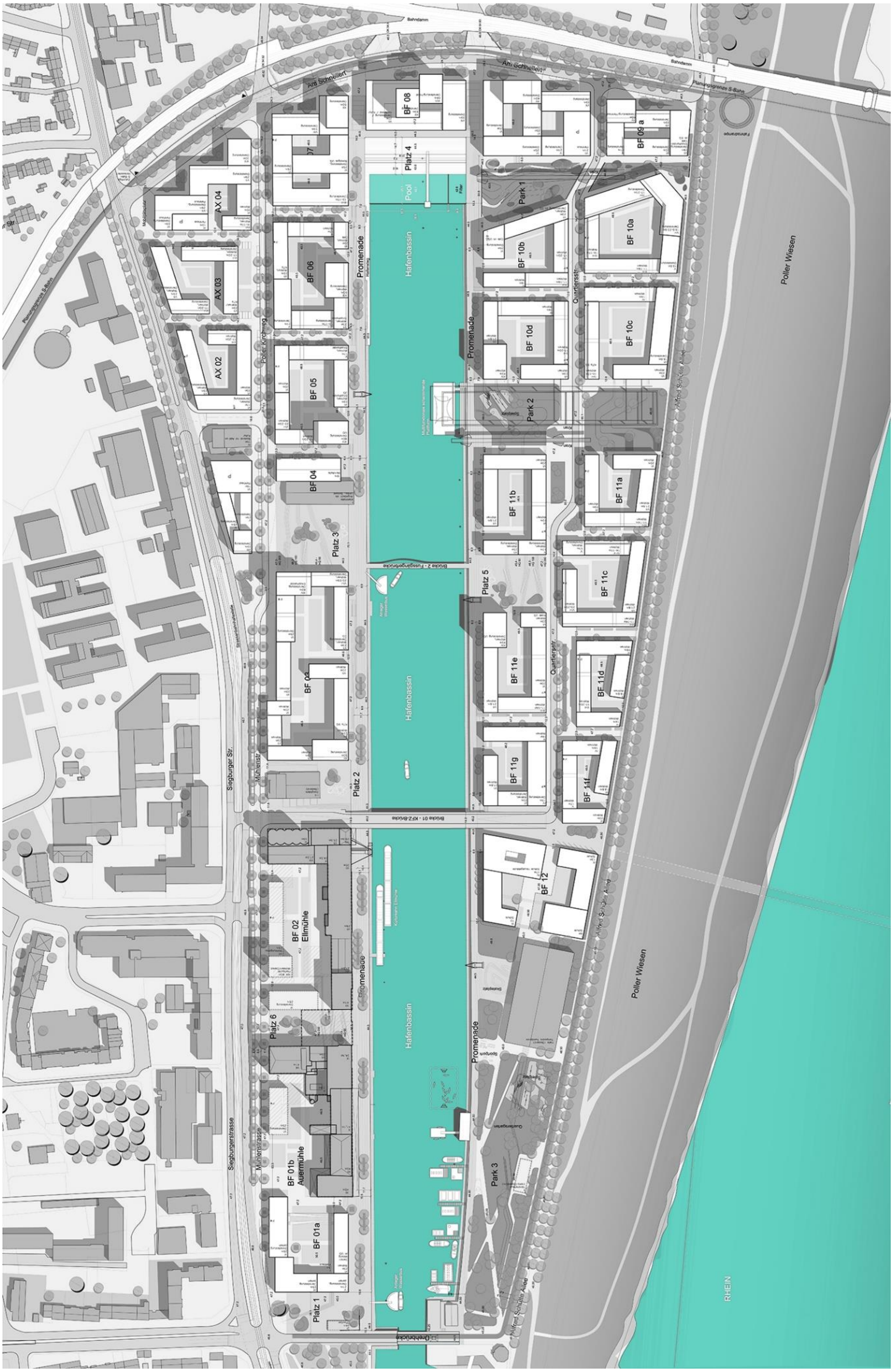


Figure 2 / Revitalization project of Deutzer Hafen. Credit: COBE [23]

1 / Introduction*

The revitalization project of Deutzer Hafen aims to transform the old industrial harbor district of Cologne into a new, vibrant neighborhood. Designed by the danish architecture firm COBE [23], the project was chosen in a competition organized by the city in 2016.

Because of its location on the Rhine river, the main focus of the project is on dealing with the river's strong tides. The promenades are made to be flooded during high tide periods, keeping the buildings and plazas dry. Additionally, rainwater is collected in a big public pool with a waterfall, which is the main attraction of the new district.

The mixed use buildings are supposed to house 5.000 people and serve as workspace for another 4.500, who will be able to access the rest of Cologne through new bicycle and pedestrian paths, including a bridge that spans directly to the city center. New public transportation routes, such as waterbuses and a train (S-Bahn) station are also planned. As well as mobility stations with bike and car sharing offers [54]. All these possibilities and easy access aim to encourage people to leave their cars at home.

Even before its construction, the project already received platinum in the DGNB (German Sustainable Building Council) [62] certification system, which is the highest possible ranking. The main aspects of the project highlighted by the council were "turning water challenges into water resources, while also creating a livable and mixed city [...] with focus on cycling, future means of transport and diverse building sites and types".

We use the future district as case study in building a decision support system for urban planning, that is composed of three parts: an agent-based model, a tangible user interface and a synthetic population.

Agent-based models (ABMs) are used in multiple fields to simulate complex systems through a set of independent agents that follow certain rules and react on an environment. In urban planning, activity-based travel demand models (ABTDM) are used to estimate the demand for travel in a region and the resulting performance of the transportation system, according to different scenarios and policy, economic, demographic or land use changes, as defined by Castiglione et. al in "Activity-Based Travel Demand Models: A Primer" [19]. They also define the focus of these models as "whether, when, and where to participate in activities and for how long.

Travel is a derived demand resulting from the need for people to engage

Figure 3 / The Deutzer Hafen district as of today. Credit: COBE [23].

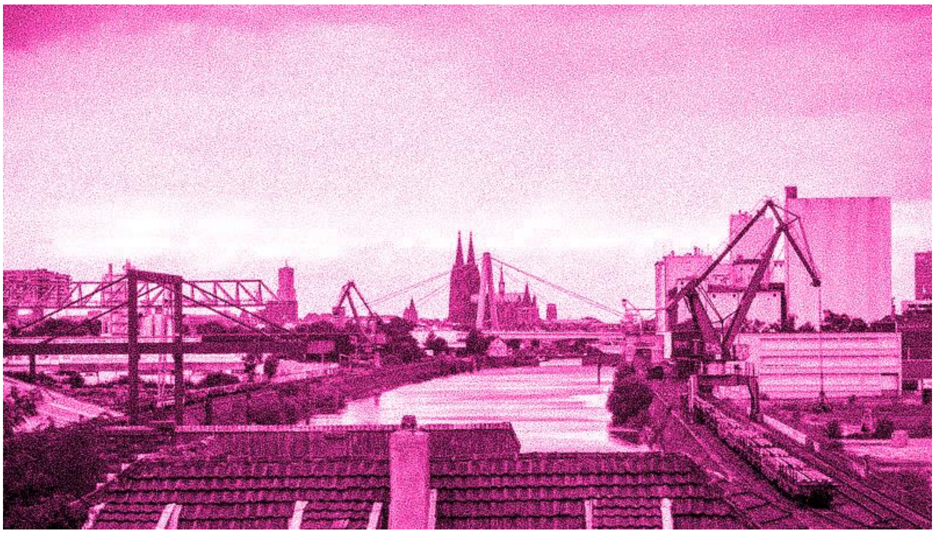
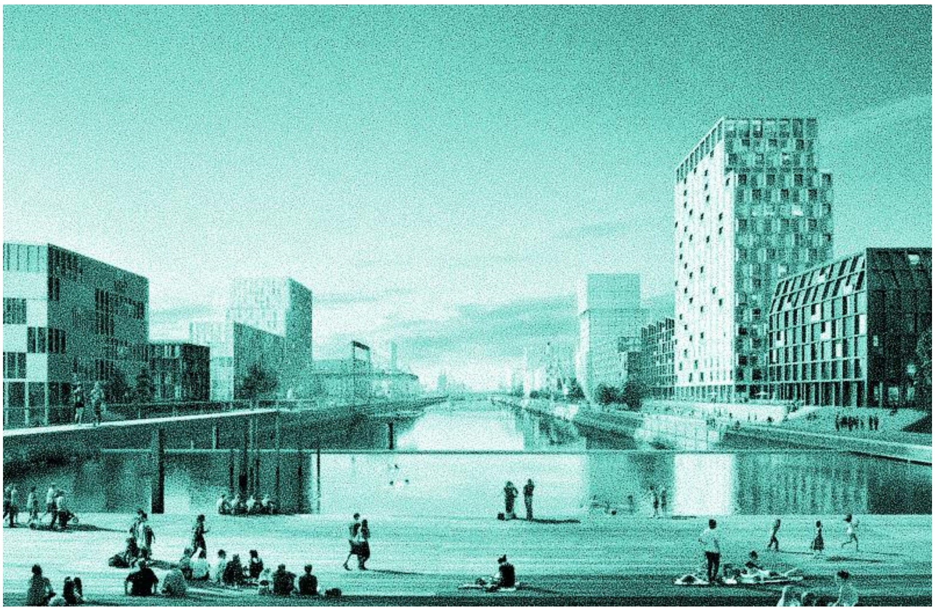


Figure 4 / Revitalization project of the district. Credit / COBE and The Beauty and the Bit [23].



in activities outside the home". This need for traveling and engaging in activities has also been connected to quality of urban spaces.

In "Life Between Buildings", Gehl [30] writes that in public spaces of poor quality, people only pass by on the way to necessary activities that they must do, like going to work or shopping. On the other hand, if the public space is of good quality, people start engaging in more optional activities, such as taking a walk or sitting on a bench. People attract more people, and so social activities, that result from the presence of others, such as just watching people passing by, also arise.

Jane Jacobs, in "The Death and Life of Great American Cities" [43], connects urban vitality to diversity in the built environment. For an urban space to be successful and safe, a diversity of people should pass by it, with different purposes, and in different times of the day.

Based on these theories, we aim to use an ABTDM to measure the urban vitality of the public spaces in the district, based on the activity and travelling patterns of the population. This is done by testing different scenarios in which we change interactive parameters of the model: the use of the buildings and the demographics of the population. We can then determine which scenarios benefit the most life in the public spaces of the district, by finding areas of interest or problematic ones.

The interaction with the model is done through a tangible user interface (TUI), that connects the digital information of the ABTDM with a physical model of the district. The TUI makes interaction with the complex system more feasible and intuitive, facilitating the participation of all stakeholders in the design process, not only specialists. In a game-like experience, the user can change the use of buildings in the model by moving tags around or adjust the demographics of the population with a slider. Visual statistics give immediate feedback to the user's actions, making complex relationships become clearer.

To simulate the activity and travelling patterns, the model needs a synthetic population, which is a virtual representation of the community of the modelled area. It is commonly built by combining census and travel or time use survey data, that may not always be up to date, because of the amount of time and resources taken to make such surveys. That is why an experimentation with a new approach is made: building our synthetic population from social media data. We are in an age with a constant flow of user generated content coming from location based social networks (LBSN) such as Twitter, Facebook and Instagram, where people share where they are, when they are and what they are doing. Mining this data can allow us to produce a sort of digital census, that is cheaper and fresher than traditional surveys.

Social media posts inside the city of Cologne are collected from Twitter and Instagram. The users collected are anonymously profiled and have their activity patterns inferred, resulting in a population that reflects a sample of the city. These profiles are then used to populate the model.

The model will enable more innovative and broader user participation in urban planning. The use of social media is a form of early user participation, for allowing planners to use a bigger and more updated set of citizen's data than traditional surveys. And the TUI table on the other hand, is a direct participation tool, with which stakeholders can be aware of their contribution and results for the city.

The ABM is made in collaboration with Maximilian Müh, who also builds the TUI table. This thesis goes into detail of how the synthetic population is created. An overall view of how the three parts work together is also given, but for more detailed information of how the TUI table is built, Müh's master thesis *Tangible Agents: a tangible user interface for an activity-based travel demand model* [55] can be referred to.

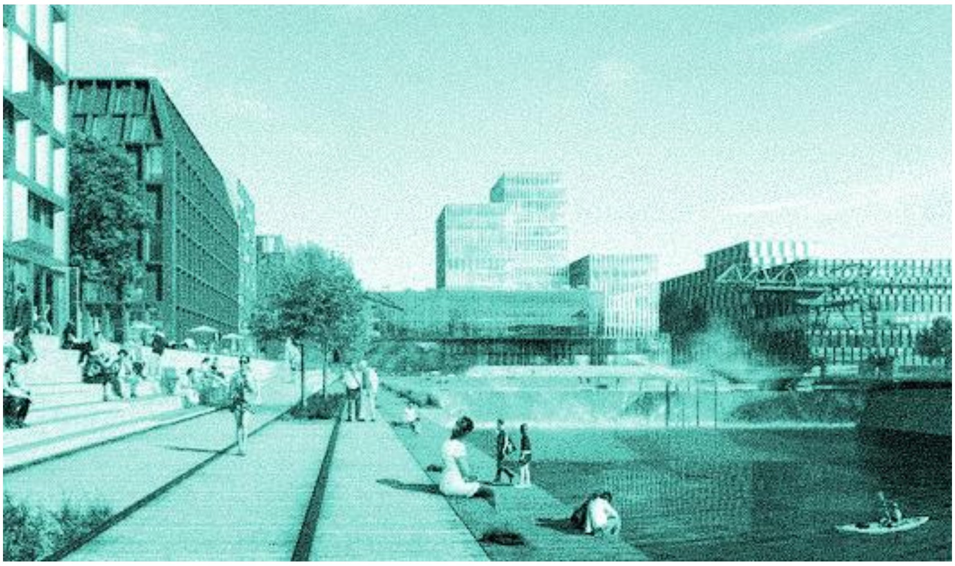


Figure 5 / Example of social activities in the new district project. Credit: COBE and The Beauty and the Bit.

2 / State of The Art

2.1 / Agent based modelling (ABM)*

In this chapter, the idea of ABM is briefly explained. Furthermore, its relationship to object-oriented programming (OOP) is highlighted. In the end, common tools and platforms for agent-based modelling in the urban context are presented and compared. A model, based on agents:

> Agent: An agent is an entity that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its objectives. [70]

> Model: A model is intended to represent or simulate some real, existing phenomenon, and this is called the target of the model. The two main advantages of a model are that it succinctly expresses the relationships between features of the target, and it allows one to discover things about the target by investigating the model. [31]

2.1.1 / A brief introduction to ABM

The concept of ABM first came up in the late 1940s with the first prototypical “cellular automata” [21], but it took until the 90s for computers to be powerful enough to support the spread of the idea. Today, ABMs are used in a variety of professions, from biology to economics [86].

The basic idea is a model that simulates *agents* in an *environment* over *time*. These agents (e.g. representing humans) behave according to a predefined *ruleset*, *attributes* (e.g. age), and *objectives* (e.g. work).

Each agent, even though behaving according to the same rules, can make different decisions, based on environment, surrounding agents, attributes, objective, and time. Agents can communicate with each other and influence others. The entirety of all agents can visualize complex phenomena.

ABMs proofed to be especially useful in space related topics in urban and geospatial studies. The first ABMs developed in this field go back to 1971, starting with segregation and other social behaviours. By now, ABMs are used to simulate all kinds of urban topics, from social behaviour, ecological and environmental phenomena to transportation systems and traffic. [21]

2.1.2 / Object-Oriented Programming (OOP)

The concept of object-oriented programming was crucial to the success of agent-based modelling. As the name suggests, a program developed with OOP consists of several objects. Objects are instances of so-called classes. A class can have methods, attributes, and variables. Variables can be specific for each object. [38]

One of the first works to include this concept was Sketchpad in 1963 by Sutherland [78]. He defined “objects”, “instances”, and “masters” which correspond to the classes.

Comparing an agent with an object shows a significant affinity. It is only natural to program a class and instantiate it to form objects (= agents). This is why most of the ABM models are programmed in languages such as Java, C++ or Visual Basic which are OOP languages [38].

In the following, an overview of tools and platforms for ABM is given.

2.1.3 / Choosing a suitable agent-based simulation platform

> The high heterogeneity and the vast amount of the available agent platforms is a fact. Hence, choosing the right or most suitable platform for a given problem is still a challenge for the developer. [50]

Alongside with the increasing popularity of ABM models, platforms and tools for modelling were developed. As mentioned, these tools heavily take advantage of object-oriented programming.

Kravari, Bassiliades et al. made a detailed survey about Agent Platforms, describing (dis-)advantages of 24 platforms [50].

One way of categorizing them is the language used to define models [80]:

- ↳ (i) Generic high-level programming languages such as Java or C++. Platforms like Repast [57] or JADE [15] are implemented as libraries in the programming environment and take advantage of the vast number of other libraries available. These platforms are suitable for large-scale, complex models but are not very approachable to non-computer scientists.
- ↳ (ii) Platforms with a specific modelling language, dedicated to a straightforward syntax for fast results. Even though they still require basic skills in algorithmics, they are much more approachable. GAMA [80] and NetLogo [2] are popular examples.
- ↳ (iii) Platforms with a graphical modelling language such as Starlogo [1] or AgentSheets [11] . These platforms are very accessible with almost no programming skills but lack the capability to build large-scale, complex models.

Even though both researchers working on the project are to a certain extent familiar to Python and programming in general, the first category platform was excluded because of the limited time of four months for the project and the high complexity of those libraries.

The last category was not suitable because of the complexity of the project, which leaves only category (ii).

NetLogo is one of the most popular platforms, both because of its easy-to-use language and the extensive and diverse model library [11] . However, it lacks some capabilities that were desired for the project such as "agentification" of GIS data, present in GAMA, which is useful for the implementation of buildings. Furthermore, NetLogo does not support the visualisation on two screens.

For these reasons, the decision was made to use GAMA (GIS Agent-based Modeling Architecture) [6].

2.1.4 / GAMA

GAMA was started in 2007 as open-source platform and has been developed and updated since then. It is a modelling and simulation development environment for building spatially explicit agent-based simulations, based on the Eclipse IDE and written in Java. The implemented language is called GAML and is an agent-oriented language, with roots coming from OOP. For being an agent-based language, it is highly intuitive and comes with several paradigms of modelling, like for example existing behaviours and skills for the agents, that do not need to be coded from zero.

It is widely applied in the domains of transport and urban planning since it supports GIS and data-driven models. It is possible to import various types of datasets, such as shapefiles, CSV files and connections to databases.

Visual representations of the model are also easy to be defined, from the appearance of the agents to charts.

Another valuable feature for this project was the possibility to modify the display of the model using a keystone correction. Since the simulation in the end is projected onto a physical model, this correction helped a great deal.

GAMA is also especially suitable for participatory, interactive simulation [81]. Recent examples in this field were integrated into CityScope [12, 35, 36, 46]. At the Institute for Applied Sciences Urban Future in Potsdam, GAMA was also integrated in a participatory tool for city development called pasymo [53]. The topics of those projects are very diverse, from mobility and transportation to air pollution, GAMA was recently used in many different fields of spatial agent-based modelling.

2.1.5 / Activity-based travel demand models

> A fundamental premise of activity-based travel models is that travel demand derives from people's needs and desires to participate in activities. [19]

The agent-based model for the Deutzer Hafen should simulate the daily life routine of its residents and the surrounding city with its citizens as an activity-based travel demand model.

An activity based travel demand model is built up on “individual persons and represent how these persons travel across the day,[…]”. [19]. Each individual (= agent) makes their own decision on how and where to travel according to their profile, the destination, and the previous mode of transportation. For example, it is very unlikely that somebody who used the car to get from home to the supermarket, will walk back home.

2.2 / Synthetic Population

An activity-based model requires the input of a synthetic population, which is a virtual community with detailed descriptions of households and people that match the sociodemographic characteristics of the modelled area. Population synthesizers are tools used by urban modelers to create synthetic populations. According to [19] these synthesizers typically use the following inputs:

- ↳ Household size;
- ↳ Household composition and life cycle (e.g.; age of householder by presence of own children);
- ↳ Number of workers per household;
- ↳ Household income category;
- ↳ Age and gender of each person; and
- ↳ Employment and student status of each person.

With this, the model is capable of considering many different types of people and households and can predict more realistic choice behaviours. Although the data requirements may vary, depending on the specifications of the model. Work and school location, vehicle or transit pass ownership are additional variables that help simulating daily patterns of travel.

Two main inputs of data are needed to synthesize a population: *control data* and *sample data*. The first one comes usually from census, and represent the marginals of the population, the demographics of that region. The second one comes from household surveys and contains the individual samples of people and households that are used to fill the population until the marginals are reached. This data can come in multiple geographic levels, that need to be combined. Most of the population synthesizers performs these two basic functions:

- ↳ combining data of different scales in the marginal controls, such as number of households in block geographies and age and income distribution in tract geographies;
- ↳ drawing samples of households and people until the sample distribution of the population is matched.

Doppelgänger [59] is an example of an open source population synthesizer, that creates populations from sets of census data in different scales. The tool allocates the households to match the smaller geographies of the age and income data, until they closely match. Normally, these marginals are not consistent and will not match perfectly. That is when the user can prioritize which marginals are more important to match and give them a bigger subjective weight.

The tool also uses a Bayesian Network for estimating missing data and sample reconstruction. A Bayesian Network is a "Probabilistic Graphical Model (PGM) that represents conditional dependencies between random variables through a Directed Acyclic Graph (DAG)" [37]. A DAG contains nodes connected by links that denote the relationship between them. One simple example can be given by the DAG below with three nodes: *rain*, *sprinkler* and *grass wet*. The grass being wet depends on if it is raining or if the sprinkler is on. The sprinkler being on, on the other hand, depends on the rain, since it will not be turned on when it is raining.

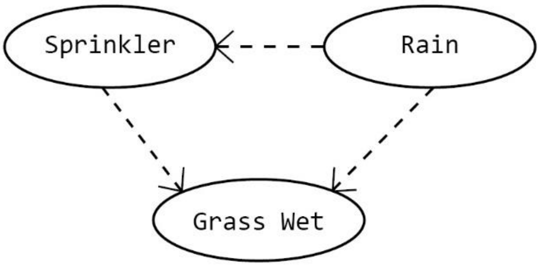


Figure 6 /
Example of A DAG.

The links between the nodes are weighted according to conditional probability, which is the probability of a node occurring when another one occurred before. It is shown by $P(X|Y)$. If the variables are dependent, $P(X|Y) = P(X,Y) / P(Y)$. If they are independent, $P(X|Y) = P(X)$. The probabilities of the whole network are calculated with the formula below:

$$P(X_1,\dots,X_N) = \prod_{i=1}^{N} P(X_i|\mathrm{Parents}(X_i))$$

In modeling the network, the DAG can be defined initially, when the random variables and their weighted links are known. But the model can also learn the structure from data samples.

Many population synthesizers like Doppelgänger use this approach to train a model and use it for predicting missing values. Given a dataset that contains for example number of people in a household and location, but no number of cars, a Bayes Net that was previously trained on a dataset with all three parameters can estimate the number of cars of the incomplete dataset.

The Bayes Net can also be used for sample reconstruction, which is the matching of the population created from the census sample to the real numbers in a determined real population. It is typically done by simply replicating the households until the desired amount for the subject area is achieved. However, having the same household with the same characteristics repeated over and over is not desirable, because the population will then lack the usual variability that exists in a real community. A Bayes Net can be used to solve this issue by capturing the relevant characteristics of the synthetic population, it is possible to use the network to create households with different sets of characteristics and increase the heterogeneity of the given population.

On the topic of synthetic populations being created from social media data, this research unfortunately found no works on it. The closest project might be Replica [66], described as "a next-generation urban planning tool that can help cities answer key transportation questions". It is a spin-off from urban tech company Sidewalk Labs [69], which in turn is owned by Alphabet, Google's mother company [13]. Replica offers public agencies and land developers "complete sense of city movement patterns" resulting in "higher confidence in critical transportation and land use decisions". They claim to do so by using de-identified mobile location data to generate travel behavior models, which are given to each people in a synthetic population created from aggregated demographic information. This data is supposed to be updated every three months, which sounds very appealing in comparison to normal travel surveys.

Although appealing, the tool also comes with concerns about citizen's privacy. News reports indicate that Replica might not be transparent about the provenance of its location data. According to this report [26] the company has been reluctant in giving the city of Portland, USA, proof that their system prevents reidentification of actual people from the data. The city is still willing to work with the company to guide policy, but they aim to test the vulnerabilities of the system in exposing identities from their unusually detailed dataset. In 2018 the company claimed using "location data from Android phones and Google apps". However, as of today, they say they no longer use any data from Google and still did not disclosure any other data source.

2.3 / Social Media as Source for Urban Design

The continuous growth in the use of online social networks presents an opportunity for usage of its big data in different fields, from social sciences to economics. Just recently, as of 2011 [65], research has emerged about the potential of using such data in the fields of transport engineering, urban planning and travel demand modelling.

These fields often rely on high resolution databases for planning, like demographic and economic attributes of people, as well as their daily travel behaviour. Such data is collected in the form of census and travel diary or time use surveys, which are very expensive and time-consuming to produce. For this reason, they are not constantly updated and can sometimes be already outdated when in use. The most recent time use survey in Germany (Zeitverwendungserhebung), for example, is from 2012/2013 [72].

Social media data, on the other hand, offers a continuous flow of information about people's activities and whereabouts, which enables it to be used as a sort of 'digital census' [87]. But it also comes with disadvantages, being the main ones bias in the population distribution in the social network and the extensive processing required to eliminate noise and extract useful information from such data.

2.3.1 / Aggregated Mobility Patterns

Much research has been done on defining aggregated mobility patterns from social media data. Although in this research the aim is to build individual mobility patterns, some of these methods can also be useful for the process about to be developed.

In [52], the urban mobility patterns of tourists and locals are defined using tweets. The users are separated into the two categories according to their profile self-declared location. Or, when no location is available, according to the number of days that they tweeted in Barcelona. The geo-located tweets are then classified into weekend or working days and plotted in a map. The pathways considered are only those that happen in one day.

Cheng et al. [90] finds daily and weekend patterns of users in a massive dataset of tweets from the whole world and compares the patterns for different cities. They found correlations in different locations, which shows that Location Sharing Services (LSS) users follow simple reproducible patterns. They also cluster users by their mobility level according to the amount and distance that they travel and, by analysing the text from their posts, find out that each group differ in the topics they talk about and terms they use. Showing that content can revel context between people and locations.

2.3.2 / Individual Mobility Patterns

Further research aims to find patterns of mobility for individual users, through methods that serve as guidance for this thesis.

In [25], commuting patterns in Beijing are identified based on Weibo (known as the Chinese Twitter) check-ins. Home and work location of users are determined based on the places they check-in the most that falls in the POI (Point of Interest) category of residence or working place. This information is then double checked by searching for key words like "home" or "work" in the posts' content. They can then establish patterns of home and work-based activity, that are later validated using an independent citywide travel logistic survey.

In [87], home location of users in Singapore is defined through call detail records (CDRs). They consider that the user lives in the area from where they used the phone the most at night, in between 7:00 and 9:00 pm. Each user is then assigned a socio-economic status (SES) according to a residential property price dataset available for that location. They compare the obtained distribution with official census data and show that it is well correlated.

Fuchs et al. [29] applies text and location analysis to detect behavioral patterns of individuals in Seattle. They categorize the tweets in 22 themes that are represented by keywords, for example, the theme “family” is associated with the keywords family, mother, father, children etc. By connecting the categories to the locations, they find places where the activities are likely performed, and find their patterns during the week. Work activities, for example, occur mostly during weekdays, although some occurrences appear on the weekend, which they say requires further investigation. Finally, the absolute and relative frequencies of different topics from each person is computed to find a topic fingerprint for that person. The users are then clustered according to this fingerprint to find people with similar interests.

Liao et al. [51] found a correlation between daily distributions of work and home locations in data collected from Twitter and from a travel diary survey in Sweden.

Zhu et al. [91] fuse a travel survey dataset with LBSN data. The travel survey accounts for the demographics and travel time of the traveller, while the location is extracted from Foursquare venues found near the coordinates from the travel survey. They use then this augmented data to build a model that predicts travel purposes from trajectory data alone (geo-coordinates and time).

An even more advanced network is shown in [88], where a model called W4 (Who, Where, When, What) trained on Twitter data can predict for example the location of a tweet (where) when only the other three Ws are present in the dataset.

2.3.3 / Privacy Issues

Even though social media posts can only be visualized and collected if they are marked as public by the author, there are still concerns about collecting and using this data without consent from the user. Approaches usually anonymize user and venues ids [52] to protect their privacy.

But users might not like the idea of their data being collected anyway. When analysing posting patterns of users on Twitter from the first to the last day they posted, Swier et al. [79] observed “that many users go through a phase of sending geolocated tweets and then stop”. They assume that one of the reasons might be because of change of attitude towards privacy. Additionally, they found a drop in the number of tweets that coincided with the release of the iPhone iOS8 system, that brought changes in how privacy and location are managed. The decline in tweets was mostly from iPhone devices, which indicates that, given the opportunity to have more control over how their phones share their data, these users opted for more privacy.

It is also possible that before the update they were not even aware of how their data was being used.

Rout et al. [73] call attention to the privacy implications of their own study, where they find Twitter users' city location from their social connections, by finding geographically local links. They stress that, even though their method is shallow, they could locate 50% of the users, which demonstrates that even if you do not publicly share your location, it could still be inferred.

Lack of awareness also exists in relation to the use of the social network itself. Although Twitter allows the use of tweets in academic research and even offer from free to paid products to do so [84], research shows that "Internet users rarely read or could fully understand website terms and conditions" and are unaware of the existence of APIs or that Twitter sells access to their data [27]. One can also argue that, even when the users is aware and gives consent to their tweets being of public view, that does not necessarily mean giving consent to them being collected and analysed.

To explore these questions from the user's point of view, Fiesler et al. [27] conducted a survey to find Twitter users' responses to the idea of the use of tweets for academic research. 61.2% of respondents did not know that tweets are sometimes used for research and when asked if they believe that researchers are allowed to use tweets, 42.7% responded that they believed they were not. When asked the reason why, 23.2% believed that Twitter's Terms and Conditions forbid it, 10.1% believed that researchers would be breaking copyright law and the majority of 60.9% believed that researchers would be breaking ethical rules. These findings confirm that many users are unaware of the consent they give to the use of their data when registering for such a social network. As of September 2020, Twitter's privacy policy [83] states clearly that tweets are public and can be collected and used by third parties:

> You are responsible for your Tweets and other information you provide through our services, and you should think carefully about what you make public, especially if it is sensitive information. If you update your public information on Twitter, such as by deleting a Tweet or deactivating your account, we will reflect your updated content on Twitter.com, Twitter for iOS, and Twitter for Android.
>
> By publicly posting content when you Tweet, you are directing us to disclose that information as broadly as possible, including through our APIs, and directing those accessing the information through our APIs to do the same. To facilitate the fast global dissemination of Tweets to people around the world, we use technology like application programming interfaces (APIs) and embeds to make that information available to websites, apps, and others for their use - for example, displaying Tweets on a news website or analyzing what people say on Twitter. We generally make this content available in limited quantities for free and charge licensing fees for large-scale access.

As of the same date, Instagram's data policy [40] also states something similar:

> Public information can be seen by anyone, on or off our Products, including if they don't have an account. This includes your Instagram username; any

> information you share with a public audience; information in your public profile on Facebook; and content you share on a Facebook Page, public Instagram account or any other public forum, such as Facebook Marketplace. You, other people using Facebook and Instagram, and we can provide access to or send public information to anyone on or off our Products, including in other Facebook Company Products, in search results, or through tools and APIs. Public information can also be seen, accessed, reshared or downloaded through third-party services such as search engines, APIs, and offline media such as TV, and by apps, websites and other services that integrate with our Products.

When asked how they feel about the idea of their tweets being used for research, respondents' answers depended mostly on the amount of tweets. The majority of respondents did not mind the tweets being used, unless if their "entire Twitter history" would be used, that is when the level of discomfort grows higher, as seen below.

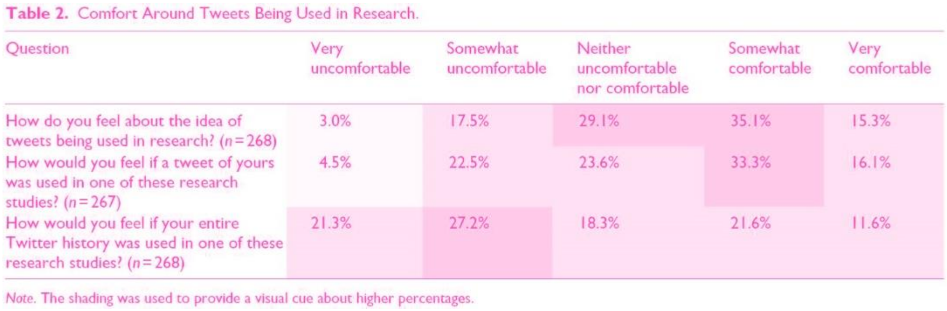


Table 2. Comfort Around Tweets Being Used in Research.

Question	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
How do you feel about the idea of tweets being used in research? (*n* = 268)	3.0%	17.5%	29.1%	35.1%	15.3%
How would you feel if a tweet of yours was used in one of these research studies? (*n* = 267)	4.5%	22.5%	23.6%	33.3%	16.1%
How would you feel if your entire Twitter history was used in one of these research studies? (*n* = 268)	21.3%	27.2%	18.3%	21.6%	11.6%

Note. The shading was used to provide a visual cue about higher percentages.

Figure 7 / Comfort around tweets being used in research. Credit: Fiesler et al.

Finally, respondents were asked if, given the choice, they would opt out of their tweets being used for research. 46.3% would not, 29.1% would and 24.6% responded that it would depend, which indicates that the context of the research could change users' minds. However, a majority of 64.9% still think that researchers should not be able to use tweets without permission. But, if asked for permission, a majority of 53.4% would give it. Some users left qualitative feedback indicating the importance of the context of the research:

> If my tweets were being used in a large scale study, I really wouldn't care. If anything was being personally picked out about me in a small study, I would care.
>
> I would want to know how it was to be used, who would see it, whether my information would be kept anonymous and how long the tweet would be kept.

The following figure shows that users are more comfortable with the idea of their tweets being used in a research when they are analysed in a large scale dataset, when their personal profile information is not used, when they are not quoted and when their tweets are mainly read by a machine instead of a human researcher:

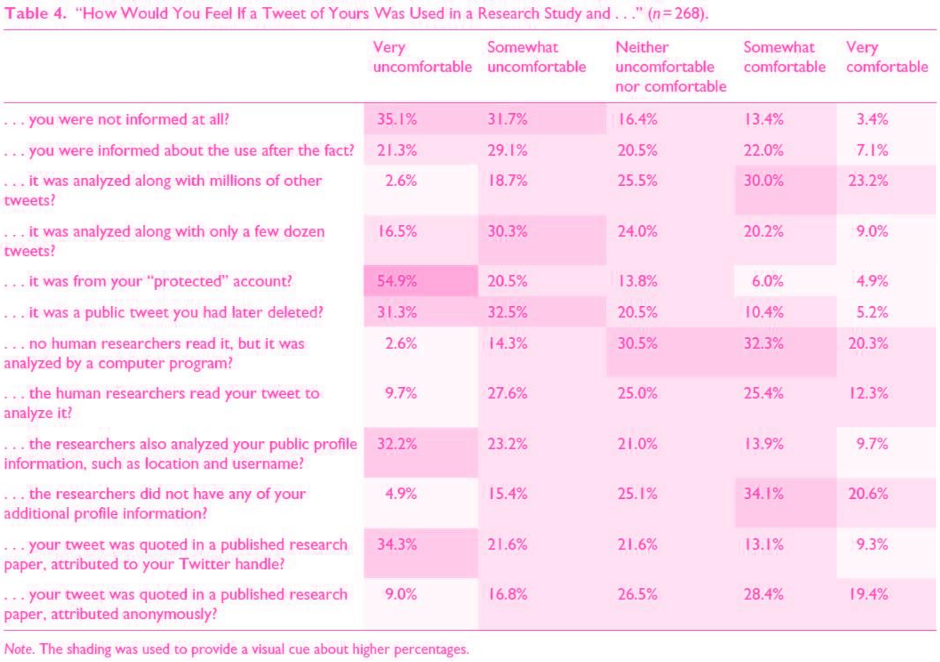


Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." ($n = 268$).

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Note. The shading was used to provide a visual cue about higher percentages.

Figure 8 / "How would you feel if a tweet of yours was used in a research study and…" Credit: Fiesler et al.

The study also notes that "there is inherent selection bias in any data collected with consent. Therefore, the only way to study people who don't want to be studied is to do so without their consent". However, even without asking users for consent, we can still use the findings of this survey as guidance to performing research in a way that would cause the least discomfort to users. That includes using data in a large scale, anonymizing identities and not sharing any quotes or other personal information.

2.4 / Tangible User Interface (TUI)

Before Steve Jobs revealed the Macintosh in the 80's, computers had very complicated interfaces and were only accessible to a few specialists. Apple was the first company to successfully create a human interface for interaction with computers. Their Human Interface Guidelines [14] would become present in our everyday life and also serve as design reference for other companies, such as Microsoft.

These graphical user interfaces (GUI) though, are bound to flat, rectangular displays, windows, mouse and keyboard [42], very different from the way in which humans interact with the physical world.

That is where tangible user interfaces (TUI) come into scene. In a framework presented by the Tangible Media Group at the MIT in 2000, a TUI was defined as "transforming human-computer interaction from abstract mousings and keystrokes into hands-on engagement" [85].

More than a decade later, another group started experimenting with TUI in the MIT. The City Science Group at the Media Lab developed CityScope, a rapid prototyping urban design tool built around a TUI [12]. As explained by Müh [55], "this interface is based on colour tagged LEGO^{TM} bricks, so called data units, which are scanned via webcam, and the projection of near real time feedback on top of them".

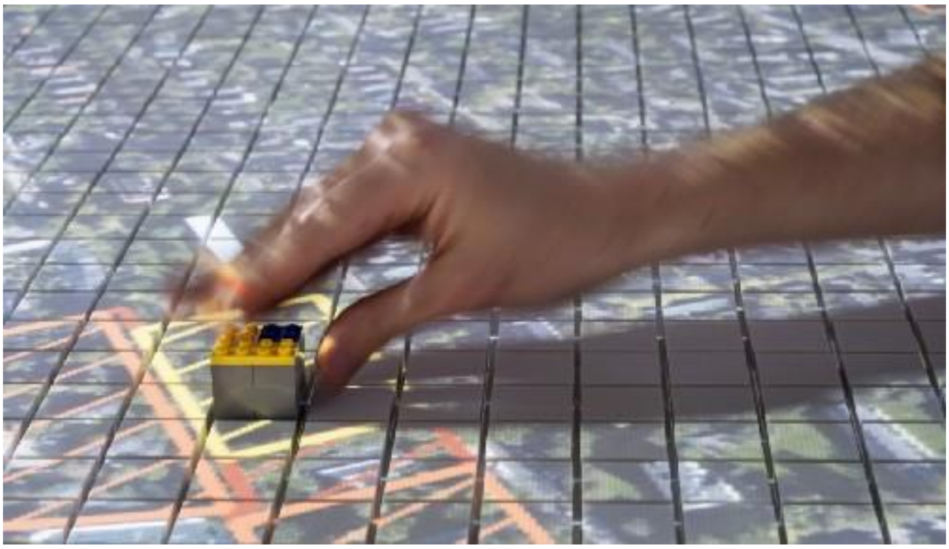


Figure 9 / LEGO blocks as housing units. Credit: Walter Schiesswohl [48].

The tool has been used in several cities and projects to simulate scenarios with the participation of all stakeholders involved. One particular project worth mentioning is *Finding Places* [58], developed as a cooperation between the City Science Lab at Hafencity University Hamburg and the MIT. The project aimed to help finding locations for accommodating the enormous wave of refugees that was coming to Germany in 2015. The tool was used in workshops with the participation of the citizens and found 161 possible locations for refugee housing. Although in the end only six of the locations were used, the project proved to be a success in building a broader acceptance of the population towards refugees, since the citizens felt included in the decisions made by the city. “This informational and educational aspect of such participatory tools is very important and could help developing cities togeter with their inhabitants”. [55]

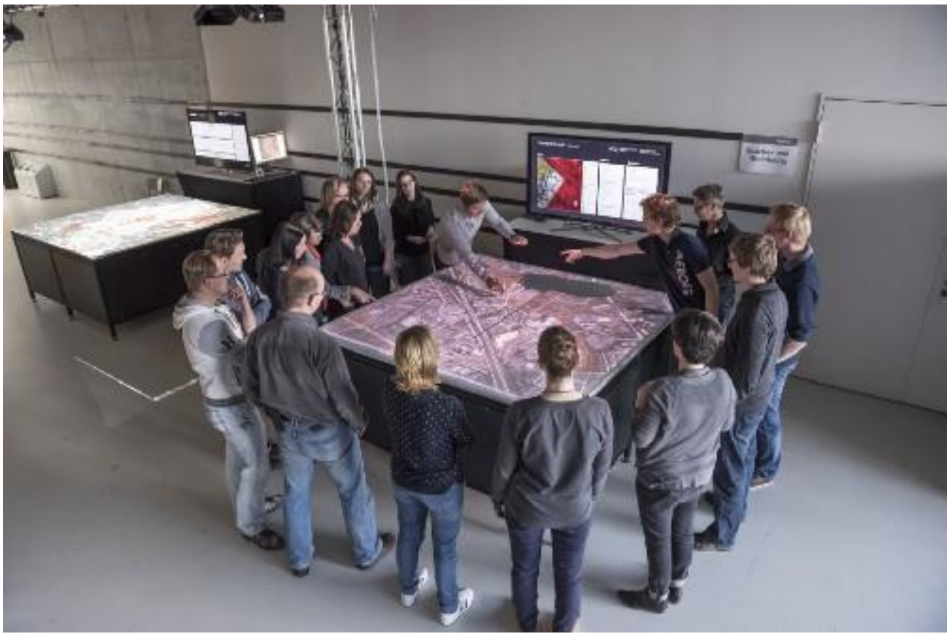


Figure 10 / Finding Places workshop. Credit: Walter Schiesswohl [48].

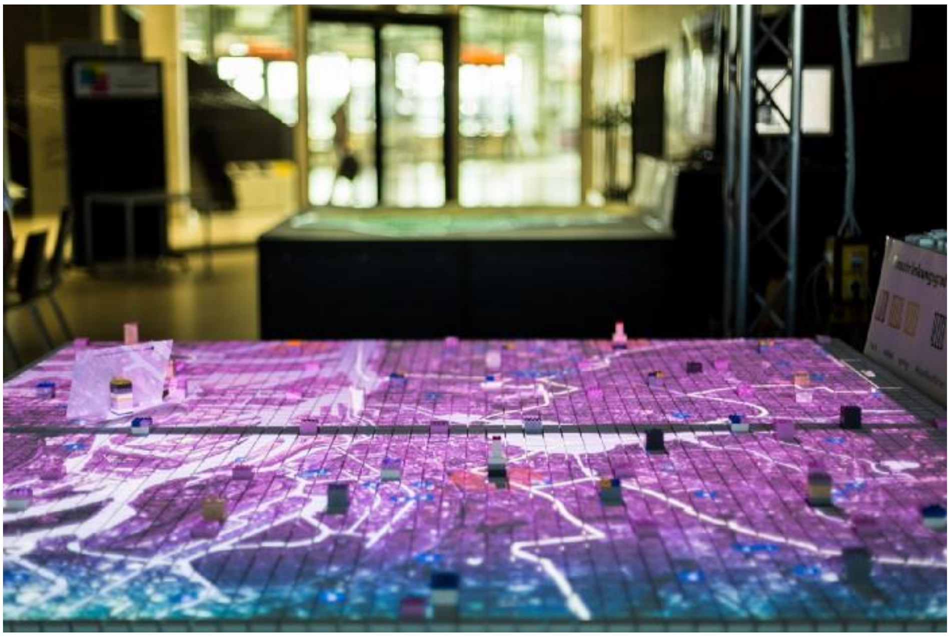


Figure 11 / The table with suggested locations. Credit: Ariel Noyman [48].

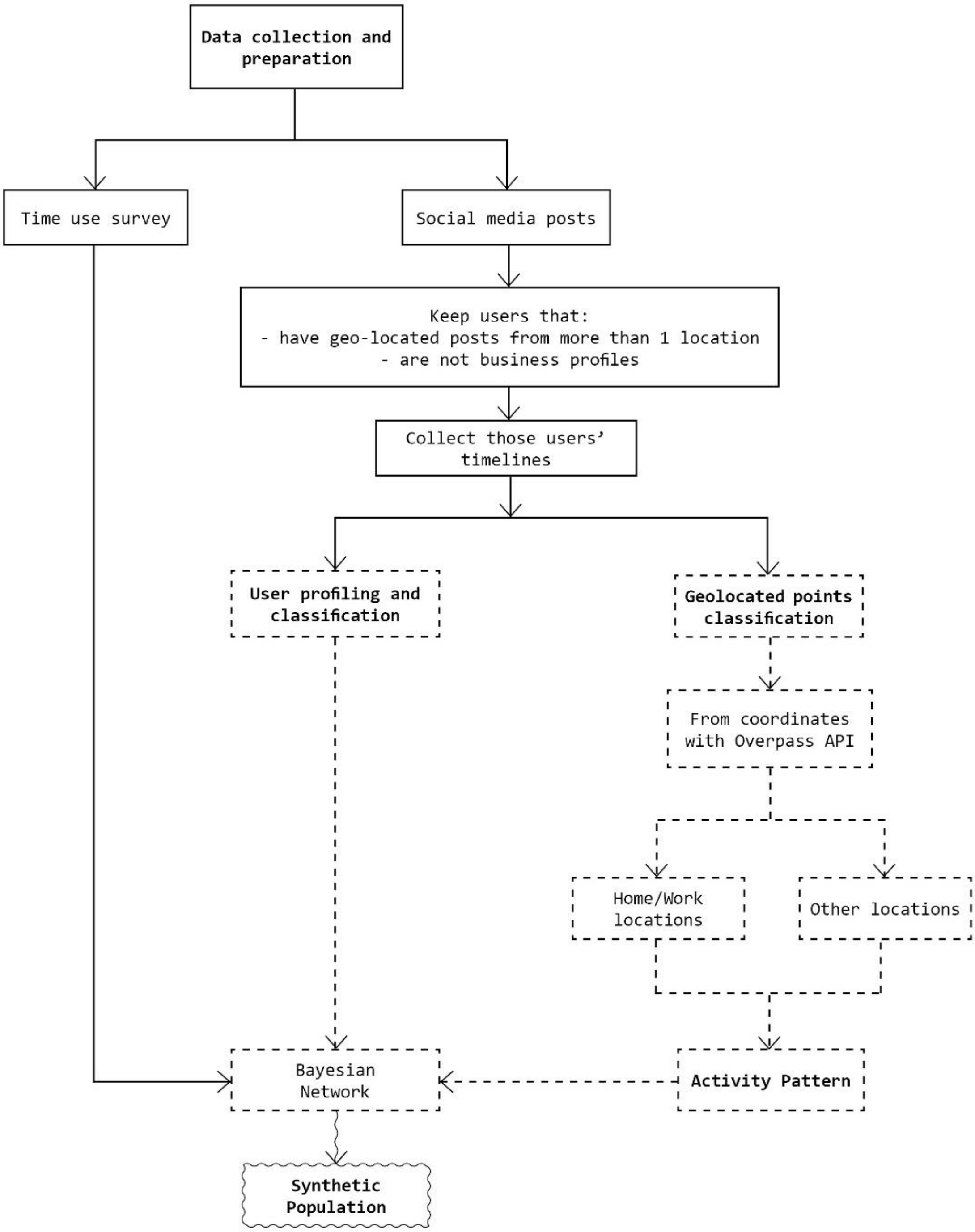

Data collection and preparation
Time use survey
Social media posts
Keep users that:
- have geo-located posts from more than 1 location
- are not business profiles
Collect those users' timelines
User profiling and classification
Geolocated points classification
From coordinates with Overpass API
Home/Work locations
Other locations
Bayesian Network
Activity Pattern
Synthetic Population

Figure 12 / Methodology diagram.

3 / A Synthetic Population from Social Media

In the next chapters the process of creating the synthetic population is explained in detail. The steps below will be followed:

- ↳ *Data collection:* Social media posts are collected from Twitter and Instagram for a period of over two months. They are the source for the activity patterns of the population.
- ↳ *Data preparation:* The social media posts are filtered to eliminate posts without geolocation, users who might be bots, business profiles etc.
- ↳ *Geolocated points classification:* After clustering spatially redundant points, each point is assigned a category according to location, retrieved from OSM's API Overpass. Home and work location for each user are determined by finding the locations from where the user generates more content.
- ↳ *Pattern mining:* Typical week and weekend day mobility patterns are inferred by finding patterns of space-time activity in posts sent in a period of time.
- ↳ *User profiling and classification:* A time use survey obtained from the open access dataset IPUMS [28] is used to train a Bayesian Network model that can predict the profile of each social media user according to their activity table.
- ↳ *Synthetic population:* Sample reconstruction is used to fit the obtained profiles to the sample size of the model, by simply cloning some categories of users until reaching the desired demographics.

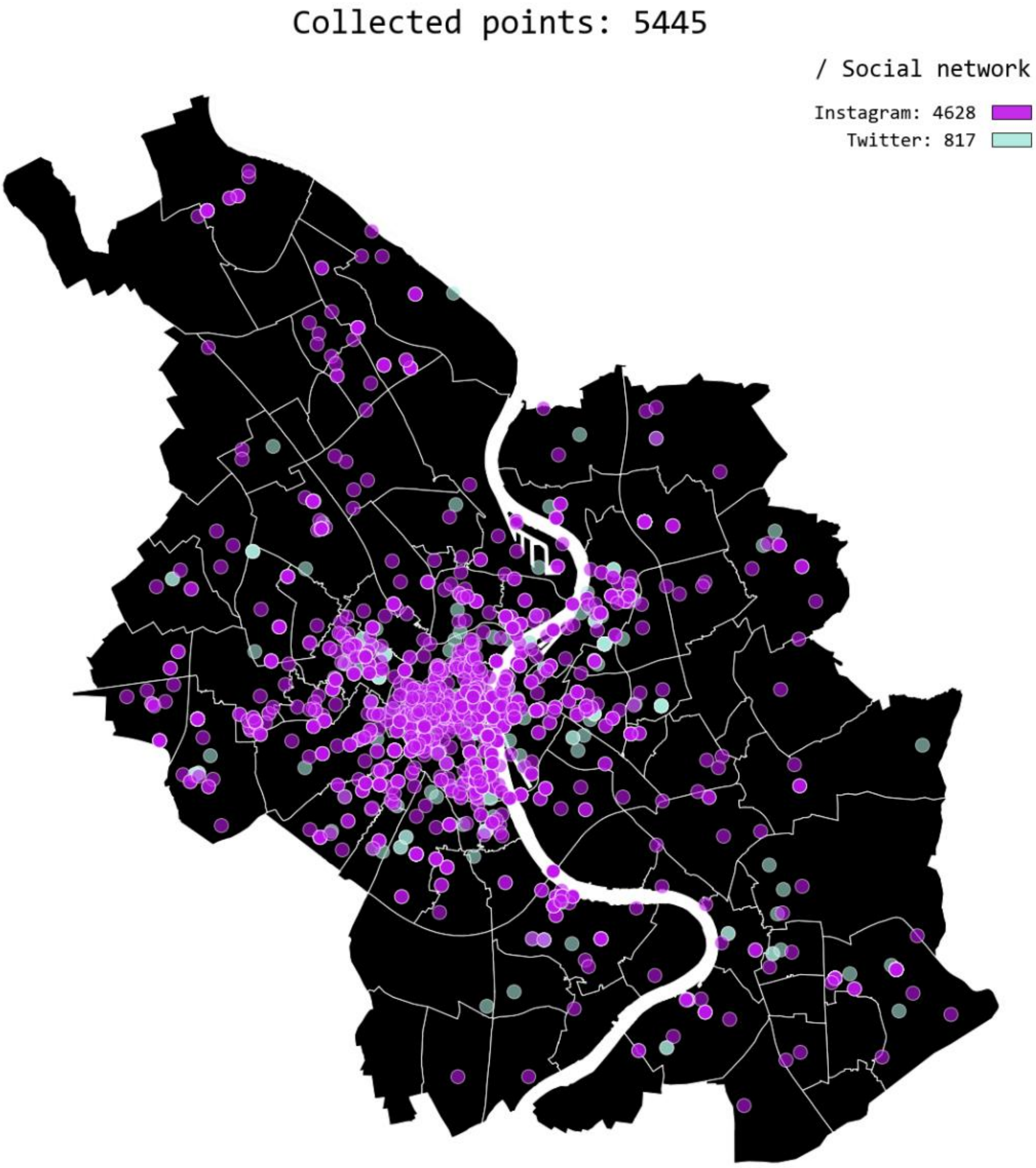


Figure 13 / Geo-located points inside Cologne collected from Twitter and Instagram.

3.1 / Data Collection

Data was collected from two different social networks: Twitter and Instagram.

To collect text messages sent from Twitter, called “tweets” it is necessary to have a Twitter profile and create a developer account. The account provides special access keys and tokens that enable connection to the network’s API (Application Programming Interface). This way the developer can access all tweets that are marked as “public” by the users, that can be filtered in many ways: by city, by coordinates, by language, by a certain time window, by a specific user, etc.

For this research, a connection with the Twitter API was open using the Tweepy [82] library inside the Python programming language. The tweets were filtered by a bounding box around Cologne. For that, it is necessary to inform the API of the coordinates of the lower left and upper right vertices that define the bounding box. Only around 1% of all tweets are geotagged with a precise location, so some tweets that are not actually inside the bounding box might end up being collected. 21.117 tweets were collected between May 27 and August 09. From those, only 1.010 (4,78%) had coordinates. With the API it is simple to choose which variables should be retrieved for each tweet. The variables were collected and saved in a CSV (Comma Separated Values) file:

- ↳ *id_str:* the unique identification number of the tweet;
- ↳ *created:* date and hour when the tweet was posted;
- ↳ *text:* the content of the tweet;
- ↳ *coordinates:* the latitude/longitude from where the tweet was posted (if available);
- ↳ *source:* the device from where the tweet was posted (iPhone, Android, Web etc);
- ↳ *id_user:* the unique identification number of the user;
- ↳ *description:* the user’s profile description text;
- ↳ *location:* the user’s self-declared location;
- ↳ *user_created:* the date and time when the profile was created;
- ↳ *followers:* the user’s number of followers.

For Instagram, a different strategy had to be used. Since the company was acquired by Facebook, their API has been changed after the Cambridge Analytica scandal [24], that involved using social media data to affect elections in the United States and other countries. Probably to avoid similar events, today the Instagram API allows only access of the user’s own account, being impossible to collect public posts from others. The solution left is to user a web scraper, basically a bot that access the Instagram website multiple times to collect data. With the scraper it is possible to filter posts by hashtags or locations. At this time the choice was to use hashtags related to the city (“köln”, “koeln”, “cologne”) instead of location, because posts tagged with the city location will always have the same coordinates from that place.

The Python library Instaloader [41] was used. It can be run even without an Instagram account, but it requires login for download of geo location of posts. Therefore, two Instagram accounts were used for scraping, the author's personal account and a new one created for this research. The use of more than one account shows to be necessary because the scraper reaches fast the limit of requests imposed by the website. This is one disadvantage of Instagram over Twitter. Because of these limitations less Instagram posts could be downloaded per day. During the first two weeks of data collection, when the algorithm was run every day, only 264 Instagram posts could be collected per day, against 902 from Twitter. But Instagram has an advantage on the percentage of geo-tagged posts. 10.700 Instagram posts with the Cologne hashtags were collected between May 29 and August 02. From those, 6.642 (62%) were geo-tagged.

The Instagram pictures are downloaded to a separate folder for each user, together with a JSON (JavaScript Object Notation) [45] file containing the information about the post. An extra JSON file is saved with information from the user's profile. Unfortunately, unlike with the Twitter API, it is not possible with the scraper to choose which information from the profile is going to be retrieved, since everything is downloaded together in the JSON file. But, for this research, no personal information from the user – such as their name – was retrieved from the files. Except for *source*, *location*, *user_created* and *followers*, the same variables extracted from the tweets are extracted from the JSON files, together with three more:

- ↳ *location_id:* the Instagram venue id from where the picture was posted;
- ↳ *location_name:* the name of the location. Can go from a city to a venue;
- ↳ *is_business:* indicates if the profile is a business account or not.

An overview of the amount of data collected can be seen in Table 1.

Table 1 / Comparison of amount of data collected between Twitter and Instagram.

	No of Posts	Posts Geo-tagged	Posts Inside Cologne
Twitter	21.117	1.010 (4,78%)	817
Instagram	10.700	6.642 (62%)	4.628

The privacy issues discussed in chapter 2.3.3 / are taken seriously into consideration. While processing the data, users are identified only by their profile id number, with no names that make them identifiable. At the end of the process, no user can be identified from the created synthetic profiles.

3.2 / Data Preparation

It is said that the work of data science is usually 80% preparation [64] and in this research it was no different. Before selecting the users, whose mobility patterns were to be defined, it is necessary to filter the collected data to find those that pass the following criteria:

- ↳ users with geo-tagged posts;
- ↳ users who post from different locations;
- ↳ users that are people, not businesses.

The filtering process starts very similar for both social networks, and differs a little at the end, because of the slight differences in the data collected.

For both Twitter and Instagram, first all posts without coordinates are discarded. Then, a shapefile with the city boundary is used to filter out all posts that were not sent from there. Lastly, all the users that posted from less than two different locations are deleted. This step already discards a good number of bots or business profiles.

Twitter has a self-proclaimed location field for each user's profile, so all users with a location different from Cologne ("köln", "koeln", "cologne") are discarded. Users with no self-proclaimed location are kept though, once they might still live in the city. Twitter also has a description text field for each profile. Businesses usually describe their activities in this field, so a list of key words is used to try and filter these profiles ("accessoires", "gegründet", "events", "freelance", "job", "handel", "online", "netzwerk", "stellenangebote", "fraktion", ".de", ".com", "uns", "wir", "angebot", "rabatt", "rabattcode", "mail", "e-mail", "kooperationsanfrage", "we", "galerien", "collective"). The list was written manually by observing the collected tweets, therefore it could be expanded as necessary.

For Instagram, cleaning out business profiles is easier, since every profile comes with a "is business account" tag. Some profiles that are business but are not tagged as such remain though, so the same keyword process used for Twitter is used to filter those using the user's profile biography.

After this filtering process, 63 users from Twitter and 100 users from Instagram remain. In the next step, all posts from each of the users' profiles are collected, to determine their individual mobility patterns.

Some profiles could not be found, likely because they were deleted by the users in between the data collection period. Another number of profiles from Instagram could not be collected because of the limitations of the data scraping process mentioned in the previous chapter. Even creating two more Instagram accounts, totalizing four, during the collection of the profiles the limit number of requests from the website was always reached very fast, even before finishing collecting one single user's timeline. Therefore, it was not possible to collect all of the Instagram users' timelines until the end of this research.

Because of these limitations, at the end 122 out of the 163 users' timelines could be collected.

3.3 / Geolocated Points Classification

For the agent-based model, each agent must have an activity table for the day, so all the geolocated points from each user's posts are categorized in activities.

Each user's unique home location is determined as the location in the "Housing" category from where the user posts the most.

In total, 103.835 posts were collected from all the Twitter users' timelines and 12.132 from the Instagram users' timelines, totalizing 115.967. The most active user had 3.247 posts and the less active had four. A more detailed distribution of tweets per user can be seen in the table below.

Table 2 / Users grouped by amount of posts collected from their timelines. The biggest group is marked.

N Posts	<= 10	11-100	101-500	501-1000	1001-2000	2000-3000	>3000
Users	4	30	38	15	9	1	25
%	3,2%	24,5%	31,1%	12,2%	7,3%	0,8%	20,4%

The data goes through the same filtering process of deleting posts without coordinates and posts outside of the city of Cologne. One extra filtering step is deleting the Instagram representative point for Cologne. Every time a user posts a picture with the location "Cologne, Germany", Instagram tags the post with the same coordinates that represent the city. This results in this same point being repeated many times in many user's sets of points. The same can happen with the posts from Twitter, because many users have connected accounts that post their Instagram content also directly on their Twitter account. After this process, the total amount of posts goes down to 9.305, being 2.668 from Instagram and 6.637 from Twitter. The new average of posts per user is 76 and the final distribution of posts can be seen in Table 3.

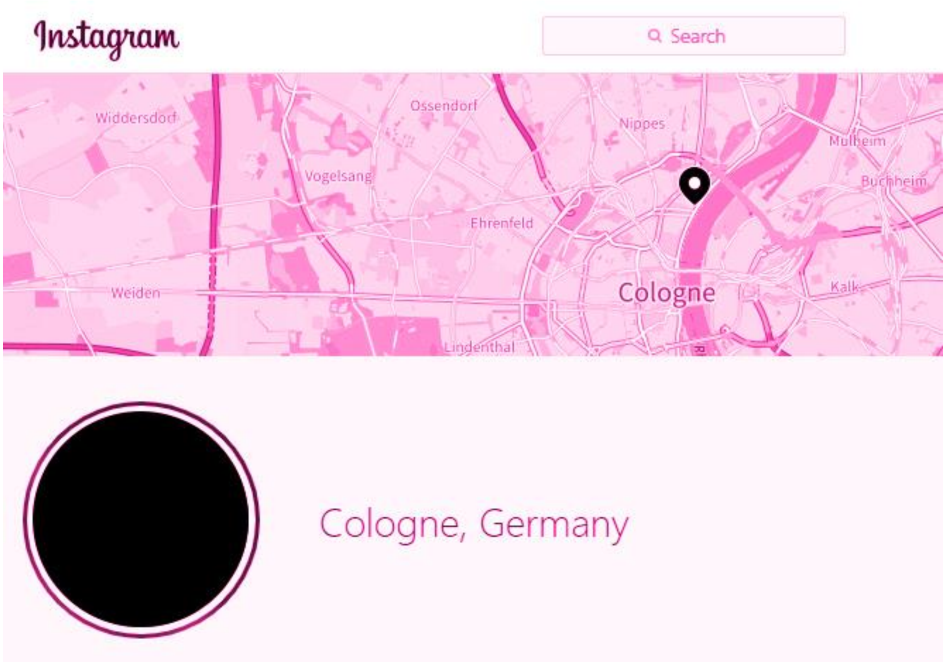


Figure 14 / Instagram's coordinates for the location "Cologne, Germany".

Table 3 / Users grouped by amount of posts left after the cleaning process. The biggest group is marked.

N Posts	<= 10	11-50	50-100	101-250	251-500	501-1000	>1000
Users	48	36	17	14	4	3	1
%	39,3%	29,%	13,9%	11,4%	3,2%	2,4%	0,8%

To exemplify the classification process, we follow the data set of one user, who we will call 'User X' and had after the filtering 225 points, that can be seen in the map below:

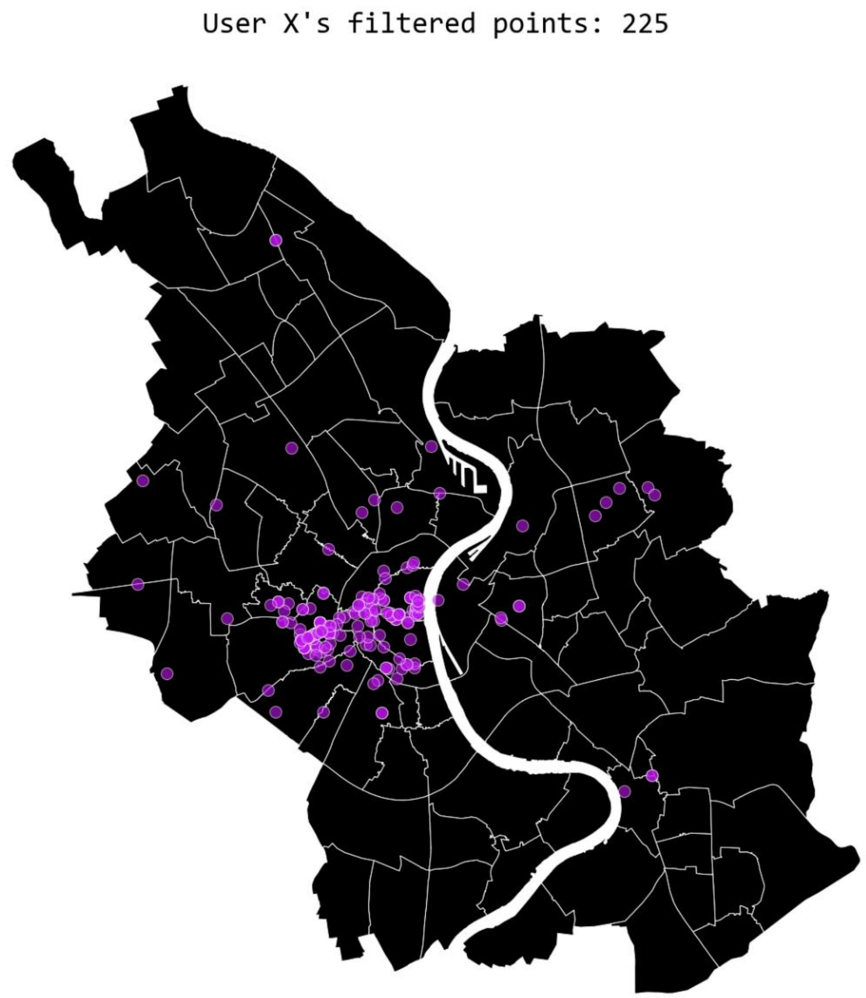


Figure 15 / User X's plotted geolocated points.

3.3.1 / Clustering of Spatially Redundant Points

Before classifying the points, it is assumed that the data contains spatially redundant points: points that represent the same place, but with slightly different coordinates values, due to the duration of time a user spent on that location or due to the accuracy of GPS positioning, which can be around 5-8 m [33]. The multiple time tags will be important later to define the mobility patterns of each user, but to avoid tagging the same place more than once, the points are clustered to reduce the amount of spatial data.

The approach used by Boeing in [17] is followed. He uses Python and its scikit-learn implementation of the DBSCAN density-based clustering algorithm [67] to cluster a data set of GPS points recorded by himself during his holiday vacation. Although mentioning that k-means is likely the most common clustering algorithm, Boeing argues that DBSCAN is superior for clustering latitude-longitude data, because k-means groups N observations into k clusters, meaning that it minimizes variance, not geodetic distance. Another reason to add is that k-means requests an input of the number of clusters to be grouped, which in the social media data set is unknown.

DSCAN takes only two parameters to group the clusters: ϵ is the maximum distance between points to be included in the same cluster, and *min_samples* is the minimum size of each cluster – any cluster smaller than this parameter is considered noise. In this case, *min_samples* is set to 1, so no point is considered as noise. ϵ *is* set to 10 m, the already mentioned GPS accuracy.

From the initial data set of 225 points from the user, 154 clusters are identified. The clusters now need to be reduced to one single point, that will be used to represent that location. For that, a function from Boeing’s paper is used. It calculates the coordinates of the cluster’s centroid and then finds the member of the cluster that is closest to the centroid. Our user had now 154 unique points, that are ready to be categorized into places.

3.3.2 / Reverse Geocoding and Tagging

To classify the points into one of the activity categories, an approach similar to the one in [79] is followed, where they use AddressBase, the definitive source of address information in Great Britain, to find the nearest address point and classify the location of tweets as residential, commercial, or other. Because no such source exists in Germany, the Overpass API [60] is used, that returns information from OpenStreetMaps.

The *around* filter finds all elements within a radius around the input coordinates. The API search takes a considerable amount of time, which is one more reason to cluster the points before. A radius of 30 m is used, to account for GPS accuracy and the average distance that separates one place from another.

The elements returned by the API are tagged with numerous keys and values, describing their location or use. To fit each point in one of the defined categories, a dictionary with tags separated by categories is used. The dictionary was created manually by Müh [55], who observed the tags while collecting OSM data for the environment of the ABM. Some tags were later added by the author, according to the list of most used tags from OSM.

Each point sometimes returns several different elements, and sometimes none. For those that return multiple elements, with multiple tags, the category is defined as the one in where most of the tags fit. Like in the example below:

```
Tags: [‘office’, ‘atm’, ‘bank’, ‘garage’]
Category: ‘DIENSTL’
```

While ‘garage’ belongs to the ‘PARKEN’ category, ‘office’, ‘atm’ and ‘bank’ belong to the ‘DIENSTL’ category, so the latter category is chosen.

If the number of occurrences of tags from different categories is the same, then one of the tags is randomly chosen, together with its category, as shown in the example below:

```
Tags: [‘office’, ‘hotel’, ‘garage’]
Category: ‘HOTEL’
```

Category	Tags
DIENSTL (Office)	office, service, post_office, atm, bank, car_rental, car_sharing, car_wash, driving_school, studio, hairdresser
EINZELH (Shopping)	commercial, residential;retail, retail, fuel, marketplace, post_office, pharmacy, vending_machine, convenience, supermarket, parking, parking_entrance, garage, garages, parking_space
GASTRO	foodservice, cafe, canteen, fast_food, restaurant, ice_cream, internet_cafe
HOTEL	hotel
KULTUR (Culture)	cultural, gathering, religion, religious, chapel, church, place_of_worship, community_centre, library, theatre, amusement_hall, arts_centre, memorial
NACHTL (Nightlife)	bar, pub, casino, gambling, nightclub, nightlife
SCHULE (School)	education, school, language_school, kindergarten, childcare
WOHNEN (Housing)	residential, apartments, dormitory, house
WOHNEN;DIENSTL	residential;office
WOHNEN;EINZELH	residential;commercial
WOHNEN;INDUSTRIE	residential;industrial
WOHNEN;ÖFFENTLICHKEIT	public;residential
HOCHSCHULE (University)	college, university
INDUSTRIE (Industry)	industrial, storage, warehouse
GESUNDHEIT (Health)	medical, hospital, nursing_home, social_facility, dentist, doctors, veterinary
ÖFFENTLICHKEIT (Public)	civic, public, fire_station, police, public_building, townhall
FREIZEIT (Leisure)	sport, stadium, windmill, bicycle_parking, boat_rental, water, forest, grass, meadow, wetland, river, swimming_pool, pond, park, farmyard, farmland, grassland, lake, garden, farm, picnic_site, square, brownfield, cemetery, recreation_ground, village_green, beach_resort, common, green, pitch, track, sports_centre

Table 4 / Dictionary of categories and tags.

140 out of 154 of the user’s cluster center points returned results and were classified inside the categories. The remaining 14 points received a ‘none’ tag. With the cluster’s center points categorized, the user’s total list

of coordinates is rearranged. Each original point is replaced by the center point of the cluster it belongs to and is also categorized as such.

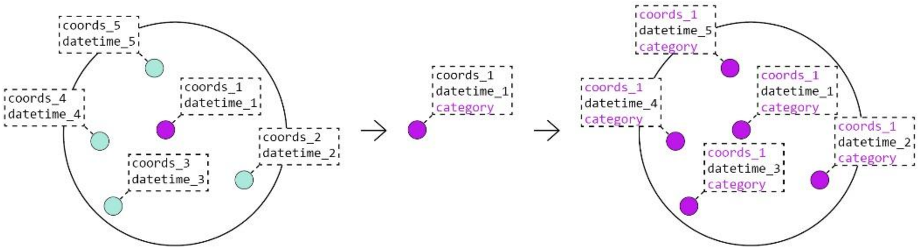


Figure 16 / Diagram of the clustering and tagging process.

This “location-based” method of classification of the posts only through their coordinates was chosen over “content-based” methods (“you are where you write about”), because, as mentioned in [73] and similarly in [89], when someone writes about an event, they might just be following it and not necessarily be there. It might also happen that someone writes about activities that they want to do, or write something about work or school, without being there. Some posts can also be too general to extract any information out of it. By using the coordinates, there is a higher chance of categorizing more points.

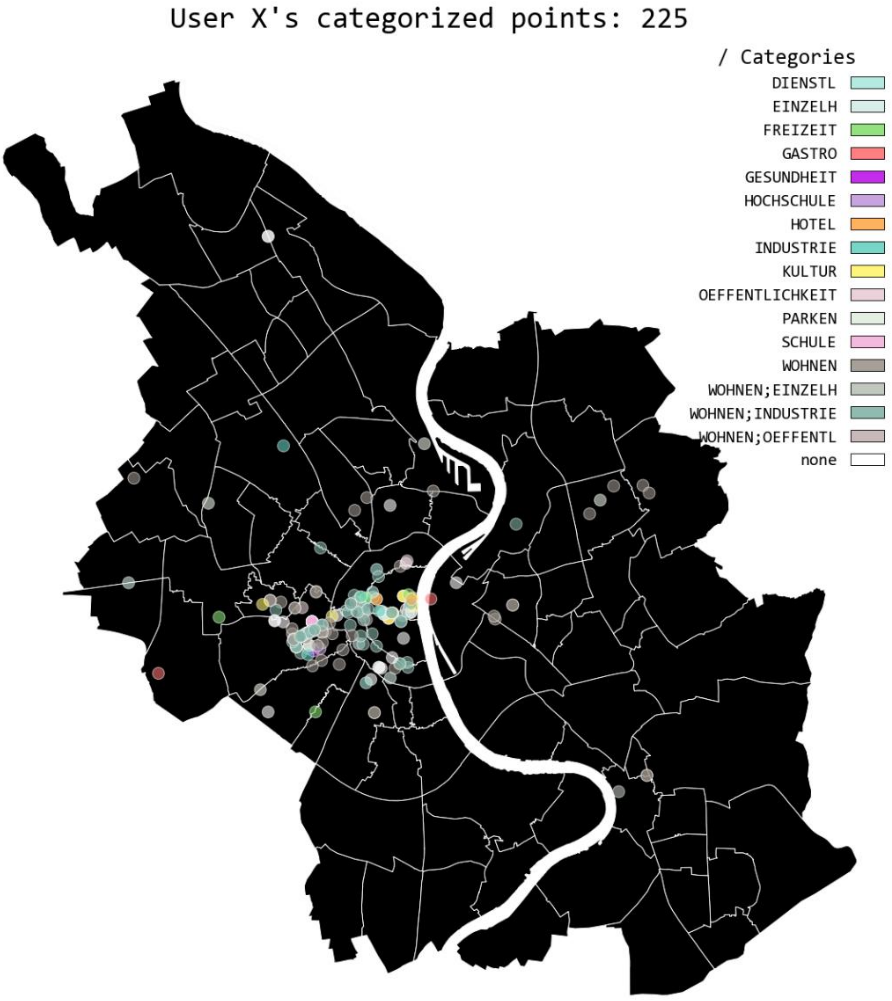


Figure 17 / User X’s categorized points.

3.3.3 / Defining Home and Work Location

Home and work location can be determined by finding the locations from where the user generates more content, like shown in [25] and [52]. With all the points categorized, the home location is defined as the most repeated point in any of the ‘WOHNEN’ categories. As mentioned by Swier et al. [79], “this is a bold assumption” that might not always be right but “seems reasonable”.

For some users, this approach is very straightforward and returns only one single point, like in the case of User X, whose home point was found in the category ‘WOHNEN;INDUSTRIE’. Others have more than one point from where the user posted an equal amount of times. In this case, the time span from the first post until the last for each point is found. The home point is then defined as the one with the longest time span: the user posted from this location for a longer period, so it can be assumed that it is more likely to be a home location, instead of a location that they might have frequented for a while but not consistently.

Users with no points categorized as ‘WOHNEN’ are discarded in the end, because it would be impossible to infer a mobility pattern without a home location.

The work or study location will be defined as simply the user’s most repeated point, that is different from the home point and belongs to any category other as ‘WOHNEN’, ‘FREIZEIT’ or ‘KULTUR’. For User X, this point was in the ‘SCHULE’ category. It is difficult to determine now if the user is a student or if they might work at a school. Later, when the geospatial data is combined with the time stamps, it might be possible to determine that from how long the user stays in this place. A student would spend less time in the school daily than a teacher.

As mentioned earlier, these are all bold assumptions and it was not possible at this time to validate this information for all the users. However, after manually inspecting User X’s profile on the social network, it was possible to determine that both home and work locations were probably correct, from two posts in which they mention they were at home and at work.

The distance from one user’s home location to their work location is calculated and saved for future use when assigning these locations for each people in the synthetic population of the model.

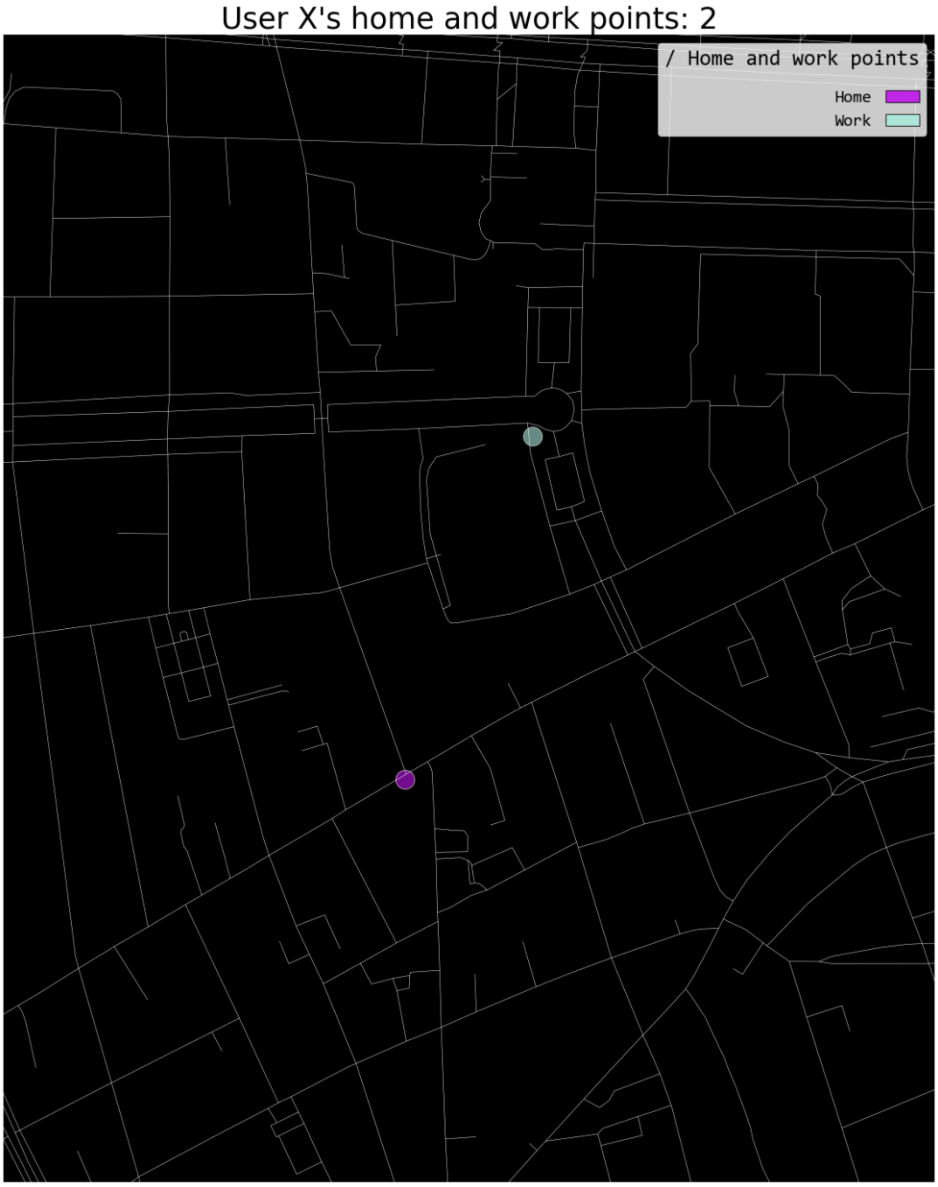


Figure 18 / User X’s inferred home and work location.

3.4 / Pattern Mining

Before deriving each user's mobility pattern, it is necessary to filter those that have enough information to do so, meaning a minimum number and variety of activities. To define these requirements, we look into two main references. In "Life Between Buildings", Gehl [30] simplifies outdoor activities in public spaces in three categories: *necessary activities*, *optional activities* and *social activities*.

- *Necessary activities* are those that a person has no choice whether doing it, they are a requirement and happen throughout the year and independently of the environment. This involves going to work or school and shopping. Most of the walking is done in this category.
- *Optional activities* are those that a person participates in if they have the desire, time, and adequate place to do so. They are highly influenced by external conditions, like the weather or the physical environment. Taking a walk or simply sitting on a bench in the park are in this category.
- *Social activities* are in a big part resultant from the other two categories. In public spaces of poor quality mostly necessary activities occur: people just want to pass by, do what they have to do and go home. In public spaces of good quality, on the other hand, the levels of optional activities arise and, together with them, the social activities. Those are the activities that depend on the presence of others in the public space, be it kids playing outside or just watching people passing by. The contacts can be active or passive, from meeting acquaintances to just seeing and hearing other people.

Castiglione et al. [19] present, on the other hand, four general categories in which activities are grouped for models: *mandatory*, *maintenance*, *discretionary* and *at-home*. Similar with Gehl's categories, they are also influenced by their priority in the daily activity pattern schedule.

- *Mandatory activities* are work and school. They are the base for activity schedules and the least flexible.
- *Maintenance activities* include shopping, going to the doctor or dropping the kids in school. They are more flexible and in some models are even grouped together with the discretionary activities, because of their similarities in schedule.
- *Discretionary activities* are the recreational ones, like eating out or visiting friends. They are the most flexible in terms of scheduling and can sometimes be replaced by at-home activities.
- *At-home activities* can be divided into working at home and other at-home activities, that are not particularly relevant to distinguish.

These two different categorizations of activities give an overview of what typically consists a daily activity schedule. The Primer categories can be distributed inside Gehl's categories. And our categories can fit in both, as seen in the table below.

It is defined then that to be considered a valid activity schedule, each user must have at least three necessary activities (one mandatory, one of maintenance and one at-home) and one optional/discretionary activity. From the total of 122 users, 83 ended up being profiled.

Gehl's categories	Primer categories	Our categories
Necessary activities	Mandatory activities	DIENSTL (Office)
		SCHULE (School)
		HOCHSCHULE (University)
		INDUSTRIE (Industry)
		ÖFFENTLICHKEIT (Public)
	Maintenance activities	GESUNDHEIT (Health)
		EINZELH (Shopping)
		GASTRO
	At-home activities	WOHNEN (Housing)
		WOHNEN;DIENSTL
		WOHNEN;EINZELH
		WOHNEN;INDUSTRIE
		WOHNEN;ÖFFENTLICHKEIT
Optional activities	Discretionary activities	FREIZEIT (Leisure)
		NACHTL (Nightlife)
		HOTEL
		KULTUR (Culture)
Social activities	Resultant of the others	

Table 5 / Our activity categories organized according to the two different categorizations.

Two different mobility patterns will be created for the user: one for weekdays and one for weekends. According to the Primer, in models time is divided in intervals of 60, 30 or even 15 minutes, in which an activity can start or end. For simplification of our model, time will be divided in intervals of 60 minutes.

For each user two tables with 24 slots corresponding to each hour are created, one for weekdays and one for weekends. Every geolocated point is appended to the slot corresponding to the time when the person posted from that location, in the appropriate table depending if the post was created on a weekday or weekend. Finally, the most repeated point in each hour is found and their corresponding activity type appended. If the points in an hour slot are repeated the same amount of times, one of the points is chosen at random.

The plots of the frequency of activity per hour on the next pages show potential in using geolocated social media posts to infer activity patterns. We can see that, on weekdays, ARBEIT occurs consistently during working hours, and so does EINZELHN. SCHULE has a peak in the morning and another one after lunch, probably reflecting the time when kids go first to school in the morning and then the time they come back from lunch. GASTRO has also peaks during lunch and dinner hours. And ZUHAUSE appears as the major activity during the whole day, which might be reflective of the period when the data was collected, during the COVID-19 pandemic, when a great part of the population was quarantining in compliance to the social distancing rules and because most commercial places were closed. The Fraunhofer Institute conducted a survey [39] in May 2020 with about 500 companies and found out that 70% of their employees were working completely from home during the pandemic, while 21% worked part from home, part in the company. Which shows that we would probably see higher occurrences of the other activities in a normal period of time. However, the pandemic may cause lasting changes in how people work, so this high proportion of in-home activities might be representative of the future.

Cui et al. mentioned in [25], that, in general, their algorithm performs better in determining home than work activities: "the poor performance usually happens to users who do not use Weibo frequently at workplace, or users who would post a microblog only when they are working overtime, which time period is not covered by the peak work event hours defined in our algorithm". That might also be an explanation for work events in strange hours in this research.

These tables represent a fraction of a typical weekday and weekend for each user, but of course, they are not complete since it is impossible for even very active users to collect posts for every hour of the day.

Additional to gaps in the timetable with no activities, some anomalies can also be observed, such as particular activities happening in unlikely hours, like GESUNDHEIT at 1:00 am or NACHTL in the middle of the day. These seemingly mistakes are probably related to the reverse geotagging done with OSM, that could sometimes retrieve wrong activity tags for the buildings.

A simple approach is made to solve these issues. First, for the model to be more dynamic, instead of keeping just one activity per hour, which would result in the agents always repeating the same pattern, with no spontaneous decisions like in real life, all the occurred activities in each hour are returned. This way, in some cases, the agent can choose randomly among these activities and add some variety to the model.

Second, a list and time windows for night activities are defined. Inside these time windows from 00:00 to 05:00 and from 21:00 to 23:00, anything that is not in the list of night activities is deleted. The list includes: GASTRO, KULTUR, NACHTL, FREIZEIT, HOTEL and ZUHAUSE. After that, ARBEIT is prioritized over other activities that might appear in the same time slot, which are deleted.

In preliminary runs of the agent-based model, we noticed that, during lunch hours, the occurrence of GASTRO activity did not seem as high as it should be. In a strategy to make the activity appear more during these hours, time windows for eating are defined, from 11:00 to 16:00 and from 19:00 to 22:00. When GASTRO appears in one of these windows next to other activities, just as ARBEIT it is chosen as the main activity over the others, that are deleted.

Finally, to fill the gaps with the closest known activity possible, each empty gap looks first for the next hour and copy its activity if there is one. If the next slot is empty, it checks the previous one and copy its activity. When both slots, previous and next, are empty, the gap is filled with ZUHAUSE.

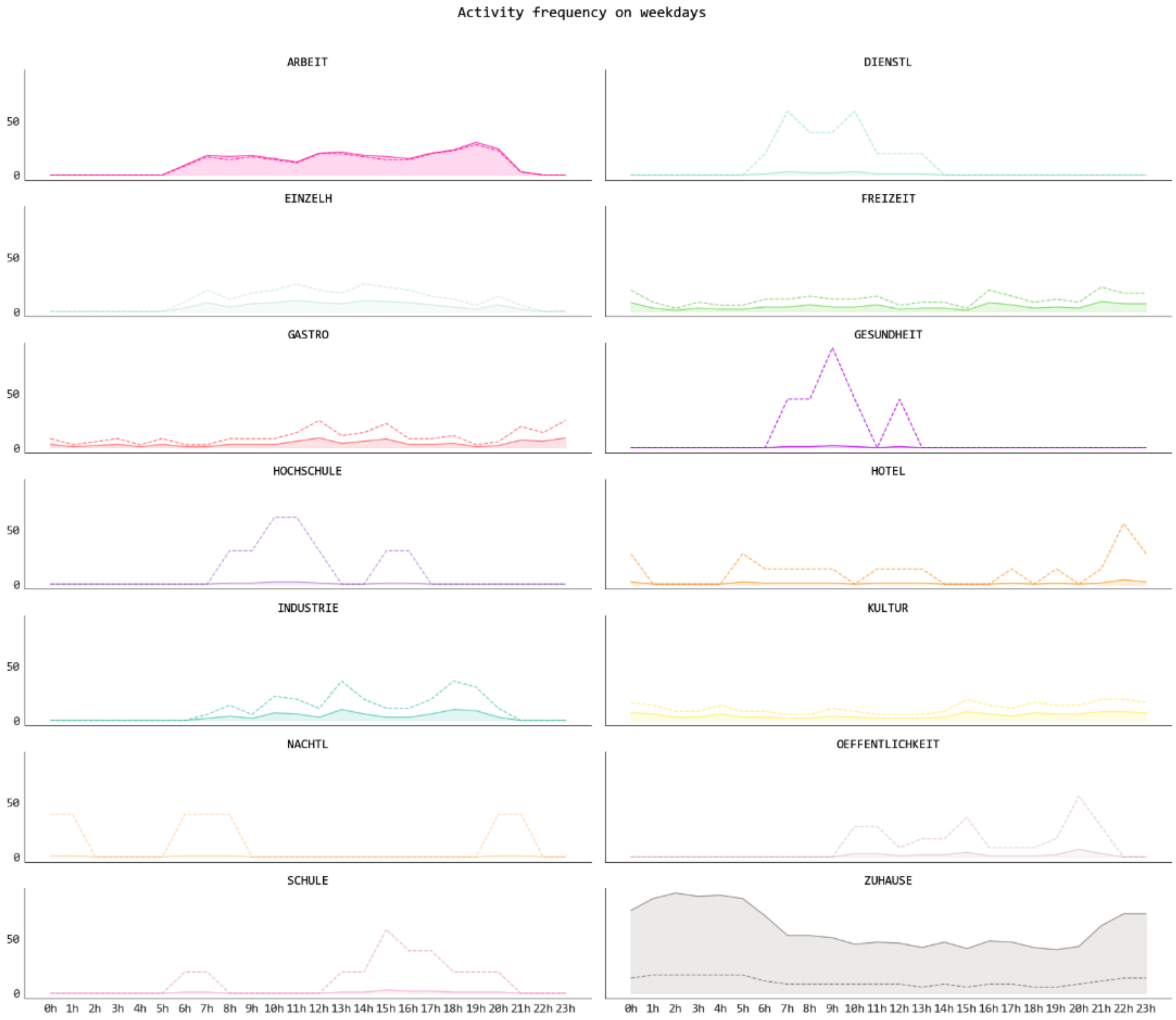


Figure 19 / Activity frequency on weekdays.

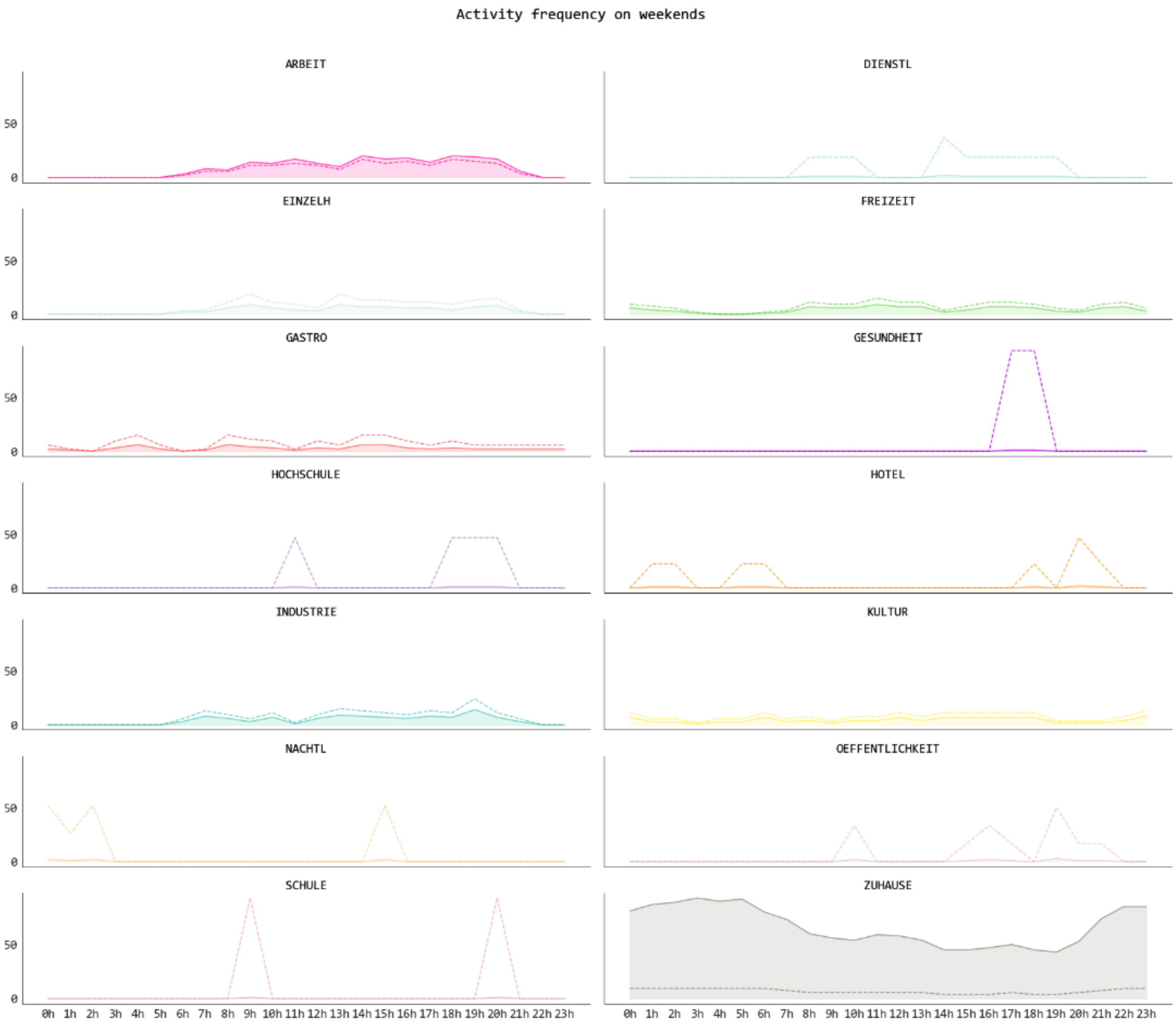
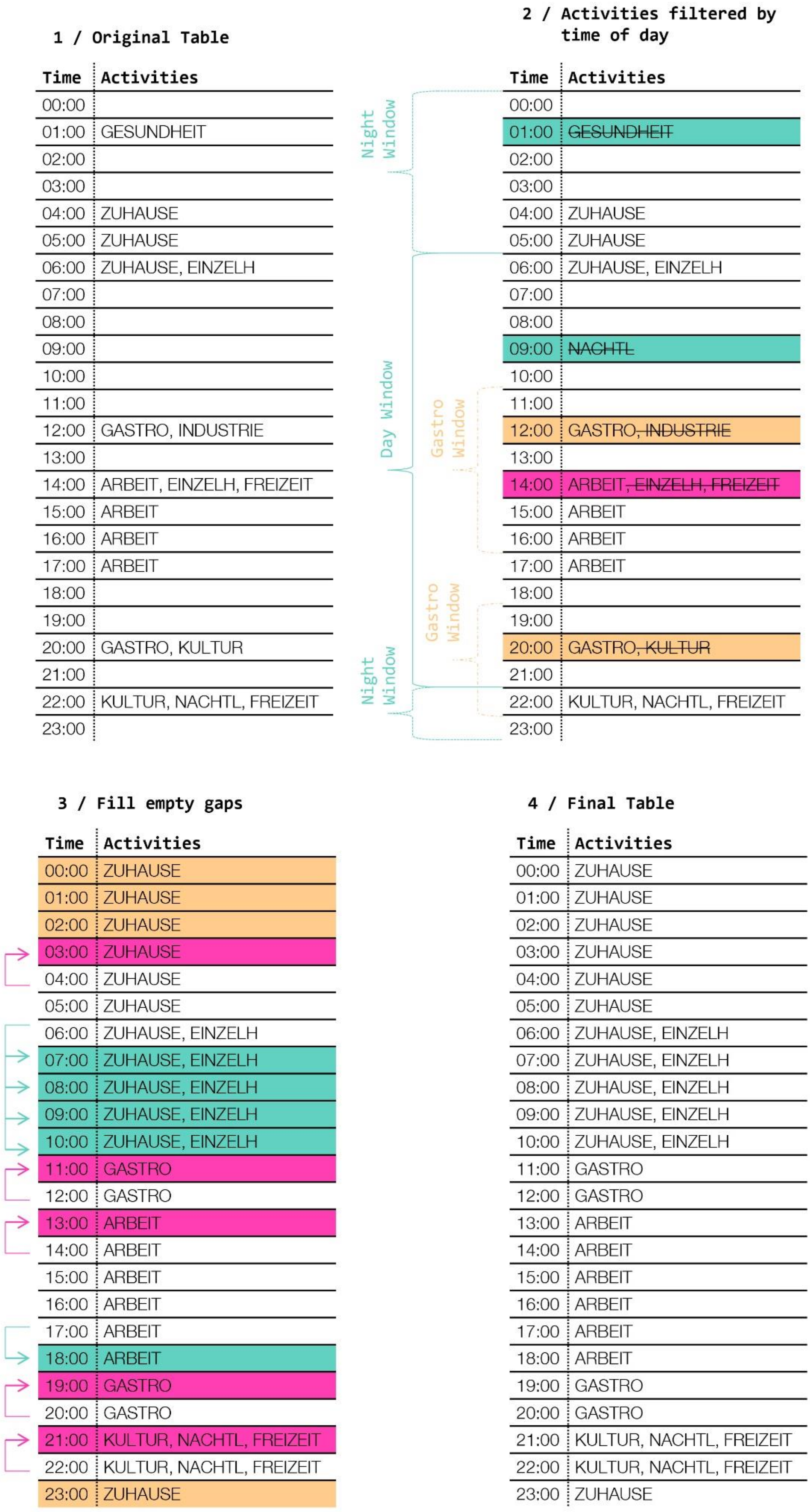


Figure 20 / Activity frequency on weekends.
- (% total of all activities)
-- (% total of each activity)



1 / Original Table

Time	Activities
00:00	
01:00	GESUNDHEIT
02:00	
03:00	
04:00	ZUHAUSE
05:00	ZUHAUSE
06:00	ZUHAUSE, EINZELH
07:00	
08:00	
09:00	
10:00	
11:00	
12:00	GASTRO, INDUSTRIE
13:00	
14:00	ARBEIT, EINZELH, FREIZEIT
15:00	ARBEIT
16:00	ARBEIT
17:00	ARBEIT
18:00	
19:00	
20:00	GASTRO, KULTUR
21:00	
22:00	KULTUR, NACHTL, FREIZEIT
23:00	

2 / Activities filtered by time of day

Time	Activities
00:00	
01:00	~~GESUNDHEIT~~
02:00	
03:00	
04:00	ZUHAUSE
05:00	ZUHAUSE
06:00	ZUHAUSE, EINZELH
07:00	
08:00	
09:00	~~NACHTL~~
10:00	
11:00	
12:00	GASTRO~~, INDUSTRIE~~
13:00	
14:00	ARBEIT~~, EINZELH, FREIZEIT~~
15:00	ARBEIT
16:00	ARBEIT
17:00	ARBEIT
18:00	
19:00	
20:00	GASTRO~~, KULTUR~~
21:00	
22:00	KULTUR, NACHTL, FREIZEIT
23:00	

3 / Fill empty gaps

Time	Activities
00:00	ZUHAUSE
01:00	ZUHAUSE
02:00	ZUHAUSE
03:00	ZUHAUSE
04:00	ZUHAUSE
05:00	ZUHAUSE
06:00	ZUHAUSE, EINZELH
07:00	ZUHAUSE, EINZELH
08:00	ZUHAUSE, EINZELH
09:00	ZUHAUSE, EINZELH
10:00	ZUHAUSE, EINZELH
11:00	GASTRO
12:00	GASTRO
13:00	ARBEIT
14:00	ARBEIT
15:00	ARBEIT
16:00	ARBEIT
17:00	ARBEIT
18:00	ARBEIT
19:00	GASTRO
20:00	GASTRO
21:00	KULTUR, NACHTL, FREIZEIT
22:00	KULTUR, NACHTL, FREIZEIT
23:00	ZUHAUSE

4 / Final Table

Time	Activities
00:00	ZUHAUSE
01:00	ZUHAUSE
02:00	ZUHAUSE
03:00	ZUHAUSE
04:00	ZUHAUSE
05:00	ZUHAUSE
06:00	ZUHAUSE, EINZELH
07:00	ZUHAUSE, EINZELH
08:00	ZUHAUSE, EINZELH
09:00	ZUHAUSE, EINZELH
10:00	ZUHAUSE, EINZELH
11:00	GASTRO
12:00	GASTRO
13:00	ARBEIT
14:00	ARBEIT
15:00	ARBEIT
16:00	ARBEIT
17:00	ARBEIT
18:00	ARBEIT
19:00	GASTRO
20:00	GASTRO
21:00	KULTUR, NACHTL, FREIZEIT
22:00	KULTUR, NACHTL, FREIZEIT
23:00	ZUHAUSE

Figure 21 / Diagram of the activity table preparation.

3.5 / User Profiling and Classification

Similar to the approach in [59] a Bayesian Network model is used to predict the missing characteristics of the user according to their activity table. Using a travel survey dataset to train the model, it can learn the connections between the activities that each person attends the most with their profile type and ownership of vehicles.

The model is built with the pomegranate Python library [63], that makes it very easy to train a Bayesian Network from data when the graph structure is unknown.

Initially, the plan was to use the latest time use survey available from Germany (Zeitverwendungserhebung 2012/2013) as training data. The data was requested at the Statistiches Bundesamt for use in an university research project, which requires that a contract is made with the university and has a waiting period of up to two weeks to receive the data. Because of the complexity and time necessary for the request to be processed, it was decided at this time to not use this data.

Instead, a time use survey from the Netherlands from the year 2005 is used. The data was obtained from IPUMS [28], the world's largest accessible database of census microdata. Part of the Institute for Social Research and Data Innovation at the University of Minnesota, it includes records from over 100 countries, that are available free of charge. Unfortunately, there is no time use survey data from Germany available, but there is from nearby countries, like Austria and the Netherlands. The dataset from the Netherlands was chosen for being more recent than the one from Austria, that is from 1992.

The dataset contains the time use diaries of 15.428 people and their demographic information, including employment status, age and vehicles per household. The diarists are categorized in our six types of people, further discussed in chapter 4.2.1 / , depending on their work or study status, age and income.

- ↳ *Retirees*: retired status
- ↳ *High School student*: student status / age under 18
- ↳ *College student*: student status / age over 18
- ↳ *Young professional*: employed status / age under 30
- ↳ *Executives*: employed status / age over 30 / income highest 25%
- ↳ *Mid-career workers*: employed status / age over 30 / income middle 50% or lowest 25%
- ↳ *Home maker*: unemployed status

The vehicles per household are separated into none, animal, non-motorized, 1 car/motorcycle and 2+ cars/motorcycles. In the training dataset, none and animal mean no vehicle, non-motorized means bike and the other two mean car.

Lastly, the places that the diarist visit in one weekday are ranked from 0 to 5, from the most to the least repeated place. Places that do not appear in the diarist's table are signalized with the number 6.

Each diarist has now a list with numerical values representing 8 different variables.

Type	Vehicle	1 Home	2 Other	3 Work	4 School	5 Services	6 Leisure
0 (Mid-career workers)	0 (none)	0	0	0	0	0	0
1 (Home maker)	1 (bike)	1	1	1	1	1	1
2 (Retirees)	2 (car)	2	2	2	2	2	2
3 (High School student)	3 (bike and car)	3	3	3	3	3	3
4 (Young professional)		4	4	4	4	4	4
5 (Executives)		5	5	5	5	5	5
6 (College student)		6	6	6	6	6	6

Table 6 / Data organization for training of the model.

To assure that the model will have the same accuracy in predicting each type of people, the number of samples for each type are equalized. The least common type was College student, with 435 samples. The model is then initially trained on 3.045 samples, being 435 of each type.

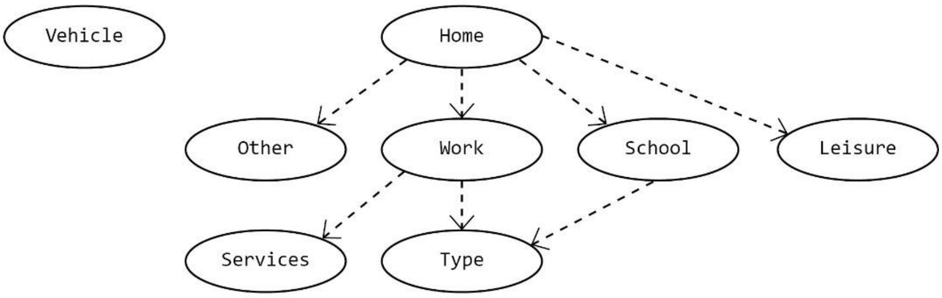


Figure 22 / Network structure of the first trained model.

As seen in the structure graph of the first model, it found that type is influenced by the amount of time that the person spends in school or work, the latter being influenced by the amount of time the person spends at home. Services are influenced by work, which makes sense, since adults are more likely to use services than students. And leisure and other activities are again influenced by the time one spends at home. The model just could not find any connection between owning a vehicle and the other variables.

Before testing the model, the data from the social media users is processed in the same way as the data from the diarists, resulting in an array of eight variables for each user. The only difference is that the first two variables, type and vehicle, are given *None* values, indicating that they need to be predicted by the model based on the other six.

```
Example array: [None, None, 0, 5, 6, 6, 1, 2]
```

For the places variables, the weekday activity table of each user is used, with each hour containing the most repeated activity in that period, as opposed to the table we export to the agent-based model, that can have optional activities in some hours, as explained in chapter 3.4 / . The activities are also classified in the groups of places from the survey, to assure consistency between the data.

Model variables	Survey places	Our places
Home	At own home	ZUHAUSE
Other	At another’s home	FREIZEIT
	At place of worship	KULTUR
	Other locations	HOTEL
	Location unknown	
Work	At workplace	ARBEIT (when work_type ≠ SCHULE or HOCHSCHULE)
School	At school	SCHULE
		HOCHSCHULE
		ARBEIT (when work_type = SCHULE or HOCHSCHULE)
Services	At services or shops	DIENSTL
		INDUSTRIE
		ÖFFENTLICHKEIT
		EINZELH
		GESUNDHEIT
Leisure	At restaurant, bar etc	GASTRO
		NACHTL

Table 7 / Categorizing activities for consistency of the data.

The first prediction results seemed reasonable for the type variable, with occurrences of all types of people, except High School students. That might be because High School students and College students have a similar activity table. However, for the vehicle variable, the results were problematic: the network predicted the same state for all users, that they all have a car, which is very unlikely. Looking again at the survey data, for all types of people, most of them actually had cars, which might be because the vehicle ownership variable is according to household, not according to each individual person. But since the interest of this research is in how each user will travel individually and which modes they will use, independently on what might be available in their household but not used by them, the source of the vehicle variable is changed.

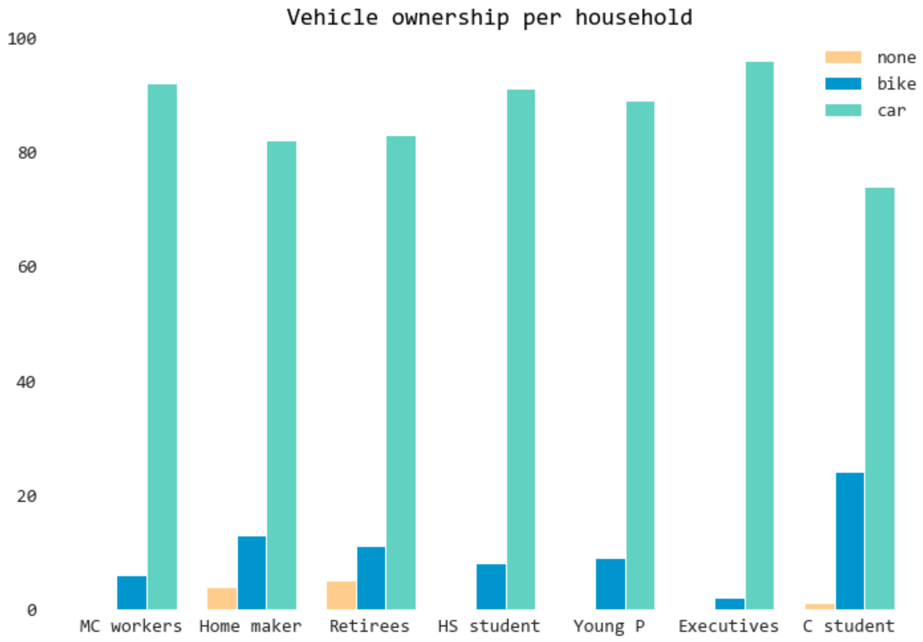


Figure 23 / Vehicle ownership according to household (%).

Additional to the household vehicle ownership variable, each diarist also has a list of travel modes they used during the day in which the diary was kept. The possible modes are travel by car etc, public transport, walk/on foot, other physical transport, other/unspecified transport. If travel by car appears in the user diary, it is considered that they have a car. Other physical transport, other/unspecified transport is considered as having a bike. Else, the user has no vehicle. The advantage of this approach is that contrary to the household vehicle ownership variable that had no value for owning both types of vehicles, now it is visible who uses both car and bike. The statistics seem more realistic: students have more bikes while the other types have more cars.

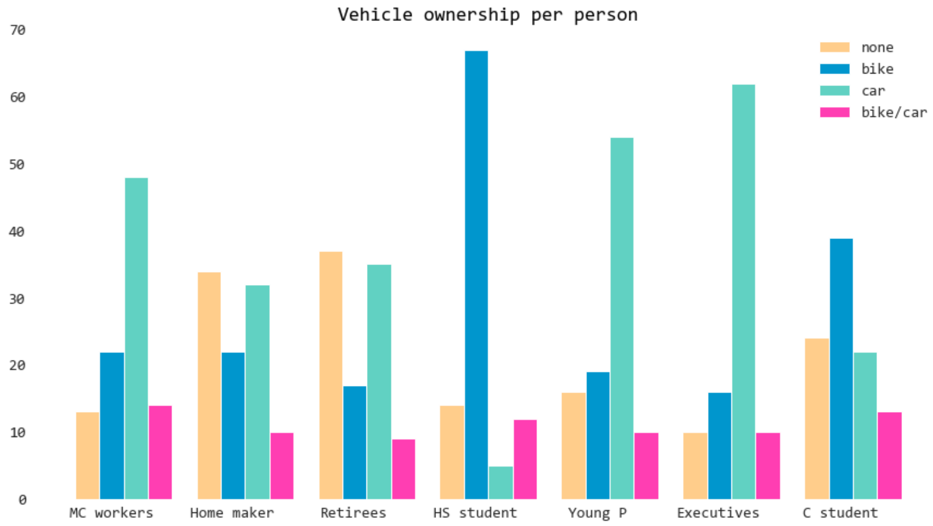


Figure 24 / Vehicle ownership according to vehicle usage (%).

When creating the set of 435 samples for each type of user, now the proportions of the vehicle variable from the whole dataset is kept, to assure that the model gets trained correctly.

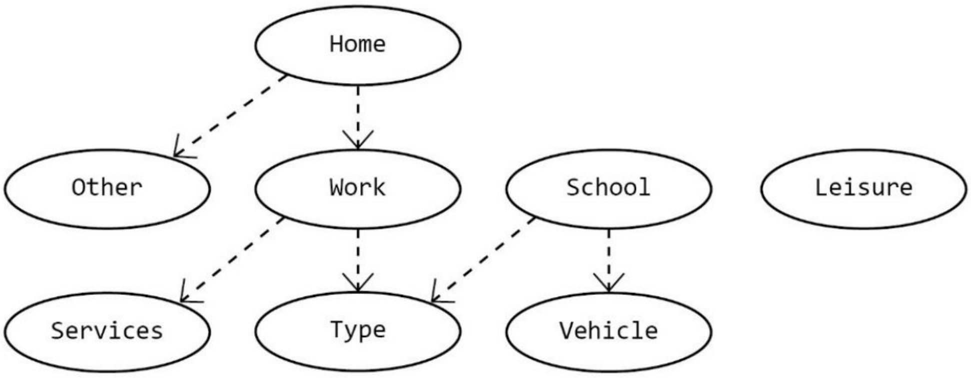


Figure 25 / Network structure of the final trained model.

The results of the changes are clear in the diagram of the final model's structure. Vehicle is now influenced by school, because the students are the ones that tend to have more bikes. Leisure activities, on the other hand, are now independent of the other variables, which could make sense, considering that all the types of people probably have leisure at some point.

This model seems to make more realistic predictions: the vehicle variable is different among users. But there is still one type of person missing, this time Home maker. That might be because, just like High School students and College students, Home makers and Retirees have similar activity tables, mostly because they spend a lot of time at home. For this reason, we decide to remove two of the types from the agent-based model. We keep Home maker and remove Retirees, and merge High School students and College students into one type Students.

Table 8 / Final distribution of people types in the profiles set.

Types	Student	Young professional	Home maker	Mid-career workers	Executive
Users	9	12	23	20	19
%	7,47%	9,96,%	19,09%	16,6%	15,77%

Although better than before, the vehicle ownership prediction still has problems. Only students have bikes and all the other types have cars. Very few users have both vehicles, and nobody has none. That might be because the size of the profiles samples is too small to make a realistic prediction for each type of person. There should certainly be some executives or young professionals that own bikes, for example. Because it is not possible at the moment to obtain more data to create more profiles, or to try and tune more the Bayes Net, we decide to attribute vehicles to each person inside the agent-based model, according to the demographics of the population and aggregated statistics about vehicle usage from Germany.

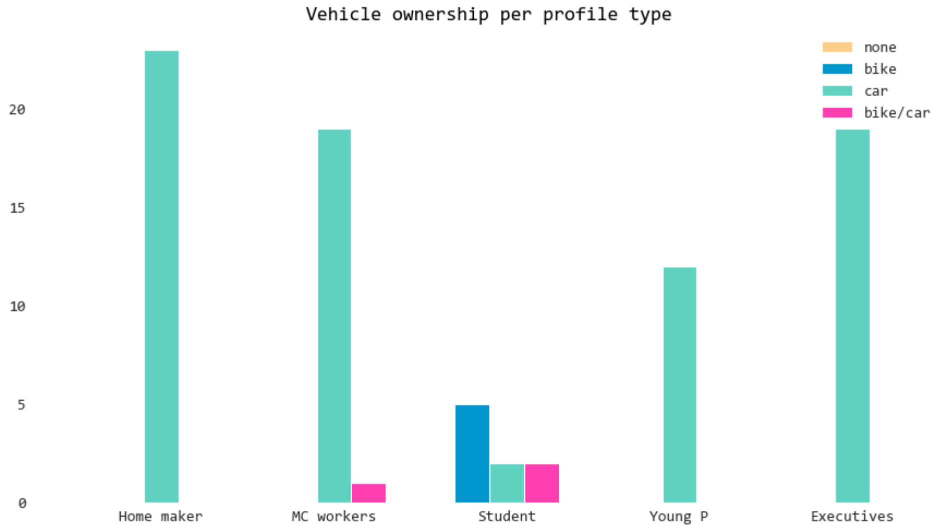


Figure 26 / Final distribution of vehicles in the created profiles.

Finally, the profiles are exported to CSV files with the following variables:

- Type;
- Work distance;
- Work type;
- Activity table weekday;
- Activity table weekend.

The sample reconstruction, which is the final step of creating the synthetic population, is done inside the agent-based model. It consists in drawing samples of people profiles until the distribution matches the demographics desired for the model. Since the demographics of the district will be an interactive parameter, that can be adjusted to visualize different scenarios, the sample reconstruction is made alongside with the running of the model, as necessary.

3.6 / Limitations and Future Development

In favour of building the profiles for the agent-based model, the social media data had, at times, to be manipulated to fit our purposes. Additionally, some technical issues might have affected the accuracy of the obtained profiles.

As shown in chapter 3.4 / , some anomalies could be found in the table of activities, such as a GESUNDHEIT check-in at 2:00 am, probably related to a wrong tag collected from OSM for that location. Because OSM is open source and updated by multiple users, many place tags might be missing, and some might be incorrect or incomplete. It might also happen that a building has more than one use, so a commercial tag from the first floor might be retrieved from a location where the user posted from their apartment on a floor above. In the future, the reverse geotagging could be proofed by building context for the activity from the user's posts, being it text analysis for Twitter and additionally image classification for Instagram. Another option would be to use a different API for the reverse geocoding. The Overpass API was chosen for being free of charge, but, when resources were available, the Google Places API [32], that retrieves places information from Google Maps, could be an option. Although the API offers some free resources, the reverse geocoding option comes with a fee per number of API calls, which was not accessible at this time.

In the same chapter, the gaps in the activity table are filled with the previous or next activity, on a simple assumption that the person might have continued the same activity for that period. Although this approach can work perfectly fine for our synthetic population for now, it is clear that it is only an approximation of the activity and mobility patterns of these anonymized individuals. To have an accurate representation of a sample of the population, much more data would have to be collected, enough for filling each user's time table, which was not possible to do in this research. It is also unknown how the pandemic affected user's behaviours on social media and how much they post. And it is also possible that the content generated by users in Cologne is just not enough right now.

Social media usage tends to keep growing, so in the future it might be easier to collect enough data to build more realistic activity and mobility patterns. In 2020, the number of social media users is estimated in 3.6 billion, with a projected increase to 4.41 billion in 2025 [71]. Different sources could also be used for collecting geolocated points, like phone location, that tracks users throughout the entire day. This type of data has been already successfully used by Replica [66], as mentioned in chapter 2.2 / . More data would also mean more users profiles, that would make a better sample of the city than the limited amount of 83 profiles we have now.

4 / Agent-Based Model*

4.1 / Environment

Before building the actual Agent Based Model, the environment, the agents will populate, must be defined, and built. It is centred on the Deutzer Hafen district and includes the surrounding city in a radius of 1,3 km around it.

In this case, the environment is defined as four different main types, including certain information about each instance of a type:

Element	Attributes
Buildings	Building use (ground and upper level)
	Building footprint area
	Number of levels
	Number of residents
	Part of Deutzer Hafen or not
Streets	Type of street
	Oneway
	Maximum speed
	Number of lanes
Parks and public spaces	Type
Public transport nodes & network	Divided in the stations and its network

Table 9 / Environment elements and their attributes.

For gathering the needed information and geometry, the open source project OpenStreetMap (OSM) was used. This process was done in Grasshopper, using the Gismo plugin [10]. This plugin streams all available geographical information from OSM to Grasshopper, given a certain point in Longitude/Latitude and a radius around it in meter (in this case 1300m).

OSM offers a vast load of information, unfortunately it sometimes lacks organisation, because of its open source approach. In general, you can divide between three types of map elements: points, polylines, and polygons. Along with this vector geometry, each element contains so called tags, always containing a key and its corresponding value. These tags describe specific features of the element. OSM has a very detailed Wiki, where each tag and key is listed[9].

After the geometry and its information is organized in a data tree in Grasshopper, it is exported as SHAPE file for further use in the GAMA platform. Therefore, the plugin BearGIS was used [4]. As spatial projection system, WGS_1984_UTM_Zone_32N was used.

4.1.1 / Buildings & population

A simple example for a key would be building. If you wanted to find all buildings in general, you would need to filter for the key building where the value is defined as yes.

```
building = yes
```

Unfortunately, sometimes buildings are defined already with the use for the key building and not just as a Boolean parameter.

```
building = school  or  building = commercial
```

This means that all elements, which

```
key = building ≠ <empty>
```

can be considered buildings. In theory, it is possible to filter for keys using Gismo, due to these irregularities in the data structure of OSM, it is not very precise though. Therefore, Gismo was only used to download the content, the filtering itself was done with a Phyton script and in Grasshopper. The building use is not always saved with its geometry though. Often, tags like café, bar, shop, or restaurant are rather saved as features of points than as features of polygons. By finding those points inside polygons and matching the features of them with those from the building polygons, this information is combined in Grasshopper. The buildings were categorized as visible in Figure 30.

The categories were later more simplified, with “PARKEN” being integrated into “EINZELH” and KiTa/KiGa integrated into “SCHULE”. This was done to support the process of creating mobility patterns.

For each building, the use for ground level and upper level was defined (see Figure 30).

This worked in four steps:

- ↳ First, land use areas such as commercial and residential were found from OSM. These were overlayed with the buildings and matching buildings were filled with that basic information for both upper and lower level.
- ↳ Second, for each building the use defined in the “building” tag, is found as described before and saved as main use of the building, which is the upper level. The use defined in the step before will be overwritten, if not the same.

↳ Third, the uses that are found as point inside building geometries are used to (i) fill gaps where no information in the first two steps was found and (ii) for all buildings with an already given upper level use, the use for the ground level is defined.

↳ At last, for those buildings that got only one use from the three steps (land use area, building geometry, points), that use was defined for both upper and lower level.

Using the *building=levels* tag, it was possible for almost 80% of the buildings to get the number of levels. For those without any information about the amount of levels, an experimental way of interpolating these number was chosen.

Using the information about area, location, ground level use and upper level use from the buildings with floor level information, a machine learning algorithm in Grasshopper was trained.

This algorithm was implemented in Grasshopper by Benjamin Felbrich at the Institute of Computational Design at the University of Stuttgart [16]. It is well documented and has several examples [8]. For this case, his example of a supervised learning algorithm that helps with fruit classification, was adapted.

The data was trained using a neural network with four so called Sigmoid layers and using backpropagation to minimize errors [16].

This algorithm was used to predict the floor levels of the missing 20%. worked for almost all buildings very well, only special structures like the Lanxess-Arena (Köln-Arena) were not predicted right, because of missing examples in the training data. For the residential buildings, good results were achieved. For a more precise result for all building types, a bigger data set would need to be created, using a bigger radius with Gismo. Since this is not the focus of this project, the results were sufficient.

At last, the buildings are categorized in the Stadtviertel of the City Cologne. This is done with statistics about households for each Stadviertel available at Offene Daten Köln [7] as JSON file including the outline geometry of the Stadviertel. This geometry was overlaid with the buildings from OSM and these buildings were categorized according to that.

The statistics about households from 2017 [7] includes information about amount and size of households, divided in one, two, three, four and more than four person households, given for each Stadtviertel as percentage. This data was used to estimate the number of residents for each building inside the radius. The dataset is combined with statistics about the inhabitants [5], also available from Offene Daten Köln from the year 2017. This data includes information about how many residents each Stadtviertel has, but also the different age groups in percentage per Stadtviertel.

Combining those two datasets, it was possible to estimate the residents per building. First, the total residential area available in each Stadtviertel was calculated.

By dividing the number of residents of each Stadtviertel with its total residential area, the average living space per resident for each Stadtviertel is calculated. With this, the number of residents per building is estimated (see Figure 27).

As visible in mentioned graphics, the number of residents per building is over proportional high in some parts at the edge or the investigated area. Especially in Humboldt and GE Südstadt the effect is visible. This is due to the fact, that the total number of residents per Stadtviertel is used for calculations, but not all the buildings of the district are included. By calculating the area of each Stadtviertel inside the area of interest, the number of residents per Stadtviertel could be calculated proportional to that. Because these citizens are likely to interact with buildings and citizens inside the investigated area, they are kept for the simulation anyway, for now.

Since most of the Stadtviertel inside the investigated radius will grow little until 2040, according to the Kölner statistische Nachrichten [49], there is no growth in the number of residents in the already existing Stadtviertel. The biggest population growth in the investigated area is going to be the newly developed Deutzer Hafen itself, as also mentioned in [49], with its almost 7000 new residents.

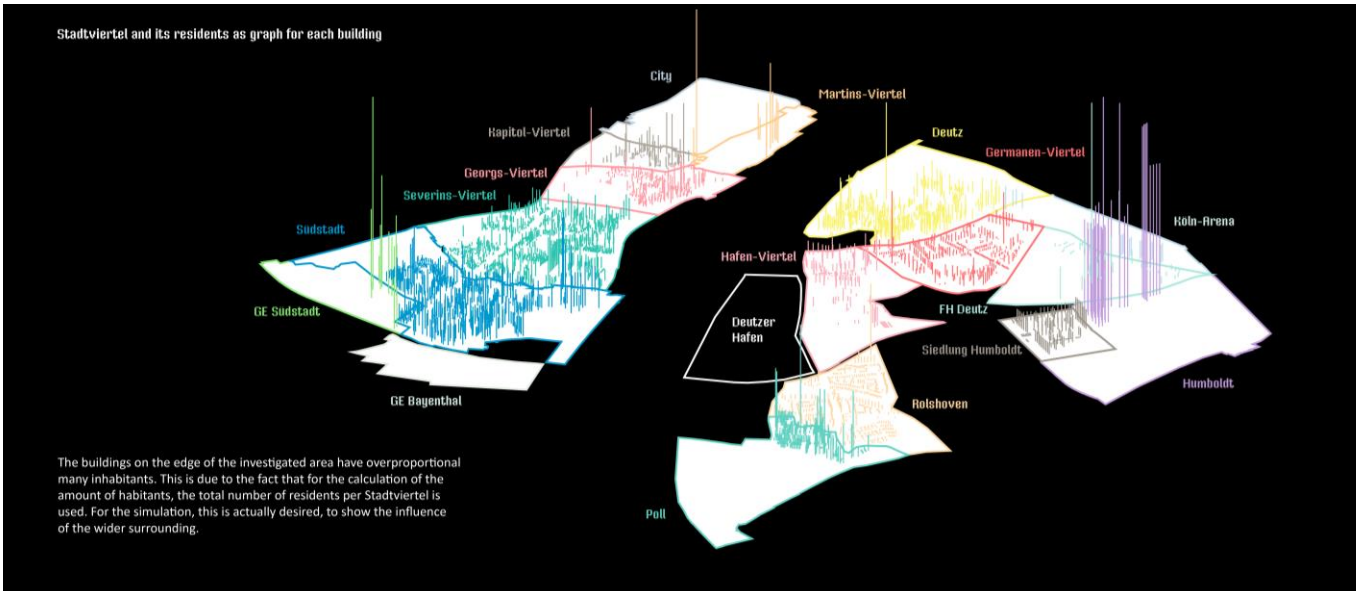
Type	Description
DIENSTL	Service, offices
EINZEHL	Shops
EMPTY	Empty/no information
GASTRO	Restaurants, cafes
HOTEL	Hotels, BnBs, hostels
KiTA/KiGA	Childcare
KULTUR	Museum, church
NACHTL	Clubs, Bars
PARKEN	Parking
SCHULE	Schools
WOHNEN	Residential
HOCHSCHULE	Higher education
INDUSTRIE	Industrial
GESUNDHEIT	Hospitals, doctors
OEFFENTLICHKEIT	Public buildings
FREIZEIT	Leisure

Table 10 /
Types of Buildings.

In the same step, a CSV file is generated, that contains information about the residents of cologne, outside the Deutzer Hafen district. This file is used later to generate the agents. The file includes the building the agent lives in and the age group, the agent belongs to. The age groups are taken from the statistics about inhabitants and used to match the proportions of each age group in each Stadviertel. Each group contains one or more types of agents that could belong to that age group. The type of each agent is then chosen randomly among the types in the list. For example, an agent in the age group of 18 to 30 can be either a Student, Young professional or Home maker.

Table 11 / Agent type per age group.

Agent Type	Age Group				
	Under 18	18 to 30	30 to 65	65 to 80	Over 80
Student	■	■			
Young professional		■	■		
Home maker		■	■	■	■
Mid-career workers			■	■	
Executives			■	■	


Stadtviertel and its residents as graph for each building

The buildings on the edge of the investigated area have overproportional many inhabitants. This is due to the fact that for the calculation of the amount of habitants, the total number of residents per Stadtviertel is used. For the simulation, this is actually desired, to show the influence of the wider surrounding.

Figure 27 / Residents per building

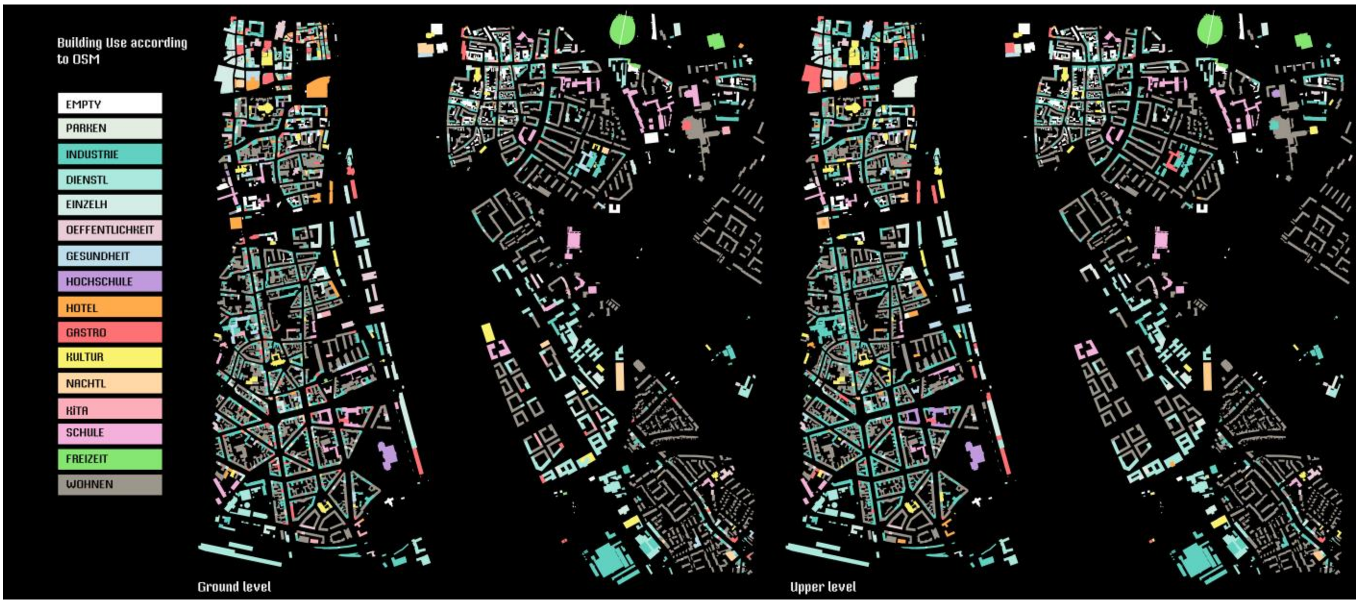


Figure 30 / Building use according to OSM

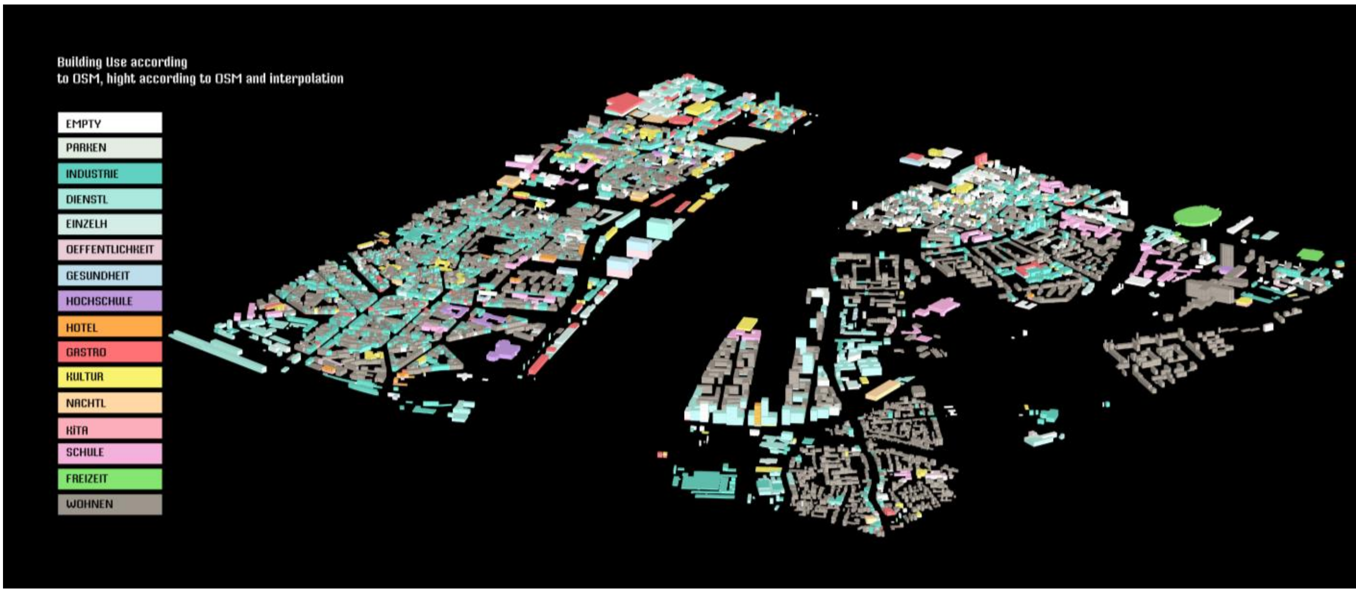


Figure 28 / Building use according to OSM with floor level

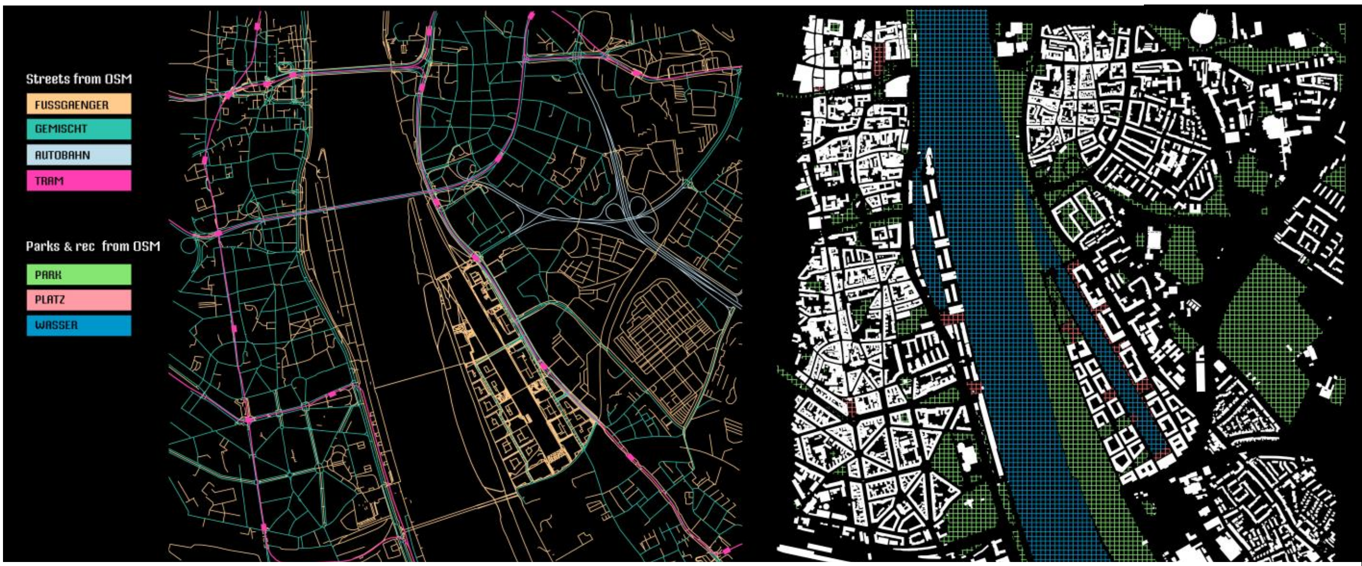


Figure 29 / Streets, public transport, parks & Rec.

4.1.2 / Streets

All streets are represented as polylines, consisting of points. The process to stream this information is like the one used for gathering the buildings information.

For the ABM model, the streets are divided in three categories:

- ↳ Pedestrian street
- ↳ Highway
- ↳ Street with pedestrians and cars (mixed)

By identifying the relevant tags and features the stream from OSM is filtered and the streets are divided in these three categories.

Especially important is the direction of the streets. It is given by the sequence of points that build the polyline.

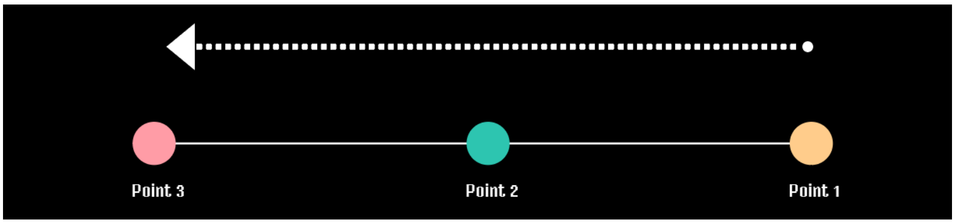


Figure 31 / Point organisation.

This organisation of points is crucial to use the Driving skill later because it gives the agent the information in which direction to move.

For this, also the information if a street section is a one-way street, is needed. In OSM this is featured as an extra tag. For streets without that information available, it was assumed that they are not one-way streets.

4.1.3 / Parks and public spaces

Public parks and spaces for recreation are found using the same logic from OSM.

4.1.4 / Public transportation network & nodes

Using the logic of OSM that was explained before, the public transportation stations were found. Since there are usually several points for one station (for different directions, different lines), the points from OSM were clustered with 100-meter maximum distance. The centre of these clusters was considered to be the actual station.

Apart from the stations, the network of public transportation was found. Unfortunately, only the tram network was available. This is used as base for the public transport.

4.2 / ABM in GAMA

4.2.1 / People Agent

The agents in the model are the future residents of Deutzer Hafen and the residents of the surrounding districts. According to Kölner Statistische Nachrichten [48], the number of residents in the surrounding area will grow little until 2040, so the existing number of 71.000 is kept for the simulation. As for the new district, 7.000 people are expected to live there in the future.

To optimize the simulation, a percentage of these numbers will be used to populate the district and the surrounding area, otherwise it would be impossible to run the model with our resources. This sample of 8% of the population will still give an idea of how the entire population would move and interact with the district. As mentioned in [19] "it may not be necessary to run a full sample of all households and people in the synthetic population in order to analyse every type of alternative scenario [...]".

Another simplification made regarding the population is that we chose to work only with individual people and not households. For such a short research time, we would not be able to establish in the model the complex relationships between members of the same household and how they affect their activity patterns. This simplification can be observed in models such as CityScope [22].

People attributes such as type and activity table are imported for each agent from the CSV files with the profiles created before. Other attributes such as living or working place are defined in the model. Every agent that lives in the district is assigned a random building of type "WOHNEN" as their living place. The agents living in the neighbouring districts are assigned an specific building as their home location, like explained in chapter 4.1 / . For all of the agents, their working place is a building with the same use as their "work type" attribute that is inside the radius of their "work distance" attribute. Vehicle usage can be related to how far one works from home [19], so assigning work places respecting this distance guarantees consistency in the profiles.

Vehicle ownership is assigned to each agent according to their type and aggregated mobility statistics from the German Ministry of Transport [18] for the year 2017. We collected mobility mode usage statistics divided by age groups and adapted them to our profile types. The types of mobility in the statistics go from *on foot* to even *airplane*, but we only use *electric bikes*, *bikes*, *motorbikes* and *cars*. We use an average of the percentages for the first two for *bikes*, and an average of the percentages for the last two for *cars*. Then, we sum the percentages of multiple age groups to get the final percentages for each profile type.

For each type of agent, we then assign a car and a bike to the percentage of that type's population according to the table. Because the assignment is made randomly, some agents will end up having both vehicles and some will have none. Unfortunately, we could not get such statistics from the Ministry of Transport.

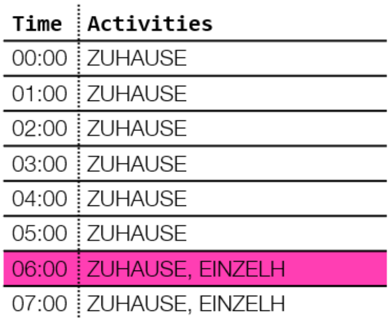
Table 12 / Percentages for ownership of bike or car.

Statistics			Distribution per Profile Type							
Age Group	Bike	Car	Student		Young Professional		Home maker		Mid-career workers/ Executives	
			B	C	B	C	B	C	B	C
0-6	4%	3%	20%	35%						
7-10	2%	2%								
11-13	3%	1%								
14-17	4%	9%								
18-29	8%	15%			59%	73%	87%	86%		
30-39	9%	13%							76%	69%
40-49	15%	18%								
50-59	20%	21%								
60-64	8%	7%								
65-74	16%	8%								
75-79	9%	4%								
>=80	3%	2%								

The five agent types were derived from the seven types observed in this CityScope project [34]. The population is created from the set of 83 profiles derived from social media, that contains a number of profiles from each type. Because the profiles set is smaller than the model's population, the profiles must be duplicated until the demographics of the model are achieved. But because vehicles are assigned randomly and every agent has some optional activities on their tables that they also choose randomly, two agents created from the same profile can still behave differently.

Each agent also follows a set of behaviours that determines when, where and how they move in the environment. In the following table an overview of attributes and the behaviours that create them can be seen.

Every hour the agent checks the next activity in their activity table and, if it is different from the activity in which they are now, they create a new trip objective. If in the next hour slot there are more than one activity, they first choose one of these at random to compare with the current activity. Let us say a person has the following activity table and it is 5:00 am and they are at home.



Time	Activities
00:00	ZUHAUSE
01:00	ZUHAUSE
02:00	ZUHAUSE
03:00	ZUHAUSE
04:00	ZUHAUSE
05:00	ZUHAUSE
06:00	ZUHAUSE, EINZELH
07:00	ZUHAUSE, EINZELH

Figure 32 / Activity table example.

People Agent

Source	Attributes	Behaviours
From synthetic profile	Type	
	Work distance	
	Work type	
	Activity table	
Defined in model	Resident of the district	
	Has car	
	Has bike	
	Living place	Find living place *{chooses at random a residential building inside or outside the district}*
	Working place	Find working place *{chooses a building with the same type as their work type and inside their work distance radius inside or outside the district}*
	Current trip objective	Create trip objective *{check if the next activity in the activity table is different from the current activity and create the next trip objective}*
	Current target	Create target {*if next activity is 'home' or 'work', create target to the living or working place. If it is a different activity, chooses at random a building from that type inside a certain radius around current position. If no building from that type is available inside the radius, choose at random one outside}*
	Possible mobility modes	Choose mobility mode *{if agent has car or bike and is at home, those options are added to the list. If agent already left home with one of these modes, that will be their mode until they return home. Exception is when agent goes to work, they can leave the mode at work and go have lunch walking, for example. But when they leave work, they must use the mode parked there}*
	Current mobility mode	
	Current place	Move *{move towards current target and update place and activity when there}*
	Current activity	

Table 13 / People agent's main attributes and behaviours.

The next hour slot contains two possible activities: ZUHAUSE and EINZELH. One of them is chosen randomly by the agent and then compared with the current activity. If the agent chooses ZUHAUSE, they will not create a new trip objective, because the activity is the same as their current one. If the agent chooses EINZELH, then they create a new trip objective.

The new trip starts in the next hour in a random minute, with a target that depends on the type of activity. If it is *work* or *home*, the target will be the working or living place. If it is a different activity, the agent will search for a place matching that activity around the area where they are now, respecting a defined radius of 400 m (which can be considered a walking distance). They then choose randomly one of the places found or, when no place is found inside the radius, a random place of that type outside the radius is chosen.

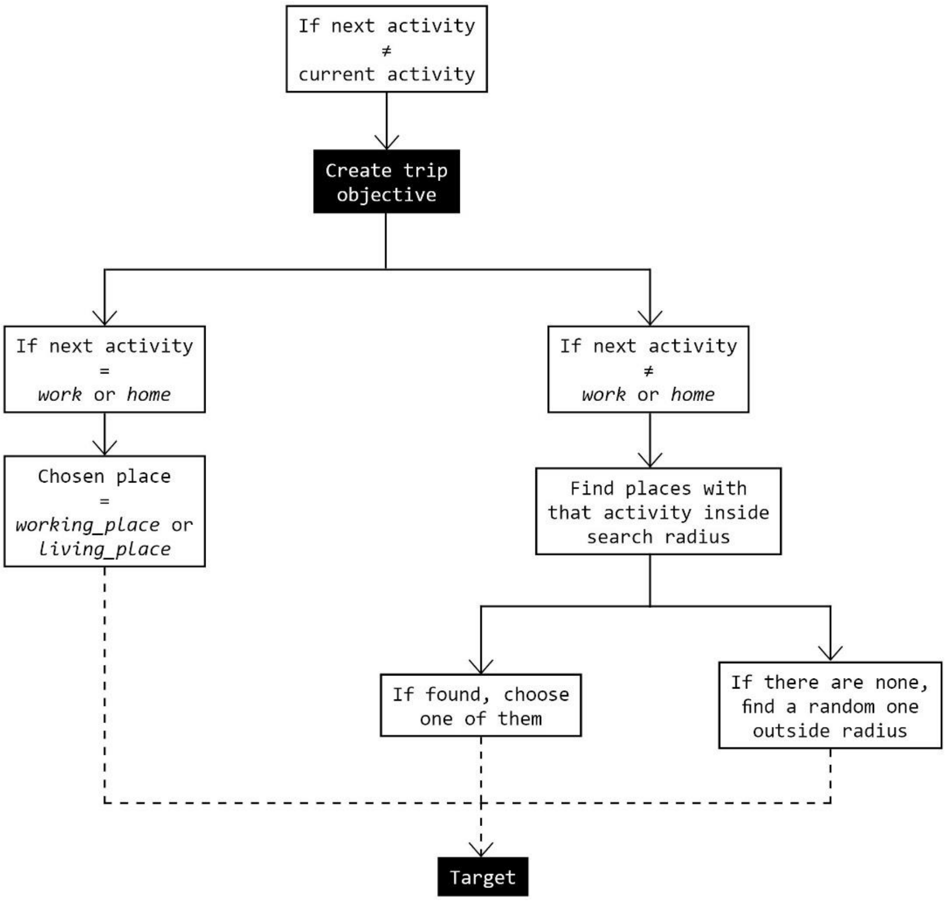

If next activity ≠ current activity
Create trip objective
If next activity = work or home
If next activity ≠ work or home
Chosen place = working_place or living_place
Find places with that activity inside search radius
If found, choose one of them
If there are none, find a random one outside radius
Target

Figure 33 / Create trip objective behaviour diagram.

With the target chosen, the agent must now choose the mobility mode to get there. Each agent will have different options of mobility modes, depending on their profile (if they have a car or not) and on how they moved around so far (if they left the house with the car, they must use the car until returning home).

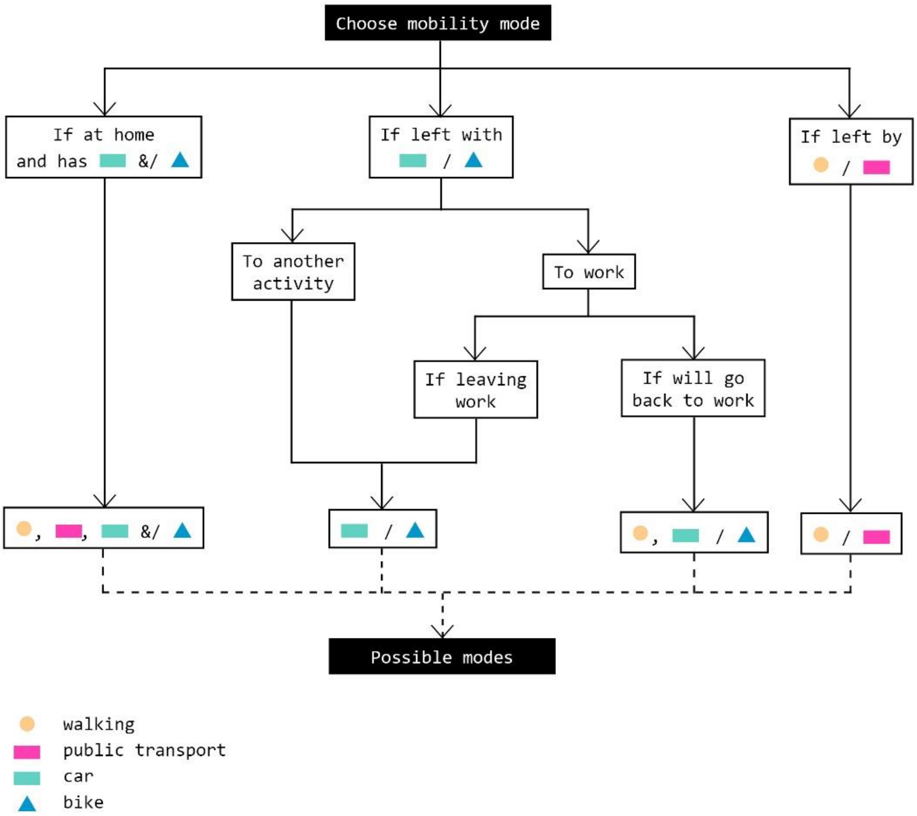


Figure 34 / Defining possible mobility modes for each trip diagram.

Once the agent has the options of possible mobility modes for each trip, one mode is defined according to three factors: travel-time, price, and difficulty. To compute these values, we follow the algorithm used in [34], where each mobility mode has pre-defined price, time and difficulty values. We slightly adapt the values to better fit our model.

Table 14 / Mobility mode types and their factor values.

Mobility Mode Type	Price / km	Waiting Time	Time / km	Difficulty / Trip
Walking	0.001	0.2	15	0
Bike	0.01	1	8	0.2
Car	0.35	3	3	0.3
Public Transport	0.1	6	3	0.5

Every trip, the factors are calculated for each mode accordingly, for example, the travel time for cars will usually be shorter, but it is cheaper and less difficult to walk. After calculating these values, each mobility mode will have a score that is weighted according to the agent's characteristics. An executive, for example, will prioritize time and difficulty over price, while for a student, they may choose the cheapest option, even if it is more difficult. For these calculations, we also use a table of weights from [34] for each type of agent, slightly adapted to our model.

Table 15 / Mobility mode weights per agent type.

Agent Type	Price	Time	Difficulty
Student	-1	-0.6	-0.1
Young professional	-0.7	-0.9	-0.75
Home maker	-0.5	-0.85	-0.9
Mid-career workers	-0.1	-1	-0.7
Executives	0	-1	-1

Given an example trip with all four mobility modes as option and comparing two different agents, a Student and an Executive, we would have the following example calculation. Table 16 shows the calculated values for a 3 km trip for each mode. These values are weighted positively or negatively, depending on the agent. In the end, each mode has a score and the agent will choose that with the highest one. The Executive chooses the car while the Student chooses the bike, as seen in Table 17.

Table 16 / Values for an example trip.

Mode	Walk	Bike	Car	Public Trans.
Price	0.003	0.03	1.05	0.3
Time	45.2	25	12	15
Difficulty	0	0.2	0.3	0.5

Table 17 / Scores per agent type for the example trip.

Agent Type	Walk	Bike	Car	Public Trans.
Student	-0.602	-0.400	-1.219	-0.584
Executive	-1	-0.953	-0.865	-1.331

When it is time to start the next activity, the *move* behaviour of the agent is activated and they will go towards their target. The path they follow will depend on their mobility mode – walking and biking can be done in most streets except for highways, driving a car is possible in less of them – and so will the speed in which they move. If the mobility mode chosen is public transportation, a different move behaviour is activated, that works parallel with the public transport agent. This behaviour is further explained in the next chapter.

4.2.2 / Public Transport Agent

To create our public transport vehicles, we again follow CityScope's approach in [34]. The vehicle agent has three attributes:

- ↳ *Stops:* the list of stops;
- ↳ *Stop passengers:* the list of people that will get out of the bus in each stop;
- ↳ *Target:* the next stop in the schedule.

A number *n* of vehicles is created at the beginning of the simulation, each one in a different stop, so there is an interval in between them. The vehicles move from stop to stop according to their schedule and embark and disembark people.

When a people agent chooses public transport as mode for a trip, they find the next stop and walk until there. They also find the stop closest to their target location and save that information. When the vehicle arrives at the stop where the people agent is waiting, the people agent 'embarks' on the vehicle and is added to the list of people who will disembark at their target stop. The people agent then moves together with the vehicle agent, in the same speed, but only the vehicle agent is visible. As soon as the vehicle arrives at the target stop of that people agent, they 'disembark' and walk the rest of the way until their target building.

This agent is of course a simplification of the public transport available in the district and represents only the tram.

4.2.3 / Building Agent

The buildings are generated using the shape files that were created in Grasshopper (see 4.1.1 /). Here, one of the big advantages of GAMA is used [80]. It is possible to create agents directly from a shape file, using the information (defined in 4.1.1 /) that is passed along with the geometry in that file. For each geometry, an agent is created including corresponding information such as lower level and upper level use.

If this use is for example "WOHNEN", the building agent will be available for people agents to become a resident. These people agents will be registered in a list in the building agent. From this list, and the living area of the building, the current occupation is calculated. This informs other people agents, which look for a living place, if they can move into that building. They should always move into the building with the lower occupation, this way we guarantee that all the buildings will be equally occupied. The building agents are separated into district buildings and buildings of the surrounding city.

4.2.4 / Urban Vitality Cell Agent

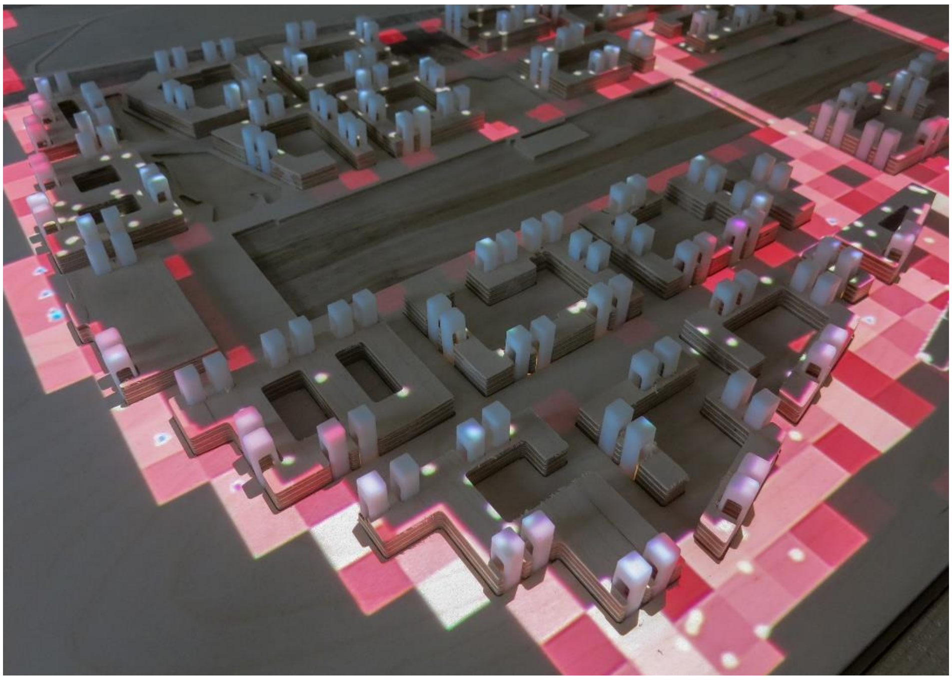


Figure 35 / Vitality grid, visualized on the interface (resolution here 25x25m)

Jane Jacobs [44] defines urban vitality as an essential factor in terms of street life. A successful and safe urban space is only possible through a diversity of people walking through a neighbourhood with different purposes and during different times of the day.

Recent works [56, 75–77] took up on Jacobs idea of urban vitality. Sung, Lee at al. [77] were the first to successfully proof the theory with a 10-year study of pedestrian activity based on surveys in Seoul, South Korea. In Italy, scientists came to similar results [56], using big data instead of surveys. At the UCL in London, Patrizia Sulis et al. [75] worked in a similar direction. These works aim to proof Jacobs theory using computational methods and keeping her idea of diversity in the built environment.

In her PhD, Sulis continued her research in the field of “Measuring urban vitality through human mobility patterns” [74]. In this work, Jacobs theory was adapted and, instead of emphasising on building diversity, the focus is directly on the diversity of people as a measurable value, that strengthens the urban vitality.

As part of this thesis, an urban vitality benchmark is implemented as a high resolution, real time heat map. This benchmark should be based on the diversity of people currently moving, their current objective, and the pedestrian flow related to time. This benchmark should also consider the mode of transportation of people, with cars giving a negative influence on the urban vitality.

Desired is a diverse, continuous pedestrian flow with little peaks or valleys in public places.

Combined with the interactivity of the model, which allows to play with the building uses of the district and the population, the relationship between those two variables and urban vitality should be explorable by the user of the model. The goal is to optimize Urban vitality for certain areas by triggering building use and population division and support the project developer with such decisions.

As mentioned, she defines three main drivers for urban vitality: (i) diversity of people (ii) different purposes and (iii) during different times of the day. This can be interpreted as diversity of agent profile, agent objective diversity and diversity in time.

Diversity itself is a measurable value and can be calculated using the Shannon entropy Index. This index was introduced in 1948 by Shannon as part of his paper “A mathematical theory of communication” [68]. It describes the variety of different species in a dataset, considering the amount of different species(i) and the number of individuals from each species(n). N is defined as the total number of individuals in the dataset.

$$\mathrm{H_S} = \sum_{i}^{S} p_i \, * \ln p_i \;\; where\; p_i = \frac{n_i}{N}$$

Equation 1 / Shannon entropy index [68]

Commonly used in ecology to describe the diversity of species, it is also applied to describe diversity in urban context as done by Cerrone, Lehtovuori et al. [20] or the diversity of people in cities [47].

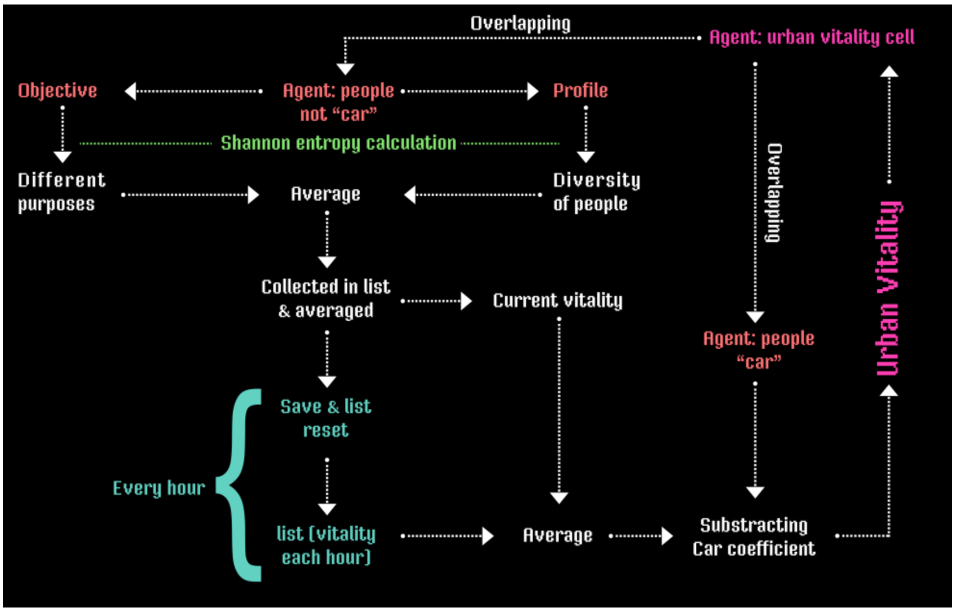


Figure 36 / Vitality Cell with time for non-interactive simulation

This grid of cells with a resolution of 50 x 50 m evaluates the urban vitality of this space in real time and gives a visual feedback. Therefore, every round of the simulation, the people agents that overlap the cell, give two information values to the cell: (i) profile and (ii) current objective. This is used to calculate the diversity of each of those three values. The first two values (profile and objective) are saved by the cell only for 60 minutes of global time.

There are five different profiles in total, as explained in the previous chapters.

Knowing this, it is possible to calculate the maximum diversity of profiles ($\mathrm{H_{profiles\_max}}$), as follows in the next equation.

$$H_{profiles\_max} = -\left(\left(\frac{1}{5} * \ln\left(\frac{1}{5}\right)\right) + \left(\frac{1}{5} * \ln\left(\frac{1}{5}\right)\right) + \left(\frac{1}{5} * \ln\left(\frac{1}{5}\right)\right) + \left(\frac{1}{5} * \ln\left(\frac{1}{5}\right)\right) + \left(\frac{1}{5} * \ln\left(\frac{1}{5}\right)\right)\right) = 1.6094379$$

Equation 2 / H $_{profiles\_max}$

H $_{profiles\_max}$ would be a group of agent people where each profile appears the same amount and each profile is present. This scenario would be the highest diversity possible.

For the objectives H $_{objectives\_max}$ can be calculated in the same way (Equation 3).

$$\boldsymbol{H_{objectives\_max} = 2.63905732}$$

Equation 3 / H $_{objectives\_max}$

Both H $_{profiles\_max}$ and H $_{objectives\_max}$ are defined in the initiation of the simulation. Each cell now calculates H $_{profiles\_current}$ and H $_{objectives\_current}$ based on the information given by the overlapping agent people. This takes all agents that passed by this cell during the last 60 minutes into consideration.

$$diversity_i = \frac{H_{i\_current}}{H_{i\_max}}$$

Equation 4 / Diversity

It is now possible to calculate the diversity, considering the maximum H and the current H values as in

Equation 4. This value is calculated in each round. From that, the current attractivity in the current hour can be calculated. Since the time equivalent of one round is flexible, one hour can last a varying number of simulation rounds R. R is counted until the next hour starts.

$$vitality_{current} = \frac{\sum_{i}^{R} \frac{diversity_{profiles_i} + diversity_{objectives_i}}{2}}{R}$$

Equation 5 / Current vitality

For Jacobs, the third factor besides different purposes and diversity of people is the continuity of people during time. This is considered by saving the vitality $_{current}$ value after an hour passed as vitality $_{per\ hour}$. This way, a list of vitality values for each hour is generated. The total vitality is then calculated as the sum of vitality $_{per\ hour}$ and current vitality, divided through the hours that passed by (T).

$$vitality_{total} = \frac{(\sum vitality_{per\ hour}) + vitality_{current}}{T}$$

Equation 6 / Total vitality

This way, only a space that constantly has a high vitality will be considered vital. Another value that is considered by the urban vitality benchmark is the traffic on the grid. The cell is counting the number of cars passing every round. This number is multiplied by the factor 0.025 and then subtracted from the vitality average. This gives the final urban vitality of the cell.

The result is a value between 0 and 1 that is translated in colours which indicate the quality of the public space as a heat map on the model. Black stands for no vitality, values in the red area indicate a bad vitality with little diversity and blue is representing high urban vitality.

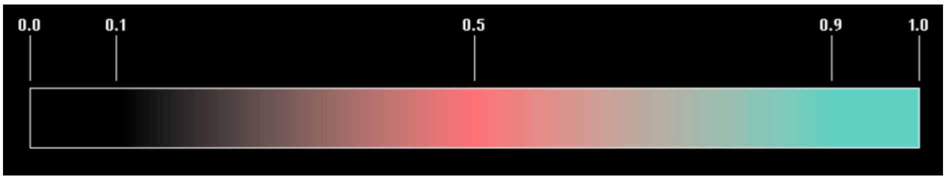


Figure 37 / Urban vitality heatmap colours

While testing the system, it turned out that, because the vitality is saved for each hour, the reactivity of the cells was not given. Therefore, for an interactive simulation where parameters might change with time, the time factor was taken out of the calculation. During the simulation, a space that constantly keeps up in the range of blue colours, is considered vital. For a non-interactive simulation, the calculation is described in Figure 36.

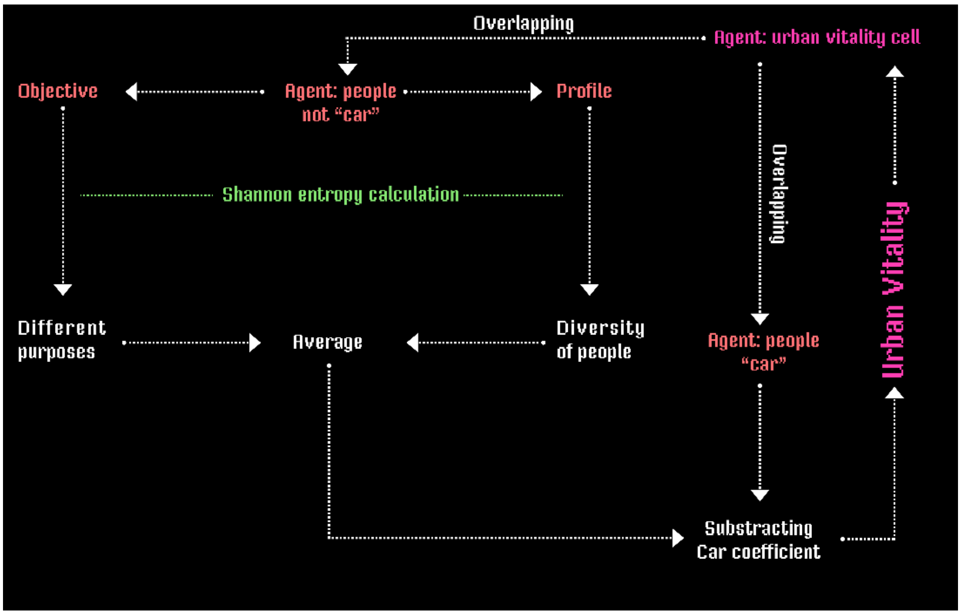


Figure 38 / Activity cell for an interactive simulation

4.2.5 / Interactive Parameters

Three main parameters allow the user to explore different scenarios in the model:

↳ *Density distribution:* the percentage of each people type can be adjusted, resulting in changes in the way they move and also in the total amount of people. Each type has a different *square meter_per_person* value, like for example College students with 30 m^2/person and Executives with 50 m^2/person. Meaning if we increase the percentage of Students, the total number of people in the district would also increase, since they each need less area. On the other hand, by increasing the percentage of Executives, who need more area, the total amount of people in the district would decrease. Executives have also a higher probability of owning a car than Students, who are more likely to ride a bike or use public transportation. Changing their percentages will also affect which modes of transportation are most used inside the district.

Agent Type	m^2/ person
Student	30
Young professional	30
Home maker	40
Mid-career workers	50
Executives	40

Table 18 / Square meters per person.

↳ *Building use:* the model starts with each building having the ground and upper level use as defined in the project. During the simulation these values can be changed for every building in the district, resulting in changes in the amount of people living in the district (if a residential building has its use changed, for example) and also in how many people visit the district and what they do there (if the amount of restaurants is increased, the district might have a higher occupation during lunch hour, for example).

↳ *Day of the week:* the agents have two different activity tables, one for weekdays and one for weekends. The user can choose to visualize each one of them in the simulation, which will create changes in the types of activities attended by the agents, reducing for example work activities and increasing leisure activities. Weekends will be particularly interesting to visualize how green areas and other public spaces are used.

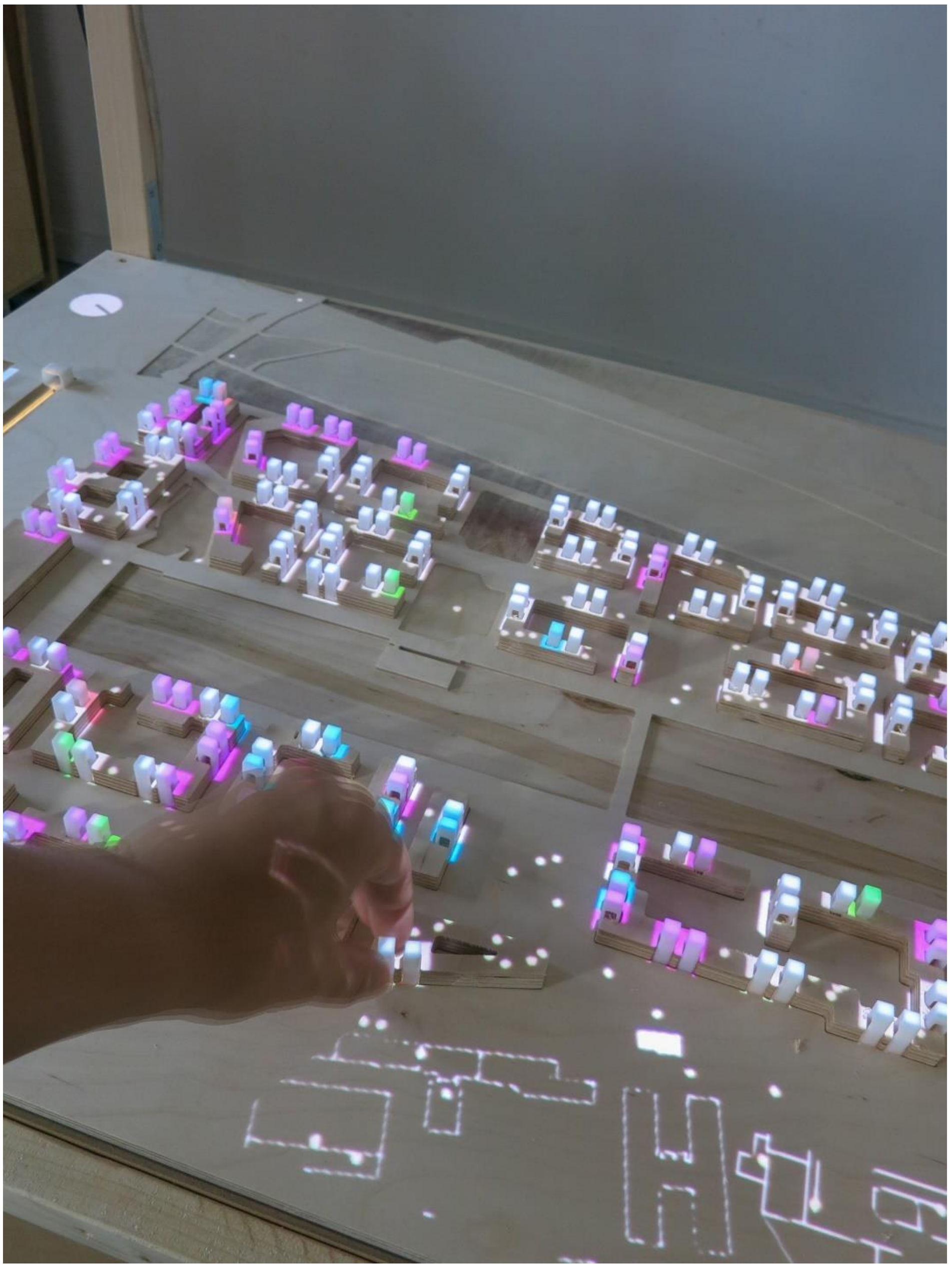


Figure 39 / Change of building uses by moving tags.

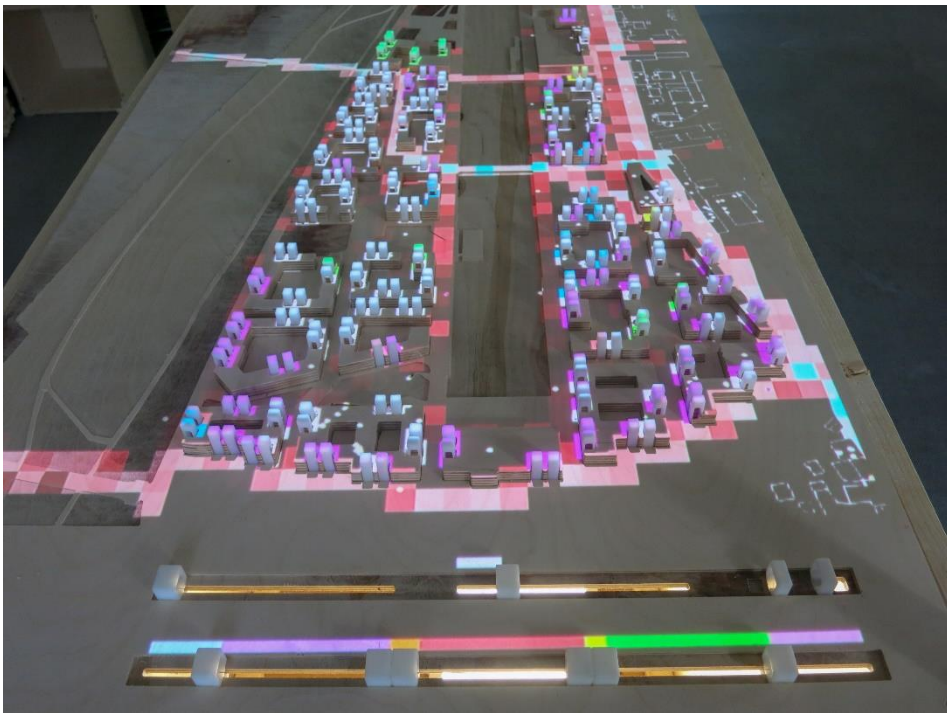


Figure 40 / Sliders for adjusting population demographics, speed of the simulation and visualizing saved scenarios. .

4.2.6 / Visualization & Data Representation

GAMA allows the user to define different displays that will show the simulation in parallel, showing different aspects of the model. In this case, two displays were defined: (i) a display that will be directly projected onto the tangible user interface and (ii) a second display that is projected by the second projector onto a wall.

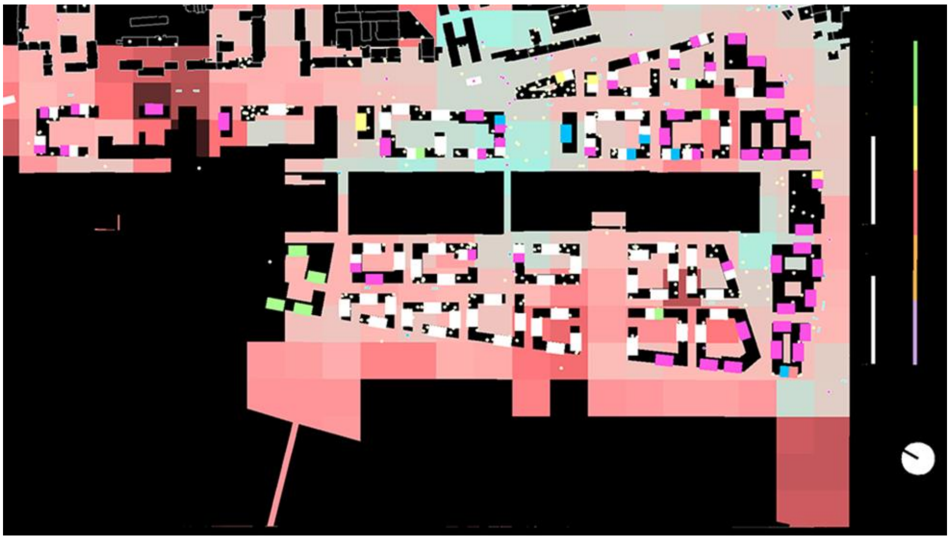


Figure 41 / Table projection.

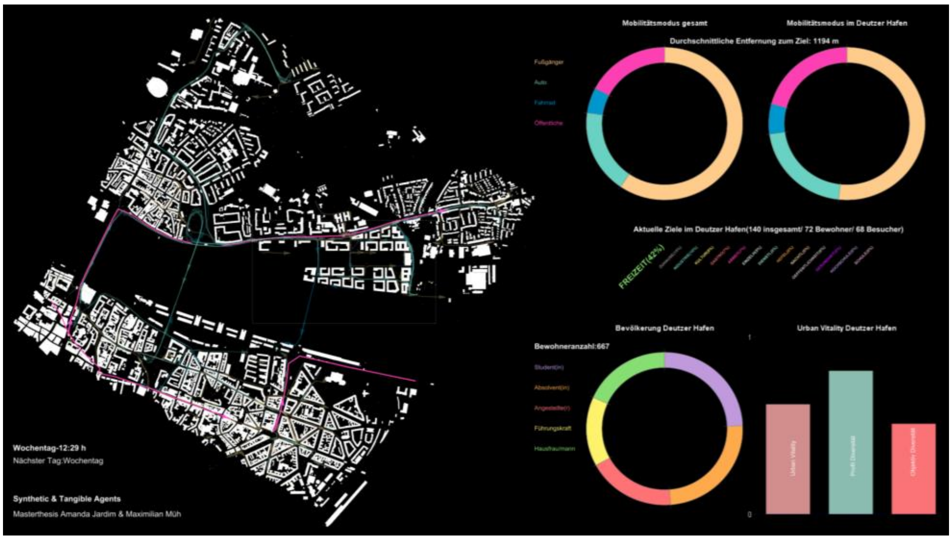

Mobilitätsmodus gesamt
Mobilitätsmodus im Deutzer Hafen
Durchschnittliche Entfernung zum Ziel: 1194 m
Fußgänger
Auto
Fahrrad
Öffentliche
Aktuelle Ziele im Deutzer Hafen(140 insgesamt/ 72 Bewohner/ 68 Besucher)
FREIZEIT(42%)
Bevölkerung Deutzer Hafen
Bewohneranzahl:667
Student(in)
Absolvent(in)
Angestellte(r)
Führungskraft
Hausfrau/mann
Urban Vitality Deutzer Hafen
1
0
Urban Vitality
Profil Diversität
Objekte Diversität
Wochentag-12:29 h
Nächster Tag:Wochentag
Synthetic & Tangible Agents
Masterthesis Amanda Jardim & Maximilian Müh

Figure 42 / Second Screen Visualization.

(i) The table projection gives information through five different real time graphics:

- 1: On the district buildings, at the location of the tags, the current building use is visualised with six different colours.
- 2: As base layer, a heat map visualizes the urban vitality (see 4.2.4 /)
- 3: On the right side, the bar chart represents the position of the sliders for population division. This is the biggest bar chart.

- ↳ 4: Above graphic 3, the lower bar chart represents the current time equivalent of a simulation step in minutes.
- ↳ 5: The third bar chart lets the user choose predefined configurations that he can activate to compare to his/her new layout. Not implemented yet, but interesting would be the possibility to save a certain layout also for comparing it later.

(ii) In the lower left corner, the current time, type of day and the following day is visualized. The second screen also gives information through six different real time graphics:

- ↳ 1: On the left side, the whole area of cologne, that was investigated, is visualized as a map. On this map, the current mobility flow through the city is shown as a trajectory map where each mode of mobility has its own colour. Combined with the two corresponding ring charts (2&3) this map helps understanding the flow of people through the city and which mode of transportation they choose.
- ↳ 2&3: These ring charts in the top right of the screen represent the cumulative mobility mode choices, one for the people in the whole investigated area and one for the district of Deutzer Hafen only. Also the average distance per trip is shown here.
- ↳ 4: This chart in the middle on the right is inspired by the idea of classic word clouds which are often used to visualize the number of appearances of a word in a dataset of words using the size and position of the word as indicator. In this case, the graphics shows the current objectives that all people which are currently in the district are heading to. The objectives are sorted from left to right according to current popularity and grow in size depending on that, with the biggest and most left objective as the most popular one at the moment. It gives also information about how many citizens are active at the moment in the district and how many of them are residents of the district or visitors.
- ↳ 5: The pie chart in the lower middle of the screen visualizes the current population divided in five types. This pie chart corresponds to the population division slider of the tangible user interface. The slider corresponds to the area occupied per type of people though, this pie chart refers to the actual number of residents. The total number of residents is also given here.
- ↳ 6: The bar chart in the lower right corner gives information about the urban vitality benchmark in the district. Unlike the heat map projected onto the model, this chart gets values of all activated vitality cells in the district and calculates the average. This is done with the vitality but also with the two values the vitality is calculated with: profile diversity and objective diversity. This makes the correlation between those three values clearer.

These statistics can be reset during the simulation, using the graphical user interface from GAMA. Later, this could be implemented as button in the TUI also.

4.3 / Test Runs

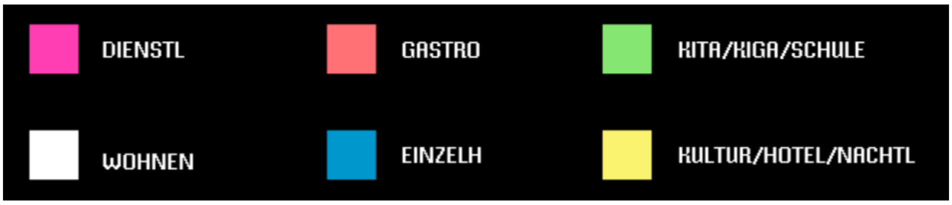


Figure 43 / Colours corresponding to building use.

Three different scenarios were used to evaluate the interactivity of the interface and the ABM:

- ↳ *Activating the waterfront*. The goal of this test was to strengthen the urban vitality in the areas around the water. An even distributed population was assumed.
- ↳ *Activating the westside parks close to the open-air stage*. Also here, the goal is to strengthen urban vitality in a certain area. An even distributed population was assumed.
- ↳ *Creating a district predominately for young people.*

As reference, the standard configuration based on COBEs design is used. The screen shots were taken always around 12:30 in simulation time. In general, tests showed that there are three kinds of zones in the district: (i) areas that are always activated, totally independent of the building program and the population. These areas are bundled especially around the central pedestrian bridge with its connection to the public transport station, the end of the pool, and the pedestrian bridge that connects the district with the city centre. (ii) Areas that are frequented constantly, typically close to the areas of the first category but with lack of diversity. These areas are easily activated during the interaction with the model. (iii) Remote areas of the district that are not frequented much and are hard to make attractive and vital.

These tests were made to show the reactivity of the model and possible uses. Scenarios like those could be played through with the model and the impact can be visualized. It also gives an idea of how results can be interpreted by the user.

4.3.1 / Reference configuration

For comparison, the layout as currently planned by COBE, is analysed. The population here is 667 (8% of the actual population). The layout is roughly divided in three parts: (i) a mixed-use area on the east side, (ii) a predominantly work related area in the south and (iii) a mainly for living used area in the west that also contains the school. The mixed-use area with its proximity to the public transport is the most vital part of the district in this scenario. Other parts of the district such as the public place with the open-air stage in the west or the head of the water pool in the south are frequented, but do not have a consistency in urban vitality. Looking at the overall statistics, it becomes clear that the district lacks of some activities in walkable distance, visible in the diversity of objectives and the fact that, compared to the whole investigated area, people tend to choose more often to not walk. The active people in the district were almost evenly distributed in visitors and residents.

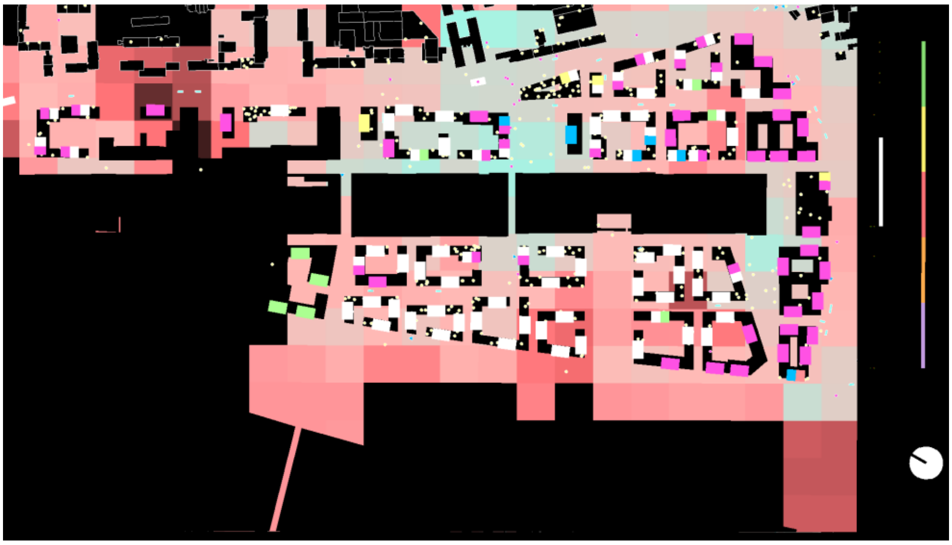
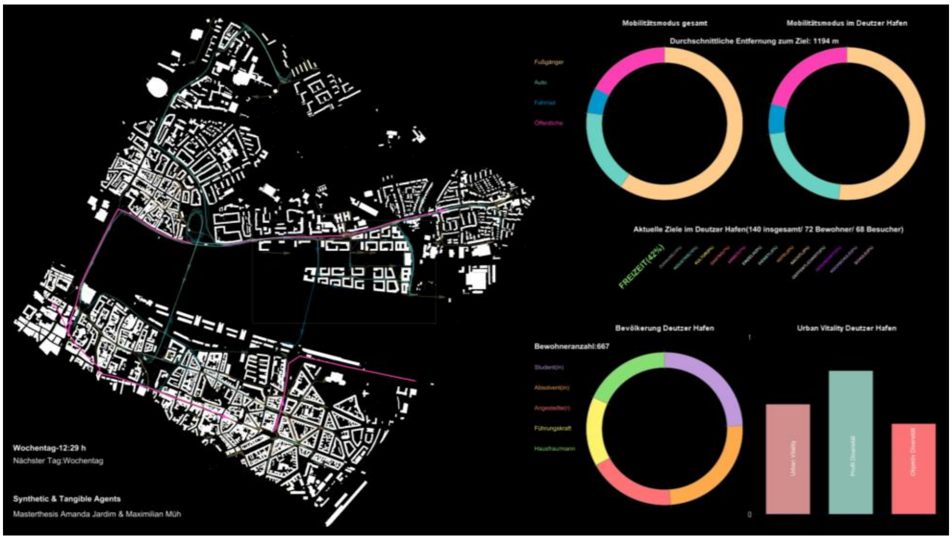


Figure 44 / reference config screen 1



Mobilitätsmodus gesamt
Mobilitätsmodus im Deutzer Hafen
Durchschnittliche Entfernung zum Ziel: 1194 m
Fußgänger
Auto
Fahrrad
Öffentliche
Aktuelle Ziele im Deutzer Hafen(140 insgesamt/ 72 Bewohner/ 68 Besucher)
FREIZEIT(42%)
Bevölkerung Deutzer Hafen
Bewohneranzahl:667
Student(in)
Absolvent(in)
Angestellte(r)
Führungskraft
Hausfrau/mann
Urban Vitality Deutzer Hafen
1
0
Urban Vitality
Profil Diversität
Objektiv Diversität
Wochentag-12:29 h
Nächster Tag:Wochentag
Synthetic & Tangible Agents
Masterthesis Amanda Jardim & Maximilian Müh

Figure 45 / reference config screen 2

4.3.2 / Activating the waterfront

This scenario was built with the idea to strengthen vitality around the water pool in the centre of the district. This is achieved by distributing daily life activities close to the pool and using the buildings further away from the water as predominately work and living related areas. As visible, the areas around the water get activated, while the remote areas calm down. The layout was developed with the original population in mind (667) and roughly matched (671). The number of visitors and residents stayed almost the same.

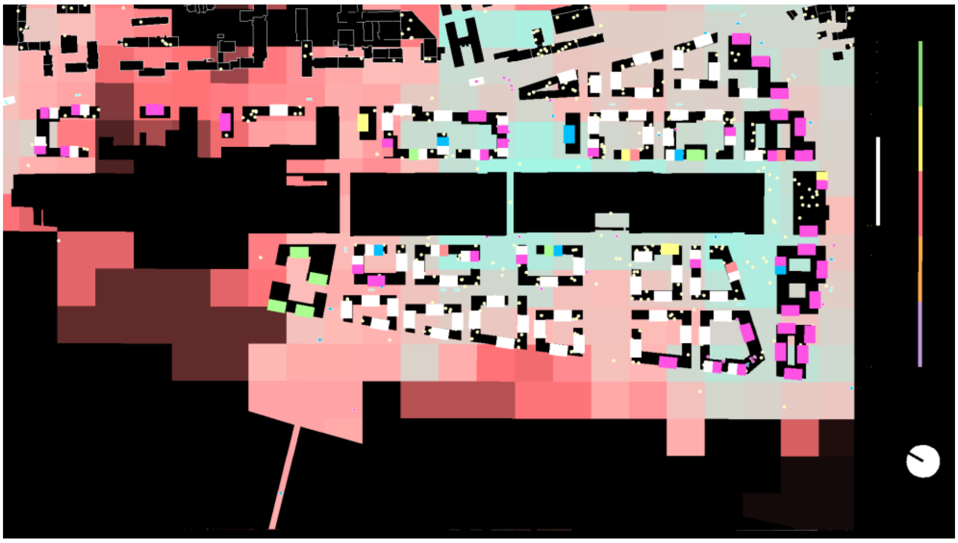


Figure 46 / Activating waterfront screen 1

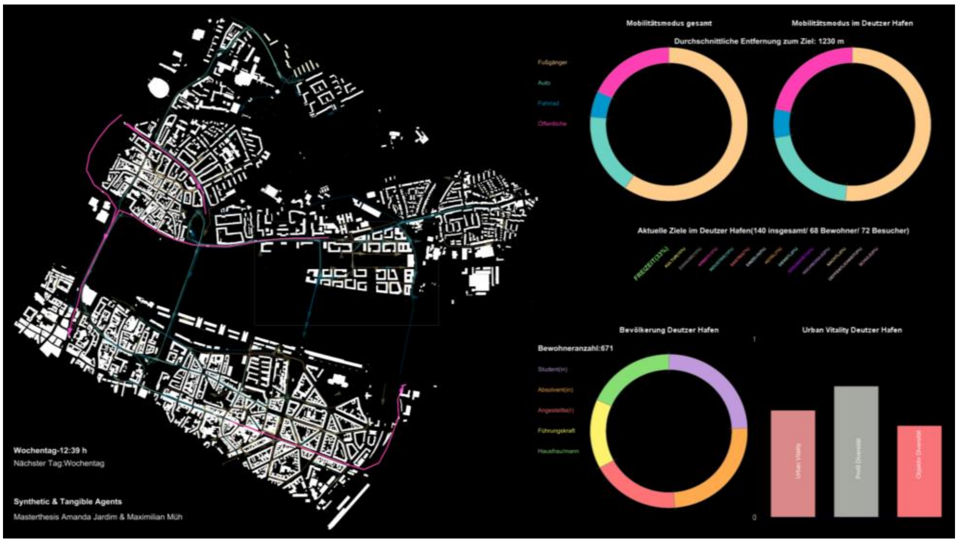


Figure 47 / Activating waterfront screen 2

Mobilitätsmodus gesamt
Mobilitätsmodus im Deutzer Hafen
Durchschnittliche Entfernung zum Ziel: 1230 m
Fußgänger
Auto
Fahrrad
Öffentliche
Aktuelle Ziele im Deutzer Hafen(140 insgesamt/ 68 Bewohner/ 72 Besucher)
FREIZEIT(33%)
Bevölkerung Deutzer Hafen
Bewohneranzahl:671
Student(in)
Absolvent(in)
Angestellte(r)
Führungskraft
Hausfrau/mann
Urban Vitality Deutzer Hafen
1
0
Urban Vitality
Profil Diversität
Objekte Diversität
Wochentag-12:39 h
Nächster Tag:Wochentag
Synthetic & Tangible Agents
Masterthesis Amanda Jardim & Maximilian Mühl

4.3.3 / Activating the westside parks

In this case, the goal was to concentrate on the west side park and the open-air stage. As visible, it was possible to strengthen that area by concentrating activities around it. The population was again kept the same. Compared to both other scenarios, the number of active people went down by around 15%. This could be because of the loss of activities on the east side, which attracted people from outside but also the other way round, residents from the district might look for activities outside.

The mobility choices were not really influenced by any decisions of building use, except taking activities completely from the district, which leads to a larger amount of non-walking traffic by the residents.

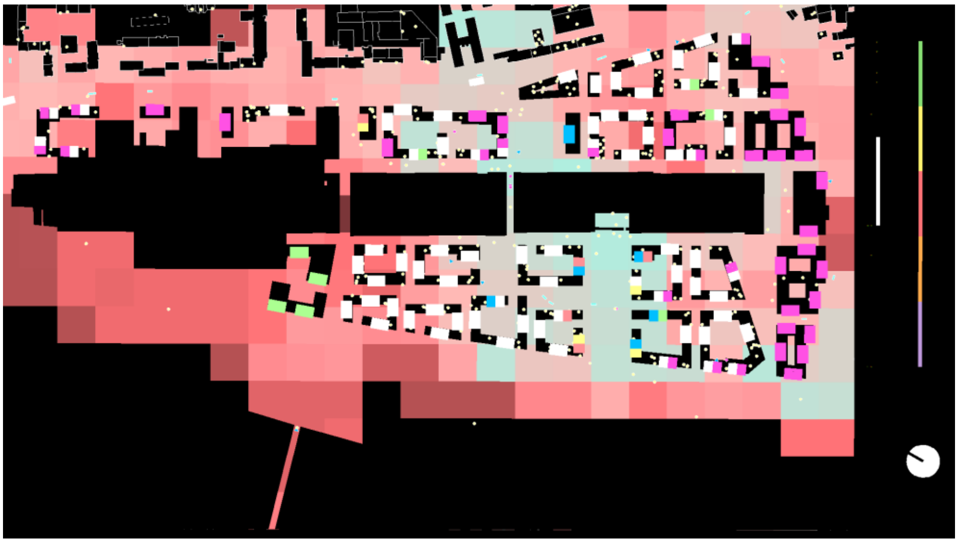


Figure 48 / Activating westside parks screen 1

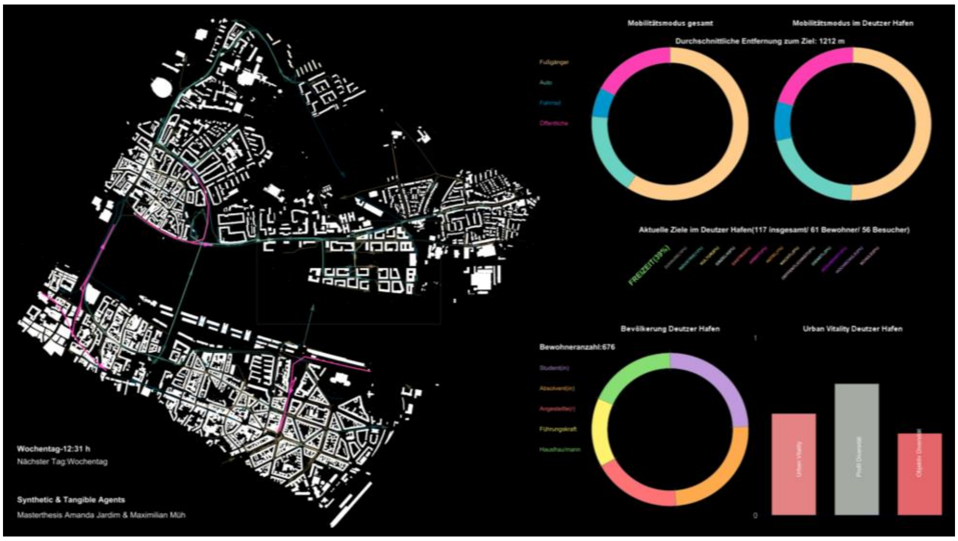

Mobilitätsmodus gesamt
Mobilitätsmodus im Deutzer Hafen
Durchschnittliche Entfernung zum Ziel: 1212 m
Fußgänger
Auto
Fahrrad
Öffentliche
Aktuelle Ziele im Deutzer Hafen(117 insgesamt/ 61 Bewohner/ 56 Besucher)
FREIZEIT(39%)
Bevölkerung Deutzer Hafen
Bewohneranzahl:676
Student(in)
Absolvent(in)
Angestellte(r)
Führungskraft
Haushau/mann
Urban Vitality Deutzer Hafen
1
0
Urban Vitality
Wochentag-12:31 h
Nächster Tag:Wochentag
Synthetic & Tangible Agents
Masterthesis Amanda Jardim & Maximilian Müh

Figure 49 / Activating westside parks screen 2

4.3.4 / Focus on young people

In the last scenario, the impact of a different, not that heterogenous population was investigated. In this case, the population mainly consisted of young people such as students and young professionals where homemakers, mid-career workers and executives were a minority. The most obvious effect is the lack of diversity, which is visible in the vitality heat map. The population of the district increased from 667 to 747 due to the fact, that students are occupying less space and live denser. The population change also has effects on the mobility choices. Compared to the standard configuration, the residents tend to choose public transport more often instead of cars.

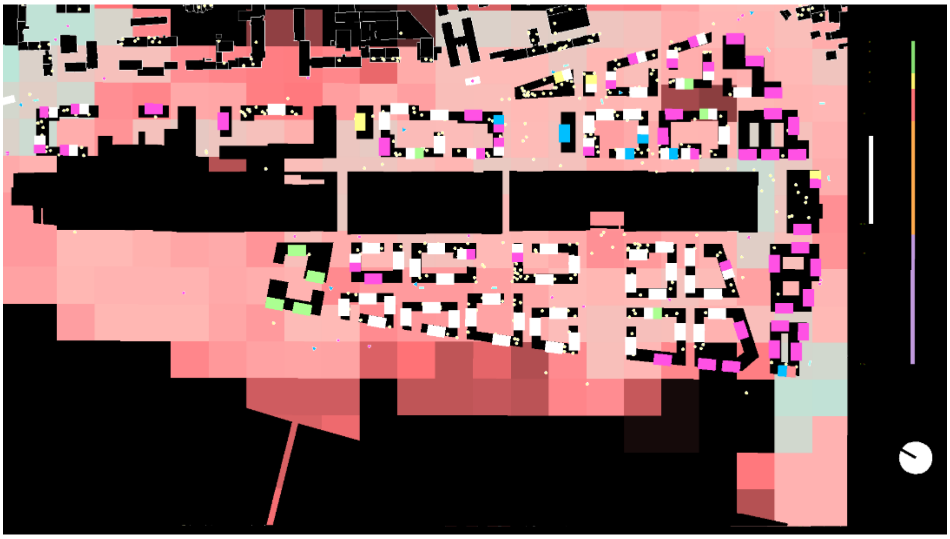


Figure 50 / Focus on young people screen 1

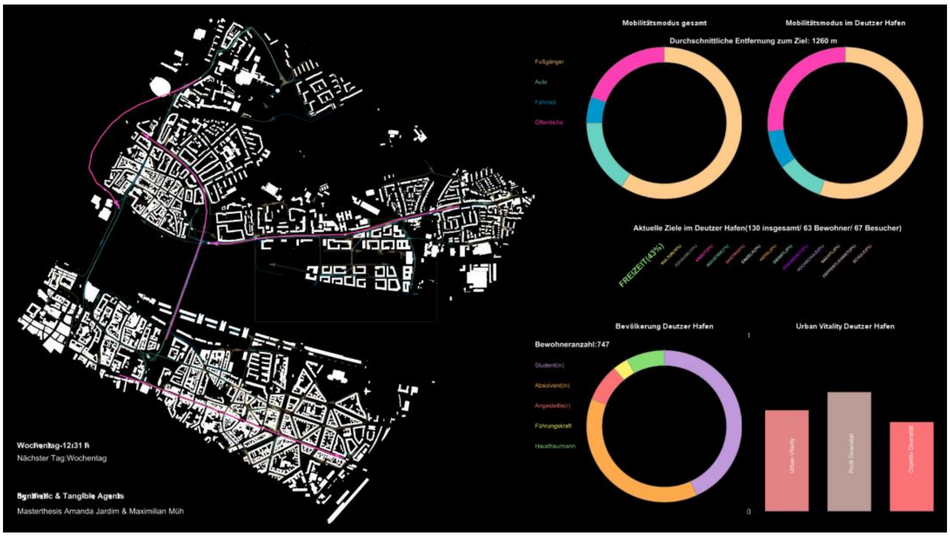

Mobilitätsmodus gesamt
Mobilitätsmodus im Deutzer Hafen
Durchschnittliche Entfernung zum Ziel: 1260 m
Fußgänger
Auto
Fahrrad
Öffentliche
Aktuelle Ziele im Deutzer Hafen(130 insgesamt/ 63 Bewohner/ 67 Besucher)
FREIZEIT(43%)
Bevölkerung Deutzer Hafen
Bewohneranzahl:747
Student(in)
Absolvent(in)
Angestellte(r)
Führungskraft
Hausfrau/mann
Urban Vitality Deutzer Hafen
1
0
Urban Vitality
Profil Diversität
Objekte Diversität
Wochentag-12:31 h
Nächster Tag:Wochentag
Synthetic & Tangible Agents
Masterthesis Amanda Jardim & Maximilian Müh

Figure 51 / Focus on young people screen 2

4.4 / Model Limitations

4.4.1 / Computational power

A major limitation was the lack of computational power to simulate the investigated area with all its 70.000 citizens. The population was reduced to 8%, to make it possible to have a fluent simulation. Higher values are possible but come with high waiting time for each round of simulation. For an interactive experience, this was not acceptable. For the presentation of this project, a high-performance computer from the university is going to be used, to improve this. As pointed out before, it might not be necessary to run the simulation with all citizens, in order to get representative results, according to [19].

This limitation might also come from the lack of experience with GAMA of both authors and the short time, in which the model was developed. The script is a "work in progress" with several parts that an expert in GAML could probably solve in more efficient ways.

4.4.2 / Effective triggers

The biggest challenge turned out to be finding the right triggers to influence the model and to make the interaction with the model possible. Since the development of the Deutzer Hafen is already very advanced, the possibilities to influence it were mainly two variables: (i) the building use and (ii) the population. Those two were chosen to be available for user interaction. While testing, it became clear though, that it was hard to really influence the urban vitality using just these variables, because they are part of a complex system where almost all variables relate to each other.

4.4.3 / Data

Another limitation is the general lack of data. For example, while it was possible to get the tram network, there was no source available for bus and S-Bahn. The only possibility would have been to manually add it. This gets visible when looking at the mobility choices.

Moreover, time-use data from Colone was not available, which is why time-use data from the Netherlands had to be used for the synthetic population.

4.4.4 / Closed ecosystem

The model as it is now, works as a closed system, meaning the influence of the surrounding of the investigated area is not considered in any decision. This is especially visible when looking at the mode of transportation, where the majority of the agents chooses "walking". This is mostly because of the lack of long-distance trips since the radius of the investigated area is 1.3 km only.

4.4.5 / Limited testing possibilities

The original idea for this interactive simulation was to learn from reactions and feedback from users, while developing it. Positioned in FabLab at the university, it would have attracted many students. Due to the current pandemic situation in the year 2020, this was not possible and not even the authors were able to test the interface in the way they would have liked to.

4.5 / Future Development

The interactive simulation was developed during this thesis to a point that shows its potential. There are several points that could be further improved though to sharpen the simulation and create a better interactivity. In the following, two main points are discussed.

4.5.1 / Mobility

At the moment, the mobility modes are chosen with a weighted means decision making model (see 4.2.1 /). The results are not really matching the current statistics from Colone though [3]. This might partly be because of lack of data and the closed system design. The way agents decide on the mobility mode needs to be reinvestigated in the future. This could be done by further adjusting the weights and values for choice making.

Shared mobility becomes also more and more import in today's cities. This should be taken into consideration in the future, possibilities to rent vehicles such as bikes or cars are spreading in the city of cologne and are also integrated in COBE's design for the Deutzer Hafen (as "Mobilitätspunkt"). In the future, this network of shared mobility points should be implemented and enable agents without car or bike to rent a vehicle there.

Furthermore, the way agents are moving should be improved. The driving skill enables the agents to use a network of lines as roads. It includes all important features for a proper traffic simulation. It was developed by Patrick Taillandier as a plugin for GAMA [61]. Agents are able to follow the concept of "right-side driving", keep distance to other agents and react to crossings and traffic lights. The necessary information was already gathered via OSM and is implemented in the GIS files used in GAMA. This feature was tested in separated simulations and could be implemented later to improve the quality of the traffic simulation. The driving skill could be used to simulate bikes, cars, and pedestrians as well.

4.5.2 / Households

The model as it is, considers each person as a household (own apartment, own vehicle). In the future, it would be good to further detail the households and divide the population into those. Households have information about members, type and number of vehicles, type of household (e.g. family, shared flat) and possibly even financial situation. This could help improving the mobility choices and can introduce shared trips (e.g. the whole family travels together).

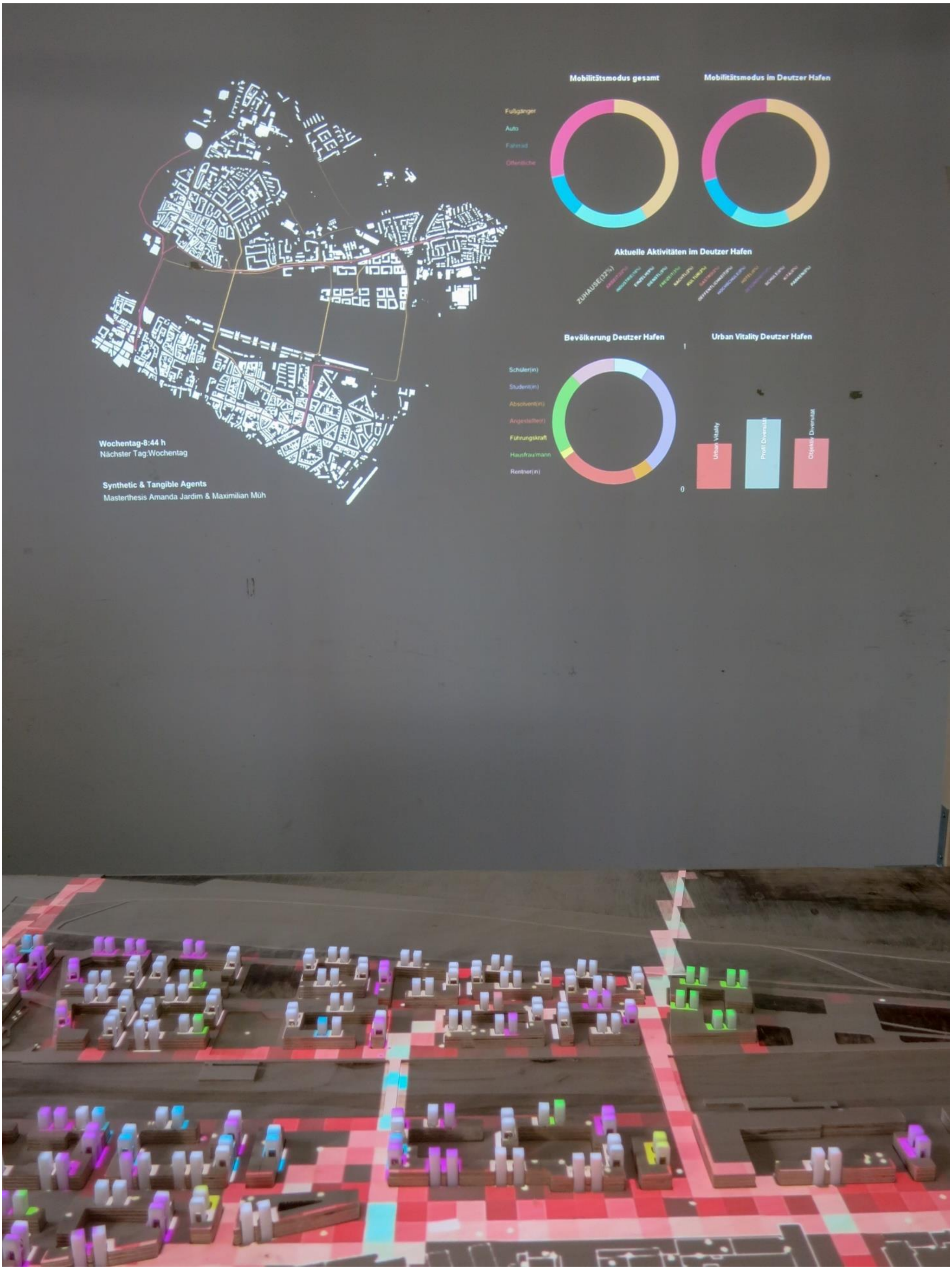


Figure 52 / Final TUI set-up.

5 / Conclusion*

Although it comes with simplifications and developments left for the future, our system showed great potential in supporting urban design decisions.

The simulation of different scenarios in the district, by changing parameters of population demographics and building use, resulted in a big impact on the calculated urban vitality of the public spaces. This response of the model not only highlights some obvious assumptions – such as areas closer to connecting bridges being more frequented – but also shows surprising results in different settings – like how easy it is to lose visitors when removing just part of the commercial buildings.

Many uncertainties still remain due to simplifications in the agent-based model and limited data and technical resources. Despite that, how the agents built from the social media profiles behave in the simulations, from where and when they go to which mobility mode they choose, was not very distant of how real people are expected to behave. In the future, it would be interesting to see how adding household information to the population or more transportation modes could affect the results. Modelling a bigger area of the city and running the model with a bigger sample of the population could also bring improvement.

Despite still not have been tested in a bigger group due to the current social distancing rules, the tangible user interface performed well between a small group of people, being easily understandable, intuitive and giving almost immediate feedback to the user when interacted with. Built with a much more affordable set-up than similar tools, it still has the potential of being reused for different projects, just by replacing the table top and keeping the rest of the hardware.

Also approachable was the concept of the agents in the simulation being based on real people from social media, as noted by a colleague who interacted with the table. Such feedback hints at the potential of making citizens feeling recognized in urban design decisions, since social media is so familiar to almost everybody nowadays. Even with the possibility of privacy concerns raising, as discussed earlier, people tend to feel comfortable in having their data used when they know how, why and for what it will be used, in which the TUI is a powerful tool in clarifying some of these questions and showing to what their data would be contributing to.

References

1. (1996) StarLogo: an environment for decentralized modeling and decentralized thinking
2. (2004) Netlogo: A simple environment for modeling complexity
3. (2018) Mobilitätswende auch in Köln in vollem Gang. https://www.stadt-koeln.de/politik-und-verwaltung/presse/mobilitaetswende-auch-koeln-vollem-gang. Accessed 14 Sep 2020
4. (2019) BearGIS. https://nicoazel.github.io/BearGIS/. Accessed 08 Jun 2020
5. (2020) Einwohner Statistik Koeln | Offene Daten Köln. https://www.offenedaten-koeln.de/dataset/einwohner-statistik-koeln. Accessed 16 Aug 2020
6. (2020) GAMA-Platform · GAMA. https://gama-platform.github.io/. Accessed 26 Aug 2020
7. (2020) Haushalte Statistik Koeln | Offene Daten Köln. https://www.offenedaten-koeln.de/dataset/haushalte-statistik-koeln. Accessed 05 Jun 2020
8. (2020) HeinzBenjamin/Crow. https://github.com/HeinzBenjamin/Crow. Accessed 05 Jun 2020
9. (2020) Map Features – OpenStreetMap Wiki. https://wiki.openstreetmap.org/wiki/Map_Features#Others. Accessed 04 Jun 2020
10. (2020) stgeorges/gismo. https://github.com/stgeorges/gismo. Accessed 04 Jun 2020
11. A. Repenning (2011) Making programming more conversational. In: 2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp 191–194
12. Alonso L, Zhang YR, Grignard A et al. (2018) CityScope: A Data-Driven Interactive Simulation Tool for Urban Design. Use Case Volpe. In: Morales AJ, Gershenson C, Braha D et al. (eds) Unifying Themes in Complex Systems IX. Proceedings of the Ninth International Conference on Complex Systems. Springer International Publishing, Cham, pp 253–261
13. Alphabet Home Page. https://abc.xyz/. Accessed 12 Sep 2020

14. Apple Computer Inc (1991) Human interface guidelines. The Apple desktop interface, [Nachdr.]. Addison-Wesley Publ, Reading
15. Bellifemine F, Poggi A, Rimassa G (2001) JADE. In: André E, Sen S, Frasson C et al. (eds) Proceedings of the fifth international conference on Autonomous agents. ACM, New York, NY, pp 216–217
16. Benjamin Felbrich (2016) Artificial Neural Networks, Stuttgart
17. Boeing G Clustering to Reduce Spatial Data Set Size
18. Bundesministerium für Verkehr und digitale Infrastruktur (2020) Mobilität in Deutschland (MiD). https://www.bmvi.de/SharedDocs/DE/Artikel/G/mobilitaet-in-deutschland.html. Accessed 10 Sep 2020
19. Castiglione J, Bradley M, Gliebe J (2014) Activity-Based Travel Demand Models: A Primer. doi: 10.17226/22357
20. Cerrone D, Lehtovuori P, Baeza JL Integrative Urbanism: Using Social Media to Map Activity Patterns for Decision-Making Assessment. doi: 10.13140/RG.2.2.24650.36802
21. Chen L (2012) Agent-based modeling in urban and architectural research: A brief literature review. Frontiers of Architectural Research 1:166–177. doi: 10.1016/j.foar.2012.03.003
22. CityScope Home Page. http://cityscope.media.mit.edu. Accessed 15 May 2020
23. COBE (2016) Deutzer Hafen. www.cobe.dk/idea/deutzer-hafen. Accessed 10 May 2020
24. Constine J (2018) Facebook restricts APIs, axes old Instagram platform amidst scandals. https://techcrunch.com/2018/04/04/facebook-instagram-api-shut-down. Accessed 17 Apr 2020
25. Cui Y, Xie X, Liu Y Social media and mobility landscape. uncovering spatial patterns of urban human mobility with multi source data. Social media and mobility landscape
26. FastCompany (2020) This startup wants to help smart cities. But they don't know where its data comes from. https://www.fastcompany.com/90465315/this-startup-wants-to-help-smart-cities-but-they-still-dont-know-where-its-data-comes-from. Accessed 01 Sep 2020
27. Fiesler C, Proferes N (2018) “Participant” Perceptions of Twitter Research Ethics. Social Media + Society 4:205630511876336. doi: 10.1177/2056305118763366
28. Fisher K, Gershuny J, M. Flood S et al. Multinational Time Use Study Extract System: Version 1.3 [dataset]. doi: 10.18128/D062.V1.3
29. Fuchs G, Andrienko G, Andrienko N et al. Extracting Personal Behavioral Patterns from Geo-Referenced Tweets
30. Gehl, Jan Life Between Buildings: Using Public Space
31. Gilbert GN (2008) Agent-based models. Quantitative applications in the social sciences, vol 153. Sage, Los Angeles

32. Google Maps Platform (2020) Places API. https://developers.google.com/places/web-service/overview. Accessed 14 Sep 2020
33. GPS.gov Official U.S. government information about the Global Positioning System (GPS) and related topics. www.gps.gov. Accessed 05 May 2020
34. Grignard A, Alonso L, Taillandier P et al. The Impact of New Mobility Modes on a City: A Generic Approach Using ABM, vol 114
35. Grignard A, Alonso L, Taillandier P et al. (2018) The Impact of New Mobility Modes on a City: A Generic Approach Using ABM. In: Morales AJ, Gershenson C, Braha D et al. (eds) Unifying Themes in Complex Systems IX. Proceedings of the Ninth International Conference on Complex Systems. Springer International Publishing, Cham, pp 272–280
36. Grisiute A From systems to patterns and back - Exploring the spatial role of dynamic time and direction patterns in the area of regional planning
37. Güney A (2019) Introduction to Bayesian Belief Networks. https://towardsdatascience.com/introduction-to-bayesian-belief-networks-c012e3f59f1b. Accessed 20 May 2020
38. Heppenstall AJ (2012) Agent-based models of geographical systems. Springer, Dordrecht
39. Hofmann J, Piele A, Piele C (2020) Arbeiten in der Corona-Pandemie – Auf dem Weg zum New Normal. Studie des Fraunhofer IAO in Kooperation mit der Deutschen Gesellschaft für Personalführung DGFP e.V.:22
40. Instagram (2020) Data Policy. https://help.instagram.com/519522125107875. Accessed 10 Sep 2020
41. Instaloader Home Page. instaloader.github.ip/index.html. Accessed 10 Mar 2020
42. Ishii H, Ullmer B (1997) Tangible bits. In: Pemberton S (ed) Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. ACM, New York, NY, pp 234–241
43. Jacobs J (1992) The death and life of great American cities, Vintage Books ed. Vintage Books, New York
44. Jacobs J, Albers G (1963) Tod und Leben großer amerikanischer Städte, 3. Aufl. Bauwelt Fundamente, vol 4. Birkhäuser, Basel
45. JSON Documentation. https://www.json.org/json-en.html. Accessed 07 Sep 2020
46. Justin Emery, Nicolas Marilleau, Nadège Martiny et al. (2017) Marrakair: une simulation participative pour observer les émissions atmosphériques du trafic routier en milieu urbain. In:
47. Kemeny T (2013) Immigrant diversity and economic development in cities: a critical review. SERC Discussion Papers. Spatial Economics

Research Centre (SERC), London School of Economics and Political Science, London, UK

48. Köln (2019) Kölner statistische Nachrichten. Bevölkerungsprognose für Köln 2018 bis 2040. Stadt Köln, Der Oberstadtdirektor, Amt für Stadtentwicklung und Statistik, Köln
49. Köln (2019) Kölner statistische Nachrichten. Bevölkerungsprognose für Köln 2018 bis 2040. Stadt Köln, Der Oberstadtdirektor, Amt für Stadtentwicklung und Statistik, Köln
50. Kravari K, Bassiliades N (2015) A Survey of Agent Platforms. JASSS 18. doi: 10.18564/jasss.2661
51. Liao Y, Yeh S, Jeuken GS (2019) From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. EPJ Data Sci 8. doi: 10.1140/epjds/s13688-019-0212-x
52. Manca M, Boratto L, Morell Roman V et al. (2017) Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. Online Social Networks and Media 1:56–69. doi: 10.1016/j.osnem.2017.04.002
53. Max Priebe, Timo Szczepanska, Leonard Higi, Tobias Schröder (2019) Partizitative Systemmodelierung als Tool für integrierte Stadtentwicklung. Projektbericht, Fachhochschule Potsdam
54. moderne stadt / Stadt Köln (2018) Integrierter Plan Deutzer Hafen. Quartiersbuch
55. Müh M (2020) Tangible Agents. A tangible user interface for an activity-based travel demand model. Master, Technische Hochschule Ostwestfalen-Lippe
56. Nadai M de, Staiano J, Larcher R et al. (2016) The Death and Life of Great Italian Cities. In: Bourdeau J (ed) Proceedings of the 25th International Conference on World Wide Web, Montreal, Canada, May 11 - 15, 2016. International World Wide Web Conferences Steering Committee, Geneva, pp 413–423
57. North MJ, Collier NT, Ozik J et al. (2013) Complex adaptive systems modeling with Repast Simphony. Complex Adapt Syst Model 1:1–26. doi: 10.1186/2194-3206-1-3
58. Noyman A, Holtz T, Kröger J et al. (2017) Finding Places: HCI Platform for Public Participation in Refugees' Accommodation Process. Procedia Computer Science 112:2463–2472. doi: 10.1016/j.procs.2017.08.180
59. Ory D (2017) A first step toward creating a digital planning laboratory is populating it. Accessed 20 Apr 2020
60. Overpass API Documentation. https://wiki.openstreetmap.org/wiki/Overpass_API. Accessed 01 May 2020
61. Patrick Taillandier (2014) Traffic simulation with the GAMA platform. In: Eighth International Workshop on Agents in Traffic and Transportation, 8 p

62. Polis Magazine (2020) Auszeichnung für den Deutzer Hafen Köln. https://polis-magazin.com/2020/06/auszeichnung-fuer-den-deutzer-hafen-koeln/?fbclid=IwAR1jHJ4o37f57w08-3ImBnYJ_H9ddNLdRN5vffz8EdSyZF4DEIwB_avJz2s. Accessed 14 Sep 2020
63. pomegranate Documentation. https://pomegranate.readthedocs.io/en/latest/. Accessed 20 May 2020
64. Press G (2016) Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. Accessed 14 Jul 2020
65. Rashidi TH, Abbasi A, Maghrebi M et al. (2017) Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transportation Research Part C: Emerging Technologies 75:197–211. doi: 10.1016/j.trc.2016.12.008
66. Replica Home Page. https://replicahq.com/. Accessed 05 May 2020
67. Scikit-learn: DBSCAN Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html. Accessed 01 May 2020
68. Shannon CE (2001) A mathematical theory of communication. SIGMOBILE Mob Comput Commun Rev 5:3–55. doi: 10.1145/584091.584093
69. Sidewalk Labs Home Page. https://www.sidewalklabs.com/. Accessed 05 May 2020
70. Siegfried R (2014) Modeling and simulation of complex systems. A framework for efficient agent-based modeling and simulation. Zugl.: München, Univ. der Bundeswehr, Diss., 2014. Research. Springer Vieweg, Wiesbaden
71. statista (2020) Number of global social network users 2017-2025. Accessed 14 Sep 2020
72. Statistiches Bundesamt Zeitverwendungserhebung (ZVE). https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Zeitverwendung/Methoden/zeitverwendung.html. Accessed 15 May 2020
73. Stumme G (2013) Proceedings of the 24th ACM Conference on Hypertext and Social Media
74. Sulis P Measuring urban vitality through human mobility patterns. UCL (University College London), UCL (University College London)
75. Sulis P, Manley E, Zhong C et al. (2018) Using mobility data as proxy for measuring urban vitality. JOSIS. doi: 10.5311/JOSIS.2018.16.384
76. Sung H, Lee S (2015) Residential built environment and walking activity: Empirical evidence of Jane Jacobs' urban vitality.

Transportation Research Part D: Transport and Environment 41:318–329. doi: 10.1016/j.trd.2015.09.009

77. Sung H, Lee S, Cheon S (2015) Operationalizing Jane Jacobs's Urban Design Theory. Journal of Planning Education and Research 35:117–130. doi: 10.1177/0739456X14568021
78. Sutherland, Ivan Edward,1938- Sketchpad, a man-machine graphical communication system. Sketchpad, a man-machine graphical communication system, Massachusetts Institute of Technology
79. Swier N, Komarniczky B, Clapperton B (2015) Using geolocated Twitter traces to infer residence and mobility
80. Taillandier P, Gaudou B, Grignard A et al. (2019) Building, composing and experimenting complex spatial models with the GAMA platform. Geoinformatica 23:299–322. doi: 10.1007/s10707-018-00339-6
81. Taillandier P, Grignard A, Marilleau N et al. (2019) Participatory Modeling and Simulation with the GAMA Platform. JASSS 22. doi: 10.18564/jasss.3964
82. Tweepy Documentation. www.tweepy.org. Accessed 10 Mar 2020
83. Twitter witter Privacy Policy. https://twitter.com/en/privacy. Accessed 10 Sep 2020
84. Twitter Developer Academic Research. https://developer.twitter.com/en/solutions/academic-research. Accessed 10 Sep 2020
85. Ullmer B, Ishii H (2000) Emerging frameworks for tangible user interfaces. IBM Syst J 39:915–931. doi: 10.1147/sj.393.0915
86. Wikipedia (2020) Agent-based model. https://en.wikipedia.org/w/index.php?title=Agent-based_model&oldid=974048486. Accessed 25 Aug 2020
87. Xu Y, Belyi A, Santi P et al. (2019) Quantifying segregation in an integrated urban physical-social space. J R Soc Interface 16:20190536. doi: 10.1098/rsif.2019.0536
88. Yuan Q, Cong G, Zhao K et al. (2015) Who, Where, When, and What. ACM Trans Inf Syst 33:1–33. doi: 10.1145/2699667
89. Zhang D, Cao J, Feygin S et al. (2019) Connected Population Synthesis for Transportation Simulation. SSRN Journal. doi: 10.2139/ssrn.3379496
90. Zhiyuan Cheng, James Caverlee, Kyumin, Lee, Daniel, Z. Sui Exploring Millions of Footprints in Location Sharing Services
91. Zhu Z, Blanke U, Tröster G Inferring Travel Purpose from Crowd-Augmented Human Mobility Data. doi: 10.4108/icst.urb-iot.2014.257173