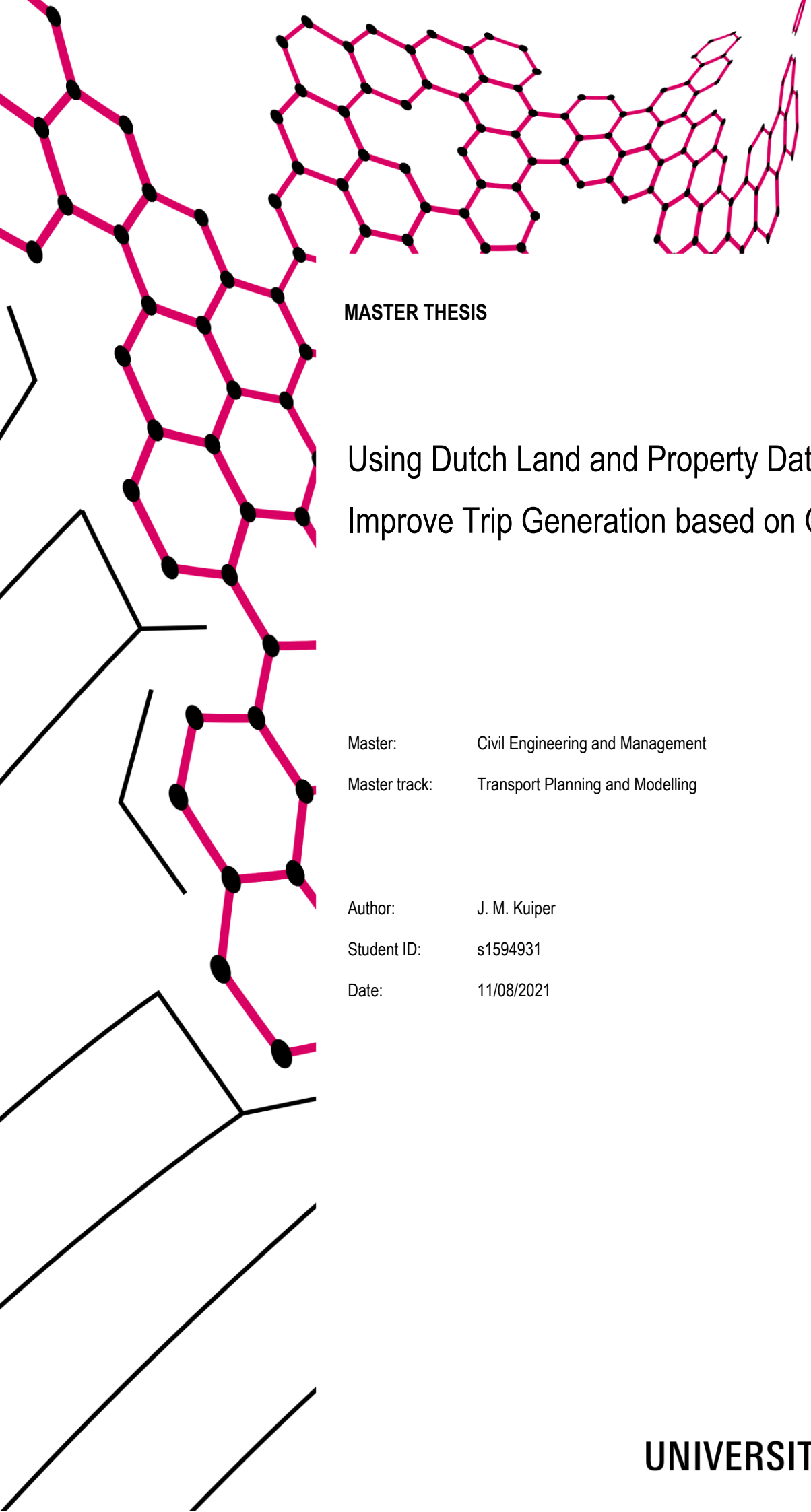


MASTER THESIS

Using Dutch Land and Property Data to Improve Trip Generation based on Open Data

Master:	Civil Engineering and Management
Master track:	Transport Planning and Modelling

Author:	J. M. Kuiper
Student ID:	s1594931
Date:	11/08/2021

UNIVERSITY OF TWENTE.

Contact information

Author:	Martijn Kuiper
Student number:	s1594931
Email:	martijn.kuiper96@gmail.com

Location thesis project:	Intern
UT supervisor:	Prof. Dr. Ir. E.C. van Berkum
Daily supervisor:	Dr. T. Thomas

University:	University of Twente
Address:	Drienerlolaan 5, 7522 NB, Enschede
Master:	Civil Engineering and Management
Master track:	Transport Planning and Modelling

Version:	Final version

Preface

Before you lies the thesis “Using Dutch land and property data to improve trip generation based on open data”, which I wrote as the last step in finishing my Masters in Civil Engineering at the University of Twente.

From December 2020 till the beginning of August 2021, I have been engaged in researching and writing this thesis. Due to the COVID pandemic, all research was conducted at home. The absence of fellow students and the challenges that rose during the conduct of this thesis project lead to hurdles I had not faced before during my study. I have learned a lot in the last few months, both professionally and personally. A special thanks to my daily supervisor Tom Thomas who was always available, always responding and who put a lot of time and effort into assisting me during the project. The regular meetings we had helped structuring my project, and the feedback always proved more than useful. Furthermore, I would like to thank Eric van Berkum for his time and effort put into my research, as UT supervisor. And finally I would like to thank my wife who made sure I did not lose my focus, and who provided helpful feedback.

After this graduation project, I will start my professional career as a BIM engineer, which I am looking forward to. I am excited to start this new learning phase, and I will look back at many joyful study years.

I hope you enjoy reading my thesis as much as I enjoyed writing it.

Martijn Kuiper

Deventer, August 2021

Summary

This thesis investigates the potential of open Dutch land and property data, so-called 'Basisregistratie Adressen en Gebouwen' (BAG) data, to be used for trip generation modelling at the urban level.

In a trip generation model (the first step of a transportation model) it is determined how many trips are produced by a zone, and how many trips are attracted by a zone. In the literature, a lot of research has been put in developing a trip generation model itself, in which through advanced analysis it is investigated how certain personal or zonal characteristics affect the travel behavior of individuals. However, the more advanced the models, the more specific the data that is required. The availability of data is a bottleneck for the application of trip generation models. In literature, there is little focus on this aspect of trip generation modelling, while in practice modelers encounter major challenges when developing a database for a trip generation model. And furthermore, developing a database quickly becomes expensive and time consuming. Not every institution or organization has the means to purchase such required data.

In the Netherlands, the Central Bureau of Statistics (CBS) is the main provider of open census data. Information on the number of residents, income, car ownership and other factors often used in trip generation studies are supplied at different administrative levels. But, in terms of aggregation level, completeness and novelty, the CBS data is lacking. For estimating work trips at the urban level, no open job data or activity data is available. Finally, for activities as shopping, sporting and other leisure, conventional open data sources are also lacking. In Dutch municipal transportation studies, sporting and other leisure activities are not even considered.

To improve the possibilities of estimating trip generation based on open data at the urban level, the use of BAG data as a source for trip generation factors and activities is researched in this master thesis. BAG is disaggregated, open data containing every building in the Netherlands including the surface area of all of its spaces and the function for which the building is constructed, such as living, industry, office, shop and sport. And for residences, different residence types are included, such as detached, multi-family and terraced. And furthermore, buildings are included for which a construction permit has been granted, which means that residences and other buildings can be included in a trip generation model that is constructed in the next few years.

To research the potential of BAG at the activity side, it has been researched what open data already is available to supply for activities and factors. Mainly for work and shopping trips, BAG data can be of added value. For shopping activities, operations have been developed that successfully identify shopping activities in BAG. Furthermore, the ability of BAG to predict trip generation factors at the household side has been evaluated. Based on BAG attributes it is possible to predict car ownership levels and the number of residents in a zone. This predictive capacity of BAG can be used in two ways; subdividing CBS District data into lower aggregated zones, suitable for estimating trip generation at the urban level, and predicting the number of residents and car ownership levels for new housing developments. Finally, in a case study in the municipality of Ede, it is showcased how BAG enables identifying bottlenecks caused by future travel demand, based on BAG predictions.

In this research, it has been found that BAG increases the possibilities of estimating trip generation based on open data, by complementing shortcomings of CBS open data at the household side, by enabling precise trip generation estimation at the activity side and by providing future land and property data that could be used to predict travel demand.

Nomenclature

Term	Meaning
A	Surface area
AFC	Automatic Fare Collection
BAG	Basisregistratie Adressen en Gebouwen [Key Register Adresses and Buildings]
BRT	Basisregistratie Topografie [Base Registration Topography]
CBS	Centraal Bureau Statistiek [Central Bureau of Statistics]
DBSCAN	Density-Based Spatial Clustering of Applications with Noise)
dens_C	Indicator for the density of the cluster (explained later)
DSA	Databestand SportAanbod [Database Sport Accommodations]
edf	Estimated degrees of freedom
eps	Maximum distance allowed between two cluster points.
GAM	Generalized Additive Models
GCV	Generalized Cross-Validation
GFA	Gross Floor Area
GP	General Practitioner
HB	Home-Based
HH	Household
IBIS	Integral Business Area Information System
ITE	Institute of Transportation Engineers
K+R	Kiss and Ride
KW	Klinkenbergerweg
max_A	Surface area of the largest VBO present in the cluster
mean_A	Average surface area of the cluster
minPoints	Minimum cluster size for cluster to be considered a cluster
N	Cluster size
NHB	Non-Home-Based
NRM	Nederlands Regionaal Model [Dutch Regional Model]
NZ	New Zealand
OD	Origin Destination
OSM	OpenStreetMap
PC4/6	Dutch postal code with 4 or 6 digits (e.g. 1234AB)
PT	Public Transport
RMSE	Root-Mean-Square Error
RQ	Research Question
SD	Standard Deviation
TAZ	Traffic Analysis Zone
UK	United Kingdom
VBO	Verblijfsobject [Accommodation object]
VGI	Volunteered Geographic Information

Table of contents

Chapter 1. Introduction

The transportation sector facing several demanding challenges. Creating a sustainable future by reducing greenhouse gas emissions and supporting multimodal development are becoming more pressing subjects within the transportation sector (Currans, 2017). If performance measures are evolving, so must the data. Transport models are the backbone for determining the transport impact of urban developments. Increasing the representativeness of transport models may lead to better decision making by policymakers. A transport model is the representation of human travel behavior. To put it simply, the modeler wants to know when and how many people are travelling from A to B, which route they take, and with what mode they travel. This could for example be measured by traffic counts or mobile phone data. In order to estimate what impact new developments - such as a new neighborhood or the expansion of an industrial area - have on traffic, the modeler needs to be able to predict travel behavior. This requires insights in travel behavior; what factors affect travel behavior, and what data can be used to model it?

Understanding trip generation is a part of understanding travel behavior, and estimating trip generation is the first step in transportation modelling. In this step, the modeler is on one hand occupied with determining how many trips people are going to undertake, and on the other hand with how many trips different entities such as grocery stores, schools libraries will attract. To model trip generation, the modeler needs to know where people live and what their activities are going to be, and what the possible destinations can be. Socioeconomic- and land use data are both valuable for determining trip production and trip attraction within trip generation modelling.

Trip generation models can support decision-making, policy developing, or other planning activities at different levels and for different institutes. However, constructing a solid trip generation model that is able to accurately predict transport demand is expensive and time-consuming. Therefore, it would be of great benefit if techniques are developed that enable accurate trip generation estimates based on open data. In the Netherlands, the Central Bureau of Statistics (CBS) is the main source of open socioeconomic and demographic data. The CBS is an autonomous administrative authority, that provides insight into social issues through reliable statistical information and data. In doing so, the CBS supports social debate, policy development and decision-making (CBS, 2020). Published socioeconomic and demographic CBS data includes attributes such as the number of households and persons, income, car ownership, densities, and so forth on different aggregated levels. Many of the aforementioned attributes are used to explain travel behavior in transportation modelling. Therefore, the CBS is a main source for transportation modelling in the Netherlands. However, the CBS contains aggregated data, and the open CSB data and the availability of CBS data differs per aggregation level and per year. These notions do affect the results of trip generation modelling, and therefore transportation modelling in the Netherlands. Although very important at the trip production side of trip generation modelling, the CBS is not the only data source used for trip generation modelling. The activity side of the trip generation modelling is where little or no open data is available. And even when data is available, the modeler may encounter the same problems; high levels of aggregation, outdated content and restricted data availability.

The Dutch government uses ten so-called base registrations to gain, maintain and update information to execute its policies (Digitale Overheid, 2020). One of these is the Key Register Addresses and Buildings (BAG). BAG is an open available geographical data source containing municipal data of all addresses and buildings in a municipality(Kadaster, 2020). Amongst the attributes included are the pitches and berths of each building object, as well as building functions and residence types as defined by the Dutch

Building Act. BAG data is disaggregated, precise data that can be actualized on a monthly or even a daily basis. The disaggregate nature of the BAG, the included building functions and surface areas, and the open availability could help overcome barriers of conventional data sources as the CBS, and therefore be of great addition to current trip generation practices in the Netherlands.

In this report, a research is described in which the potential of the BAG within trip generation modelling is examined. This potential increases when the BAG is able to discern attributes that are relevant to model trip generation. A number of challenges arise by doing so; all kinds of different data sources are used with different units. A trip generation model should not collide with existing travel behavior theory, as well as taking into account the availability of data. While BAG itself does not contain as many attributes as the CBS, the present attributes, such as building functions and surface area, can be a basis to estimate or predict other attributes relevant to trip generation.

The report is structured in the following way. In Chapter 2, a literature study is presented in which general trip generation definitions are explained (Paragraph 2.1) and in which the impact of data use within trip generation is reviewed (Paragraph 2.2). Next, in Chapter 3, the research problem is described and research questions are formulated. In Chapter 4, it is examined what open data is available to supply for activities and trip generation factors, based on which it is concluded how BAG can fill in the blanks. In Chapter 5, methodologies are developed to analyze the potential of BAG for trip estimates and in Chapters 6, 7 and 8 the results are presented. At the end conclusions are drawn and results discussed in Chapter 9 and 10 respectively. The research is finalized by providing recommendations for further research in Chapter 10 as well.

Chapter 2. Literature study

In this chapter, scientific literature relevant for the research is reviewed, which provides a knowledge base and identifies a research gap. The literature review starts with building a theoretical cadre in Paragraph 2.1. This will enable the reader to understand the common definitions and concepts used within transportation modelling, and more specific, trip generation modelling. Subsequently, the focus shifts towards the use of data within trip generation modelling in Section 2.2.1. Next, the use and aggregation levels of zones in Dutch municipal transportation models are described in Section 2.2.2 and finally, the availability and aggregation levels of the main Dutch open data source, the CBS, are evaluated, also in Section 2.2.2. Based on this information, research gaps are identified at the end of this chapter.

2.1 Theoretical cadre: Trip generation modelling practices

In this Paragraph, the scientific literature has been reviewed on general trip generation modelling practices. Relevant concepts, definitions, models, and theories related to the research topic are described which will create a theoretical cadre for the remainder of the thesis.

The classic transportation model

The classic four-step transportation model is a model structure that has predominantly been used since the 1960s as the basis for transportation modelling. The modelling is an iterative process in which four sub-models succeed each other, resulting in a network to which trips of preselected modes (in most cases private vehicle and public transport) are assigned to transportation networks. The four sub-models are trip generation, distribution, modal split and assignment. With zones and networks as the basis, population data and employment, educational, recreational and shopping data are used to estimate the produced and attracted transport activities in each zone (trip generation). In the next step, the travel activities are assigned to zones, and the distribution over the study area is determined. In the modal split model, the mode for each travel activity is determined, and at last, each activity is assigned to the network (Ortúzar & Willumsen, 2011).

Trip generation definitions

In the first step of the classic transportation model, the trip generation stage, the total number of trips that each zone in a study area produces and attracts is predicted. A prediction can be made individual or household-based, or based on the properties of each zone such as population, number of cars and/or employment. Discrete choice modelling is also amongst the solution space; determining how many trips or journeys is a person going to undertake (Ortúzar & Willumsen, 2011). Depending on the scope of a project, trip generation models are applied on different levels; strategic, tactical and operational (Zenina & Borisov, 2013).

Within a transport model, movements of people are mostly represented by trips. A trip is a one-way movement on a network from its origin to its destination. Trips are either Home-based (HB) trips or Non-home-based (NHB) trips. HB trips include the home of the trip maker in either the origin or the destination of the trip, an NHB trip has neither (for example a trip between two workplaces). Different origins and destinations can both produce and/or attract trips, see Figure 2.1. Trip generation refers to the total amount of trips generated by (mostly) all households within a zone, be it HB or NHB trips. Trips (and journeys) have found to be more representative of real travel behavior if characterized by purpose (for example work, school, shopping), time of day (AM peak, off-peak, 24 hours), and household type (based on for example income, car ownership and household size) (Ortúzar & Willumsen, 2011).

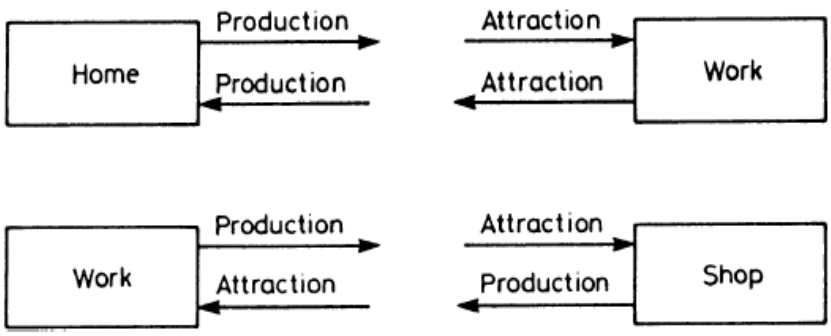


Figure 2.1: Trip production and trip attraction (Ortúzar & Willumsen, 2011)

Trip dimensions

Most current trip generation studies represent travel behavior as trips (Moeckel et al., 2015) (Currans, 2017). Travel behavior can also be represented as tours or activities. The difference between these models is the depiction of travel activity in the model (Figure 2.2). Trip-based models only consider each separate trip whereas tour-based models consider the connection between trips. Activity-based models model travel in the context of activities, trying to represent travel theory as best as possible (TNO, 2007). Activity-based modelling has gained more attention over the past decade. Although the execution of the activity-based model varies widely in different studies, the central idea is the same; representing travel behavior by predicting what activities and travel is conducted by the individuals of a household. When, to where, how long, with whom the trips are conducted are examples of problems the model needs to solve. The focus of the activity-based model is not the aggregation of the total number of trips, but on activity making, including everything that it depends on (Rasouli & Timmermans, 2014) . It is expected that the momentum of the activity-based models will resume the coming years. A barrier to overcome is, for example, the need for more data, and the sheer complexity of developing activity-based models can be considered a barrier to. From count data and mobile data (Section 2.2.1), contextual data such as the purpose of the activity cannot be deducted. This is underlined by Moeckel et al. who mention that implementing an activity-based model is a significant undertaking. The authors follow the principles of 'agile development', which means preserving an operational model and focusing on modules that need the most attention for improvements (Moeckel et al., 2015). While the substantiation of activity-based modelling may be best supported by current travel demand theory, it is still the outcome of trip generation models that matters. In a comparative analysis carried out by Chang et al., outcomes of trip-based generation models were compared with an activity-based model, in which the classic category-type, trip-based model showed the overall best performance. The authors mention that the outcome can be data-specific, and that some methods may better replicate observed patterns. But it is the outcome of the forecasts that (mostly) matters. Seemingly sophisticated methods are not a guarantee for better performance within trip generation (Chang et al., 2014). Considering the current major barriers for activity-based modelling, and the current large-scale application of trip-based generation models in transportation studies, the literature on trip-based generation modelling will be explored in the following sections.

Uni-modal and multi-modal modelling

In 2017 in the Netherlands, 42% of total trips made were by car, from which 29% as car driver and 13% as car passenger. Furthermore, 6% of trips were made by public transport (PT), 26% by bicycle and 23% walking (KiM, 2019). Within transportation modelling, different modes can be considered, depending on the aim of the model. Traditionally, traffic impact assessments are more car-centric, as transportation policies were developed aiming at providing infrastructure for the car. However, current transportation policies increasingly aim at alternative and sustainable transportation modes such as PT and cycling. To predict the impact of these policies, transportation models should encompass all modes including walking, cycling, PT and freight (Cooley et al., 2016). The lack of sufficient data for multi-modal modelling and the higher complexity in data collection result in a major barrier for multi-modal

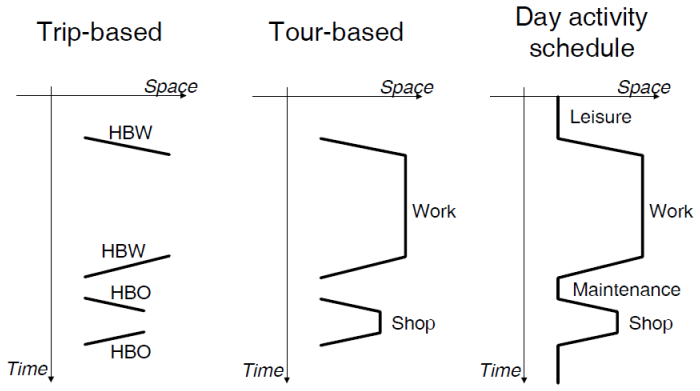


Figure 2.2: Representation of travel as trips, tours or activities in transport modelling (TNO, 2007).

modelling. According to Gruyter, most studies focusing solely on vehicle trip generation and do not even mention the word vehicle. (Gruyter, 2019)

Trip generation modelling approaches

Through the years, many trip generation methods have been developed and used. Regression and cross-classification models are predominant, but other approaches such as discrete choice models or activity-based models are on the rise (Ortúzar & Willumsen, 2011) (Moeckel et al., 2015). The cross-classification or category analysis method is (currently) the most predominant approach for trip generation. Based on for example survey data, trip rates are estimated for household- or person categories with different attributes such as income or car ownership for different trip purposes. No prior assumptions about the relationship between independent and dependent variables need to be made, and these relationships can differ between household- or person categories (Ortúzar & Willumsen, 2011). A disadvantage of the method is the large amount of data needed. Based on the number of chosen attributes and the levels within these attributes, a total number of possible categories is created. For each category, at least 30 survey samples are required (Moeckel et al., 2015). And furthermore, no effective way is available to choose amongst attributes. There are ways to enable multiple household attributes to be included while reducing the need for data. When combining multiple attributes and levels, categories are created that do not make sense. Moeckel et al. developed an algorithm to aggregate redundant categories based on the average standard deviation of each category, and decreased the total of aggregations from 67 million to 200 (Moeckel et al., 2015).

Regression models are based on establishing relationships between characteristics of a zone or household and trip generation. An advantage is that any characteristic that is thought to impact trip generation can be added by the researcher. However, multicollinearity between certain characteristics, such as car ownership and income, can lead to biased conclusions (Shi & Zhu, 2019). For zonal-based regression models, the unit of the zone is critical as regression models will only explain inter-zonal variety. Larger zone units might diminish inter-zonal effects. Moreover, the number of trips a person undertakes is assumed to be continuous, while in reality it is a discrete process (Chang et al., 2014). To overcome the several limitations of both category analysis as regression models, more sophisticated methods have been developed. For example, discrete choice methods allow for discrete processes to be modelled. These methods are used a lot in mode choice models, among others.

Thus, a transport model is required to be able to predict the impact of demographic or land use changes to transportation networks. Although new model developments such as activity-based modelling are on the rise, the classic four-step, trip-based transportation model is predominantly used for transportation analyses. The development of these models can be a timely and costly endeavor. Extensive data gathering is needed for both the development of the model, as for the socioeconomic and land use data to which these models are applied. In Paragraph 2.2 the scientific literature on the challenges of gathering (open) data for trip generation modelling has been reviewed.

2.2 Trip generation data gathering

Now that the theoretical cadre is built, the procedures and barriers of gathering data for trip generation modelling are reviewed from the literature. First, the use of trip factors and trip rates and the methods of gathering trip rate data are described, together with a review of the impact of aggregation levels on model outcomes and the development of zoning systems in Dutch transportation models (Section 2.2.1). This paragraph is ended with a description of the presence of open socioeconomic and land use data in the Netherlands (Section 2.2.2).

2.2.1 Trip generation model data

Trip generation factors and trip rates

In transport modelling, socioeconomic and land use data are often used to explain trip generation. Characteristics of residents may for example explain trip production, whereas land use data is used to determine both trip production as trip attraction. Trip rate studies are occupied with establishing relationships between trip volume and characteristics that are thought to be explanative. Some trip generation studies consider only socioeconomics or land use, whereas others include both (Shi & Zhu, 2019). For personal trip productions, income, car ownership, family size, age, driver's license possession, family members, household structure, level of employment, disadvantaged groups, value of land, residential density and accessibility are factors that are often considered and sought to include in trip generation studies (Shi & Zhu, 2019) (Moeckel et al., 2015) (Roorda et al., 2010). For personal trip attractions, employment numbers and/or roofed space available for different land uses are the most used variables (Ortúzar & Willumsen, 2011) (TNO, 2007). These factors are included because they can explain trip generation, and perhaps more important, because they are able to be measured and represented by data. Unobservable factors, such as cycling culture, can affect trip rates, even in comparable sites in terms of land use and type of location (Miller et al., 2006). To estimate the trip generation for a study area, both trip rates are needed that are applicable to the study area, as well as socioeconomic and/or land use data to which the trip rates can be applied.

Although numerous studies have analyzed trip rates, the usability of this knowledge is questionable. Even when using equivalent methods to determine trip rates for comparable neighborhoods, major differences will occur within trip generation, due to the large and random variation that is inherent to trip generation. Therefore, borrowing trip rates from other studies is possible, but the uncertainty of the rates should be taken into account, instead of just copying the mean (Miller et al., 2006). Milne and Abley compared trip rates from the United Kingdom and New Zealand. Although comparable trip rates were found, differences were also present. NZ residential trips rates were found to be higher than UK trip rates. Half of the trip rates of different comparable land uses were found to be similar (Milne & Abley, 2009). Thus, trip rates from studies in different countries cannot blindly be adopted.

Thus, to estimate trip generation in a Dutch environment, trip rates that are applicable in the Netherlands should be gathered. In a discussion paper originating from 2008, the use of transport models by Dutch authorities is reflected upon. One of the findings is that the amount of travel data based on which trip rates can be estimated has been reduced over the years. This applies to both surveys as traffic counts. Often, in practice, incorrect information is added to the modelling process. And often, uncertainties within traffic demand are not accounted for. The authors highlight the importance of transportation modelling in decision making (Schoemakers & Geurs, 2008).

Trip generation data gathering

The gathering of data on people's travel patterns has been developed and changed over the years. Since the 1930s, travel diaries have been used to increase the knowledge of people's travel patterns (Axhausen & Rieser-Schüssler, 2013). Household travel surveys are a major source for trip generation data. Along with trip data, contextual data of the trip makers can be gathered, which is a major source for modelling trip generation. A significant disadvantage of the conventional household survey is the problem of underestimation, as people tend to underestimate their travel volume (Thomas et al., 2018). Another

disadvantage is the high cost of conducting a large-scale survey, especially when a high level of detail and up-to-date data is required (Saadi et al., 2017). Several countries (including the Netherlands) conduct yearly household travel surveys to gain insight in the travel behavior of its citizens. The surveys reveal a vast amount of information about travel patterns on a regional or national scale, such as trips per day, trip motive, social status and modal split. The high level of detail also requires large sample sets to be able to determine trip rates for different household categories (Gruyter, 2019).

Despite some of the advantages of household travel surveys, inaccuracies within reported travel times and distances and other biases led to a shift towards passive tracking in the late 1990s (Thomas et al., 2018). The large-scale growing use of the smartphone and its GPS technology in the 21st century presents new opportunities for precise automatic trip and mode detection, and reveals the vast underreporting that occurs within self-completed trip-based diaries (Thomas et al., 2018). Shi and Zhu argue that in the nearer future, mobile phone signaling data will become increasingly important for trip generation research. The fact that almost everyone carries a mobile phone, the increasing availability and sharing of data and the unmatched depiction of someone's movement support the authors claim (Shi & Zhu, 2019). Shi and Zhu used provider data to passively analyze trajectories that are recorded when phone users carry out an action like navigation, texting or calling, resulting in detailed trip generation numbers for different zones. The results seem promising, but the scarcity of research on mobile phone data in trip generation, the lack of current data availability, and the inability of the data to link trips to socioeconomic characteristics (due to privacy) are major barriers. Thomas et al. also noticed the increasing potential of the mobile phone. They developed an application that uses GPS to automatically detect trips and used modes. The application revealed to be promising to reduce the underreporting that occurs with conventional travel surveys. Although promising, disadvantages remain. As concluded by Shi and Zhu as well, mobile data lacks contextual data. And the automatic detection of trip and modes reveal to be biased. Especially for shorter trips, it is difficult to determine the correct mode (Thomas et al., 2018).

More precise trip detection can be achieved by traffic counts. However, daily variation of counts, misclassification of travel modes and errors are disadvantages of traffic counts. A comprehensive set of counts is needed to compensate for these uncertainties. And not every count location should be given the same priority. Traffic counts can be an indication of trip generation, route choice and depending on what is counted, even mode choice and destination choice. Therefore, the count locations are quite significant. And the way the count data is gathered can differ by for example duration, equipment and the period of data gathering. This all may affect the liability of each count location (TNO, 2007).

Trip generation data gathered from surveys or mobile applications often provides information on the use of different modes, whereas traffic counts are more car centered. For public transportation, (automated) counts are a major source of data, contributed with survey data. Automatic fare collection (AFC) is a new data source that has great potential to measure origin destination flows. This data source is however not error free. It is not possible for every trip to infer the correct origin or destination. Not everyone uses the AFC system and fare evasion is also a problem. The data is promising but it should, for example, be supplemented by survey data to correctly represent public transport passenger flows. Other data sources could be automatic passenger count systems or farebox data, however, these data sources are prone to similar uncertainties and the qualities of each of the data sets should be properly examined (Egu & Bonnel, 2020).

With the above-mentioned trip generation data gathering methods, it is theoretically possible to map current traffic flows. However, a transportation model is necessary to predict changes to traffic flows due to the developments such as a new neighborhood, infrastructural changes, behavioral changes, mode developments. Therefore, the above-mentioned data sources are used to establish trip rates (by methods as described in Paragraph 2.1), and for calibration and validation of a model. As mentioned before, trip rates based on household surveys may be biased due to underreporting. Mobile phone data may solve this problem but the current scarcity of research on this subject may present barriers. Traffic counts are very precise but lack contextual data, which are needed to estimate relations between different trip

generation factors and trip generation. Thus, when trip rates are included in a study, the uncertainties of the data should be taken into account.

2.2.2 Land use and socioeconomic data

Impact TAZ size

Little attention is present in the literature towards the availability of socioeconomic and land use data in countries, and the impact of the quality of available data towards the outcome of trip generation models. Some attention however is given to the impact of the development of zoning systems, or Traffic Analysis Zone's (TAZs). Here the aggregation level of the zone could be a compromise for the reflection of reality, and therefore impact traffic analyses. The availability of socioeconomic and land use data and the development of TAZs within trip generation have therefore overlapping problems.

The development of zoning systems is heavily dependent on the experience of the modeler. There are no strict rules attached to the process. A practical consideration is that the zones should be compatible with units used within available census data. In addition, there are a number of considerations that can be taken into account to develop a zone system. Zones should be as homogenous as possible. If in a certain region population density largely varies over space, averaging the population of the region by making it a zone will lead to wrongful estimations. Zones do not have to be equally sized and the centroid of each zone should be easily determined (Ortúzar & Willumsen, 2011). The effect of zoning development on modelling outcomes is scarcely described in the literature. However, some implications of zonal development on modelling outcomes can be found. If intra-zonal trips are not considered in trip generation modelling, a low number of zones can largely affect model outcomes, when for example considering total vehicle miles travelled (Ding, 1998). If a low-detailed road network is used in a model with a large number of zones, traffic that would travel on lower-level roads in a real situation are detouring on higher-level roads in the model. And the other way around, if zones are too large while analyzing traffic activity on low-level roads, biases will occur (Jeon et al., 2012). Therefore, zoning systems should be developed according to the level of detail to which transportation effects are modelled.

Zoning in traffic models Dutch municipalities

To get an idea of the development of zoning systems in the Netherlands, the zoning systems developed in the model of available technical reports of traffic models developed at municipalities were analyzed. The municipalities of Utrecht, Ede-Wageningen, Hilversum and Harderwijk have published technical reports describing the specifications of the models in relatively high detail (Royal Haskoning, 2018), (Goudappel Coffeng, 2018b), (Royal Haskoning, 2015), (Oranjewoud, 2009). Although the models were developed by different municipalities and engineering firms, the way in which the zoning system is created is very similar.

In the municipal traffic models, different model areas are defined: a study area which is the municipality itself, and influence areas and peripheral areas which model external traffic. Input of traffic from

Figure 2.3: Zoning system of traffic model Harderwijk, zones (small blue lines) are based on aggregated PC6 regions and align with districts and quarters (thick blue lines) (Oranjewoud, 2009)

Figure 2.4: Different aggregation levels of CBS data in the city of Enschede, from left to right: districts, Quarters, PC4, PC6

external areas is determined by regional traffic models, the Dutch Regional Model (NRM) or adaptations of it. Different zoning systems are used in the model areas, with the most detailed zoning in the study area.

In the traffic models of Ede-Wageningen, Haaksbergen and Hilversum, zones are developed based on quarter and district-level data (administrative units as used by the CBS) and aggregations of PC6 regions (Figure 2.3). A zone is defined as an area with a certain logical coherence for which the necessary data, such as inhabitants and jobs, are known. Several starting points are mentioned for aggregation. PC6 areas must have common boundaries and access to the areas must, as far as possible, take place via the same roads. Additional zones are added in places where municipalities have planned future developments. The process of creating the zones is described in a fairly similar way; however, the number of zones that are finally being used differ enormously. In the study area of traffic model Ede-Wageningen, with a population of 130,000, 891 zones were developed. The Hilversum model has 289 zones with a population of 90,000. The model description of Utrecht does not describe how the zoning was developed.

Thus, from the literature it can be concluded that the zoning system significantly affects modelling outcomes. The number of intra-zonal trips will increase with larger zones, and the TAZ size must align with the level of detail that is demanded. From the model reports of the Dutch municipalities it can be concluded that the homogeneity of the zones, the connection of the zones to the network, and the connection between zones and available census data formats are kept in mind during development of the zonal scheme.

Overall, the process of developing transportation models can be a technical and resourceful endeavor. Acquiring and transforming data for trip generation modelling can be a great part of it. The municipalities from which the technical reporting of their traffic models was obtained have outsourced this process to engineering consultants. Not every institute or agency has the means to frequently develop or update a transportation model that could improve decision-making. Therefore, it could be of great added value when parts of the transportation modelling process could be carried out based on open data. However, finding techniques for collecting data for trip generation modelling that are less expensive is a major issue within transport planning (Caiati et al., 2016). To estimate the potential of using open data for trip generation modelling data in the Netherlands, the availability of open census and land use data has been reviewed in the next section.

Open socioeconomic and land use data

Census data is often the main source of socioeconomic and land use data used for the zoning system in a trip generation study. The CBS which provides the census data provides for the number of households, persons, income, car ownership, densities, and so forth on different aggregated levels for different years. It has a multitude of data attributes that are relevant for and used in trip generation studies to estimate traffic demand.

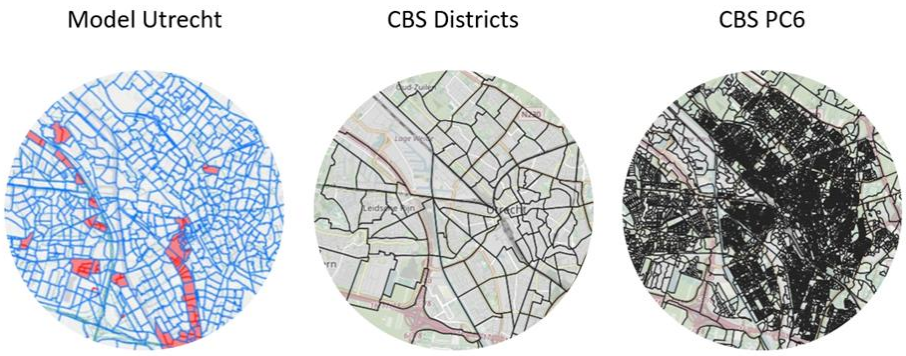


Figure 2.5: TAZ size used in transportation model of Utrecht compared to CBS administrative units

The CBS data is provided in different formats (Figure 2.4). Between these formats, the level of aggregation, the unavailable areas due to privacy restrictions, and the availability of open data differ. In terms of aggregation levels, the PC6 format is the only data type that is properly able to provide for data in such a detail that the diversity in the socioeconomic and land use data is accounted for. When considering traffic demand estimates at a city level, PC4 and Quarter data are far too aggregated. District data can account for some variety in socioeconomic and land use data throughout a city, but the intra-zonal variety of this format is still likely to affect the results of a traffic demand analysis in a city at large. Developing a zoning system at this level might only be suitable for analyzing traffic estimates at high-level roads. Thus, in terms of aggregation, the PC6 format is the only suitable format to provide data needed for trip generation modelling at city level. In Figure 2.5, the TAZ size of the transportation model of the city of Utrecht is compared to the size of PC6 regions and Districts in Utrecht.

The CBS has put restrictions on the availability of certain data attributes in areas in which the number of occurrences of a certain data attribute is less than five. Mainly at the PC6 level, the impact of this restriction is significant. For example; around 450.000 PC6 areas are present in the Netherlands. In 44.000 areas the number of residents is restricted, in 69.000 areas the average household size is restricted, and in 232.000 areas, almost half, the number of areas with 0 to 14 residents is unavailable.

Table 2.1: Overview of the CBS data formats and its limitations

	PC4	PC6	Quarters	Districts
Aggregation	Very large	Very small	Very Large	Medium
Privacy limitations	Negligible	Manifold	Negligible	Negligible
Availability	-2 years	-5 years	-2 years	-2 years
Land use data	Limited	Limited	Limited	Limited

And furthermore, the availability of all attributes within the data formats differs. For PC4, Quarters and Districts, the CBS takes about two years to completely register all data attributes. But only five or more years old PC6 data is made publicly available. Within five years socioeconomic differences, as well as new building developments may significantly affect the data. Overall, the availability of land use data (destination data) in the CBS is limited to the number of facilities per sector type. An overview of the differences between the CBS formats is presented in Table 2.1.

Open CBS land use data is limited. In the Netherlands, BAG is an openly available land use data set that contains disaggregate data of the location, shape, size, function, building year and address of all buildings in the Netherlands (Kadaster, 2020). The completeness and the disaggregate nature of BAG are very promising, but it has not been used as a source for trip generation modelling. Another land use data source is OpenStreetMap (OSM). OSM geodata is a comprehensive disaggregated data set that contains point features, line features and polygon features that depict places, facilities, infrastructural objects and roads, buildings and more. The input and modification of the data rely entirely on volunteers, called volunteered geographic information (VGI), from which OSM is one of the most well-known examples (OSM, 2020). The availability of data is massive and incomparable to other open databases,

but the open nature and the voluntary basis of OSM mean that there are areas that are incomplete or included with errors.

The most important indicator for predicting working trips is the number of jobs at a working place. Work trips are responsible for more than a fifth of all generated trips on an average working day. Based on the number of employees within a certain sector, the transportation modeler is able to estimate how many trips a working location is going to generate. Employment data at a level at which it enables trip generation estimations at a city level is not openly available in the Netherlands. This provides large limitations for trip generation estimation based on open data. For educational trips, the number of students at a school or university is an important trip generation factor. Open data of the number of students at PC6 level is available in the Netherlands.

2.3 Research gap

In Paragraph 2.1, current practices within trip generation modelling have been reviewed. Through the years, a research shift has been made from trip-based models to tour- and activity-based models. Models that consider travel behavior as tours or activities need additional, hard to obtain, contextual data. Activity-based models need a multitude of resources and expertise to develop. Regardless of the research shift, trip-based models are still predominantly used within transportation studies. Therefore, it is still valuable to improve trip generation practices, especially for policy makers or institutes that do not have the means to develop these complex models.

Two key elements of determining trip generation estimates for a study area are the development of a trip generation model and the development of a zoning system containing socioeconomic and land use data. The first element, the development of trip generation models, has received much attention in the literature. The impact of the availability of (open) socioeconomic and land use data and the development of a zoning system based on the results of trip generation estimates have hardly been considered. This might be a result of the variety of data that is openly available in different countries. But finding new techniques for collecting data for trip generation studies that are less expensive is a major issue in transport modelling. More research on the use of open data in trip generation could be of great additional value.

Both the trip generation model and the socioeconomic and land use data will affect trip generation estimates. Trip rates are established mathematical relations between trip volumes and characteristics that are thought to be explanative. Adopting trip rates from other countries can lead to erroneous model results. In the Netherlands, census data from the CBS is limited in terms of aggregation levels and actuality. Recent housing developments might not be included in the open available data, and some variables can only be acquired on a higher aggregated level. OSM data is a potential source for open land use data, but the accuracy and reliability of the data are often questionable due to its voluntary nature. BAG data is an openly available disaggregated land use data set. The non-availability of land use and socioeconomic data, and data to estimate trip rates is a bottleneck for trip generation studies. Furthermore, it is not clear what data is exactly required and what data is available. A synthesis on the availability of open data, and on the requirements that should be put onto a trip generation database would be a major step forward.

Considering the lack of techniques to collect data for trip generation that are less expensive, and the large limitations of open socioeconomic and land use data that is available in the Netherlands, it is sought to develop a new technique to use BAG as a data source to develop the required socioeconomic and land use data attributes as complete and accurate as possible. This may lead to improved trip generation estimates, and it could make trip generation studies in the Netherlands more accessible and less expensive. Ultimately this could lead to improved policy-making.

Chapter 3. Problem definition

In the previous chapter, a foundation has been laid using the research that is conducted on the topic of this research. By reviewing the literature it could be estimated that available open data is not able to produce trip generation estimates. Based on the identified research gaps in Paragraph 2.3, a detailed problem description is presented and subsequently a research goal and research questions are formulated.

3.1 Problem description

Data gathering for trip generation modelling can be a resource consuming and time-intensive occupation. Not much research has been put into developing techniques to acquire less expensive and high-quality data for trip generation. The impact of the quality and availability of socio-economic and land use data on trip generation estimates is significant. Wrongful aggregation of data may lead to erroneous traffic estimates. The TAZ size that is used in Dutch municipal transportation studies is between PC6 and District level. If open data is used for trip generation estimations at the municipal level, it should be able to provide required attributes at this level of detail.

The CBS provides many useful attributes but the available data formats are either too highly aggregated, or too limited in their data availability in terms of the publication date and privacy limitations. Data at PC6 level is required to create a zoning system that is sufficiently able to provide detailed attributes. Land use/activity data provided by the CBS is too scarce and employment data, an important indicator for estimating working trips are not open available at the required level.

To solve the aforementioned problems, and to contribute to the identified research gap (Paragraph 2.3), the potential of BAG data to be used as an open data source for trip generation modelling is researched. BAG contains open, disaggregated and monthly updated land and property data that contains the function and the surface area of every building in the Netherlands. Examples of BAG functions are office, industry, shop and jail, but also multi-family house, terraced house and more. BAG data could have an significant potential in estimating trip generation at the destination side. And because of the actuality and disaggregate nature of the data, and the combination of its functions and surface areas it is expected that BAG data could solve the limitations of PC6 data of CBS, and perhaps account for required socioeconomic attributes. Furthermore, BAG might solve the issue of unavailable employment data. The presence of functions of different working sectors in BAG combined with the surface areas may provide logical attributes to estimate employment data at the local level.

3.2 Research goal and questions

As concluded in the previous sections, open data availability could increase the possibility of institutes without an abundance of resources to carry out transport analysis when it could contribute to decision-making. Moreover, the availability of socioeconomic and land use data, and therefore the quality of the zoning system, will significantly affect the quality of transport analysis. Therefore, the following research goal has been formulated:

> Increasing the possibilities of trip generation estimates based on open data in the Netherlands by researching the potential of BAG to be used as a data source for trip generation modelling

BAG data is land use data that doesn't contain data attributes such as residents per household, income or household structure; attributes relevant for trip generation estimates. However, BAG data includes attributes explaining housing type and surface area. The hypothesis is that both housing type and surface area can account for socioeconomic data that are needed to determine trip generation. If this holds, BAG data could remedy the deficiencies of CBS open data. The same holds for employment data and activity data, attributes present in BAG which could account for these unavailable trip generation factors.

In order to reach the research goal, some research questions are formulated which need answering. Each question will be accompanied by an brief explanation.

Research question 1: What trip generation factors and activities need to be discerned in BAG in order to enable trip generation estimates?

Before the potential value of BAG within trip generation modelling can be estimated, it needs to be determined what data attributes BAG needs to represent. These attributes are necessary for the modelling process to estimate the required trip generation estimations. The number and type of socioeconomic and land use attributes that BAG needs to represent depend on the following questions:

A. What are important trip generation factors on the production and attraction side?
B. What activities need to be distinguished?
C. What open data is already available to provide for these factors and activities?

A trip generation modeler will establish relationships between the number of trips generated by an entity and its socioeconomic or land use characteristics, the so-called trip generation factors, such as income, car ownership, or the number of employees per square m2. Learning about what trip generation factors are needed to estimate trip generation will be a first step to determine to what extent BAG could be used as a data source.

The same holds for the activities that are relevant to discern in a trip generation model. What motives should be discerned in a transportation model? What activities belong to which motive? And how can these motives be linked to functions in BAG? BAG already includes several attributes that provide for information about the use function of a building, such as shop, or lodging. But it has to be determined what destinations need to be discerned in a model to estimate trips at a realistic level. For example, a supermarket as entity generates many trips compared to other shopping destinations. But more important is the trip generation per roofed space; when this is approximately equal to other shopping destinations, there is no need for distinction.

If open data is already available to provide for these factors and activities, there is no need for BAG to supply this data. For every aspect, it will be reviewed whether open data is available, and how this data could be put to use. A complete synthesis of the availability of open data might proof valuable, and even contribute to the research goal.

Research Question 2: To what extent can BAG provide for trip generation factors and activities required for trip generation modelling?

During RQ1 it is determined what socioeconomic attributes, and what activities must be provided for by BAG when estimating trip generation. Now it needs to be determined to what extent and how accurate BAG is able to provide for these attributes. The attributes already present in BAG will be the starting point for including others. The process of estimating values in BAG based on present attributes will introduce uncertainties that need to be analyzed.

Research Question 3: How does BAG improve trip generation outcomes based on open data and how does this impact transportation modelling outcomes?

When it is clear to what extent BAG is able to be used as data source for trip generation estimates, it can be determined how the use of BAG within trip generation can affect trip generation outcomes. By comparing the developed BAG data with conventional data sources, the added value of BAG can be estimated. This evaluation should lead to the conclusion of whether the research goal has been fulfilled.

3.3 Scope

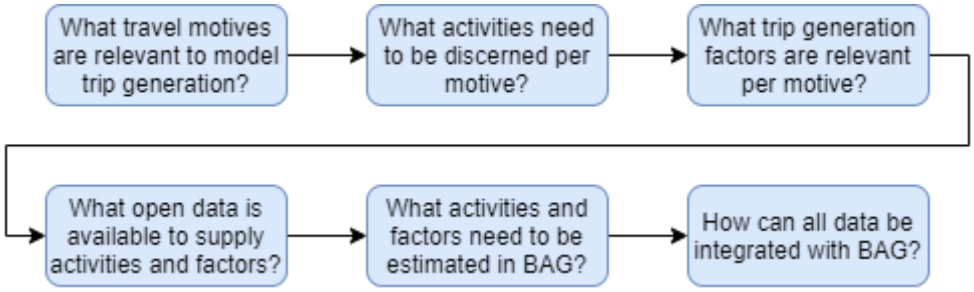
Within the research, activities and trip generation factors are considered that are relevant for an average, 24 hour working day. All research is carried out with the aim to enable trip generation modelling at the urban level. This has implications for considered aggregation levels, and the level of detail to which data sources have to comply.

Modelling work-related trips is a significant part of trip generation modelling, especially on an average working day. BAG has large potential for estimating work trips at the activity side, because of the included functions such as office, industry and healthcare. Within the research, the possibilities of BAG to predict work trips at the activity side has been explored. It was not possible to fully research this subject within the time-span of the thesis project. Therefore, up till Chapter 4, work trips are considered. Thereafter, the potential of BAG related to work trips is not further included in this thesis project.

Chapter 4. Data

In this chapter, it is described what data is needed to model trip generation at a municipal level and what open data in the Netherlands is available for this purpose, to eventually determine what open data is not or marginally present. The next step is to determine how suitable BAG data is to fill in the open data gaps. But first, it should be determined what a database should provide to estimate trip generation at a municipal level, or more specific, what activities and what trip generation factors are needed to determine trip generation. In this chapter, it is systematically determined what data is needed, what is already available, and what BAG should provide.

In order to achieve this goal, the following questions need to be answered:



As explained in Chapter 1, within person trip rate studies and modelling practices trip generation models are mostly structured by travel purpose, or motives. Per motive, trip rates are identified (often based on travel surveys) that explain how many trips a person in a model is going to make with the stated motive, in a given time. Determining what motives are relevant to model trip generation is a first step in determining what data is required.

Within motives, activities could generate relatively fewer or more trips per factors like square meter, job, and unit. It would improve the outcome of trip generation estimates if activities that significantly differ from other activities within a motive could be discerned in a trip generation database. And it should be determined what trip generation factor is best suitable to estimate the trip generation of a motive or an activity and whether square footage is suitable for shopping trips or the number of employees.

After these steps, it is clear what activity locations and trip generation factors are required to determine trip generation. This can be used to determine what open data is already available, and what BAG has yet to provide. And finally, all data sources should be integrated with each other. BAG data contains relevant attributes to be able to do so. Thus, BAG data is used for:

- Estimating activity locations and trip generation factors for which no recent, detailed and complete open data is available
- Connecting all available open data sources

First, a detailed description of BAG data is presented. Then, the above questions are answered to finally determine what BAG should provide to enable trip generation estimates based on BAG data. The results generated in this chapter are used to answer the first research question.

4.1 BAG data

In this section, BAG data and its use are described in detail, including how it can be retrieved, the definitions of each attribute used within the research and the process of making BAG data suitable to be used as a data source for trip generation estimates.

The primary open data source analyzed in this study is BAG data (or simply, BAG). BAG is disaggregated, monthly updated, open available data. BAG contains locations, functions and surfaces of all buildings, or rather, all use spaces in the Netherlands. In addition, each location has an address including street, house number and postal code. The location, functions and surface areas being present lead to BAG being potentially a very suitable data source for estimating trip generation. The functions can tell something about the type of activity at a certain location, and in addition, the surface area can be used to estimate the amount of trips generated on an activity. In turn, the available address data makes it possible to link BAG with other data sources. If open data is already available for an activity, the available address data will make it possible to link this open data to buildings or use spaces in BAG.

4.1.1 Availability

BAG data is made openly available in XML format from the webpage of the Kadaster (Kadaster, 2021a). For this research project however, BAG has been retrieved from another web service, the webpage BAG GeoPackage, made available by the developers of Geoparaat (GeoParaat, 2021). There are two differences between both data sources: the format in which the data is available and the presence of historical data. The GeoPackage format enables direct use of the data in a GIS environment. And the BAG database from the Kadaster is a historical database, in which records of buildings and use spaces are present that have ceased to exist. The developer of the BAG GeoPackage webpage acknowledges the demand for a snapshot of BAG data for applications, in which only current BAG records are present. Because of the available data format, and the absence of abundant historical data, the BAG data from Geoparaat is used within the research project. Each month, a new GeoPackage of BAG data is made available.

In June 2021, BAG XML data could be retrieved from the following webpage:

https://www.kadaster.nl/zakelijk/registraties/basisregistraties/bag/bag-producten

In June 2021, a BAG GeoPackage could be retrieved at the following webpage:

https://geoparaat.baasgeo.com/bag/

The BAG snapshot of the 1st of January 2021 is used within the thesis project.

4.1.2 Discrepancy between BAG functions and actual activities

Usage functions used within BAG do not necessarily represent the actual activity at the place. Legally, a permit to carry out other activities than stated by BAG can be handed out, without changing the function of the object in BAG (Kadaster, 2021b). However, since 2018, changes in use of a space are registered in BAG, even when a permit is not needed. The question is to what extent BAG functions comply with actual activities at BAG objects. It would be a challenge to estimate to what extent this problem could affect results. It is not possible to exactly determine what activities are actual, and what activities are only permitted. The expectation however is that a significant majority of BAG functions do represent actual activities.

4.1.3 BAG data structure and attributes

From the BAG snapshot database, three data sets are used within the research; BAG_VBO, BAG_Building and BAG_Address. The key attribute of BAG used within the research is BAG VBO.

VBO can be translated as ‘accommodation object’. In this report, an accommodation object will be referred to as VBO. A VBO is defined as follows by law:

Smallest unit of use located within one or more buildings and suitable for residential, commercial, or recreational purposes that is accessed via its own lockable entrance from the public road, a yard, or a shared traffic area, and can be the subject of property law legal transactions, and is functionally independent.

In Table 4.1 the attributes present in VBO that are relevant for the research are presented. In Appendix A, all attributes from BAG_VBO, BAG_Building and BAG_Addresses used within the research can be viewed.

Table 4.1: BAG_VBO Attributes

BAG_VBO		
Attribute	**Type**	**Description**
BAG VBO	Point	Smallest unit of use located within one or more buildings and suitable for residential, commercial, or recreational purposes that is accessed via its own lockable entrance from the public road, a yard, or a shared traffic area, and can be the subject of property law legal transactions, and is functionally independent.
Identification	Character	Unique identifier of the VBO
Address identification	Character	Unique identifier of the address of the VBO
Building identification	Character	Unique identifier of the building in which the VBO is present
Surface area	Integer	Surface area of the VBO
Status code	Integer	VBO status: 0. Unrealized 1. Shaped 2. Out-of-use 3. In use 4. In use (unmeasured) 5. Withdrawn
Meeting	Logical: TRUE / FALSE	VBO for meeting people for art, culture, religion, communication, childcare, catering on the spot and watching sports
Jail	Logical: TRUE / FALSE	VBO for a coercive stay of people
Healthcare	Logical: TRUE / FALSE	VBO for medical examination, nursing, care or treatment
Industry	Logical: TRUE / FALSE	VBO for the commercial processing or storage of materials and goods, or for agricultural purposes
Office	Logical: TRUE / FALSE	VBO for administration
Lodgings	Logical: TRUE / FALSE	VBO for providing recreational or temporary accommodation to people
Education	Logical: TRUE / FALSE	VBO for teaching
Other	Logical: TRUE / FALSE	VBO for functions other than mentioned in which the stay of people plays a subordinate role
Sport	Logical: TRUE / FALSE	VBO for practicing sports
Shop	Logical: TRUE / FALSE	VBO for trading materials, goods or services
Living	Logical: TRUE / FALSE	VBO for living

4.1.4 BAG data operations

Selecting useful BAG data

Before BAG is suitable to be used as a source for trip generation data, all abundant BAG data should be eliminated. Besides VBOs and buildings that are currently in use, the database contains records of buildings and VBOs that have yet to be built, for which licenses are requested but not permitted, or that will be demolished. And furthermore, some records are incomplete or contain illogical values for attributes. It could be that registration of the BAG entities was erroneous, or that information was simply not gathered. In either case, these records cannot be included in further analyses as they might contribute to uncertain outcomes, if not removed. Ideally, BAG VBOs and buildings included in the analysis should:

- Be currently in use
- Contain (logical) surface area values
- Be used for activities to which travel activities occur

To achieve these goals, the following operations have been carried out. VBO objects with the function *other* have been removed from the BAG data. In these spaces, the stay of people has no primary role, therefore these objects are irrelevant for trip generation. To only include VBO objects currently in use all records with a VBO status code other than 3 or 4 have been excluded. And finally, VBOs *living* with a surface area below 14 or above 2.700 are removed. By law, it is not permitted to develop properties with a VBO *living* smaller than 14 m^2. VBO *living* with a surface area smaller than 14 m^2 have not been registered properly. Removing VBO *living* larger than 2.700 m^2 is for the same reason. A square footage of above 2.700 is assumed not to be plausible. The same values are used by the CBS for publishing statistics for residential properties in the Netherlands.

In Appendix B, the exact attributes and attribute values that are used to carry out the above-mentioned tasks, and other data operations are listed.

Joining BAG_Building and BAG_Address to BAG_VBO

The key data element of BAG used within the research is BAG_VBO. Therefore, BAG_Building and BAG_Address data are joined to BAG_VBO data, based on the *Building Identification* and *Address identification* attributes present in BAG_VBO and the *Identification* attributes in BAG_Building and BAG_Address. By doing so, each VBO is accompanied by the PC6 zip code of its location, and it is possible to estimate in what residence type a VBO *Living* is present. In Figure 4.1 the geometry data of BAG_VBO and BAG_Building is displayed. As can be seen, some buildings contain one VBO, whereas others contain multiple. Dependent on what activity type is present, how the building is constructed, and how the VBOs have been registered, one or multiple VBOs are present in one building. A building with an office, a shop and several residences on top will have separate VBOs for each of them. Each residence is separately registered. An office can have one or multiple VBOs, dependent on how the space is divided. The same holds for shops or other functions. There are supermarkets registered as one VBO, and supermarkets in which the shopping space and the warehouse are separately registered

Figure 4.1: BAG VBOs and Buildings

VBOs with multiple functions

The major advantage of using BAG as a database for trip generation is the presence of VBO functions and surface areas. VBO functions already reveal important information about destinations. Coupled with its surface area, this can act as a great basis for trip generation estimation. Most VBOs in BAG contain one function (97,6%). However, 2,3 % of the objects contain two functions or more, in total 220.000 VBOs. The VBOs with double functions need to be analyzed further, since 220.000 objects are not trivial. A possible solution is to divide the surface area of each VBO equally by the number of functions. As this is a simplistic solution, it might be worth analyzing combinations of VBO functions.

In Table 4.2, combinations of different VBO functions are listed. For each prevalent combination (more than 10.000), spatial analysis in QGIS has been carried out in order to determine what specific destination types contain multiple VBO functions, and where these VBOs occur. By plotting the specific VBO combinations on a map, it might become clear if specific activities can be linked to certain VBO combinations. For example, it has been found that almost all VBOs with both functions *living* and *industry* are farms. Based on this analysis, a definitive function can be given to these VBOs.

Table 4.2: Multiple functions BAG

VBO function combination	Number of occurrences	Findings based on spatial analysis	Definitive function and division of surface area
VBO == …. AND VBO == Living (170.000)			
Healthcare	40.000	Mostly nursing homes and assisted living. Job location for healthcare employees. For production of trips, maybe considered homes, but it is likely the trip generation for these specific places is far less. There are cases of hospitals including both functions.	½ Living ½ Healthcare
Other	10.000	Mostly homes	Living
Lodging	10.000	Mostly homes, often no sign of a possibility to rent a room	Living
Shops	20.000	Often seem to be homes close to centers, which have received the VBO function shopping as well, likely to facilitate short term shopping locations near centers.	Living
Meeting	10.000	Mostly homes	Living
Industry	65.000	Almost all cases are farms. Sometimes homes at industrial areas. Quite useful for agricultural employment data.	Living with a square footage of 100m^2 Agriculture
Education	6.000	Often small buildings near schools. Almost always residences.	Living
Office	12.000	Small offices at home..	Living
Sports	256	Negligible	Delete
Jail	45	Negligible	Delete
Other VBO combinations			
Industry AND office	19.000	Mostly companies at industrial areas that also include office spaces. Industrial surface area assumed to be the largest.	¾ Industry ¼ Office

4.1.5 Summary statistics BAG data

The BAG_VBO snapshot from the 1st of January 2021 contains a total of 9.849.513 records. After selecting useful BAG data (Section 0), 8.574,744 records remain. Of these records, 7.877.086 have a living function, with an average surface area of 119 m^2.

In Figure 4.2, four histograms are displayed which the occurrence of VBOs with corresponding surface areas for different residence types. The difference in the volume of detached and semi-detached houses, and the amount of terraced and multi-family houses in the Netherlands are evident in the graphs. The differences of average surface area per housing type are as expected, with low averages for multifamily houses, and high averages for detached houses. Contrary to what the graphs suggest, there are also houses above 350 m^2. For the clarity of the graphs, the houses with a surface area above 350m^2 are not presented.

In Table 4.3, summary statistics of all other VBOs are presented. VBO *industry* is the most prominent VBO, with both a high total surface area as occurrences. VBO *lodging* has relatively many records with a low average surface area.

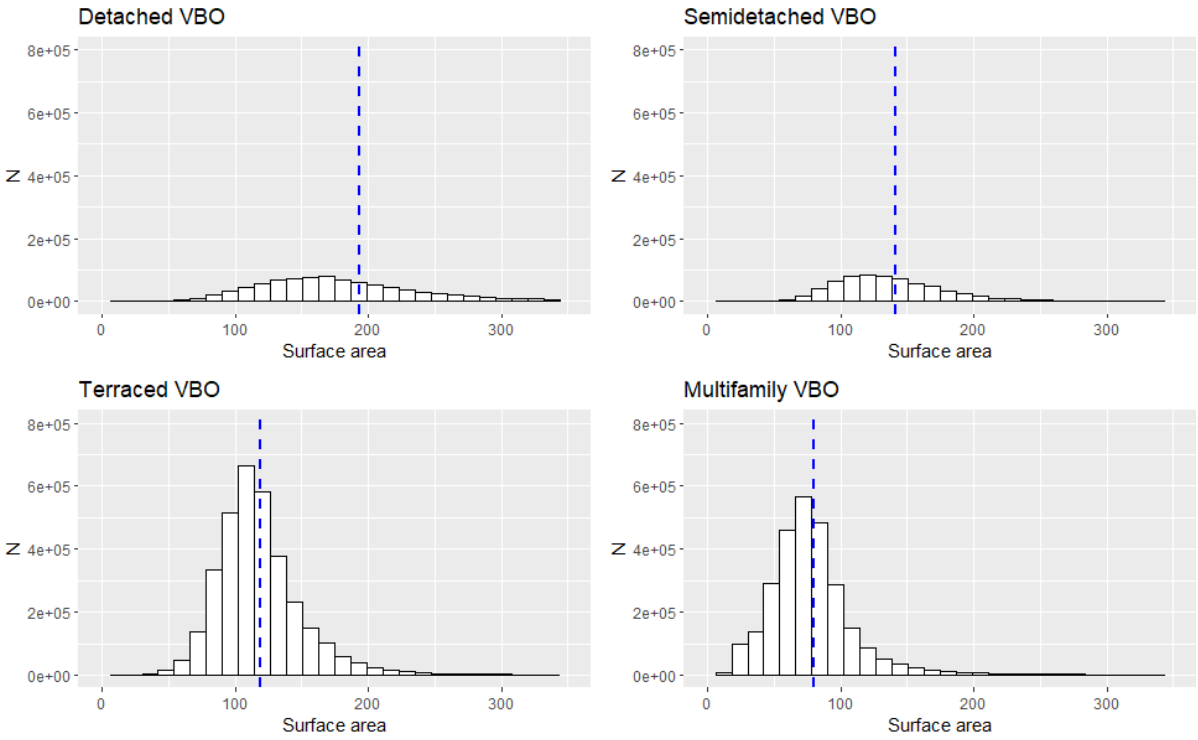


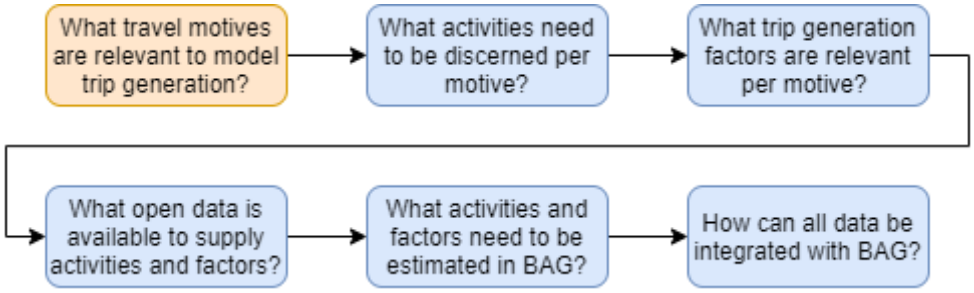
Figure 4.2: Four VBO types surface area statistics

Table 4.3: Properties of VBO objects

VBO	Records	HA total	Average surface area (m^2)
Meeting	60830	2961	487
Industry	198476	21155	1066
Healthcare	22535	1699	754
Lodging	126336	1420	112
Education	13313	3087	2319
Office	92462	5931	641
Shop	123830	4568	369
Sport	9450	980	1037

4.2 Motives

In this paragraph, a description is given of the brief, simple process to determine what motives need to be discerned in a transportation model.



To determine what is a logical structure for a traffic model, two types of sources were consulted. First, the use of motives in current practices has been reviewed, by consulting the technical reports of transportation models of seven different-sized cities in the Netherlands. The most important motives should be included in these studies. However, current practices are not necessarily best practices. Therefore, data from Dutch national travel surveys of the past four years has been examined. For an average working day, the average number of trips for a total of thirteen motives are calculated. Based on information from both the studies and the surveys it is determined what motives are relevant to include in a transportation model.

In Table 4.4, the motives, trip generation factors and modalities of seven transportation models of cities in the Netherlands are displayed. The trip generation factors will be referred to in Section 4.3.

Table 4.4: Motives and trip generation factors transportation models Dutch cities

Study area	Model year	Modalities	Motives	Trip generation factors	Source
Ede-Wageningen	2018	Passenger car & freight	Work Business Shops Other College	Labor force Jobs Residents Jobs retail Student places college Student places university Degree of urbanization Car ownership	(Royal Haskoning, 2018)
Harderwijk	2009	Passenger car	Work Business Shops Other	Inhabitants Jobs Jobs retail	(Oranjewoud, 2009)
Hilversum	2015	Passenger car, freight, public transport, bicycle	Work Business Education Shopping Other	Degree of urbanization Residents jobs retail jobs other car ownership labor force student places	(Royal Haskoning, 2015)
Utrecht	2018	Passenger car, freight, public transport, bicycle	Work Business Shops Education Other	Residents, Number of households Jobs retail Jobs other (industry, distribution, mixed, retail, entertainment, services, offices, high-end), Degree of urbanization Parking capacity	(Goudappel Coffeng, 2018b)
Woerden	2020	Passenger car, freight, bicycle, public transport	Work Business2Business Business2home Shops Education Other	Jobs Labor force Jobs industry Jobs other Residents Jobs retail Student places (VO/MBO/HBO/WO) Residents aged 0 – 34 Degree of urbanization Households Car ownership	(Royal Haskoning, 2020)
Twente	2011	Passenger car, freight, bicycle, public transport	Work Business Shops Education Other	Jobs Labor force Residents Jobs retail Residents aged 0-34 Student places Households Car ownership	(Goudappel Coffeng, 2011)
Rotterdam Den Haag	2018		Work Business2Business Business2home Shops Education Other	Jobs Labor force Residents aged 0-34 Student places Residents Jobs retail Car ownership	(Goudappel Coffeng, 2018a)

The use of motives in the transportation model reports are almost the same with only some slight differences present. The modalities considered in the studies minimally affect the use of motives. Harderwijk did not include educational trips. Ede-Wageningen only included college trips. All other studies do consider educational trips. Furthermore, some studies considered one business motive, whereas other divided it into business to home and business to business.

Based on the use of motives in existing transportation studies, the following motives seem reasonable to include when considering trip generation:

- Work
- Business (to home / business)
- Education
- Shopping
- Other

In Figure 4.3, the average number of trips on an average working day in the Netherlands are presented, based on travel survey data from 2016 till 2019. The presented rates include all travel modes. The motives work, education and shopping are, as expected, significant motives and responsible for more than half of the trips made per day. Surprising is the low number of professional visits and business trips, considering the presence of the business motive in all transportation models (Table 4.4). The motives sports / hobby, other leisure and visiting / staying over facilitate almost a quarter of daily trips. These trips can all be assigned to specific activity locations: sport accommodations such as the gym or a football club, recreational activities such as dining and terraces, and residences for visiting trips. Nonetheless, these trips are all combined into one motive in all municipal transportation models, together with the remaining motives bringing / picking up persons, services / personal care and other. It can be argued that when data is present of activity locations of motives and more motives are considered in transportation models, the quality of trip generation estimates will increase.

Based on the use of motives in current transportation models, the results from Dutch national travel survey data, and the scope of the research, the following motives are considered for researching the potential of BAG for trip generation estimates:

Motive	Trip rate	Percentage
Work	0.66	22%
Professional visits	0.06	2%
Business	0.04	1%
Bringing /picking up persons	0.23	8%
Bringing / picking up goods	0.05	2%
Education	0.43	14%
Shopping	0.52	18%
Visiting / staying over	0.22	8%
Walking / touring	0.15	5%
Sports / hobby	0.27	9%
Other leisure	0.20	7%
Services / personal care	0.12	4%
Other	0.03	1%
Total	2.97	100%

Figure 4.3: Trip rates based on ODIN and OVIN (2016-2019) on an average working day

- Work
- Business (to home / business)
- Education
- Shopping
- Sports/hobby
- Other leisure
- Services / personal care
- Visiting / staying over
- Bringing / picking up persons
- Other (Walking / touring, Other)

The motives, from which the activities and trip generation factors will be further analyzed, have preliminary been connected to VBOs, see Figure 4.4. The motives are all assumed to be home-based trips with the activity-end of the trip presented (except for professional visits / business). For each motive not every specific activity has yet been considered, but the diagram reveals a first indication of the most likely connections between trip activities and VBOs in BAG. As can be seen, trip motives and VBO functions are broadly consistent. For work trips BAG even enables distinguishing between sectors. And other motives such as shopping and education can be linked one on one to specific VBO functions with a reasonable certainty that all activities within the motives are also located within the connected VBOs. For other motives, such as services / personal care, the question is to what extent the activities within the motive are related to VBOs *healthcare*. For example, it might be that a gas station is an activity that should be distinguished as part of the motive services / personal care, which is included as VBO *shop* in BAG. But for now, the major links are visualized.

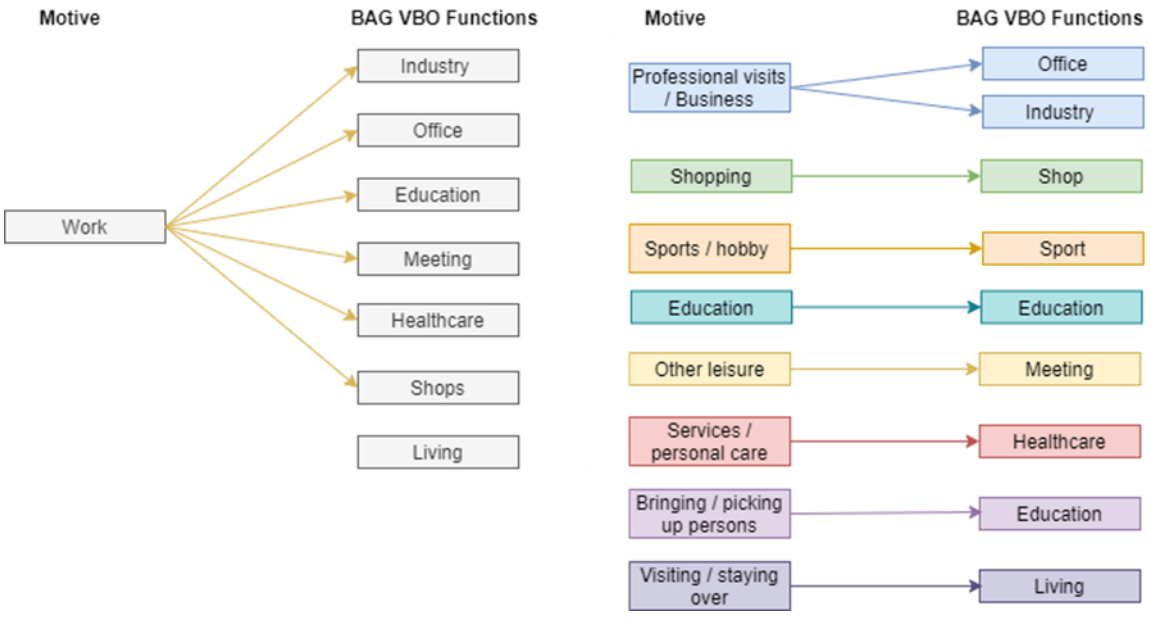
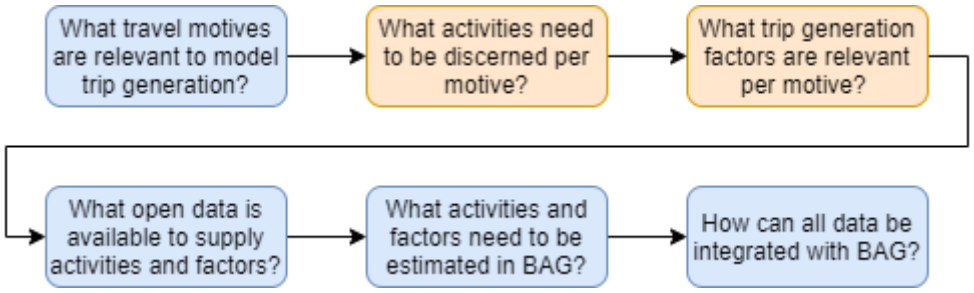


Figure 4.4: Most significant links between trip motives and BAG VBO functions

4.3 Activities and factors

Now that it has been determined what motives are reasonable to distinct, the type of activities per motive and its specific trip generation characteristics at the activity-end are examined. It is also useful to review the use of explanatory variables, or trip generation factors, to predict trip generation per motive or activity. The trip generation factors by which the trip generation of an activity is best determined, dictates what data is required.



For every motive, the use of trip generation factors to predict trip generation at the activity side is examined from four sources:

- Dutch municipal transportation models (Table 4.4)
- CROW publication 381 (CROW, 2018)
- ITE common trip generation rates (ITE, 2021)
- Comparisons of NZ and UK trip and Parking Rates (Milne & Abley, 2009)

Within person trip rate studies and modelling practices, trip generation models are mostly structured by travel purpose or motives. Per motive trip rates are identified (often based on travel surveys) that explain how many trips a person in a model is going to make with the stated motive, in a given time. For example, a trip rate of 0.1 with the motive 'healthcare' could mean that a person with certain characteristics will make 0.1 (out of 1) trips per day to a healthcare destination. A healthcare destination could be a general practitioner (GP), a pharmacy or a dentist. It would increase the quality of a trip generation model if present land use data could distinguish these destinations. A hospital will attract a different number of visitors compared to a GP and this will affect transport model outcomes. However, there are hundreds of possible destinations thinkable, and (open) available data of trip generation numbers of each destination, and the actual location of these destinations is not present. Therefore, destinations should be listed which would significantly improve trip model outcomes if distinguished in present land use data. Per motive, destinations are subject to the following questions:

- What destinations deviate in terms of trip generation?
- What destinations are prevalent?
- What destinations have similar trip characteristics?
- What destinations are attended on an average working day?

A supermarket attracts many visitors and is prevalent within cities. The impact of this activity is evident, and the distinction of the land use attribute grocery store will contribute to the trip model outcomes significantly. Including for example a tourist office within the land use attributes might be more of a discussion. Or a stadium, which at times can attract many visitors, but not consistently on an average working day. And furthermore, not in every city a stadium is present. The question might be asked what the added value of distinguishing stadiums in BAG will be, if easily added manually. Considering the similarity of trip characteristics of destinations, if multiple destinations within a motive attract comparable numbers of trips, distinguishing each specific destination would be redundant.

To estimate what destinations will be included, for each motive(group) destinations are listed. These destinations are retrieved from CROW publication 381 (CROW, 2018). The CROW is a Dutch knowledge institute for transport, infrastructure and mobility. Publication 381 of the CROW contains trip rates and parking rates for numerous activities. With these trip rates, direct comparison of the trip attraction characteristics of the destinations is possible. The large degree of uncertainty of the CROW rates, as well as the fact that the traffic generation numbers only apply to automobile traffic, should be kept in mind. The fact that these rates apply to automobile traffic is not a problem. The rates are only included to enable comparison between activities, and partly determine whether it is necessary to distinguish a specific activity. The results of this process can be applied to both unimodal and multimodal studies. As an indicator of the prevalence of the destinations, the number of each destination in the cities of Enschede (160.000 citizens) and Hengelo (80.000 citizens) have been included.

For each motive, the use of trip generation factors to predict trip generation at the activity side is listed and the activities provided by the CROW are analyzed based on trip rate, prevalence and representation on BAG. Then, it is decided for each motive what activities need to be distinguished:

Education

Use of trip generation factors:

Source	Explanatory variables used
Municipal transport models	Number of students
ITE	Number of students / GFA*
CROW	Number of students
UK and NZ	NA

GFA = Gross Floor Area

Activities as provided by the CROW:

Education	Traffic generation per day All visitors	Share of Students	Prevalence	BAG
Primary school	NA		Enschede: 57 Hengelo: 31	Education
Secondary school	12,8 (per 100 students)	5,3 (42%)	Enschede: 20 Hengelo: 16	Education
College	13,3 (per 100 students)	9,0 (68%)	Enschede: 1 Hengelo: 0	Education
University	23,6 (per 100 students)	10,3 (44%)	Enschede: 1 Hengelo: 0	Education
ROC	11,7 (per 100 students)	5,4 (46%)	Enschede: 6 Hengelo: 6	Education

The education motive is responsible for 14% of daily trips, therefore important. Above, the most relevant educational destinations are listed. Primary school destinations are unique in the sense that most visitors are brought and picked up by their caretakers. The traffic generation numbers of the CROW are for this motive mostly irrelevant, considering the large proportion of cyclists to these locations. For these locations, it can be assumed that the number of students is a highly explanatory variable. The difference in traffic generation between activities can be explained by the difference in car use amongst students, but also the amount of other staff present. The traffic generation per student, which is relevant for the motive education, can be estimated by multiplying the share of student car trips per education type by the total traffic generation rate. The share of students is also supplied by the CROW.

Education destinations that should be distinguished in a trip generation model:

- Primary school
- Secondary school, ROC
- University, College

Shopping

Use of trip generation factors:

Source	Explanatory variables used
Municipal transport models	Number of employees
ITE	GFA
CROW	GFA
UK and NZ	GFA / Number of employees

Activities provided by the CROW:

Shopping destinations	BAG	Traffic generation per day per 100 m² GFA	Prevalence
Supermarket • Neighborhood (<600m2) • Full-service supermarket (>1000m2) • XL supermarket (>2500m2)	Shops	 63,6 92,3 104,3	Enschede: 20 Hengelo: 18
The inner city (100.000-175.000)	Shops	29,1	-
District / village shopping center • Small • Average • Large	Shops	 47,7 49,3 49,9	Enschede: 6 Hengelo: 3
Shopping boulevards	Shops	20,8	Enschede: 0 Hengelo: 0
Home decor boulevards	Shops	7,9	Enschede: 1 Hengelo: 2
Outlet centers	Shops	23,1	Enschede: 0 Hengelo: 0
Construction stores	Shops	27,4	Enschede: 5 Hengelo: 6
Garden centers	Shops	13,7	Enschede: 2 Hengelo: 3
Furniture stores / home decor (outside home decor boulevards)	Shops	16,7	Enschede:4 Hengelo: 3
White good shops	Shops	84,7	Enschede: 1 Hengelo: 2
Wholesale	Shops	33,3	Enschede: 11 Hengelo: 11
Recycle store	Shops	16,6	Enschede: 6 Hengelo: 2

Shopping is an important motive with 18% of daily trips. Above, the most important destinations with the motive Shopping are listed. Some activities, such as the boulevards and centers, contain clustered destinations. A transportation model is (often) not developed at a level that (walking) traffic between the shops at these activity groups is of interest. Therefore, a destination within one of these centers or boulevards could relatively attract fewer trips, as visitors may combine certain shops. Thus, inner cities, home decor boulevards and shopping centers need to be distinguished from the Shops buildings in BAG. Home decor boulevards specifically have a relatively large surface area and not distinguishing this destination may lead to overestimation of trips to these locations. Supermarkets have a large trip rate per square meter and are prevalent. Wholesale destinations are often in large-scale buildings, are prevalent, and compared to for example supermarkets or shopping centers, the trip rate is low. Not distinguishing wholesale destinations could lead to a vast overestimation of traffic surrounding wholesale destinations. All other destinations are in the same trip rate range or not prevalent (white goods shops).

Shopping destinations that should be distinguished in a trip generation model:

- Supermarkets
- City centers
- Village shopping centers / district shopping centers
- Home decor boulevards
- Wholesale
- Other

Sports / hobby

Use of trip generation factors:

Source	Explanatory variables used
Municipal transport models	NA
ITE	Dependent on the activity; number of acres, number of holes, number of fields, number of courts
CROW	Mostly GFA, other dependent on the features of the activity
UK and NZ	NA

Activities provided by the CROW:

Sports/hobby destinations	BAG	Traffic generation per day per 100 m^2 GFA	Prevalence
Gym (around 750 m^2)	Sport	30,8	Enschede:15 Hengelo: 10
Fitness centre (often largen than 1500 m^2)	Sport	31,6	Enschede: 5 Hengelo: 4
Sporting hall	Sport	12	Enschede: 15 Hengelo: 6
Swimming paradise (covered)	Sport	10	Enschede: 1 Hengelo: 0
Ice skating hall	Sport	4,8	Enschede: 1 Hengelo: 0
Sporting fields	NA	NA	NA
Dance studio	Sport	18,7	Enschede: 6 Hengelo: 2

Sports and hobby destinations are responsible for 9% of daily trips. Above, the most important locations are listed. The trip rate of the gym and fitness center significantly stand out. Furthermore, gyms are prevalent in the cities of Enschede and Hengelo. The other destinations have trip rates in the same bandwidth, or are not that prevalent.

Sports / hobby activities that should be distinguished in a trip generation model:

- Gym / fitness center
- Other

Other leisure

Use of trip generation factors:

Source	Explanatory variables used
Municipal transport models	NA
ITE	Mostly GFA, others dependent on the features of the activity; such as number of seats, number of pools
CROW	Mostly GFA, others dependent on the features of the activity
UK and NZ	NA

Activities provided by the CROW:

Other leisure	BAG	Traffic generation per day per 100 m^2 GFA	Prevalence
Library	Meeting	8,2	Enschede: 4 Hengelo: 2
Theatre	Meeting	11,1	Enschede: 5 Hengelo: 2
Cinema	Meeting	16,4	Enschede: 2 Hengelo: 1
Museum	Meeting	-	Enschede: 2 Hengelo: 2
Casino	Meeting	14,8	
Bowling	Meeting	-	
Snooker	Meeting	-	
Children play hall	Meeting	3,7	
Restaurant	Meeting	-	Enschede: Dozens Hengelo: Dozens
Cafe	Meeting	-	Enschede: Dozens Hengelo: Dozens
Hotel	Lodging	-	Enschede: 9 Hengelo: 2

The other leisure motive is responsible for no more than 7% of daily trips. The activities as provided by the CROW are not all accompanied by trip generation numbers. It can be seen that certain activities, restaurants and cafes are prevalent compared to others. Being able to assign specific trip generation numbers to these activities could increase the quality of trip generation outcomes. However, there are no traffic generation rates available for these activities. And according to the CROW, traffic generation at these activities is very difficult to estimate. Therefore, it makes no sense to distinguish these activities in a trip generation model

Other leisure activities that should be distinguished in a trip generation model:

- Other leisure

Services / personal care

Services / Personal care	BAG	Traffic generation	Prevalence
Barber	Shop	NA	Enschede: Dozens Hengelo: Dozens
Beaty salon	Shop	NA	Enschede: multiple Hengelo: multiple
Medical practice	Healthcare	23,4 (per treatment room)	Enschede: 20 Hengelo: 20
Pharmacy	Healthcare	125,4 (per unit)	Enschede: 18 Hengelo: 16
Physio practice	Healthcare	13,9 (per treatment room)	Enschede: 25 Hengelo: 20
Dentist	Healthcare	29,6 (per treatment room)	Enschede: 18 Hengelo: 18
Hospital	Healthcare	7,0 (per 100m2)	Enschede: 1 Hengelo: 1

The services / personal care motive is responsible for only 4% of daily trips. Especially for services, it is difficult to find CROW activities that can be assigned to this motive. Other, logical destinations are the medical destinations that are listed above. The trip generation factors are varying per activity and comparison of trip rates is therefore not possible. As can be seen, healthcare is a BAG VBO function, so the medical activities can be distinguished in BAG. Often medical practices, pharmacies, physio practices and sometimes dentists are located close to each other. Even when trip rates are varying per activity, distinguishing these activities in BAG would not drastically improve trip generation estimates.

An exception to this are hospitals. Although not prevalent in cities, hospitals are mostly huge buildings. Wrongful trip estimates could drastically affect transport estimates surrounding hospitals. Examples of service activities that are not included in BAG, but that could be thought of are banks, garages, gasoline stations and car washes. Most of these activities are represented as shops in BAG, just like barbers and beauty salons. A possibility to deal with this is to assign part of the motive services / personal care to BAG VBOs *shops*. For this motive it could however be worth it to consider gas stations as an extra separate activity, considering Dutch citizens on average refuel more than once a month.

Services / personal care activities that should be distinguished in a trip generation model:

- Other
- Hospital
- Other medical activities
- Gas stations

Bringing / picking up persons

The CROW does not define trip rates for specific bringing and picking up activities. However, both at primary schools and childcare centers mostly bringing and picking up persons occur. The CROW has developed special calculation tools for municipalities to deal with bringing and picking up traffic at these activities, primarily intended to develop so-called kiss and ride (K+R) parking places. Other locations where K+R parking places can be found are stations.

Bringing / picking up person activities that should be distinguished in a trip generation model:

- Primary school (number of students)
- Childcare (number of pupils)
- Station (unit)

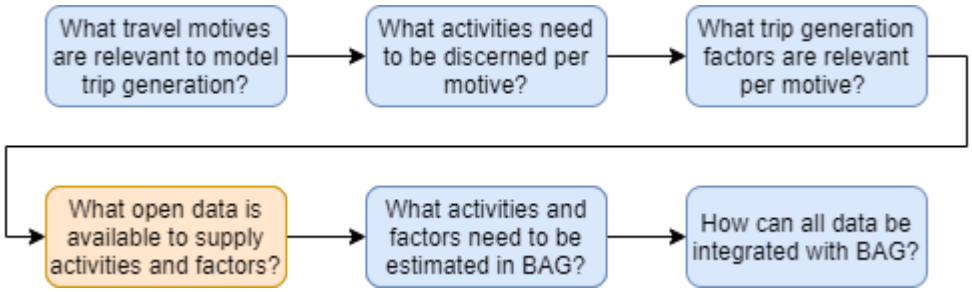
Work & Business

Source	Explanatory variables used
Municipal transport models	Number of employees
ITE	GFA
CROW	GFA
UK and NZ	Number of employees / GFA

The motives work and business are merged together because the activities are partially similar (Figure 4.4) and the trip generation factors are the same.

For the use of GFA as a trip generation factor BAG is not suitable; multiple types of businesses should be distinguished, whereas BAG only manages to name some general VBO functions with most important *office* and *industrial*. And furthermore, the work motive is about commuter trips to and from work. It is very likely that the factor employees for this is better able to estimate how many trips a work activity generates within this motive compared to GFA, this is also endorsed in the literature. The traffic generation figures of CROW are mainly used to predict traffic generation for new developments. The size of the building to be developed is often better known beforehand than the expected number of employees. This is probably the reason why CROW and ITE use GFA. Therefore, instead of trying to distinguish a number of specific BAG work destinations in order to link certain trip generation figures to them, it is better to estimate the number of employees per activity based on BAG functions.

4.4 Available open data



In this part, open datasets are described that provide data for trip generation factors, or data for activity locations that are needed to estimate trip generation at the municipal level. These datasets can have one of two purposes:

- Providing recent, complete, open and detailed data of either trip generation factors or activity locations that can be used to determine trip generation estimates at the municipal level
- Providing open data that is outdated, incomplete or highly aggregated that can be improved by using BAG

If open data is available that is complete and detailed enough to estimate trip generation for a certain activity, then it is not necessary to investigate the potential of BAG to facilitate the same data. It can be assumed in advance that these attempts make the outcome more uncertain than the original. But, if complete, detailed and recent data is not available, and an incomplete, outdated or too aggregated alternative exists, BAG estimations could improve the imperfections of the data.

Below, all open data is that was found to be available and could supply one of the above factors or activities is presented. For each data source it is briefly described what is included in the data and how it could be used as a data source for the activities listed in Section 4.3. Furthermore, the link between BAG and the data is mentioned. In Appendix C, the data attributes are described more extensively and it is mentioned how the data can be retrieved.

4.4.1 Open data sources and connection to motives and activities

DSA

The Database Sport Accommodations (DSA) is an open database in the Netherlands containing information of around 28.000 sport facilities. The use of the database for non-commercial and commercial purposes is permitted, provided that reference is made to the database. In addition, it is permitted to further develop, transform and build products based on the data, provided that these products are available under the same conditions.

The DSA provides for activities and trip generation factors for the motive Sport / Hobby.

All thinkable sport facilities are present in the data, including outdoor facilities such as football, tennis or even skiing. Of every facility additional information is present, such as the number of indoor or outdoor accommodations and the number of sporting halls. These are all useful factors to determine trip generation. The activities as identified in Section 4.3, gym / fitness center, are also included. The DSA includes much more activities than mentioned by the CROW. It could be worthwhile to investigate how the DSA could be best put to use in a trip generation model

Thus, all required activities for the motive Sport / Hobby can be found in the DSA. But, the GFA of activities is not present. Of each record in DSA, the exact address is available. Through the address, DSA can be linked to BAG and the GFA of an activity, for example the gym, is available.

Education Data

The Dutch educational institute DUO publishes annual open figures of educational institutions in the Netherlands. The following educational institutions are included here:

- Primary school
- Secondary school
- ROC
- College
- University

Of each education accommodation, among others, the following data is available:

- Address
- PC6 postal code
- The number of students
- The number of employees

The education data provided by DUO can supply all activities and trip generation factors for the motive education. The number of students is the best trip generation factor to determine the trip attraction of education activities for the motive education. No data from BAG is required.

The education data can supply for activity and trip generation factor data for the activity bringing/picking up persons: Primary school.

Childcare data

DUO publishes data being updated twice a week of locations of childcare accommodations in the Netherlands with detailed information. Of each accommodation, among others, the following data is available:

- Address
- PC6 postal code
- Number of childcare places

The childcare data can supply for activity and trip generation factor data for the activity bringing / picking up persons: Childcare. No data from BAG is required.

BRT

Another geographical base registration in the Netherlands containing open available data besides BAG is the base registration topography (BRT). The BRT contains topographical files that are made available at different scale levels. The BRT is a detailed and national covering geographical database describing the physical environment of the Netherlands. Among others, roads, railroads, heights, buildings and area functions are included. The prerequisites of objects included in the database are that the object is long-lasting and defining for the environment. Among others, the following buildings are included in the BRT data:

- Church
- Synagogue
- Mosque
- Abbey
- Other religious buildings
- Gas station
- Hospital

Of each building, among others, the following attributes are known:

- Raw outline of the building
- Function

BRT data can supply activity data for three motives.

For the motive other leisure, no specific activities need to be distinguished. Almost all leisure activities are registered as VBO *meeting* in BAG, while activities from other motives are not registered as *meeting*. Therefore, trips with the motive other leisure can be perfectly connected to VBO *meeting*. However, religious buildings are also registered in BAG as VBO *meeting.* These buildings, especially on a working day, hardly generate trips, while being among the largest buildings in a city. If VBO objects in BAG would be considered as other leisure activity, trips would be largely overestimated at these locations. BRT can be used to identify VBO objects in BAG as religious buildings, and restrict them from trip generation estimations.

For the motive bringing / picking up persons, BRT could supply data for the activities Hospital and Station.

For the motive services / personal care, BRT could supply data for the activities Hospital and Gas station.

BAG and BRT data do not share common attributes to enable an easy connection between the databases. However, both datasets are geographical. Based on a spatial overlay of the required data from BRT, the building functions from the BRT can be transferred to BAG. Then, to VBOs in the BAG Building, the function attribute can be added.

IBIS data

IBIS (Integral Business Area Information System) data is open geographical data containing information of business areas in the Netherlands. For each business area, among others the work location type, environmental classifications, and total surface areas are available.

For the motive shopping, IBIS data could supply data for the activity Wholesale. Wholesale shops are mostly located in industrial areas. By identifying VBO *shop* in industrial areas, wholesale shops are located in BAG.

OSM data

OSM is an open-source project aimed at developing freely available and editable maps. The input and modification of the data rely entirely on volunteers, which means that the reliability of the data is questionable. Among numerous other attributes, the OSM data contains the location of supermarkets.

For the motive shopping, OSM data could supply data for the activity Supermarket. Because of the unreliability of OSM data, it is first sought to identify supermarkets in BAG data without OSM. If this is not possible, the OSM data will be used. Then it will also be determined how to identify supermarkets in BAG based on OSM data.

NRM employment data

NRM data is the only dataset described in this section not containing any open data. The data is included because it is the only source of employment data that could contribute to the research. The NRM employment data that is used originates from 2016 and is used in the NRM transportation model, a model which most of the municipalities mentioned in Table 4.4 use to determine external transportation demand for the study area.

At PC4 level, the number of employees for the following sectors are available:

- Industry
- Retail
- Services

- Government
- Agriculture
- Other

NRM data could provide data for the motives work and business. It should be researched to what extent BAG is able to estimate employment numbers. The BAG VBO functions as presented seem promising to estimate employment rates for different employment sectors. Thus, further analysis is needed to estimate employment data.

CBS data

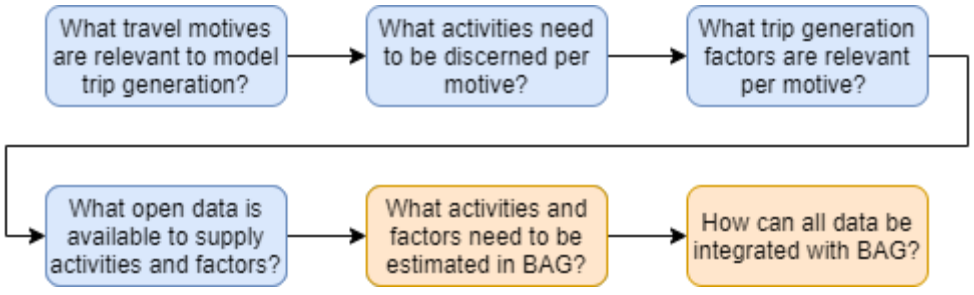
Most trip generation factors present in the municipal studies determine the number of trips at the home side (Table 4.4) are all retrieved from CBS data. Based on the municipal studies, the most relevant trip factors at the home-end are:

- Residents
- Car ownership
- Labor force
- Residents aged (0-34)
- Households

In the literature study (Chapter 3), the limited open availability, the differences in aggregation levels and the privacy limitations of CBS data have already been described. In Appendix D, the specific attributes and datasets of CBS data used in this research are presented. Car ownership is only openly available at District level, which is too aggregated for modelling at municipal level. The number of households, residents and residents aged 0-34 and 25-65 (approximation of labor force) are available at either District level or outdated and privacy restricted PC6 level (2016). An example of the privacy limitations of PC6 data is the following:

In total, 455.618 PC6 areas are present in PC6 2016 data. In 232.895 areas, the number of residents aged 0-14 is not available, because less than 5 residents are present. In 284.612 areas, the number of residents aged 14-24 is not available. Determining the number of residents aged 0 to 34 would be problematic. Thus further analysis is needed to determine how BAG could improve the availability of above-mentioned trip generation factors. This will be further explored in Chapter 7. For now, the focus will remain on activities and trip generation factors at the activity-end of trips.

4.5 Activities and factors to be estimated in BAG



In the previous sections, it has been described what open data is available, and what activities and trip generation factors are relevant to model trip generation. In Paragraph 4.1, the attributes of BAG data are described, which is the most important data source within this research. In Paragraph 4.2, trip generation motives are defined for which in Paragraph 4.3 it is determined what activities per motive are necessary to distinguish, and what trip generation factors can be used to calculate the trip generation per activity.

In Paragraph 4.4, available open data sets are mentioned in which attributes are present that can either provide for activity location data, or trip generation factor data. It is also described how these data sources could provide for activities and trip generation factor data. In this paragraph, an overview of what data is available, what motives are covered, and what further analysis is needed to be able to estimate trip generation based on BAG is made. This leads to concrete steps to be taken to estimate the potential of BAG to model trip generation, which is the goal of the research.

In Table 4.5, for each motive and activity (column 1), the data that is available to identify the location of the activity (column 2), and the trip generation factor required to estimate trip generation for the activity (column 3) are presented. In column 4 it is concluded whether the available data sufficient, how the data can be integrated with BAG, and whether additional analysis is needed. It is a comprehensive overview of data needed and present for trip generation estimates, that makes clear which data is not yet available. From this, it can be concluded in what way BAG can be used to improve the estimation of trip generation with open data.

Table 4.5: The link between motives, data and the use of BAG

Motive, activities and trip generation factor	**Activity location data**	**Trip generation factor data**	**How to use BAG?**
Work / business Activities: No specific activities defined based on CROW activities. Trip generation factor: Employment numbers	**BAG VBO**: -Industrial -Office -Shop -Meeting -Healthcare -Education **Education data:** -Address -PC6	**NRM**: Employment data at PC4 level -Industry -Retail -Services -Government -Other **Education data:** Employment numbers -Primary school -Secondary school -ROC -College -University	The number of employees per educational institute and the location can be retrieved from education data. Other employment data is available from the NRM. NRM data is however not open and too aggregated (PC4). It should be researched to what extent BAG is able to estimate employment numbers. The BAG VBO functions as presented seem promising to estimate employment rates for different employment sectors. Thus, further analysis is needed to estimate employment data.
Education Activities: -Primary school -Secondary school -University, College, ROC Trip generation factor: Number of students	**Education data:** -Address -PC6	**Education data**: Number of students -Primary school -Secondary school -ROC -College -University	The education data provides detailed enough information for both activity locations and student number data. It is even possible to distinguish more education institutes than necessarily needed. There is no need to include BAG data for this motive.
Shopping Activities: -Supermarkets -City centers -Village/district shopping centers -Home decor boulevards -Wholesale -Other Trip generation factor: GFA / employment	**BAG VBO:** -Shop **IBIS** -Industrial areas **(OSM**: -Supermarkets)	**BAG VBO**: -Surface area	The BAG VBO *surface area* is equal to the gross floor area of an activity, if all VBOs representing the activity are identified and summarized. It could therefore perfectly be used to estimate trip generation. The VBO *shop* attribute is not able to distinguish the necessary activities. Further analysis is needed to enable trip generation estimates for shopping activities in BAG. Additional data could also enable identifying activities (IBIS, wholesale activities). A possible alternative for activity data is OSM. If some activities are not

			possible to distinguish with BAG, OSM data can be used. However, there are no matching attributes present in BAG and OSM. A spatial join operation would be required to combine OSM and BAG data.
Sport/Hobby Activities: -Gym / fitness center -Other Trip generation factor: Dependent on the activity	**BAG VBO**: -Sport **DSA** -Address -PC6	**BAG VBO:** -Surface area **DSA** -Sport type -Number of indoor accommodations -Number of outdoor accommodations -And more	The DSA data is a complete and useful data source to estimate trip generation for sport trips. Both the location of each sport activity as several trip generation factors are present. The only useful trip generation factor not present in DSA data is GFA. This attribute can be retrieved from BAG data by matching the address of the sport activity from DSA with the corresponding BAG VBO address. The DSA data enables distinguishing even more sporting activities than strictly necessary.
Other leisure Activities: -None specific	**BAG VBO:** -Meeting **BRT**: -Church -Mosque -Synagogue -Other religious building	**BAG VBO:** -Surface area	The VBO *meeting* attribute will be the main data source to locate and calculate trip generation for other leisure activities. All CROW leisure activities are present as VBO *meeting* in BAG. In BAG, religious buildings are registered as VBOs *meeting.* These are large, prevalent buildings that do almost not generate trips on average working days. Based on BRT data, VBOs can be selected that have a religious purpose, to avoid surreal large trip estimates surrounding these activities.
Services / personal care Activities: -Hospital -Other medical activities -Gas stations -Other Trip generation factor: GFA	**BAG VBO:** -Healthcare -Shops **BRT:** -Hospital -Gas station	**BAG VBO:** -Surface area (excl. gas station)	With both BAG and BRT data combined, all relevant activities can be found. BRT includes attributes that enable to distinguish hospitals from VBO *healthcare* objects. And BRT includes gas stations which are unrecognizable in BAG, due to the not-walled characteristics of gas stations. No additional BAG analysis is needed to specify more activities. The BAG VBO *surface area* suffice for each location, except for gas station to which it cannot be applied.
Visiting / staying over Trip generation factor: Households / residents	**BAG VBO**: -Living	**CBS Districts:** -Residents	Residents data at district level as provided by the CBS is too highly aggregated. Further analysis with BAG is needed to estimate the number of residents at an acceptable aggregation level.
Bringing / picking up persons Activities: -Childcare -Primary school -Station -Hospital	**Childcare data:** -Address -PC6 **Education data: Primary** -Address -PC6	**Childcare data:** -Childcare places **Education data: Primary** -Student places	For this motive, BAG data only required to estimate the GFA of the hospital. For all other activities and trip generation factors, open data is available.

Trip generation factor: Childcare places, number of students, GFA	**BRT:** -Hospital -Station		

Concluding

Based on the synthesis in Table 4.5, the following can be concluded. For work and business trips, analysis with BAG is required to estimate employment numbers in BAG at an acceptable aggregation level. For shopping trips, BAG VBO *surface area* is a useful trip generation factor to estimate shopping trips. However, no activity data is openly available to distinguish shopping activities. Additional analysis in BAG is needed. If necessary, OSM data could be used as a backup. For all other motives, existing open data, or currently present attributes in BAG provide required activity and trip factor data to estimate trip generation.

Chapter 5. Methodology

In this chapter the research methodology to answer research question 2 and 3 is laid out. In the former chapter, the basis has been laid to develop a method for research question 2. It is clear what activities should be distinguished in BAG, namely, shopping activities, part of the most important motive beside work. Furthermore, a method needs to be develop that increases the availability of trip generation factors at the home end by using BAG. And finally, a case study is developed to determine how BAG improves trip generation estimates compared to conventional open data sources, and how this impacts surrounding traffic intensities, answering research question 3.

.

5.1 Activity analysis

To answer a part of RQ2, methods have been developed to distinguish the shopping activities estimated in Chapter 4 in from VBO *shop* in BAG.

- Supermarkets
- City centers
- Village shopping centers / district shopping centers
- Home decor boulevards
- Wholesale
- Other

During the conduct of the research, the exact methods for distinguishing these activities in BAG are developed. Here, the main elements of the analysis are generally described.

To identify the shopping activities in BAG, a combination of surface area-, location- and cluster analysis is used to identify specific destinations. City centers, shopping centers and home decor boulevards are clustered activities. By analyzing the clustering of VBO *shop*, together with the clustering size, shape and surface areas, it is sought to develop rules that identify what VBOs belong to what shopping activity. Then, wholesale activities are identified by selecting VBO *shop* present in industrial areas with IBIS data (Appendix C). When these two activities are identified, it is sought to identify supermarkets based on its surface area. The hypothesis is that supermarkets could stand out from other VBO *shop* by their large surface areas. If this is not successful, it should be determined how supermarket data from OSM could be linked with BAG. In Figure 5.1 a schematic overview of the described analyses is given.

5.2 Trip factor analysis

To answer a part of RQ2 it is sought to predict trip generation factors relevant at the household side with BAG attributes. With regression analysis, it is estimated how well BAG is capable of predicting trip generation factors based on two BAG attributes:

- VBO *surface area*
- VBO *residence type:* Detached, semi-detached, terraced and multi-family

Dependent on the aggregation level of the trip generation factors determined by the CBS data, regression will be carried out at District or PC6 level. PC6 data enables constructing separate prediction models for each residence type. A sufficient number of PC6 areas can be found in which one residence type is uniformly present. In Figure 5.2, an example of a uniform residence type is presented. In Chapter 7, the exact trip generation factors that are analyzed will be described. For each factor, regression models are estimated and results are validated by comparing prediction outcomes for different cities.

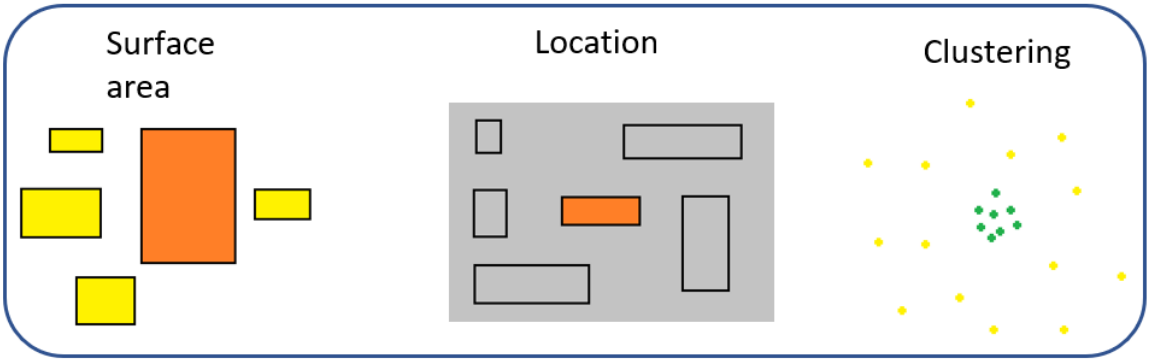


Figure 5.1: Identifying specific activities in BAG

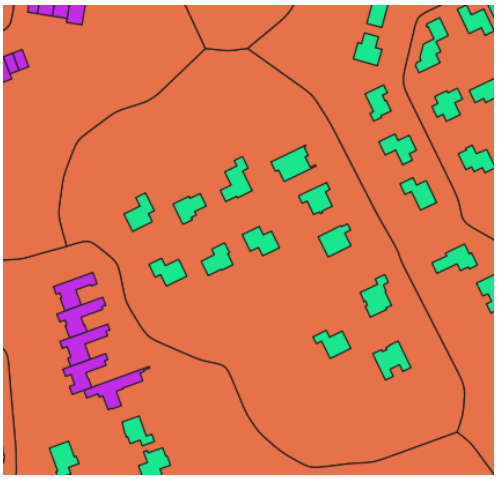


Figure 5.2: Uniform housing category in a PC6 area

5.3 Case study

To answer RQ3 – How does BAG improve trip generation outcomes based on open data and how does this impact transport modelling outcomes?– a case study is developed for the town of Ede. In a district in Ede, a new residential area is being developed. Trip generation estimates for the new developments are largely underestimated by the municipality of Ede. To display the potential of BAG for trip generation modelling, trip generation estimations at the household side are calculated based on four data sources:

- Conventional open data
- Data accessible by the municipality of Ede
- Current BAG data
- Future BAG data

To estimate the impact of using BAG data, traffic intensities generated by the district on surrounding roads are determined and compared with the traffic intensities of other data sources.

Chapter 6. Shopping activities in BAG

In Chapter 4 it has been decided what activities need to be distinguished in BAG, to enable precise trip generation estimates. In this chapter, through spatial analysis, operations and by combining different open data sources, it is sought to find ways of distinguishing specific activities in BAG. The following activities need to be distinguished for the motive Shopping:

- Supermarkets
- City centers
- Village - / district shopping centers
- Home decor boulevards
- Wholesale

To distinguish specific activities in BAG, two types of operations have been used:

- Conducting analyses on attributes already present in BAG to generate additional information, based on a combination of attributes
- Combining BAG data with other open data to provide for attributes not present in BAG

To identify specific activities for the motive shopping, a certain order of analyses will be used. First the clustered shopping activities are analyzed; the shopping centers, city centers and home decor boulevards. The clustering of VBOs *shop*, the size and shape of the clusters, and the surface area of the VBOs present in the cluster should provide information to decide what type of shopping activity is present. Then, wholesale activities are located by finding VBOs shop in industrial environments. Finally the remaining VBOs *shop* are analyzed to determine whether it is possible to locate supermarkets based on BAG attributes. In short, a couple of rules are designed based on which, with a certain uncertainty, it can be predicted to what specific shop activity a VBO belongs. These rules are developed based on shops present in the city of Enschede. Then, the rules are validated by applying them to VBOs *shop* in the municipalities of Hengelo and Veenendaal. For the research, the clustering analysis is carried out in a GIS environment. Because the rules are developed during this research, the visualization of the data is necessary to first manually insert activity locations, and then to visually check to what extent the designed rules are able to predict the activities. However, all data operations carried out in GIS can also be carried out in R (open source programming language). This means that once the rules have been developed and tested, all operations can directly be carried out in R. This increases the usability of the developed methods because all data operations can be automatized. This is ultimately a logical condition for the development of a trip generation database with open data.

6.1 Analysis of clustered activities

In the city of Enschede, 1340 VBO *shop* are present covering a total of 51,4 ha with an average surface area of 384 m^2.

6.1.1 Developing rules

DBSCAN

The main tool of analysis for distinguishing the clustered activities city center, shopping center and home decor boulevard, is the DBSCAN. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an algorithm that groups together points that are close to each other, based on the distance between the group and the size of the group (Daszykowski & Walczak, 2009). The algorithm is also able to identify points that are on the border of a cluster, so partly in lower-density areas. A DBSCAN is deemed to be of great use for distinguishing city / shopping centers and boulevards because of the ability of the algorithm to recognize centered or clustered points, and because of the ability to include 'border' points. An advantage of the DBSCAN as opposed to other methods such as k-means clustering is that no fixed number of clusters must be specified in advance. Furthermore, the algorithm is easy to use and requires only two parameters to be estimated.

To execute a DBSCAN, the values of two parameters have to be estimated:

- *eps*: The maximum distance allowed between two cluster points. This parameter specifies how large the distance between two points can be in order to be considered neighbors.
- *minPoints*: The minimum cluster size for concentrated group of points to be considered a cluster. For example, if set at 8, at least 8 points have to be within a distance of *eps* to each other in order to be considered a cluster.

For all three clustered shopping activities (city center, shopping center and home decor boulevard), the same parameters are established. To establish the parameters, a couple of operations have been carried out. First. the city center, all shopping centers and a home décor boulevard in the city of Enschede have been manually added to a point Shapefile layer in QGIS (Figure 6.1). The CROW uses the following definitions for shopping centers:

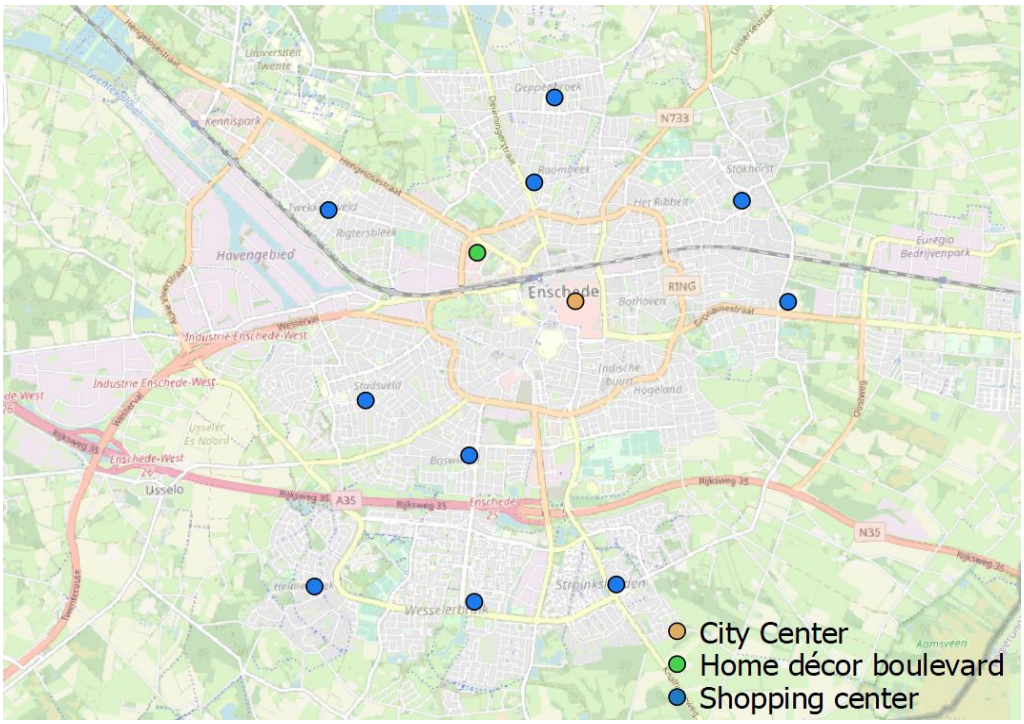


Figure 6.1: Centered shopping activities in the city of Enschede

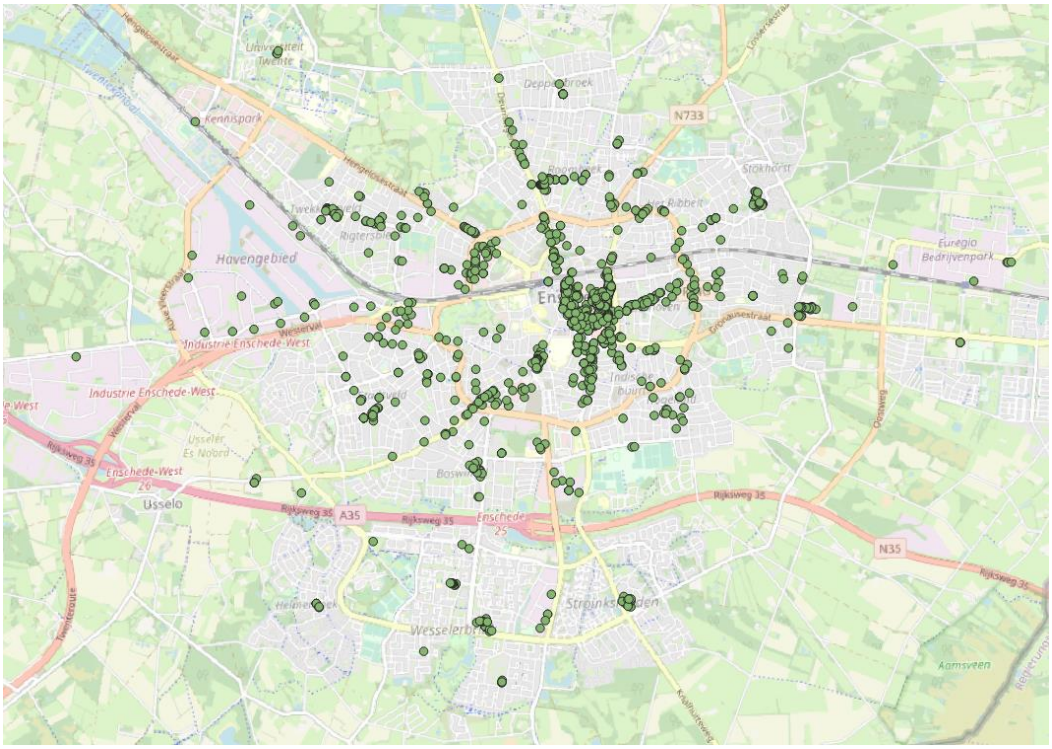


Figure 6.2: All VBOs *shop* in the city of Enschede

Supportive shopping areas (mainly for targeted shopping) that complement the downtown or main shopping area of a core in a municipality.

This definition leaves a lot of room for interpretation. The shopping centers as shown in Figure 6.1 are labeled as shopping centers because they are called shopping centers, because there is at least one fairly sized supermarket present (1.000 m2 or more), because there is almost always a large central parking area and a mix of usual shopping center stores such as drugstores, butchers, greengrocers, and hairdressers. And often the shopping centers are indoor affairs, but this is not a prerequisite. In Figure 6.2 all VBOs *shop* in the city of Enschede are presented.

So, the points presented in Figure 6.1 should ideally be identified as clusters in the DBSCAN carried out with the input presented in Figure 6.2. Based on trial and error, the parameters *eps* and *minPoints* have been determined by trying to identify all centers present in Figure 6.1 as clusters in the data from Figure 6.2. The main point of departure was that as many centers as possible were identified, without having a lot of grouped stores that are not shopping centers being identified as shopping centers. The changing of the parameters has different implications.

The value of *minPoints* cannot be too low. Too many (coincidentally) grouped shops would be labeled shopping center, resulting in a higher difficulty of identifying actual shopping centers later on. A too high value of *minPoints* results in small shopping centers being overlooked. The value of *eps* cannot be too low because even in shopping centers, and especially outdoor shopping centers, a fair distance between shops could be present. A value too low would split up identified centers. The value of *eps* cannot be too high either as clustered points could expand beyond centers. Including some random shops that are not actually part of a shopping center will not have too large implications for trip generation estimations, but it could affect operations that will be used to identify what cluster belongs to what activity.

In Table 6.1 and 6.2, the impact of fluctuations in both parameters on the number of clusters and the total number of shops in all clusters is displayed.

The *minPoints* parameter largely affects the number of identified clusters. Especially for lower numbers, the number of clusters identified largely increases. The number of shops in clusters increases as the number of clusters increases. However, the size of each identified cluster remains the same.

Table 6.1: Sensitivity *minPoints* parameter in Enschede (*eps* = 120)

minPoints	**Number of clusters**	**Nr. of shops in clusters (1340 in total)**
7	26	1107
9	22	1068
10	21	1059
11	19	1017
12	16	989
13	17	961
15	15	929

The *eps* parameter does mostly affect the cluster size. With a gradual increase of *eps*, the cluster size gradually increases. But for the *eps* value of 80, the number of shops included in clusters strongly decreases. In Figures 6.3 and 6.4 the implications of adjusting the *eps* value are visualized. Shops outside the home decor boulevard are included in the cluster for a high value of *eps*. And for a low value of *eps*, shops in the city center of Enschede are put into multiple clusters.

Table 6.2: Sensitivity *eps* parameter in Enschede (*minPoints* = 11)

eps	**Number of clusters**	**Nr. of shops in clusters (1340 in total)**
80	19	891
100	20	979
110	19	991
120	19	1017
130	19	1039
140	18	1044
160	18	1075

Figure 6.3: Shops outside home decor boulevards in Enschede included in cluster (*eps* = 160)

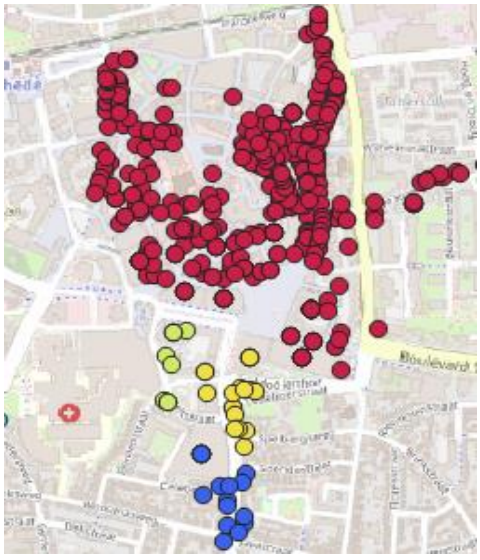


Figure 6.4: City center of Enschede divided into multiple clusters (*eps* = 80)

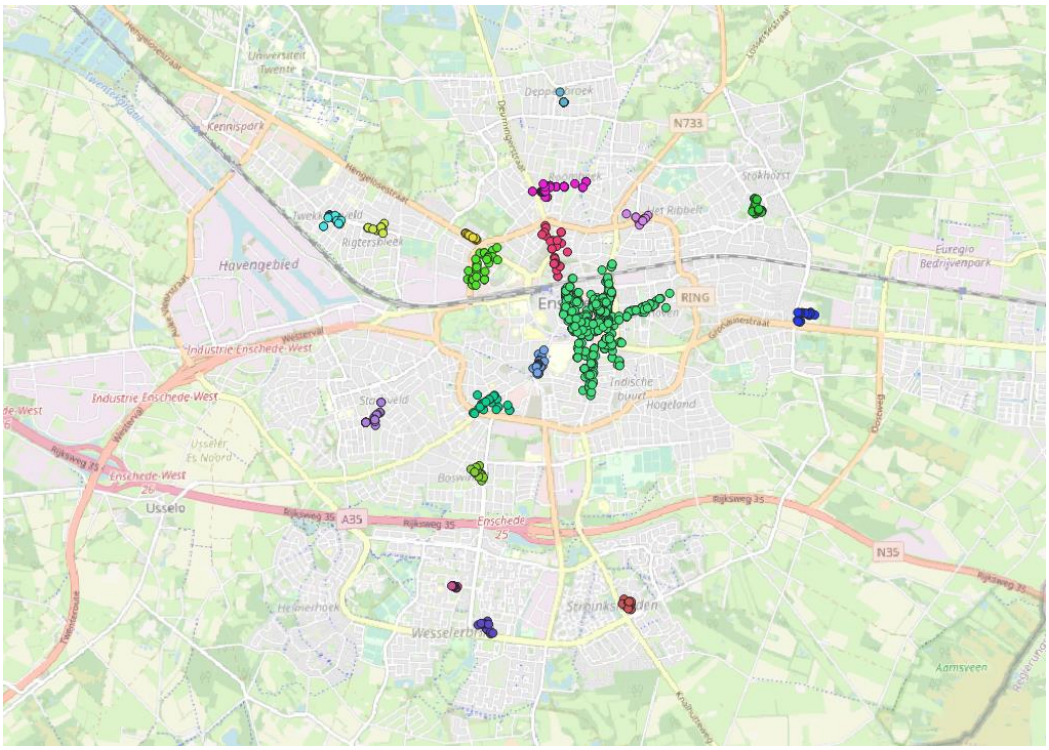


Figure 6.5: Results of DBSCAN (*eps* = 120, *minPoints* = 11*)* VBOs *shop* in the city of Enschede

To identify as many centered shopping activities as possible, to minimize the number of incorrect identified clusters, and to include the correct number of shops per cluster, the following DBSCAN parameters have been found to be most optimal for Enschede:

eps: 120 meter

minPoints: 11 points

In Figure 6.5, the results of the DBSCAN are presented. Of the 12 centered shopping activities (Figure 6.1), 11 were identified. The used parameters lead to the identification of 19 clusters in total. One small shopping center has not been identified. Altering the parameters to include this shopping center leads to a large increase of identified clusters. Ideally, only the shopping activities labelled in Figure 6.1 come out of the analysis, but this is not achievable by only using a DBSCAN. Therefore, the attributes of the VBO's from Figure 6.5 need to be analyzed further. In Figure 6.6 the properties of the clusters are presented, with:

- N; the number of VBOs in the cluster
- Type; the kind of centered shopping activity (CC = city center, SC = shopping center, HB = home decor boulevard, NO = unlabeled clustered shopping activity). This attribute has been manually added to the clusters
- A; the total surface area of all VBOs in the cluster
- mean_A; the average surface area of the cluster
- max_A; the surface area of the largest VBO present in the cluster
- dens_C; an indicator for the density of the cluster (explained later)

Based on the above-mentioned VBO attributes, rules are developed that determine what shopping activity is present at what identified cluster. First, the activities that are relatively easy to identify based on the values in Figure 6.6 are distinguished, the city center and the home decor boulevard. Then, with the remaining clusters, rules are developed for identifying shopping centers.

CLUSTER_ID	N	type	A	mean_A	max_A	dens_C
1	571	CC	172137	301	10096	8.3
14	33	HB	45172	1369	7887	26.7
19	20	SC	14663	733	4415	13.5
17	26	SC	11898	458	3156	11.2
7	61	SC	14708	241	2927	-4.8
13	32	NO	11715	366	2280	44.1
2	20	SC	3977	199	2058	18.9
11	33	SC	7990	242	1762	-10.3
5	22	SC	5405	246	1719	19.2
8	30	SC	7341	245	1609	8.6
9	27	NO	7946	294	1476	42.3
6	21	NO	8114	386	1423	25.0
3	21	SC	5580	266	1418	10.9
4	53	SC	9669	182	1400	10.0
10	12	NO	2700	225	901	40.4
18	12	NO	2177	181	773	43.4
12	12	NO	2077	173	508	21.2
16	11	NO	1211	110	179	5.7

Figure 6.6: Features of clustered shopping activities in the city of Enschede

City center

The city center is the main shopping area of a city. In the city center, most VBOs *shop* within a city should be present. In Figure 6.6 it can be seen that the N value of cluster ID 1 stands out by far with 571 VBOs. Therefore, N alone could be a certain indicator for assigning the activity city center to a cluster.

Proposed rule: the cluster with the largest value of N is the city center.

Home decor boulevard

Home decor boulevards are shopping areas with generally large accommodations. Many of the bedding, furniture, lamp or mixed-use stores take up a lot of space. A logical indicator for identifying this activity is the average area of the VBOs in the cluster. When looking at the values of mean_A in Figure 6.6, one value stands out by large. This cluster represents the home decor boulevard visualized in Figure 6.5. In the Netherlands, the average home decor store has a surface area of 1.100 m^2 (Locatus, 2020).

Proposed rule: clusters with a mean_A larger than 900 m^2 are home decor boulevards.

Shopping centers

For the city center and home decor boulevard, it has been established how these activities can be distinguished from other clustered VBOs *shop*. Cluster 1 and 14 can be removed from consideration of other shopping centers. For remaining clusters it needs to be determined how shopping centers can be distinguished from unlabeled clustered shopping activities.

A first prerequisite for a shopping center being a shopping center is the presence of a medium to large-sized supermarket. These shops are relatively large compared to other shops present. The max_A value of each cluster could therefore be a logical indicator to determine whether a supermarket is present in a cluster. The cluster data in Figure 6.6 has been sorted on decreasing values of max_A. The bottom-four clusters are indeed no shopping centers and at these clustered VBOs in Enschede no supermarket is present. Manual checking shows that the max_A value of 1400 from cluster 4 is the surface area of a supermarket. In all other shopping centers as well, the largest surface area is that of a supermarket. Thus, based on the max_A value, clustered activities can be identified in which no supermarket is present, and in which no other large shop is present. According to the CROW, medium-sized supermarkets have a surface area between 1.000 and 2.000 m2, and XL supermarkets reaching from 2.500 to 4.000 m^2 (CROW, 2018).

- Proposed rule: clusters with max_A below 1.000 m^2 are no shopping centers

However, there are still three other clusters (6, 9, and 13) present in which either a supermarket or another shop with a large surface area is present that are not shopping centers. In Figure 6.7, four clusters are shown, where the top two are no shopping centers, and the bottom two are. Clusters 7 and 3 are both more densely grouped and in clusters 13 and 6, the shops are more stretched out and not all shops are adjacent. They are close enough to be identified by the DBSCAN as clusters, but the shape of the cluster significantly deviates from clusters representing shopping centers. If somehow these shape differences can be quantified, shopping centers can be distinguished from other shopping clusters that are not shaped in the same way.

To quantify the differences in shape of the clusters, an indicator dens_C has been developed that quantizes the density of a cluster which enables comparing the clusters on the grouping of the shops.

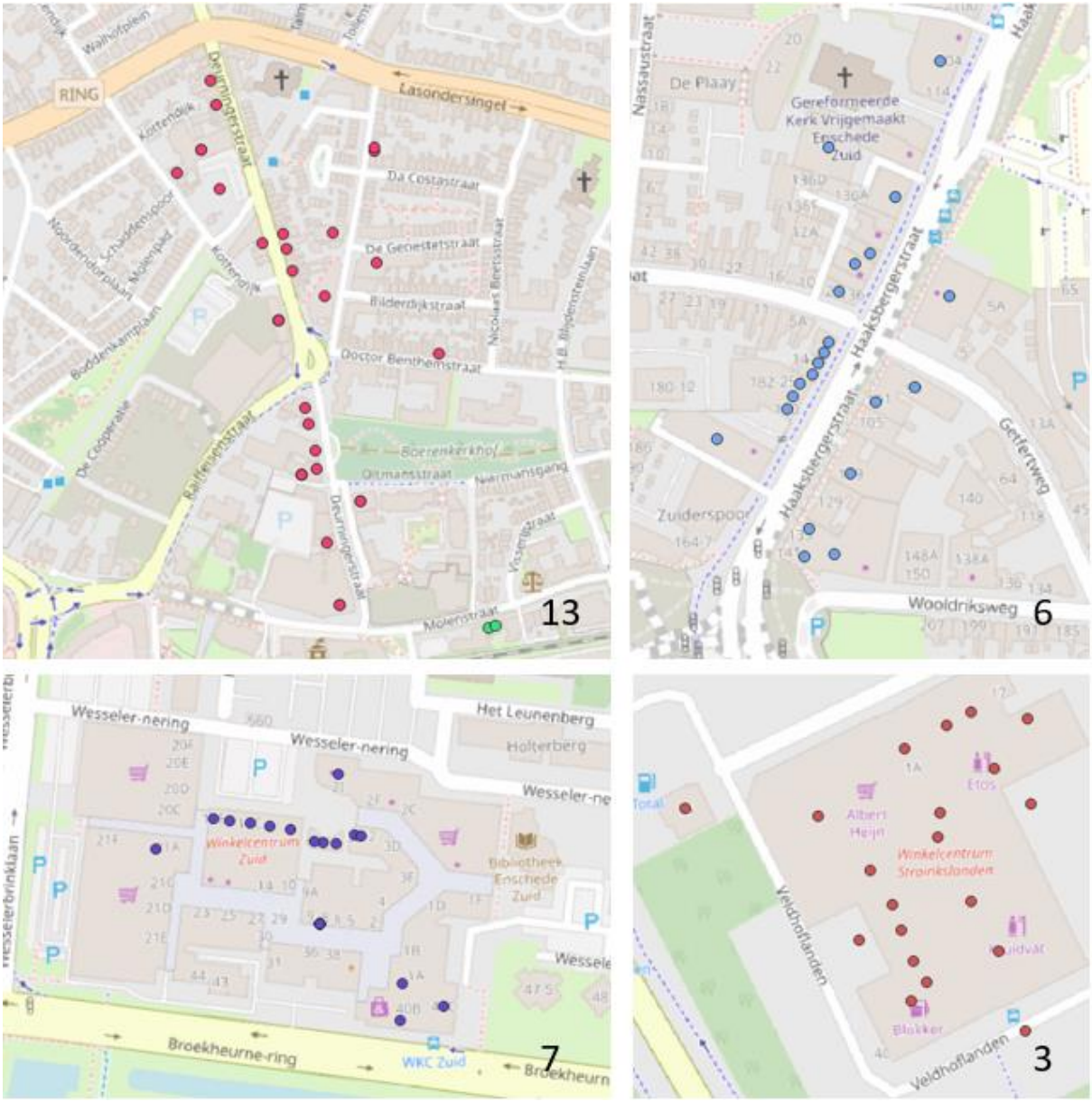


Figure 6.7: Four clusters of VBO *shop*. Top-two are ordinary clustered shopping locations. Bottom-two are shopping centers.

In Figure 6.8, a schematic overview of the distance relation between two VBOs *shop* in BAG is illustrated with:

- nnd: The nearest neighbor distance between VBO 1 and VBO 2
- A1 and A2: The surface area of respectively shop 1 and 2
- d1 and d2: The distance between the VBO location and the shop boundary of VBO 1 and 2
- d: The distance between shop 1 and 2

In BAG, the shape or outline of a VBO space is not included. Concerning spatial characteristics, the point location and the surface are known. Assuming perfect rectangular shops and perfect centered VBO points, the distance between nearest neighbors shop 1 and shop 2 can be calculated by:

$$d = nnd - d1 - d2$$

$$d = nnd - \frac{1}{2}\sqrt{A1} - \frac{1}{2}\sqrt{A2}$$

By calculating d for x nearest neighbors of a shop in a cluster, and averaging these distances, a number $dens\ shop$ is developed indicating the density around this shop.

$$dens\ shop = \frac{1}{x}\sum_{i=1}^{x} d$$

To determine the overall density of a cluster, the $dens\ shop$ values of all shops in the cluster are averaged.

$$dens\ C = \frac{1}{N}\sum_{i=1}^{N} dens\ shop$$

with N the cluster size. A numerical value is created indicating the density of each cluster. It is important to subtract the distance between a VBO and the border of its space. When ignored, shopping centers with relatively large shops will appear relatively low dense, when using the proposed indicator. The indicator was also tested using only nnd instead of d. The performance of the indicator is more effective when subtracting the assumed distances form the VBO location to its boundary despite the fact that by no means every shop is rectangular shaped.

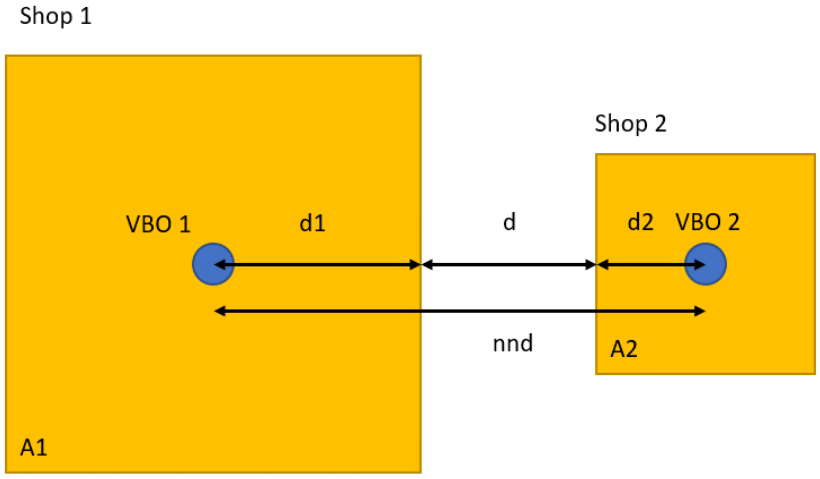


Figure 6.8: Distances between shops related to VBO location and shop size

By fixating the value of x for all clusters, the $dens\ C$ value becomes suitable for comparing the clusters. If including for example the neighboring distances of all shops in a cluster, large clusters will automatically have a higher value of $dens\ shop$, without being necessarily lower-dense. If the value of x is too low, clusters such as cluster 13 and 6 (Figure 6.7) could appear to be high dens according to the proposed indicator, while overall having an elongated shape. For distinguishing shopping centers from non-shopping centers, the value of x has been put at 7. This value leads to the greatest differences between shopping centers and non-shopping centers, and is therefore most suitable.

CLUSTER_ID	type	dens_C
11	SC	-10.3
7	SC	-4.8
8	SC	8.6
4	SC	10
3	SC	10.9
17	SC	11.2
19	SC	13.5
2	SC	18.9
5	SC	19.2
6	NO	25
9	NO	42.3
13	NO	44.1

Figure 6.9: dens_C values

In Figure 6.9, the $dens\ C$ value of all remaining clusters are presented, sorted according to $dens\ C$. Clusters 6, 9 and 13, the non-shopping centers, have the highest $dens\ C$ value, with 25, 42.3 and 44.1. The values of the shopping centers are considerably lower, and the differences between the highest SC (cluster 5), and the lowest non-SC (cluster 6) are sufficient. Thus, by determining the $dens\ C$ value of all clusters, an indicator is created that enables identifying clusters that are too spread out to be shopping centers. Based on the max_A value and the dens_C value, all shopping centers have been identified in Enschede. No additional selection rules need to be developed.

Proposed rules to separate non-shopping centers from shopping centers:

- clusters with max_A below 1.000 m^2 are no shopping centers
- clusters with dens_C value above 22 are no shopping centers

6.2 Validation of cluster analyses

To test the proposed rules for distinguishing clustered shopping activities, the rules are applied to VBOs *shop* in a mid-sized, Hengelo (+/- 80.000 citizens), and in a smaller city, Veenendaal (+/- 60.000 citizens). In Table 6.3, the clustered shopping activities present in Hengelo and Veenendaal are listed.

Table 6.3: VBO *shop* statistics Hengelo and Veenendaal

	Hengelo	Veenendaal
VBO *shops*	976	637
Average surface area	384	487
Shopping center	5	2
Home decor boulevard present	0	1
City center	1	1

For a geographical overview of all VBOs *shop*, and the locations of the shopping facilities in both cities, see Appendix E.

DBSCAN

A DBSCAN with the proposed parameters *eps* = 120 m and *minPoints* = 11 is carried out on the VBOs in Hengelo and Veenendaal. The results are presented in Figure 6.10 and 6.11. For both cities, the DBSCAN led to the identification of all present clustered shopping activities. In Hengelo, six clusters have appeared that are no shopping centers, three in Veenendaal. The values of the parameters are accurate for identifying clustered shopping activities in Hengelo and Veenendaal.

City center
Proposed rule: the cluster with the largest value of N is the city center.

For Enschede and Hengelo, this rule holds. Both centers stand out by far concerning the cluster size.

Home decor boulevard
Proposed rule: a cluster with a mean_A larger than 900 m^2 is a home decor boulevard.

In Hengelo, no home decor boulevard is present. In Veenendaal, one cluster is present with a mean_A value of 1.148 which is indeed a home décor boulevard. For both Hengelo and Veenendaal, the proposed rule holds.

Shopping centers
Proposed rules:

- clusters with max_A below 1.000 m^2 are no shopping centers
- clusters with dens_C value above 22 are no shopping centers

The data in Figure 6.10 and 6.11 has been sorted to max_A. In Hengelo, all clusters that are no shopping centers have a max_A equal to or below 700. In Veenendaal, one cluster that is not a shopping center can be identified based on the max_A rule. All shopping centers in both cities have a max_A value of above 1.300. Therefore, the max_A rule holds. However, not all clusters have correctly been identified yet.

After applying the max_A rule in Veenendaal, cluster 4 and 7 remain of the non-shopping centers (type is 'NO'). Both values for dens_C are above 22, and the values for the shopping centers are below 22. In Hengelo, no type 'NO' clusters need to be considered, but cluster 4 poses a problem, with a dens_C value of 31.6. And furthermore, the dens_C value of cluster 12 is close to 22. Cluster 4 will be further analyzed and the 22 value of dens_C might need reconsideration.

In Figure 6.12, cluster 4 is visualized. The cluster is not an ordinary shopping center, but more of a shopping area. It includes one actual shopping center with two supermarkets, a parking area and additional shops; the building on the right. But at the left, two additional supermarkets with a central parking area are present. In-between, a multitude of shops is present that connects both places and turns it into a large shopping cluster. The spread-out shape results into a high dens_C value of the cluster.

CLUSTER_ID	N	type	A	mean_A	Max_A	dens_C
1	357	CC	92072	258	3509	14.5
11	43	SC	10786	251	1946	7.7
9	26	SC	8087	311	1874	18.1
3	21	SC	7443	354	1746	5.2
12	11	SC	3912	356	1390	21.5
4	34	SC	9532	280	1385	31.6
10	27	NO	6601	244	700	32.6
5	11	NO	2413	219	699	47.9
8	13	NO	2120	163	528	53.0
6	26	NO	3666	141	458	30.8
7	11	NO	1691	154	351	20.0
2	19	NO	1538	81	283	28.9

Figure 6.10: DBSCAN VBO *shop* Hengelo (*eps* = 120, *minPoints* = 11)

CLUSTER_ID	N	type	A	mean_A	Max_A	dens_C
6	16	HB	18369	1148	3150	29.9
1	358	CC	108276	304	3006	10.4
2	28	SC	6658	238	2165	10.8
4	28	NO	8516	304	1618	27.4
5	17	SC	4327	288	1598	20.6
7	15	NO	5504	367	1326	50.5
3	39	NO	7421	190	582	26.3

Figure 6.11: DBSCAN VBO *Shop* Veenendaal (*eps* =120, *minPoints* = 11)

To identify this shopping center, the parameters of the DBSCAN should be altered, which will negatively affect other results and is therefore not a viable solution. The unique composition of shops in this area, and the elongated shape of the cluster that contains more than only one shopping center, make that the proposed rules for this particular cluster are not successful.

Thus far, the proposed rules for identifying shopping centers have resulted in identifying six out of seven shopping centers in Veenendaal and Hengelo. However, setting the 1.000 m^2 for the max_A rule and the 22 for the dens_C value based on the shops in Enschede has been arbitrary to some extent. The 1.000 m^2 has been backed up by data from the CROW, but the value of dens_C is only based on the data from

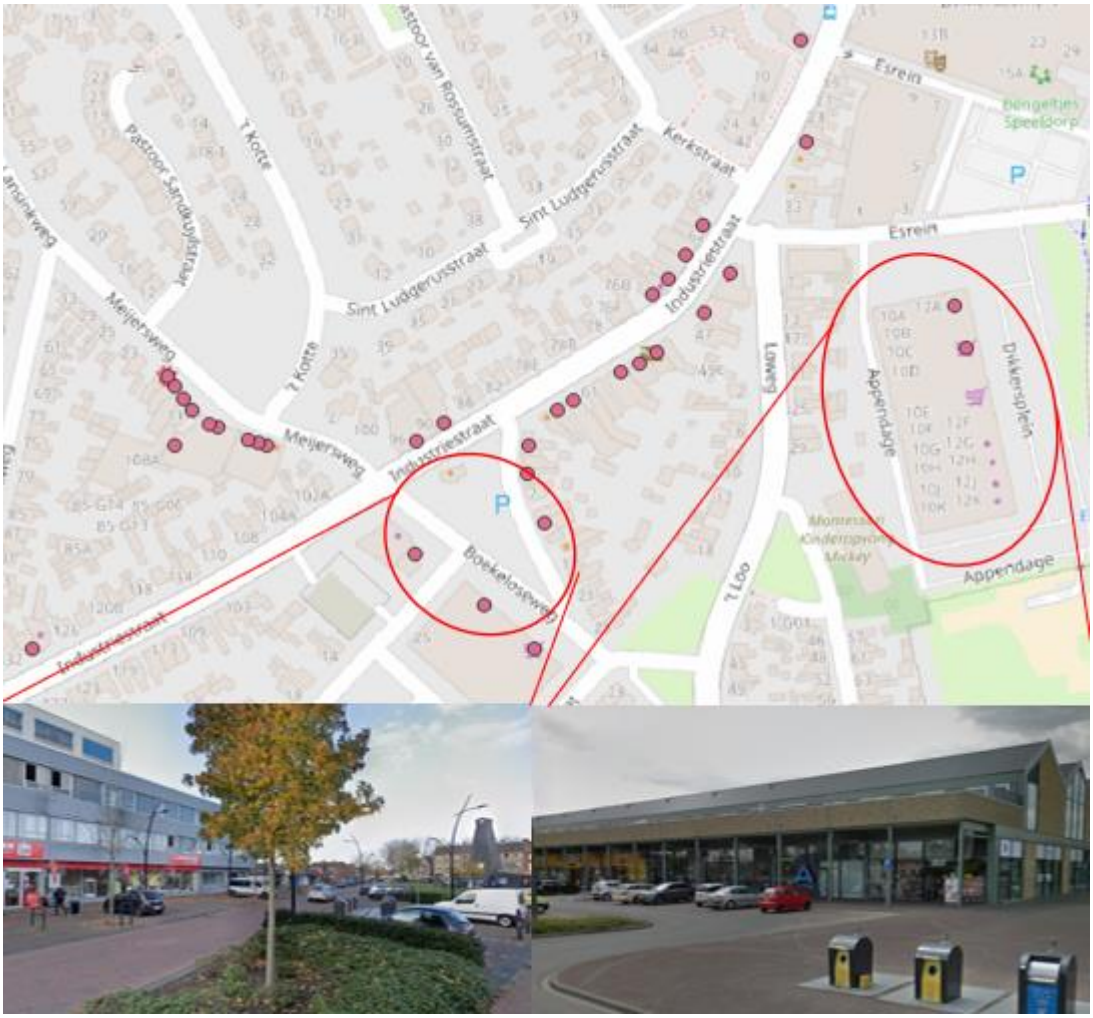


Figure 6.12: Cluster 4 in the city of Hengelo

Enschede. To increase the validity and to improve the values, data of clusters in all three cities is visualized.

In Figure 6.13 and 6.14, the max_A values and the dens_C values of all clusters are presented. It can be seen that both rules are able to identify a majority of non-shopping centers, but that not a single rule is completely effective. Based on Figure 6.13 it could be argued that 1.000 m^2 is a bit on the low side. However, supermarkets below 1.300 m^2 could be present in shopping centers in other cities, and the presence of the dens_C rule is worth avoiding this risk. In Figure 6.14, some overlap is present between the dens_C value of shopping centers and non-shopping centers. Notable is that the three non-shopping center clusters with the smallest dens_C values have a very low surface area. After applying the max_A rule, the situation in Figure 6.15 arises. Besides the outlier of cluster 4 (Hengelo) of the shopping center values, a clear boundary can be made between shopping centers and non-shopping centers. The most optimal value to for the dens_C value is set at 23, instead of the earlier 22.

The definitive rules for identifying clustered shopping activities are:

- A DBSCAN with parameters *eps* = 120 m and *minPoints* = 11 to identify potential clustered shopping activities
- The cluster with the largest value of N is the city center.
- A cluster with a mean_A larger than 900 m^2 is a home decor boulevard.
- Clusters with max_A below 1.000 m^2 are no shopping centers
- Clusters with dens_C value above 23 are no shopping centers

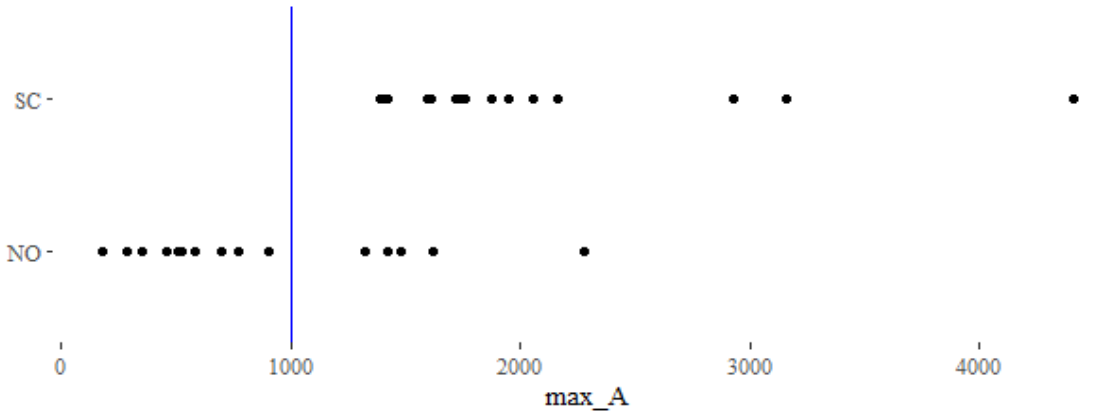


Figure 6.13: max_A values of all clusters

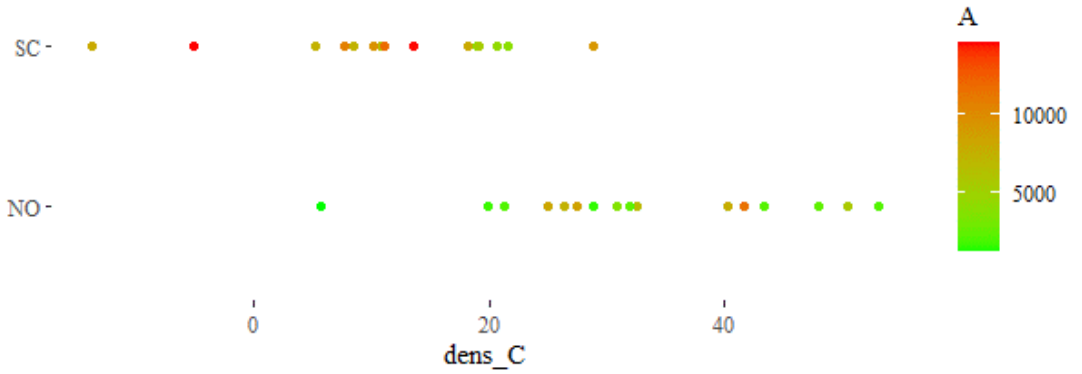


Figure 6.14: dens_C values of all clusters (colored by surface area)

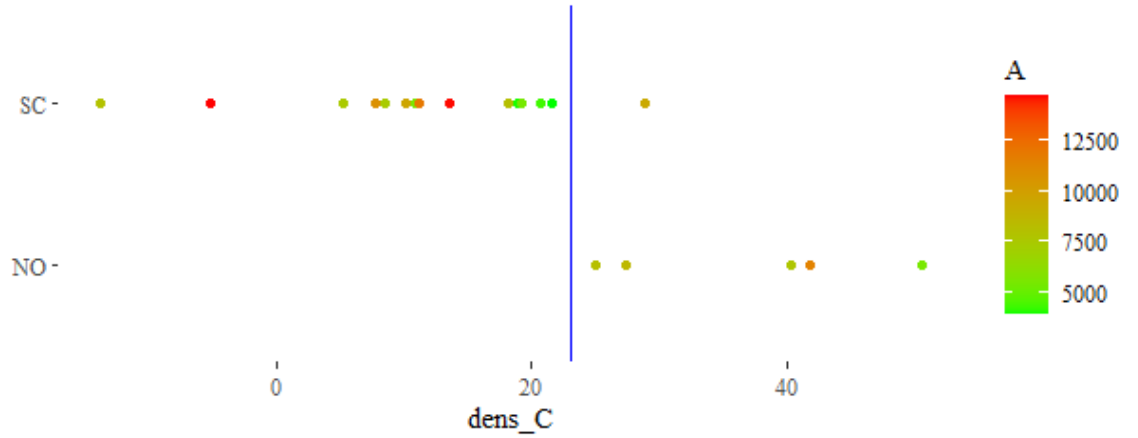


Figure 6.15: dens_C values after applying max_A rule (Optimal value = 23)

6.3 Wholesale and supermarket activities

In this paragraph, the results of identifying wholesale activities and supermarket activities in BAG, combined with IBIS and OSM data will be presented.

Wholesale

The CROW describes wholesale activities as trade in food, building materials or office supplies in mostly a large-scale establishment in an industrial area on the outskirts of a city, where one can exclusively make purchases when in possession of a pass (CROW, 2018).

With IBIS data, through spatial joins, it is possible to select VBOs *shop* that are present in industrial areas. According to the definition of the CROW, these shops could be wholesale activities. In this paragraph it will be determined how well wholesale activities are distinguished when combining IBIS and BAG data. For two industrial areas, one in Enschede and one in Hengelo, it is determined to what extent BAG VBO within IBIS industrial areas are wholesale activities. All data is gathered by manually looking up each VBO *shop* in Google Maps. As it being time-consuming labor, it was not possible to include more than two industrial areas.

Through a spatial join in a GIS environment, all VBOs *shop* in an industrial area in Enschede and Hengelo were selected. In Table 6.4, the total number of VBOs *shop* per city can be compared with the VBOs *shop* in industrial areas. The number of VBOs *shop* in industrial areas is relatively low, but when considering the surface area, in both cities around one-fifth of VBOs *shop* surface area is present in industrial areas. This corresponds with the definition from CROW, mentioning that wholesale activities are large-scale. The large amount of shops present in industrial areas makes it worthwhile reviewing these activities. In Figure 6.16, the location of the VBOs *shop*, all industrial areas and the areas of analysis are presented.

Table 6.4: Statistics VBO *shop* Enschede and Hengelo at industry

	VBO *shop* total		**VBO *shop* at industry**	
	N	m^2	N	m^2
Enschede	1340	514.000	64 (5%)	86.000 (17%)
Hengelo	978	485.000	93 (10%)	98.000 (20%)

In Table 6.5, the results of reviewing all activities within industrial areas Twentekanaal Zuid (Hengelo) and Havengebied (Enschede) are presented.

Table 6.5: Summary statistics industrial areas and VBO *shop*

City	Hengelo	Enschede
IBIS industrial area	Twentekanaal Zuid	Havengebied
BAG VBO *shop*	41	17
BAG VBO *surface area*	66.000 m^2	17.000 m^2
Nr. of wholesale	13	8
Nr. of non-wholesale	27	9
Surface area wholesale	39.000 m^2 (60%)	11.000 m^2 (65%)
Surface area non-wholesale	27.000 m^2	6.000 m^2
Activity types wholesale	Food / non-food, construction and industry, catering, bricks, car parts, mounting materials, office supplies	Office supplies, auto parts, iron ware, food
Activity types non-wholesale	Car dealers, car repair, car other, empty stores, gas station	Machinery rental, kitchen shop, car dealer, gas station, trailer rental, camper shop

In both industrial areas 60 to 65% of VBO *surface area* belongs to wholesale activities, while making up half of the number of VBOs. In both cities, no wholesale activities have been found outside the industrial areas. A reasonable majority of the shop area within IBIS industry areas are thus wholesalers. It could be possible to distinguish wholesale activities by analyzing the properties of the VBOs within the IBIS areas, based on for example VBO *surface area*. To do so, more data must be gathered, as two areas are not sufficient. The non-wholesale activities seem to be shops that do not generate too many trips per square meter as well. A viable option could be to treat all VBOs *shop* inside IBIS areas as wholesale activities. With no wholesale activities outside the IBIS areas and a 'success rate' of 60 to 65%, combining IBIS and BAG data to find wholesale activities provides an acceptable solution.

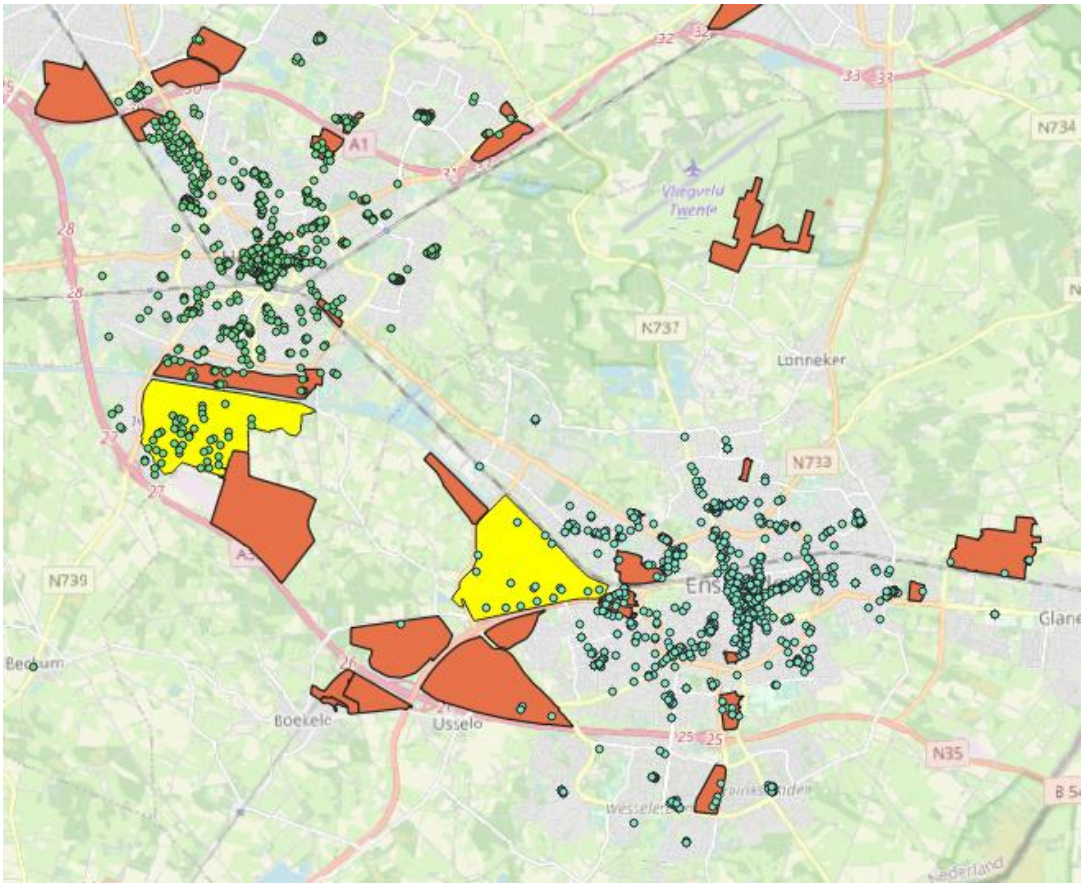


Figure 6.16: IBIS industrial areas and VBO *shop* in Hengelo (North-West) and Enschede (South-East)

Supermarket

The hypothesis was that, when shopping centers, city centers, home decor boulevards and wholesale activities were selected, it was possible to identify supermarkets from the remaining VBOs *shop* in a city, based on the surface area of the remaining VBOs. Supermarkets have a relatively large surface area between 1.000 to 3.500 m2 according to the CROW (CROW, 2018). However, during the research it was found that by no means it was possible to identify supermarkets based on BAG attributes only. Too many VBOs *shop* were present in the same surface area range that had not been identified yet, and that are no supermarkets.

Therefore, OSM data is required to identify supermarkets outside city centers and shopping centers. Through OSM data, the location of supermarkets (that are available in OSM) can be retrieved. However, no surface area is included, which should be retrieved from BAG. BAG and OSM share no common data attributes and a spatial join is required, where the challenge lies in the fact that BAG VBO and OSM points of interests are not perfectly aligned. The closest VBO *shop* near an OSM supermarket is not in all cases the actual supermarket. Developing operations and testing results for combining OSM and BAG data requires time-consuming handiwork for which unfortunately no time was left during the conduct of this research project. Furthermore, the completeness and reliability of OSM should be reviewed before it can be stated as useful. In the recommendations on future research and application (Chapter 10) it is suggested how BAG and OSM can be combined to select supermarkets.

6.4 Concluding

For distinguishing shopping activities in BAG, several operations have been developed that can be automatized and through which with a significant certainty shopping activities can be given a specific function. Based on a DBSCAN, in three different cities, 24 of 25 clustered shopping activities were identified. In each city, the city center was correctly identified, as were the home décor boulevards present in two cities. Of the shopping centers, 16 of 17 centers were identified. By combining IBIS data with BAG data it was possible to find all wholesale activities in two cities. Of the distinguished wholesale activities, 60 to 65% were identified to be actual wholesale activities.

Chapter 7. Trip generation factors in BAG

In this chapter, part of research question 2 will be answered. The capability of BAG to predict trip generation factors relevant at the home-end is analyzed. In Chapter 4, trip generation factors were identified that are often used in practical transportation studies. Those factors will be focused on in the analysis. The trip generation factors:

- Residents
- Car ownership
- Labor force
- Residents aged (0-34)
- Households

In BAG, the number of households is available. As a first step in determining which factors need to be predicted by BAG, the distribution of the number of residents and cars per household based on CBS District 2018 data (Appendix D) are plotted in Figure 7.1. As can be seen, a household is not a good predictor for the number of cars or the number of residents present. On average, 1,1 cars are present in a households with a standard deviation (SD) of 0,35, which is a relative spread of 32%. Concerning residents, 2,3 are present in households on average, with an SD of 0,35, a relative spread of 27%. So both for car ownership and residents, it can be of benefit to further analyze the predictive ability of BAG for these factors.

The labor force is the number of people able to work, which can be approximated by the number of residents aged 25 to 65. This factor is used to determine the number of work trips at the home-end. The residents aged 0 to 34 are used to model trips for the motive education. The average number of residents aged 0-25 per HH (household) is 0,61, with an SD of 0,18 (30%). In Figure 7.2, the number of residents aged 0 to 24 and 25 to 65 at district level are expressed as a percentage of the total number of residents, and plotted. The number of residents aged 0 to 34 could not be retrieved from CBS Districts, therefore residents aged 0 to 25 is used. On average, 27,4% of residents at district level is aged 0 to 24, with an SD of 5,4%. And on average, 52,5% of residents is aged 25-65, with an SD of 7,0%. The dispersion in the spread of these factors is much lower. The condition for these values is that the number of residents must be available.

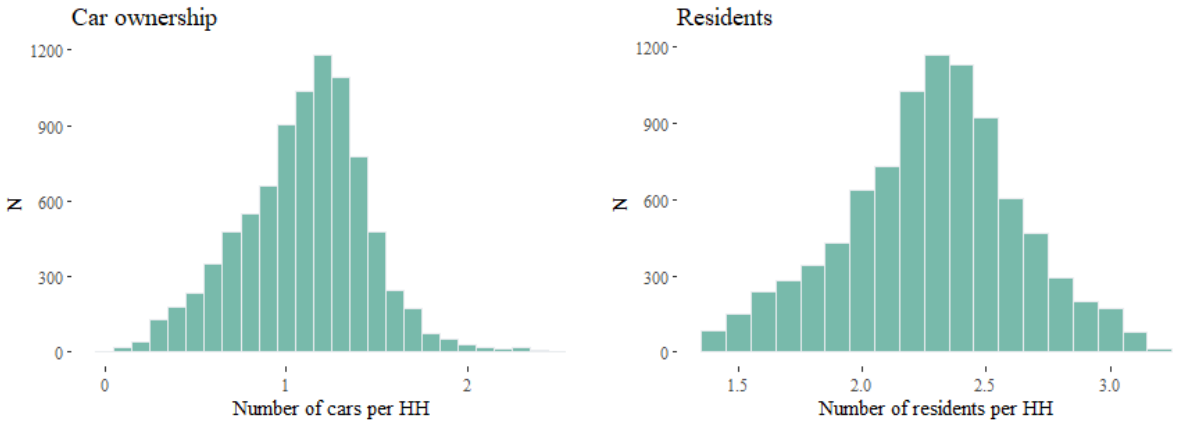


Figure 7.1: Distribution Car ownership and Residents per HH

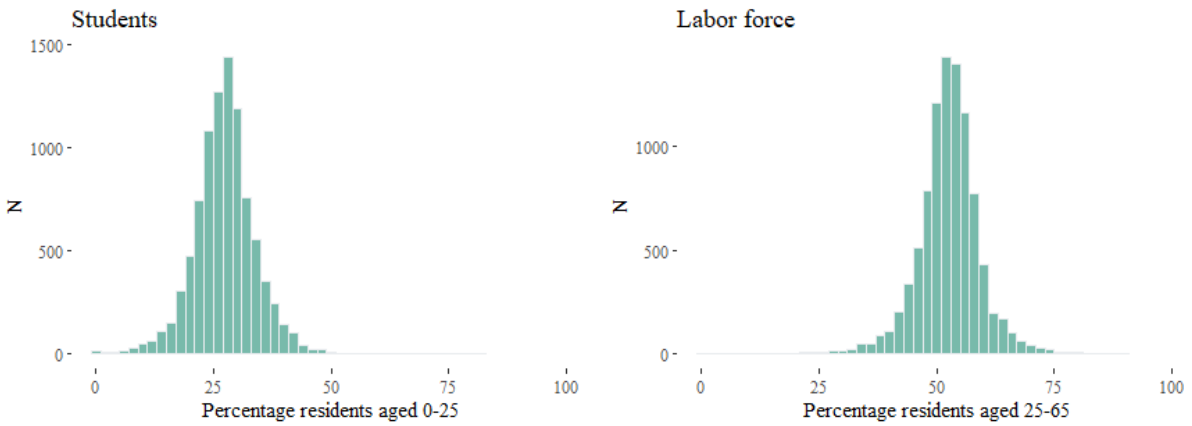


Figure 7.2: Distribution percentage residents aged 0-25 and 25-65

Thus, if it is managed to predict the number of residents in BAG, the number of students and the labor force can already be approximated much better in BAG. Therefore, the current most important trip generation factors are the number of residents and car ownership, which will be analyzed in this Chapter. In Paragraph 7.1, a regression analysis is carried to predict car ownership levels, in Paragraph 7.2, the predictive ability of BAG to determine the number of residents is analyzed.

7.1 Car ownership regression analysis

In this paragraph, the process and results of estimating car ownership, or the number of cars per household, based on BAG attributes is described. The goal is to determine how precise BAG predictions could be, and how predictions deviate in different types of cities. From CBS District 2018 data, car ownership levels per household at district level are available for the whole of the Netherlands. Eventually, the results of this analysis can be used for two purposes:

- Distributing highly aggregated car ownership data into smaller zones
- Predicting car ownership for new housing developments

7.1.1 The regression data

Data used (Appendix A and D) in this analysis are:

- BAG VBO *living*
- CBS Districts 2018
- CBS linking table

The hypothesis is that both the VBO *living* surface area (HH surface area) as the VBO *living* residence type are important factors to determine the average number of cars per household. The regression analysis will be carried out at District level. For the analysis, only BAG VBOs *living* with one function have been included. The average household surface area per district of VBO *living* is the predictor variable to estimate car ownership levels at District level. If VBOs with multiple functions are present, the average calculated surface area will be biased, because it cannot be determined to what extend what part of the surface area belongs to VBO *living*. For estimating the regression coefficients, every district in which the difference in the number of households between BAG and CBS is larger than 5% has been removed. In districts with larger differences, for example large new construction development could have taken place, through which the average residence surface area could have significantly changed. And furthermore, only districts with a population density larger than 100 residents per km2 have been included. For this analysis, the main interest lie for the determination of car ownership levels in urban areas. Outside urban areas, other factors could play a role when considering car ownership, and for this study, the interest in car ownership levels in rural areas is marginal (See Figure 7.3).

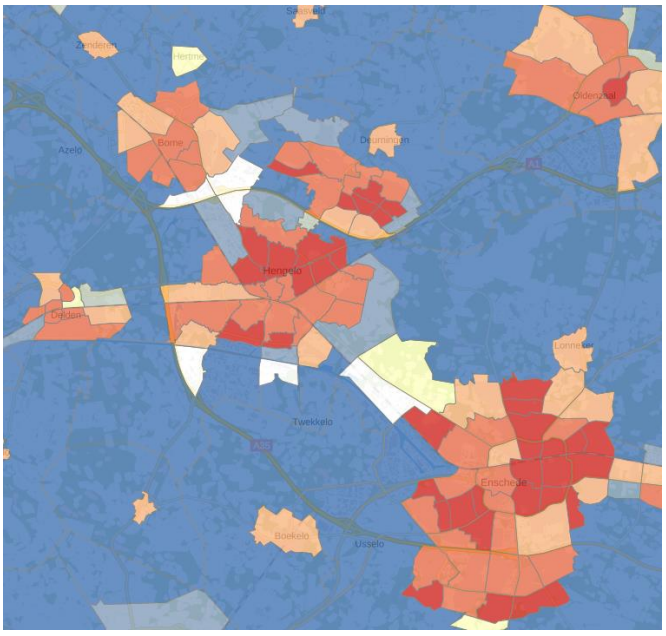


Figure 7.3: Population density and rural areas (blue and white)

The CBS linking table (Appendix D) has been used to aggregate VBO *living* surface area and VBO *residence type* to district level, and to merge BAG data with CBS district data from 2018.

In Table 7.1, the summary statistics of data used for the car ownership regression analysis are presented. In the analysis 4.419 of 16.772 CBS districts are included. Furthermore, required data of other attributes is available for all 4.419 included districts. The most important variables in the analysis are the CBS number of cars per HH, and the VBO average HH surface area.

Table 7.1: Summary statistics car ownership regression data

	Records	Mean	SD	Min	Max
VBO *surface area*	4419	114299.9	97248.8	5835	1255248
VBO *surface area* detached	4419	19342.17	26498.27	0	267659
VBO *surface area* semi-detached	4419	12640.53	16299.16	0	142650
VBO *surface area* terraced	4419	59125.57	59390.23	0	668055
VBO *surface area* multi-fam	4419	23191.63	38746.1	0	636812
CBS number of residents	4419	2212.823	1943.718	105	28450
VBO N	4419	983.1466	885.3572	50	13895
CBS number of cars per HH	4419	1.111224	0.279495	0	2.4
CBS surrounding address density	4419	1497.23	1412.501	20	12259
CBS income per earner	1307	25.69067	4.816421	14.3	80.3
VBO average HH surface area	4419	124.091	29.93409	23	372.7586

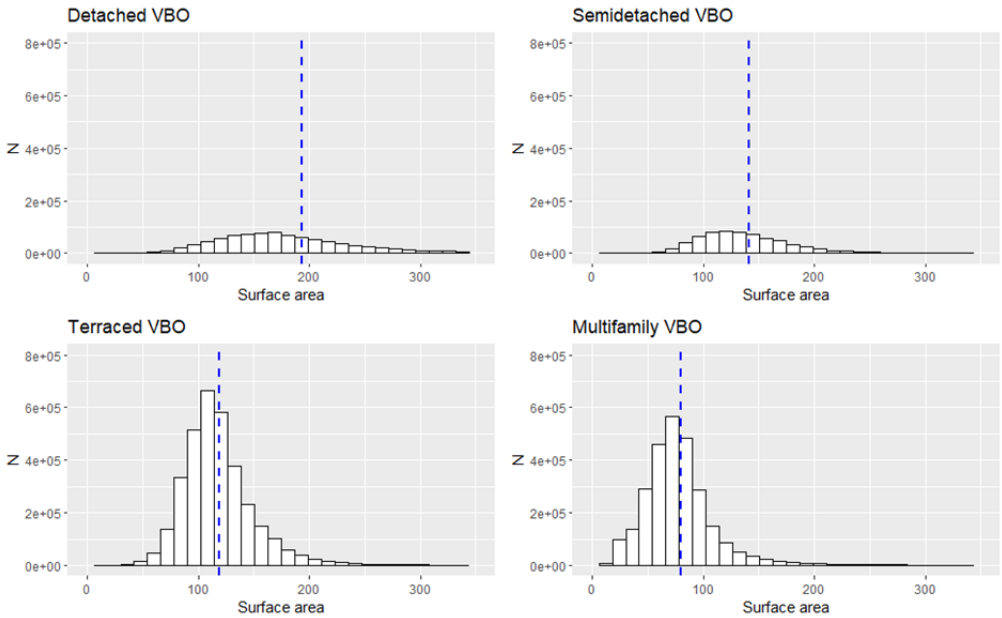


Figure 7.4: Distribution household types over surface area

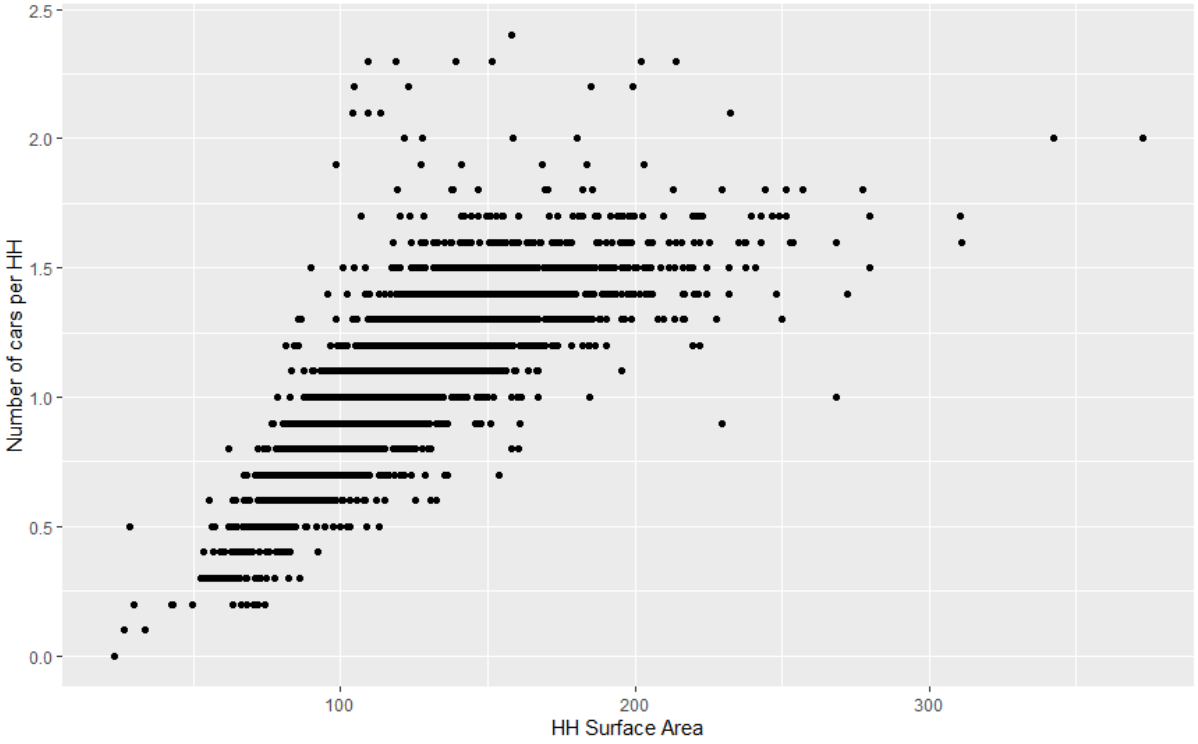


Figure 7.5: Scatterplot of car ownership and HH surface area

7.1.2 Considered variables

The goal of the analysis is to predict car ownership levels as accurate as possible based on BAG attributes. When it is possible to estimate car ownership levels with a certain precision at district level, it is also possible to estimate car ownership levels at a model zone level. At district level, there is enough variety in car ownership levels and residence size and types to be able to estimate relations between these characteristics based on a regression (Table 7.1). The main BAG attributes of interest are the VBO *living* surface area and the BAG residence type. The underlying assumption of the surface area of a house being a logical predictor for car ownership is that the owner of a large house could have the means and the space to buy and store one or more cars, whereas the owner of a small house might lack the means for a car, or even the necessity to use one. The housing type could give a little more information on the possibilities and the needs of the owner to possess one or more cars. When considering the residence types multi-family, terraced and (semi-)detached, house size largely accounts for residence type. This can be seen in Figure 7.4. The mean size of a multi-family building is below 100 m2, whereas detached houses are almost 200 m2 on average. However, a lot of overlap of house size between residence types does occur. Therefore, it is expected that when residence type is added as a predictor, it is possible to estimate car ownership even slightly better.

Thus, the main attribute from BAG to estimate car ownership is VBO *surface area*. At district level, this data is present as VBO average HH surface area (Table 7.1). As a first step in predicting car ownership based on BAG, the number of cars has been plotted against HH surface area (Figure 7.5). The scatter plot indicated a positive relationship between both variables. A slightly curved point cloud is revealed. The higher the car ownership levels, the broader the range of HH surface areas at which the car ownership levels occur. The correlation between both data sets is 0,77, so the coherence between the variables is fairly strong. Around 4.400 points have been plotted, resulting in a lot of overlapping points due to the way car ownership data is available, per decimal. To visually analyze the data, it might be helpful to add a bit of random variance to the car ownership levels. This reduces the overlapping of the

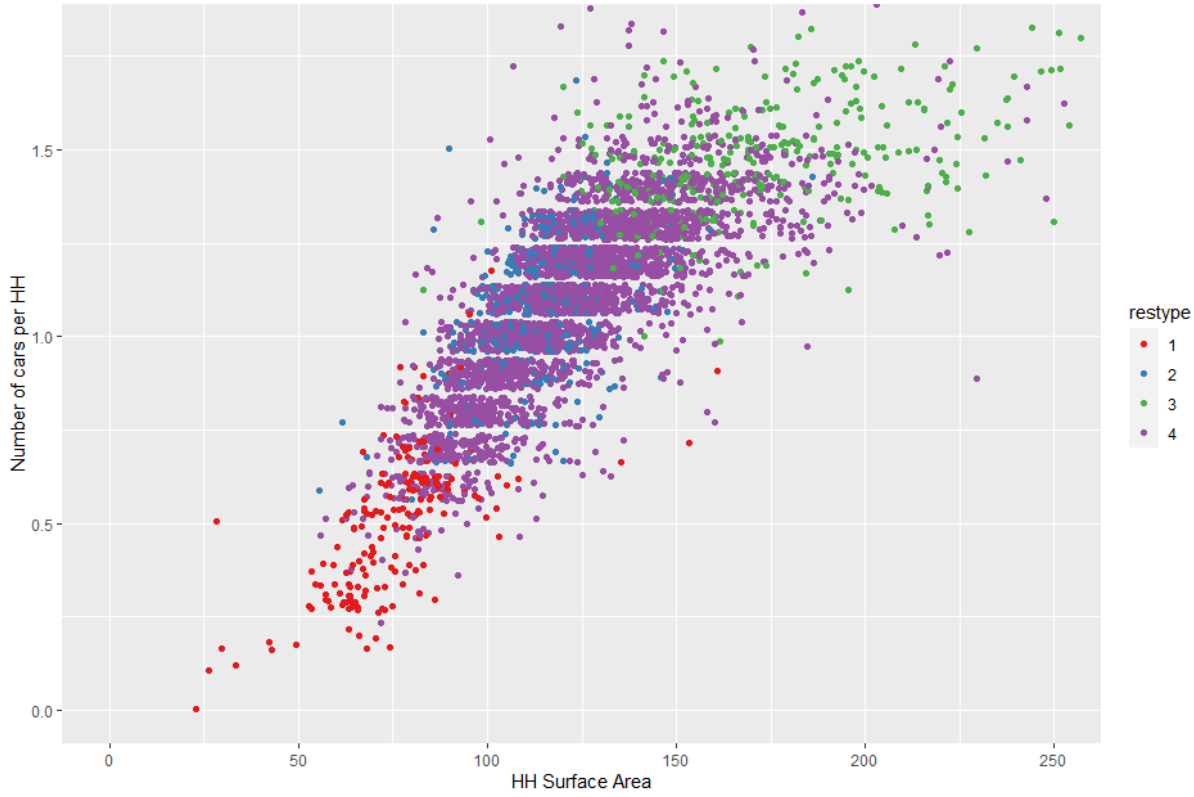


Figure 7.6: Number of cars, HH surface area and dominant residence type

points. Therefore, in the scatterplots of Figure 7.6 and 7.7, some random noise has been added to the car ownership levels data.

In Figure 7.4, the number of cars per HH are plotted against HH surface and different points are colored by the dominance of a residence type in the district. The color of the points refers to the following district characteristics:

1. More than 70% of the VBO surface area is multi-family
2. More than 70% of the VBO surface area is terraced
3. More than 70% of the VBO surface area (semi-)detached
4. No dominant residence type is present in the area

Several things can be deducted from Figure 7.6. The districts in which multi-family houses or (semi-) detached houses are dominant are extremes in the scatter plot, both in terms of car ownership, as in terms of HH surface area. It seems that terraced dominant districts are a bit more present in the left part of the scatterplot, what could indicate that terraced houses have relatively more car ownership than other residence types with the same HH surface area. Despite the aforementioned hypothesis, the districts in which one residence type is dominant do not seem to have an evident difference in car ownership compared to other districts with the same comparable surface areas. The extremes as shown in Figure 7.6 are as expected, based on HH surface area. It is therefore expected that using residence type as an additional predictor to estimate car ownership will not result in a large improvement, but a slight improvement of the prediction is likely.

Another variable that is often related to car ownership is urban density. In Figure 7.7, the surrounding address density from CBS district data has been added as a colour to the plot. Numbers 1 to 5 are indicators for the surrounding address density (or urban density), reaching from more than 2.500 addresses per km^2 to less than 500 addresses per km^2. Based on the figure, it cannot be concluded to what extent the urban density could improve the estimation of car ownership. However, it can be concluded that in general, for houses with the same surface area, a decreased urban density means an

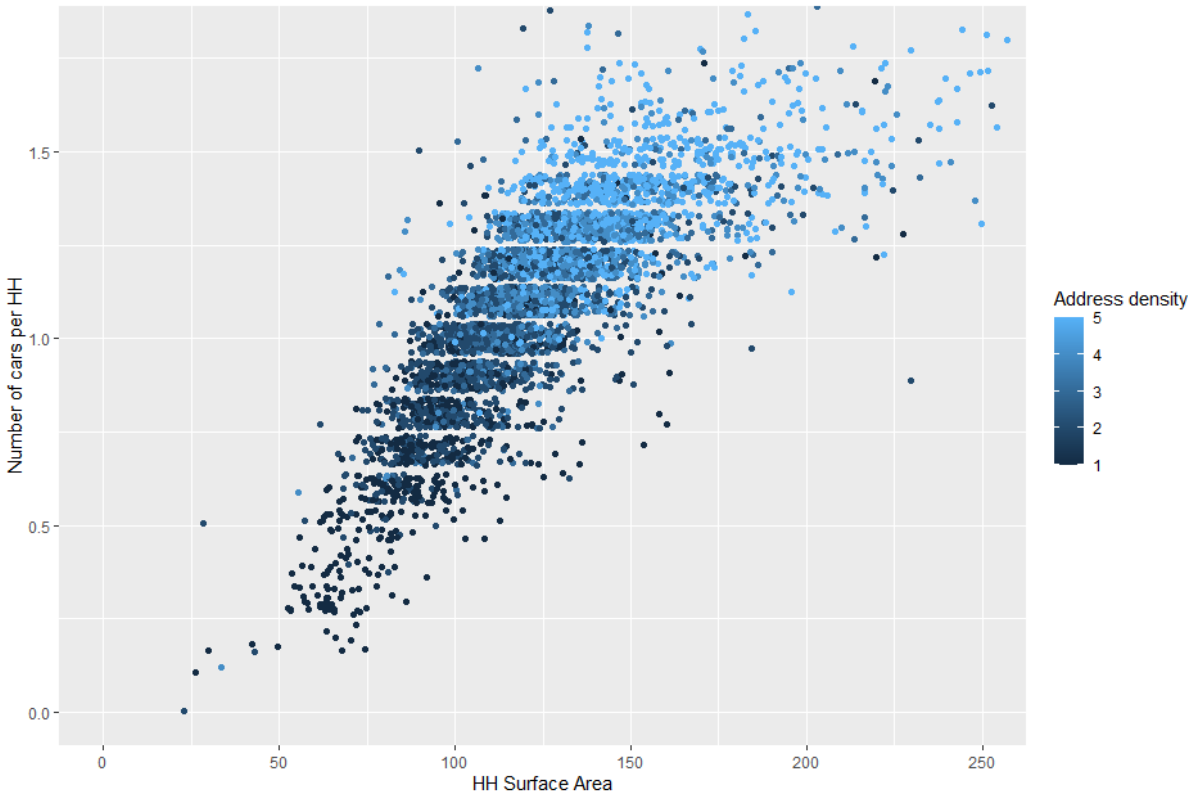


Figure 7.7: Number of cars, HH surface area and surrounding address density

increased car ownership level. It is likely that including the variable urban density will decrease variance of the estimation. The similarities with the plot in Figure 7.6 are apparent. HH surface area and surrounding address density are negatively corelated with each other by 0,69. Thus, adding urban density as a predictor for estimating car ownership might increase the explained variance, but the increased precision of the estimation will likely not be too large, as HH surface area partly explaines urban density. It is a logical given that districs with larger houses have a lower address density. And the closer you get to the center in an urban environment, the more compact the houses will generally be.

A final variable that has been be considered is income. An increase of financial means will increase the potential of owning at least one, and perhaps even more cars. A problem however is the availability of income data by the CBS. Data on income is available only in neighbourhoods with a minimum number of households due to privacy reasons. Table 7.1 shows that income data is available in 1307 of the 4419 districts. 1307 neighborhoods is still quite a lot, but the problem is that the household size in these districts is relatively low. Districs with larger houses simply have fewer households. As a result, the neighborhoods with income data available are not representative. Therefore, this variable is not included in the regression analysis.

7.1.3 Regression analysis

To estimate car ownership based on BAG, the average HH surface area at district level has been determined to be the most significant predictor from BAG. The variables housing type and urban density will be added to determine how much the estimation can be improved. First, a regression analysis is carried out with HH surface area as a single predictor. The best linear fit is sought after, and it is determined how well HH surface area is able to predict car ownership. Subsequently housing type and surrounding address density are added as a categorical variable and respectively a simple linear variable to estimate whether, and how much car ownership predictions are improved.

Single linear regression

The slightly curved shape of the scatter plot indicates that a curved function might be able to fit the data best. A simple linear, quadratic, cubic and logarithmic function have been fitted to the data, of which the cubic function was found to be the best fit, with an R-squared value of 0,68. Furthermore, a weighted and non-weighted regression have been carried out, based on the sample size of each data point; the number of BAG VBO *living* present. Although the weighted regression achieved slightly better results, the non-weighted regression is assumed to predict car ownership more realistically. With the non-weighted regression, districts with a relatively small number of houses present are just as important as large districts. It is expected that in smaller districts with a high average VBO *surface area* (Figure 7.6) more detached and semi-detached houses are present.

In Figure 7.9, the residuals of both a simple linear fit and a cubic linear fit have been plotted against the models fitted values. The residual plot of the simple linear fit strongly indicates a curved relationship between the car ownership and HH surface area. Both for the lower and higher car ownership levels, the model largely overestimates its predictions. The residual plot on the right shows an almost linear fitted line. The lower values are slightly underestimated, whereas higher values are slightly overestimated. In

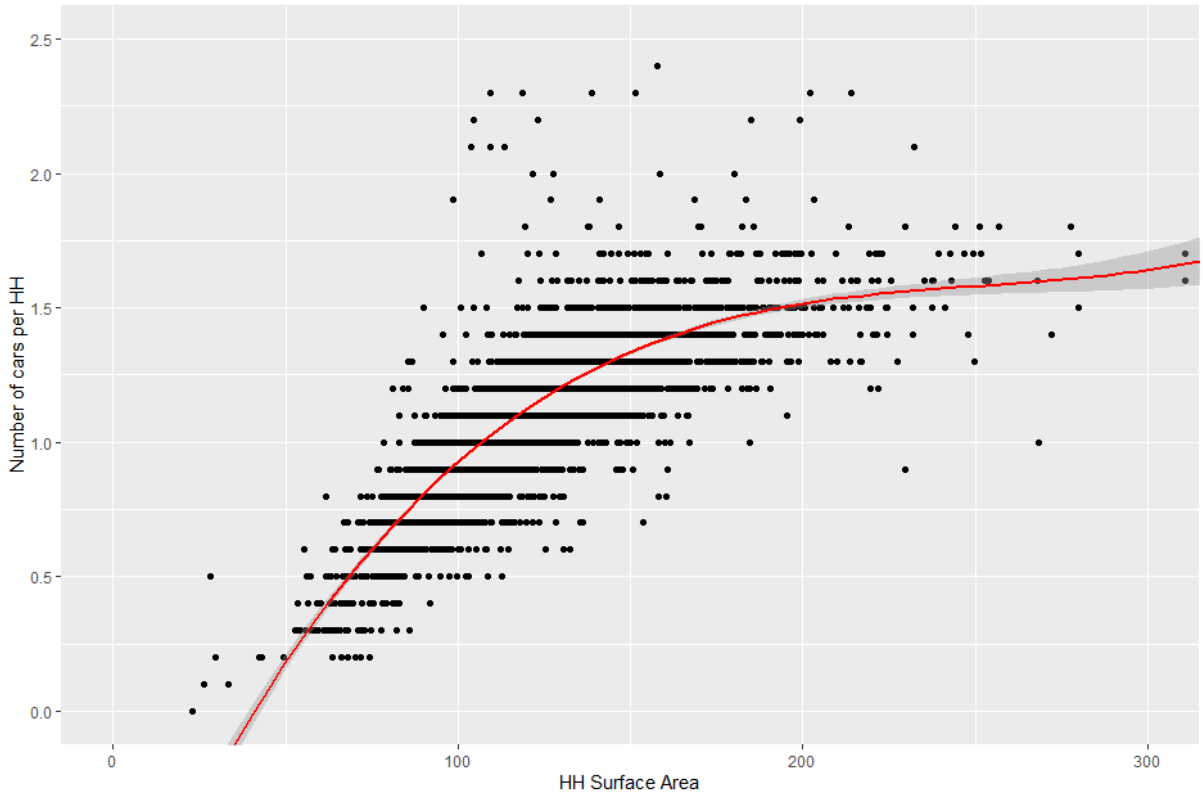


Figure 7.8: Cubic regression line predicting car ownership based on HH surface area

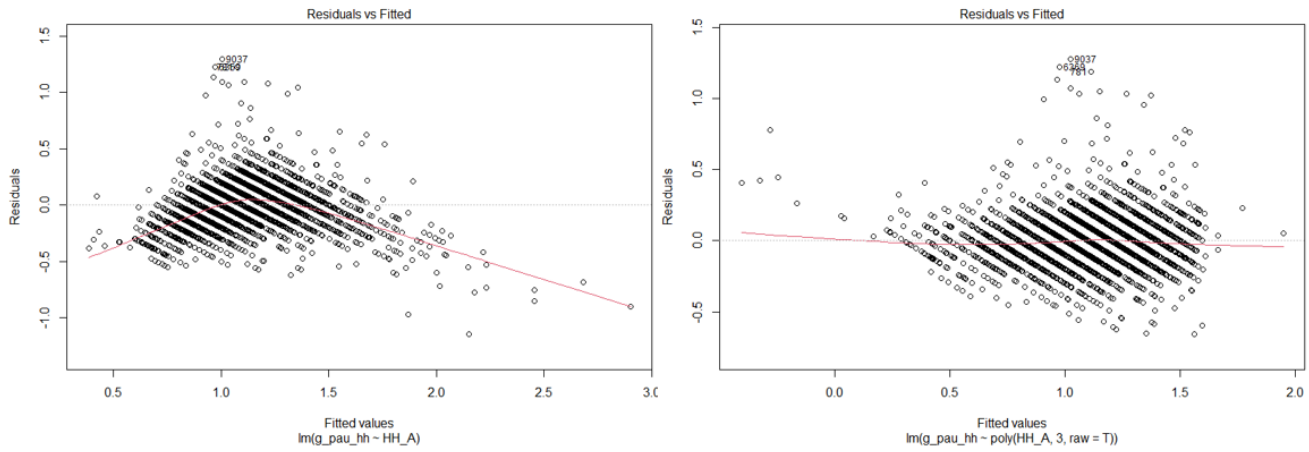


Figure 7.9: Residuals vs. Fitted comparison between simple linear fit (left) and cubic linear fit (right)

general, it can be said that the cubic fit is more precise. It is not likely a better fit can be achieved for the variable HH surface area. In Figure 7.8, the cubic regression line has been plotted.

A significant regression equation was found with a p-value < $2{,}2e^{-16}$ and an R-squared value of 0,68. On average, the model deviates from actual car ownership level with a residual standard error of 0,16. The following regression equation has been developed:

Fit 1: HH Surface area

$$CarOwnershipHH = -1{,}02 + 0{,}03 \cdot HH_A - 1{,}12 \cdot 10^{-4} \cdot HH_A^2 + 1{,}48 \cdot 10^{-7} \cdot HH_A^3$$

Based on the P-values of the coefficients it is determined whether the variables used for the regression can be included or must be rejected. A significance level of 0,05 must be reached. The P-values of all coefficients are far smaller than 0,01 and therefore statistically significant.

Although residual plot indicates a good fit, still a large amount of the variance is not explained by household surface area.

Multiple linear regression

To try to decrease the variance of the prediction, more variables are added to the regression analysis. First, residences types are included, and subsequently, it is analyzed whether urban density might improve predictions.

In the process of adding multiple variables, the fit of HH surface area has been reevaluated. It has been found (based on the R-squared values) that the cubic fit is under any circumstance resulting in the best predictions. Urban density is included through the variable CBS surrounding address density (Table 7.1), which is a continuous variable indicating the address density per square kilometer within 1 kilometer of a district. The variable is treated as a simple linear variable. Residence type has been included by introducing three categorical variables:

1. Multi-family dominant
2. Terraced dominant
3. (Semi-) detached dominant

If one of the above variables is equal to 1, more than 70% of the HH surface area in the district belongs to the specific residence type. If none of above variables in 1, the district contains mixed residence types. The regressions were performed in the following compositions with the following results:

Fit 2: HH Surface Area + Residence type

A significant regression equation was found with a p-value < $2{,}2e^{-16}$ and an R-squared value of 0,70. On average, the model deviates from actual car ownership level with a residual standard error of 0,15. The following regression equation has been developed:

In districts with no dominant residence type present:

$$CarOwnershipHH = -7{,}7e^{-1} + 2{,}6e^{-2} \cdot HH_A - 1{,}0e^{-4} \cdot HH_A^2 + 1{,}3e^{-7} \cdot HH_A^3$$

In districts with mainly (semi-)detached residences:

$$CarOwnershipHH = (-7{,}7e^{-1} + 1{,}1e^{-1}) + 2{,}6e^{-2} \cdot HH_A - 1{,}0e^{-4} \cdot HH_A^2 + 1{,}3e^{-7} \cdot HH_A^3$$

In districts with mainly multi-family residences:

$$CarOwnershipHH = (-7{,}7e^{-1} - 1{,}3e^{-1}) + 2{,}6e^{-2} \cdot HH_A - 1{,}0e^{-4} \cdot HH_A^2 + 1{,}3e^{-7} \cdot HH_A^3$$

In districts with mainly terraced residences:

$$CarOwnershipHH = (-7{,}7e^{-1} + 1{,}8e^{-2}) + 2{,}6e^{-2} \cdot HH_A - 1{,}0e^{-4} \cdot HH_A^2 + 1{,}3e^{-7} \cdot HH_A^3$$

The three included categorical variables affect the height of the intercept in the equations. In districts with mostly detached residences present, car ownership levels are relatively 0,11 higher. In districts with mostly multi-family residences present, car ownership levels are relatively 0,13 lower. For terraced-dominant districts, car ownership only slightly deviates from mixed-districts car ownership levels.

Adding the residence types resulted in a slight improved fit, with an increased R-squared value of 0,02 and a decreased residual standard error of 0,01 compared to Fit 1.

Fit 3: HH Surface Area + Residence type + Address density

As a third variable, address density is added to the regression equation. This led to a significantly improved fit with a p-value < 2,2e-16 and an R-squared value of 0,74. On average, the model deviates from actual car ownership level with a residual standard error of 0,14. Thus, besides the BAG attributes, urban density contributes to a reduced variance in the car ownership estimations. Therefore, when predicting car ownership levels based on BAG attributes only, it may be expected that in cities with different levels of urban density, estimated car ownership levels on average will deviate from actual car ownership levels.

In the next section, the results of Fit 2 will be explored for different types of cities.

7.1.4 Car ownership predictions

To estimate the predictive capability of Fit 2 under different circumstances, the predictions are examined in four different cities: Rotterdam, Haaksbergen, Groningen and Wassenaar. Rotterdam and Haaksbergen are included based on their differences in urban density, and Groningen and Wassenaar are included based their political orientation. Rotterdam is one of the largest cities in the Netherlands, with around 620.000 residents. Haaksbergen is a small, rural village with around 25.000 residents. In Wassenaar, 44% of residents voted for the VVD, a large right-wing party in the Netherlands (often called the ‘car party’ of the Netherlands). In Groningen, the largest party in the city council is GroenLinks, a left-wing party with sustainable aspirations. Groningen has a large university which could also be the reason Groenlinks is as large as it is, due to its popularity among students.

According to CBS Districts 2018, the average number of cars per HH in the Netherlands is 1,10. According to the estimated BAG fit, an average of 1,11 cars per HH is present, a difference of 1% with an Root-Mean-Square Error (RMSE) value of 0,20. In Table 7.2, statistics of predicted and actual

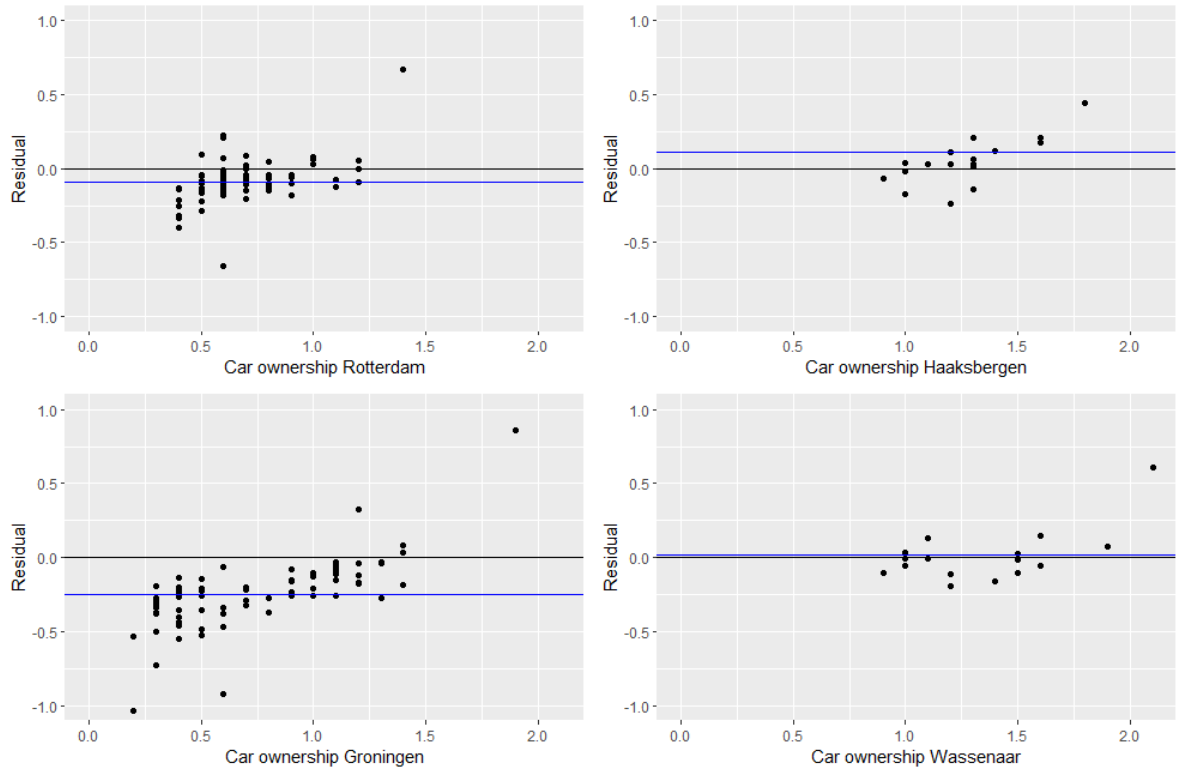


Figure 7.10: Residual plot number of cars per HH at District level in four Dutch cities

averages of the four cities are presented. In Figure 7.10, the residuals of each estimation in a city have been plotted.

In the city of Wassenaar, estimations are relatively least spread out, and the number of cars per HH is best estimated. The average number of cars per HH in Wassenaar is highest. Compared to Groningen, car ownership levels are significantly higher, which was expected due to the political orientation of the residents being more left-winged. In Groningen, the number of cars per HH is most underestimated, with a large average of -0,25 cars per HH. The average HH surface area in Groningen is 113 m^2, which is much larger than the average HH surface area in Rotterdam, which is 92 m^2. Groningen and Rotterdam both have a similar number of cars per HH, while in Rotterdam the average residual is -0,09. Thus, in Groningen, the HH surface as a predictor for number of cars per HH is less effective. The differences between Rotterdam and Haaksbergen are obvious as well. The number of cars per HH in Haaksbergen are much higher on average. The number of cars is slight underestimated by the fit.

Table 7.2: Statistics cars per HH four cities

City	Actual cars per HH	Predicted cars per HH	Average residual	RMSE
Rotterdam	0,69	0,77	-0,09	0,18
Haaksbergen	1,33	1,22	0,11	0,20
Groningen	0,73	0.98	-0,25	0,35
Wassenaar	1,36	1,34	0,01	0,17

In three of the four cities, Rotterdam, Haaksbergen and Wassenaar, the model predictions are reasonably well, with RMSE values between 0,17 and 0,20. In Groningen, the models largely overestimated the number of cars per HH. It might be that both the political orientation as the number of students present in the city make that the model was not able to capture the lower values of Groningen.

7.1.5 Concluding

A model has been developed that enables predicting the number of cars per HH at District level. The main predictor variable is the average HH surface area of residences at District level. The model is able to predict the number of cars per HH with an RMSE value of 0,2 cars per HH. In different cities, predictions on average can deviate. In practice, it could be considered to adjust predictions for cities in which the average residuals deviate.

The created model can be used to subdivide large zonal data such as District data into smaller zones, or to predict the number of cars per HH for new housing developments. For new housing developments, the estimated prediction errors need to be seriously considered. For subdividing large zones, the predictions of the models can be used without hesitation. The prediction based on District level could lead to biased results when drastically changing the level of aggregation. By no means, the developed model can be applied to individual households. However, for estimating the number of cars per HH in zones that are perhaps two, three times smaller than the average district, results should remain trustworthy.

7.2 Residents

In this paragraph, the process and results of estimating the number of residents on BAG data attributes are described. The goal is to determine how precise BAG predictions of residents could be, and how these predictions deviate in different cities. These results can help estimate how BAG could contribute in estimating the number of residents for new developments and for smaller zones. Considering the use of residents as a trip generation factor in trip generation models, resident can be included in different ways, such as the total number of residents per zone, residents per HH or residents per age group. In this analysis, the total number of residents per zone is considered. The hypothesis is that both the surface area as the residence type from BAG are important factors to determine the average number of residents per household. By multiplying this number with the number of households in a zone, the total number of residents in a zone is achieved.

The regression analysis will be carried out at PC6 level, with CBS PC6 2016 data (Appendix D). At PC6 level, it is possible to find enough zones in which one residence type (terraced, multi-family, detached, semi-detached) is present to enable separate regression analyses for these residence types. Then, based on the established relationship between the BAG attributes and the number of residents, predictions are made at District level, and compared with CBS Districts 2020 data, the most recent CBS residents data. By reviewing the results in different city types, it is estimated how reliable BAG predictions are in different environments. Eventually, the results of this analysis can be used for two purposes:

- Distributing highly aggregated resident data into smaller zones
- Predicting resident data for new housing developments

7.2.1 The data

Data used for the prediction are (Appendix A and D):

- BAG VBO *living*
- CBS PC6 2016
- CBS Districts 2020
- CBS linking table

For the regression analysis, the average number of residents per HH (household) at PC6 level will be predicted by the average HH_A (household surface area) and the residence type. For the analysis, only BAG VBOs *living* with one function have been included. To each VBO record, CBS PC6 data is joined

based on the PC6 postal code present in both data sets. The residence types terraced and end-of-terraced are merged into one housing type. The assumption here is that, on average, no more or less residents live in a corner house than in a terraced house. Because BAG data originates from 2021 and PC6 data from 2016, the housing market could have evolved significant in many PC6 areas. Therefore, PC6 areas with a housing difference between BAG and CBS larger than 3 houses have been removed from the analysis. Of each residence types, the number of records is presented in Table 7.3.

Table 7.3: Division VBO *residence* type

Type	Records
Detached	929.744
Semi-detached	680.296
Terraced	3.444.591
Multi-family	2.753.233

To enable regression analysis for the different residence types, PC6 areas have been selected in which one residence type is uniformly present. In Table 7.4 the properties of the data that is used in the analysis is presented.

Table 7.4: Summary statistics regression data residents per HH

Type	PC6 Records	Housing mean	HH_A mean	HH_A SD	HH_A median	Res. mean	Res. SD
Detached	13.724	10,45	202	59	193	2,55	0,55
Semi-detached	3.718	12,32	132	35	128	2,49	0,59
Terraced	104.327	17,08	115	25	113	2,47	0.51
Multi-family	61.650	24,37	77	20	76	1,66	0,43
All	**183.492**	**18,94**	**109**	**44**	**104**	**2,21**	**0,63**

From the statistical summary in Table 7.4, the following can already be deducted. The household surface area averages for detached, semi-detached and terraced residences are significantly different, while their residents mean are close. Thus, for these residence types, predicting the number of residents based on the average HH_A should result in different outcomes. For multi-family residences, both the average

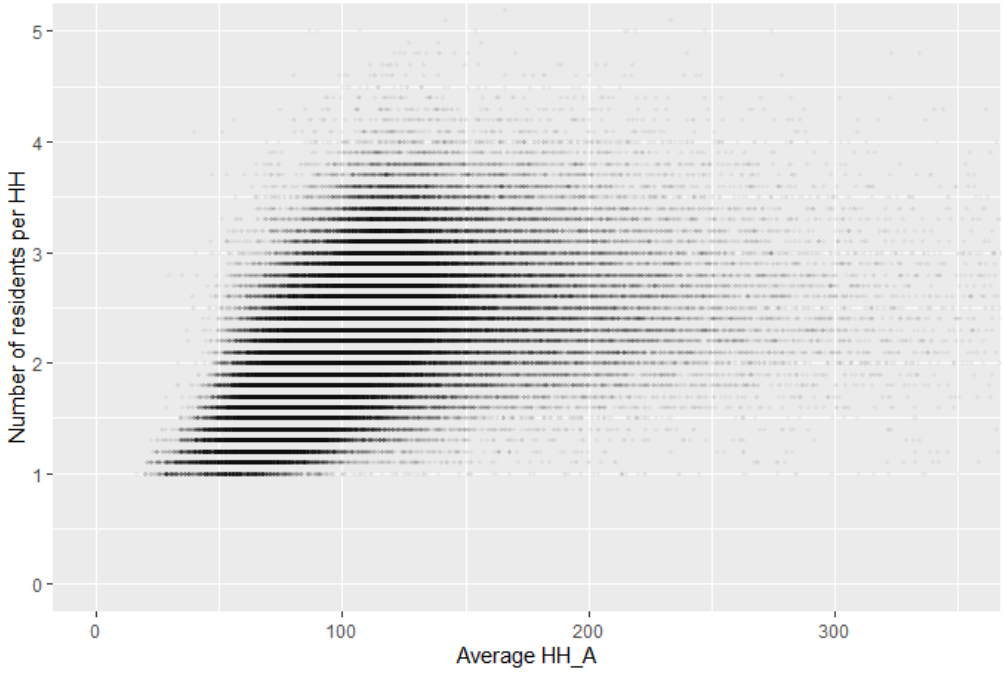


Figure 7.11: Scatterplot Number of residents per HH and Average HH_A at PC6 level

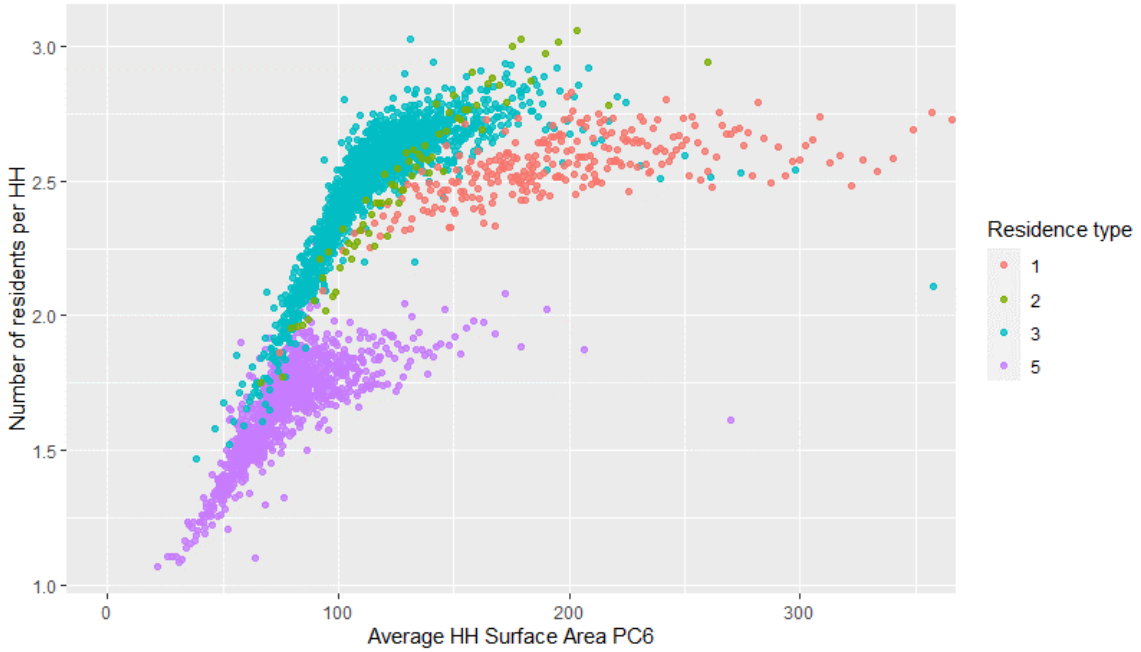


Figure 7.12: Scatterplot number of residents per HH and average HH_A, binned (50) (1 = detached, 2 = semi-detached, 3 = terraced, 5 = multi-family)

HH_A as the number of residents is lower compared to the other. It cannot be stated based on this data, whether predictions based on multi-family residences would behave differently compared to the other.

After regression analysis at PC6 level, predictions have been made at District level. CBS 2020 District data contains the most recent open resident data available in the Netherlands. BAG data originates from the 1st of January 2021. With the CBS link table data, BAG prediction results are aggregated to District level. The differences in number of residences will be accounted for during prediction, by introducing a correction factor per District. Only districts with a population density larger than 100 residents per km^2 have been included. For this analysis, the main interest for the determination of the number of residents in urban areas lie in which a density of 100 residents per km^2 or more is present.

7.2.2 Regression analysis

In Figure 7.11, the number of residents per HH and the corresponding average HH_A of all uniform PC6 areas have been plotted. With a correlation of 0,47, the coherence between these two variables is relatively weak. For lower values of Average HH_A, a positive relationship with the number of residents per HH is observed. For values above 150 m^2, the number of residents per HH seems to stay at the same level, with perhaps a slight decline reaching 300 m^2. However, with this large number of records and the significant spread in the data, it is difficult to estimate the exact shape of the relationship with the two variables. Furthermore, it would be interesting to learn how the residence type affects the relationship between both variables.

Therefore, the data of the four residence types has been equally binned by 50 data points per bin. This reduces the variation in the data, and the shape of the relationship between Average HH_A and the number of residents per HH becomes more explicit. In Figure 7.12, the result of this process is presented. As is clearly visual, the residence type affects the relationship between HH_A and number of residents per HH. A few things can be deducted from the figure. In PC6 areas with terraced and semi-detached houses, a higher number of residents per HH is present compared to PC6 areas with detached houses, at equal average HH_A values. Furthermore, the average number of residents in multi-family houses at PC6 level with an average HH_A value of 100 or more are much lower compared to the other residence

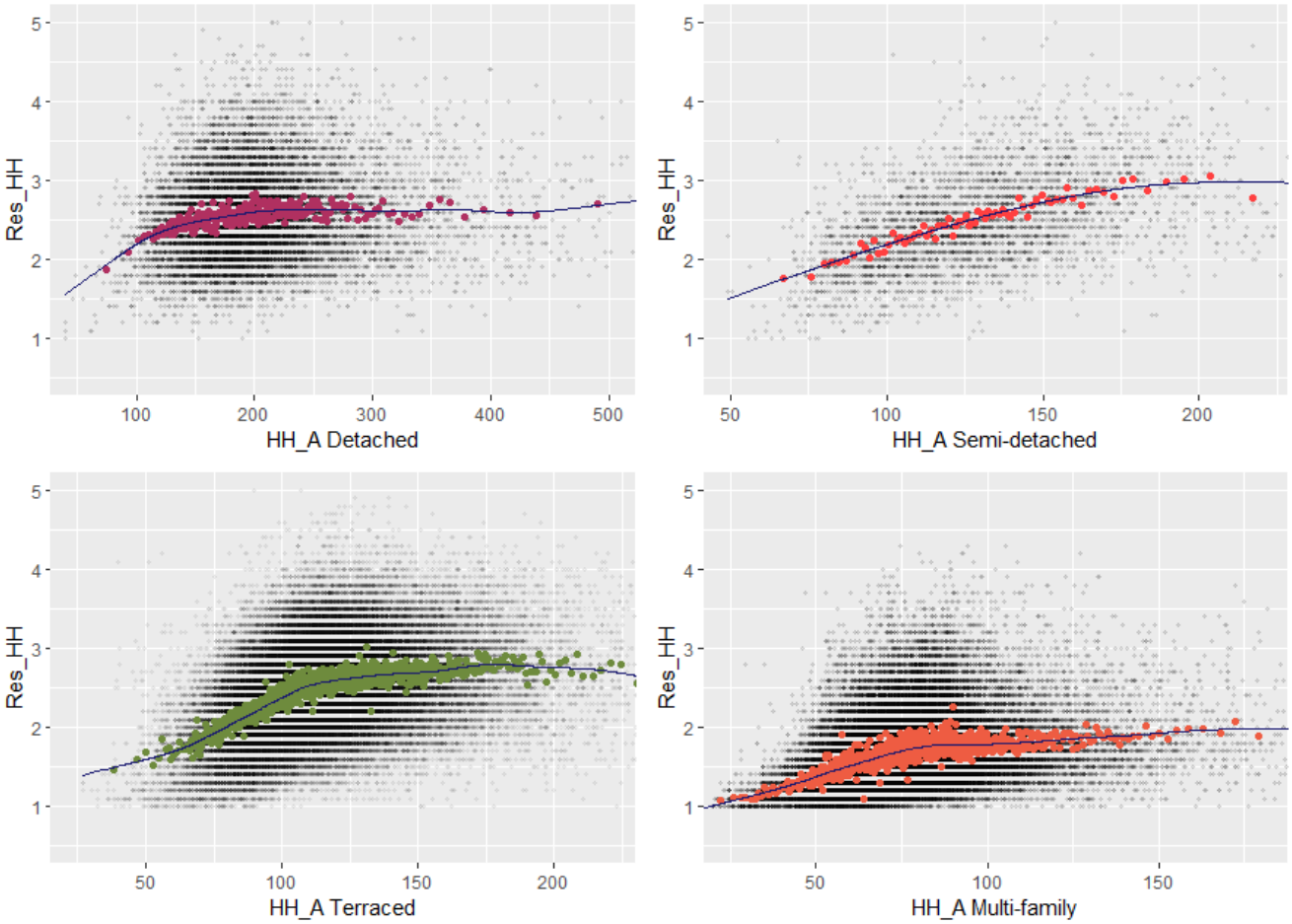


Figure 7.13: GAM fits for predicting the average PC6 number of residents per HH (RES_HH) based on the average HH surface area (HH_A) by residence type at PC6 level

types. Based on Figure 7.12, it can be determined that including residence type will increase the quality of predictions.

To predict the number of residents per HH at PC6 level, four separate models have been fitted for the different residence types. In Figure 7.13, the data for each residence type is plotted, including the bins (each averaging 50 data points) and the estimated fits (when comparing the plots, keep in mind the deviating x-axis scales). The fits have been developed with the goal of predicting the number of residents as good as possible. This resulted in the use of generalized additive models (GAM). For GAM, no relationships between the variables have to be assumed a priori. When looking at the bin plots in Figure 7.13, the data follows different patterns that do not strongly indicate a specific fit shape. Several fits have been explored up to fourth-degree polynomials. Especially at the lower and higher-ends of the HH_A variable, the number of residents were consistently over- or underpredicted. The residuals are not normally distributed, which is a prerequisite for using ordinary least regression models. Therefore, GAMs have been applied, leading to the best predictive fits, by being able to follow the pattern revealed by the data.

A GAM combines multiple different base models to create a single model that is able to smoothly follow the patterns in the data, while balancing overfitting and underfitting. Overfitting is prevented by penalizing the wiggliness of the fit, the extent to which a model changes direction. The method used for determining wiggliness for the estimated fits in Figure 7.13 is the Generalized Cross-Validation method (GCV). This method divides the analysis data repeatedly into a large modelling subset and a small validation subset and tests overfitting. The GAMs in the plots are developed on all data for each residence types. The binned values have been added to visually estimate the goodness of the fits. As can

be seen, the developed models are well-fitted to the data. In Table 7.5 the statistics of the models are presented.

Table 7.5: Statistics of the estimated GAMs

Model	edf	R-squared	RMSE
GAM1	9,1	0,03	0,53
GAM2	4,1	0,27	0,50
GAM3	15,5	0,18	0,46
GAM5	12,4	0,11	0,41

The edf (estimated degrees of freedom) value is an indicator for the complexity of the model. The GAM2 model is the least complex, following a relative constant path through the data, whereas the GAM3 model is the most complex, changing form and direction more often. The low R-squared values are expected, as a large spread can be observed in the number of residents per household at PC6 level. On average, the residuals of the GAM5 model deviate least from its actual values, with a RMSE value of 0,41. For detached houses, the highest RMSE is noted. The lower performance of the prediction of detached houses is not unexpected. As can be seen in Table 7.5, the average number of houses per PC6 area for detached residences is around 10, much lower compared to terraced and multi-family houses. That means that for terraced- and multi-family residences, variance in the number of persons per household is averaged out more in the data used in the regression analysis, compared to values for detached residences.

For now, the large observed variance in the prediction of the developed models is not a cause for concern. The models will be used to predict the number of residents in model zones of a higher aggregation level. To test the predictive ability of the models, estimations have been made at District level.

7.2.3 Residents prediction at District level

Predictions are made as follows; all VBO *living* records with one VBO function are gathered. For VBO *living* objects with multiple VBO functions, it cannot be determined to what extend the included surface area belongs to what function. Then, at PC6 level, the average HH_A for each residence type is calculated. To these values, the models are applied and the total number of residents is summed at PC6 level. Then, with the CBS linking table, the PC6 data is aggregated to District level. Some PC6 areas are present in two or more districts. These areas will be assigned to one randomly picked District. This affects the reliability of the predictions to some extent, but not too much. In the Netherlands, 445.000 PC6 areas and 13.808 Districts are present, resulting in 32 PC6 areas per District on average.

After aggregating the predictions to District level, the number of residents will be corrected by the percentual difference between the number of BAG houses and the number CBS District 2020 houses. There are two reasons these numbers deviate. First, only houses with one VBO function are included in BAG, and second, the BAG data originates from 2021. When the difference in houses are corrected for, the number of residents predicted by BAG and the actual number of residents provided by the CBS can be compared.

According to CBS Districts 2020, a total of 17.389.915 residents are present in the Netherlands. With the estimated BAG models, a total of 17.353.682 are predicted, a difference of 0,2%. The average number of residents per HH predicted at District level is 2,22, with a RMSE value of 0,33. As expected, the uncertainty of the predictions are smaller at District level than as found at PC6 level (Section 7.2.2).

To estimate the predictive capability of the estimated models under different circumstances, the predictions are examined in four different Dutch cities: Enschede, Barneveld, Lochem and Utrecht. In Table 7.6, statistics of predicted and actual values per city are presented. In Figure 7.14, residuals of the estimated number of residents per HH at District level are presented per city.

Table 7.6: Statistics residents four cities

City	Actual residents	Residents predicted	Average residual residents per HH	RMSE residents per HH
Enschede	146.930	148.613	- 0,04	0,27
Lochem	12.190	13.577	- 0,28	0,31
Barneveld	43.540	40.328	+ 0,13	0,27
Utrecht	353.380	307.219	+ 0,23	0,33

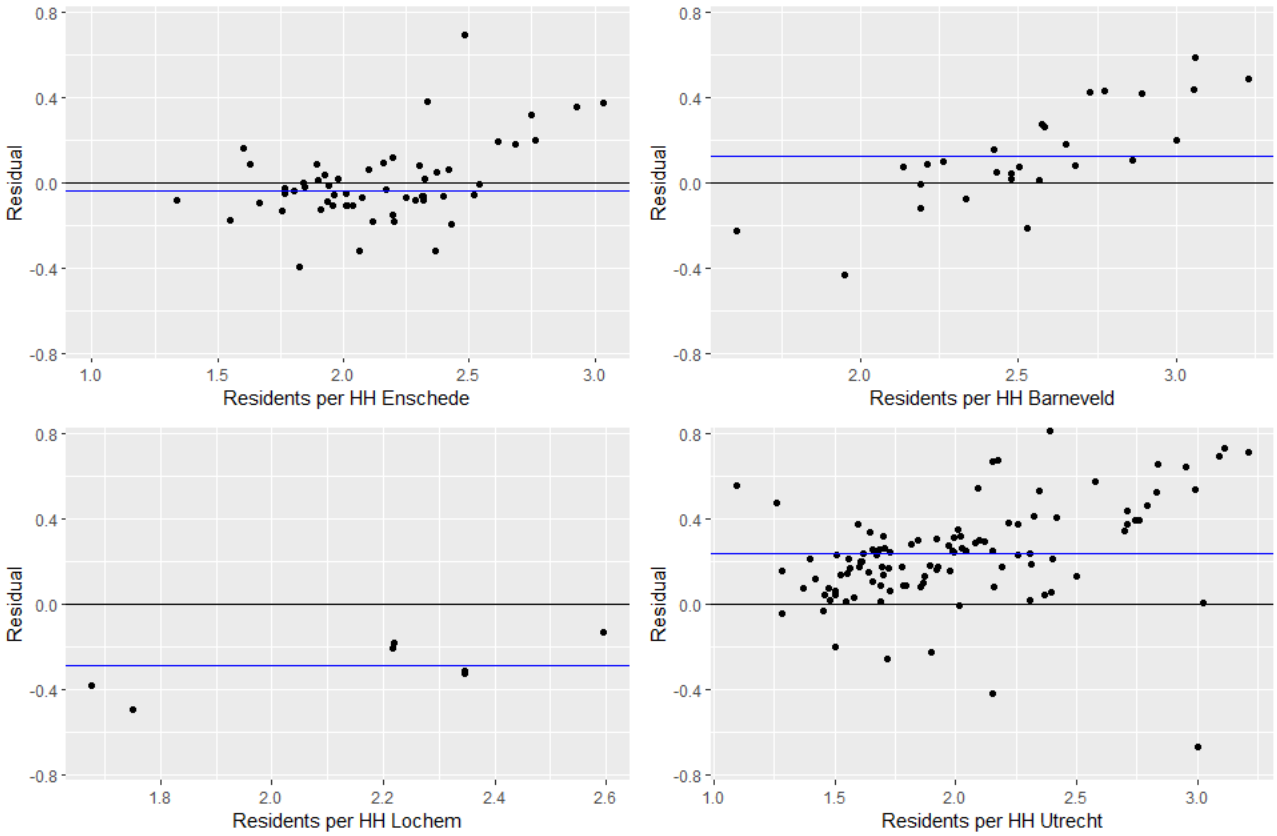


Figure 7.14: Residual plot residents per HH at District level in four Dutch cities.

Lochem is a municipality with a relatively large amount of residents aged 65 or above. Barneveld on the other hand is one of the youngest municipalities, with almost 35% of the residents aged between 0 and 25 years old. Utrecht is one of the large cities in the Netherlands, with a large population in a dense environments. In terms of age or city size, Enschede does not stand out in particular.

In the city of Enschede, predictions are relatively the least spread out, and the total number of residents is best estimated. In Barneveld, the number of residents per HH is overall underestimated, with 0,13 residents per HH on average. The large number of residents between 0 and 25 indicate in Barneveld indicate the presence of many or large families, leading to relatively more residents per square meter HH. In Lochem, the number of residents per HH is overestimated with 0,28 residents per HH. The high-age levels of the city indicate the presence of many one- or two-person households, leading to relatively fewer residents per square meter HH. In Utrecht, the number of residents is largely underestimated, with 0,23 residents per HH. In the large city, space is scarce and the residences are relatively small with an average surface are of 97 m^2, compared to a 119 m^2 national average. Despite the smaller residences, the city is attractive for families to reside. This leads to an underestimation of 46.000 on a total of 353.000 residents in the city of Utrecht.

7.2.4 Concluding

Four models have been developed based on four different residence types, that enable predicting the number of residents in a PC6 area, using GAM. The main predictor variable is the average HH surface area of residences at PC6 level. The models have been built based on data from 2016, and have been applied to data from 2020. When correcting for the number of residences, the models were able to accurately predict the total number of residents in the Netherlands, with only a difference of 0,2%. That means that using the models based on 2016 CBS data for current and near-future predictions will not affect prediction uncertainties.

At PC6 level, predictions on average deviate with an RMSE of 0,4 to 0,5 residents per HH. At District level, this is reduced to 0,33. Thus, at higher aggregation levels predictions become less uncertain. Furthermore, it was found that for different types of cities, predictions on average can deviate. In practice, it could be considered to adjust predictions for cities in which the average residuals deviate. The created models can be used to subdivide large zonal data such as District data into smaller zones, or to predict the number of residents for new housing developments. For new housing developments, the estimated prediction errors need to be seriously considered. For subdividing large zones, the predictions of the models can be used without hesitation.

Chapter 8. Trip generation based on BAG: a case study in Ede

In order to answer RQ3, to display how the use of BAG data could affect trip generation estimates, and to understand how this affects traffic, a case study is carried out in the municipality of Ede. In the east-part of the town Ede, an entire neighborhood is being developed, with approximately 1.200 new residences being constructed in a couple of years. According to the municipality, the trip generation in that part of the city is fairly underestimated. BAG data is up-to-date and it is even possible to obtain data for residences for which a construction permit has been applied for. These residences will be put into use sometime within four years . The actuality of BAG data should solve the trip underestimations of the municipality of Ede, and the residences that are already registered in BAG but not yet built should even make it possible to estimate the number of trips in the near future with reasonable accuracy. For these residences, the type and surface area of the residences are already known, which are data attributes of which it was found during answering RQ2 that the number of residents and car ownership per household can be determined. And besides determining how the trip generation underestimation of Ede can be solved, the results of estimates based on BAG will be compared to trip generation estimates based on the most recent conventional open data source: CBS PC6 2016. Comparison of both results can lead to the conclusion to what extent BAG increases the potential of open data to be used for trip generation modelling.

First, the characteristics of the case study are briefly described. Then, the method to determine the trip generation and the traffic on surrounding roads are explained after which the obtained results are discussed.

8.1 Veluwse Poort

In Figure 8.1, the district Veluwse Poort is displayed in the town of Ede in which the new residences are being developed. The district is adjacent to the Klinkenbergerweg, the main distributor road on the east side of Ede. The residences currently built are all concentrated in the north part of the district. In Figure 8.2, BAG VBO *living* currently in use and VBOs for which a permit has been granted are presented. The construction of residences on the east-side of the district is almost finished, whereas progress at the westside has to be made. At the 1st of January 2021, 793 VBO *living* were present in BAG with the status *in use*. Furthermore 164 VBO *living* were present with the status *formed*, which means that in the nearer future, a residence is being finished. If all projects are finished, a total of 1.200 residences will be present which means that on short notice, another 250 VBO *living* will be additionally registered in BAG. The growing number of residences naturally has a large impact on the traffic situation in the immediate vicinity of the district. The addition of 1,200 houses in a place like Ede (75.000 citizens) is a relatively large change.

8.2 Estimating car trip generation and surrounding traffic intensities

The purpose of the case study is to illustrate the potential of BAG by estimating trip generation for the development in Veluwse Poort. The estimated trip generation will be compared with prognoses based on the most recent available conventional data (which the municipality of Ede has access to), and the most recent open data source. First it is described how the trip generation will be estimated and subsequently the case study is developed.

For the case study, the motorized traffic generated by the households in the area will be analyzed. A possible way of estimating trip generation of the households is applying trip rates published by CROW publication 381 (CROW, 2018). These rates are developed with the purpose of predicting trips of new developments, for which for example it is not known how many cars will be present in the area. The CROW makes a distinction between housing types, which is an indication of the income of the owners, and therefore of the number of cars they might possess. Thus, the basis of the developed trip rates by

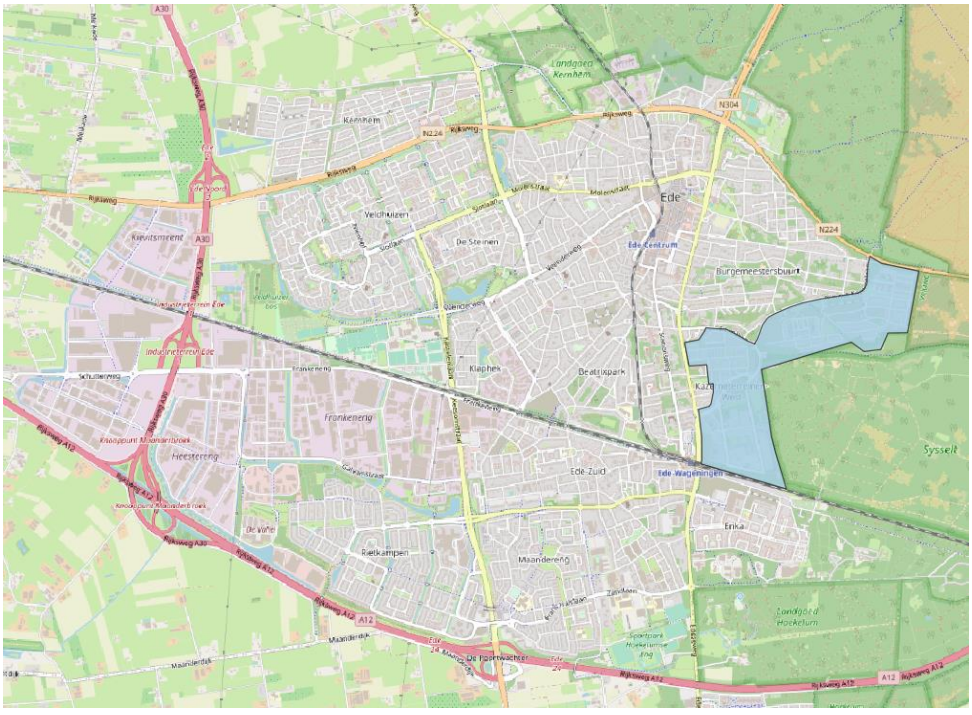


Figure 8.1: District Veluwse Poort in Ede indicated in Blue

the CROW, and the most important factor, is car ownership. Therefore, the trips generated will be based on car ownership. There is already data present of car ownership in the district, and based on BAG, future car ownership can be determined. In the Netherlands, a rate of 6,1 trips per car per day can be applied to predict the number of motorized trips per residence to estimate the trip generation of a household (Brouwer et al., 2014). For this case study, this trip rate is assumed to be linear. The formula to determine the number of generated trips in a zone is:

$$Trips\ generated = CarOwnership_{HH} * Number\ of\ HH * 6{,}1$$

To make a relevant comparison between BAG data and data available to Ede, it is sought for to use the most recent CBS data possible that the municipality of Ede could use. This is CBS PC6 data from 1 January 2016 for the number of households, and from CBS Districts 2019 data the car ownership on district level can be obtained. In the traffic model of the municipality of Ede, zones have been created by merging PC6 zones. In order to accurately estimate trip generation at the urban level, it is important to work with zones that can reflect the diversity in trip generation factors at a low aggregation level. The district level is too highly aggregated for this purpose (Chapter 1). For this research, PC6 data from 2020 is not available, therefore CBS districts 2020 data is used. To transform the district data into smaller zones, the following steps are taken:

1. The new housing developments in the district Veluwse Poort are divided into 3 zones.
2. For each zone the traffic generation is calculated based on BAG data.
3. The traffic generation for the entire district is calculated based on the most recent CBS data; CBS districts 2020 for the number of inhabitants, CBS districts 2019 data for car ownership.
4. The calculated generation based on CBS districts data is divided over the 3 zones based on the ratio of the calculated BAG generation per zone.

A condition for making a trip generation model on urban level is therefore the availability of data with an acceptable aggregation level. Data at district level does not suffice. The most recently available open data that can be used is CBS PC6 2016. A trip generation estimate will also be made based on this data. Comparisons with BAG results will indicate to what extent BAG improves the potential of trip generation estimates with open data. Before the generation is calculated, the zones will be developed and the surrounding roads will be analyzed.

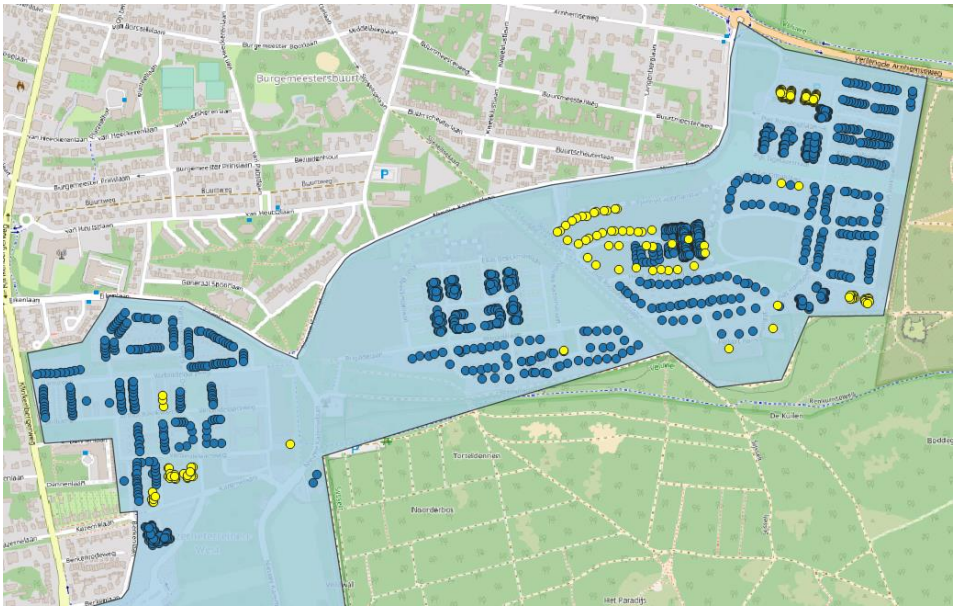


Figure 8.2: BAG VBO *living* current in use (Blue) and permit granted (Yellow)

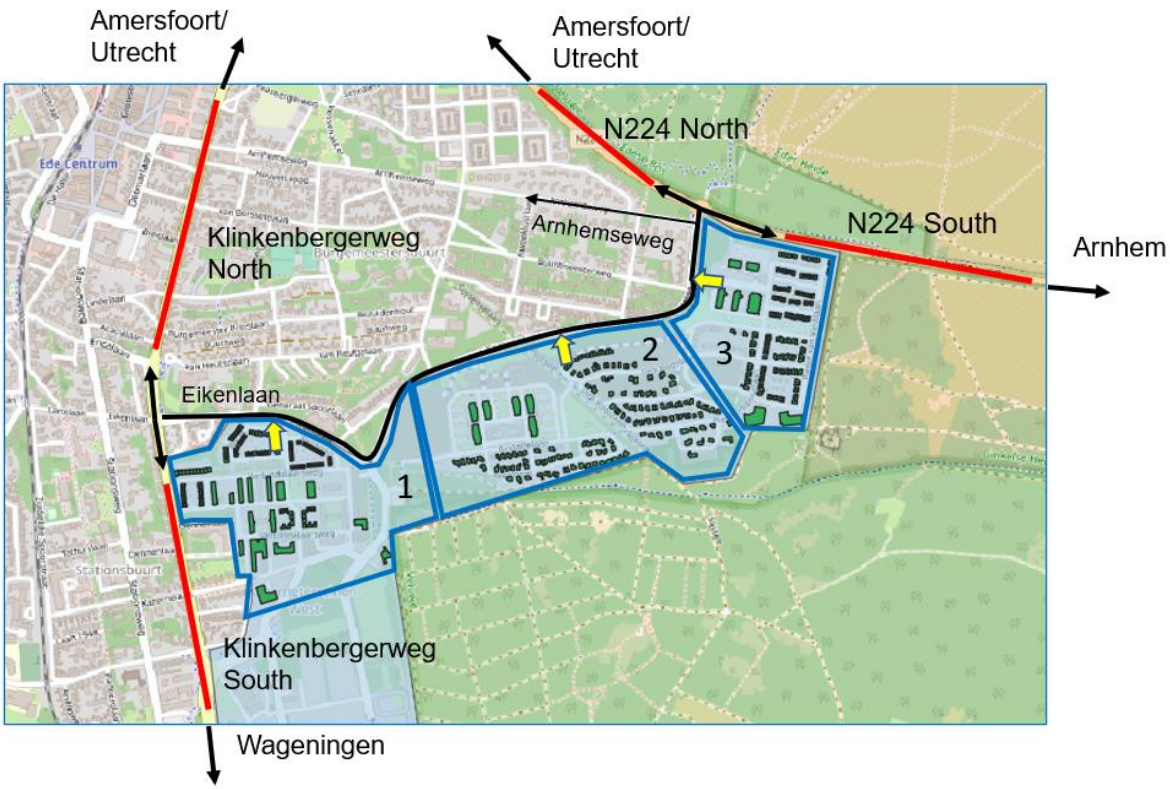


Figure 8.3: Zones and direction of outgoing traffic district Veluwse Poort

8.2.1 Trip generation estimations

In Figure 8.3, the developed zones for the case study are displayed, along with traffic directions which will be discussed later. As can be seen, the residences have been divided three zones (highlighted in blue). The current entry roads to the district were the most important rationale for dividing the residences amongst the zones. The yellow arrows are the locations at which residents can access the road transportation network of Ede. It is assumed that all residents in each zone go to the corresponding entry. The entry and exit locations of the zones are retrieved from the development plans of Veluwse Poort. The plans include an infrastructural change that will lead to more entry locations, but that is not implemented yet. Therefore, the current state of the surrounding infrastructure is used for the case.

The trip generation of Veluwse Poort will be estimated four times, based on different data sources:

1. Estimations based on open data: CBS PC6 2016 (Number of households) and CBS districts 2019 (Car ownership)
2. Estimations based on most recent data available for Ede: CBS districts 2020 (Number of households) and CBS district 2019 (Car ownership)
3. Current BAG estimations: BAG VBO *living* with status 3 and 4 (*in use*)
4. Future BAG estimations: BAG VBO *living* with status 2, 3 and 4 (*in use* and *formed*)

The number of VBO *living* objects is equal to the number of households in each zone. To determine car ownership levels, the regression equation based on HH surface area estimated in RQ 2 will be used:

$$CarOwnershipHH = -1{,}02 + 0{,}03 \cdot HH_A - 1{,}12 \cdot 10^{-4} \cdot HH_A^2 + 1{,}48 \cdot 10^{-7} \cdot HH_A^3$$

In Table 8.1, the household and car ownership data retrieved from different sources are presented, and the number of trips generated based on these values is calculated.

Table 8.1: Total trips generated at district level Veluwse Poort

Data	Number of HH		Car ownership		Trips generated
Open data	105	CBS PC6 2016	1,1	CBS Dis. 2019	740
Ede data	600	CBS Dis. 1/1/2020	1,1	CBS Dis. 2019	4.000
BAG current	793	BAG 1/1/2021	1,1	CarOwnershipHH	5.300
BAG future	957	BAG 1/1/2021	1,1	CarOwnershipHH	6.400

The car ownership level retrieved from the CBS and the car ownership determined with the function CarOwnershipHH are the same. The differences in the number of households and the resulting trips generated are large. In open data, 105 households are present in the District, whereas currently 793 are present. The consequence is that a number of trips is generated that is seven times larger. In BAG, almost 200 more households are registered than in the most recent CBS data. And in the near future is can be seen that 164 more residences will be built, leading to an increase of trips generated by more than 50% compared to CBS estimations.

Table 8.2: BAG trips generated per zone

Zone	BAG Current	BAG Future
1	1.650 (32%)	1.780
2	1.420 (28%)	1.840
3	2.000 (40%)	2.550

In Table 8.2, the number of trips generated by BAG data per zone are presented. In each zone, the number of cars per households is calculated, based on which the number of trips generated per households is estimated. Next, the trips of all households in a zone are summed. In zone 3, future developments will have the biggest impact on trips generated. In zone 2, the smallest car ownership levels are present, with an average of 0,9. Zone 1 and 3 have levels just above 1,1. Based on the percentages of BAG Current estimations, trips generated by CBS data will be divided over the zones. In the next section, the trips generated in Veluwse Poort based on different data sources will be applied to the surrounding road network, to estimate the impact of the use different databases.

8.2.2 Route choice and traffic intensities

In Figure 8.3, the main directions of outgoing traffic are visualized. For ingoing traffic, the same reversed routes apply. West of the district, the Klinkenbergerweg is present, the main distributor road for the east of Ede. By accessing this road, all other parts of Ede can be reached, and regional access is gained by continuing to the north or the south along the Klinkenbergerweg. North of the district, the N224 is present, which is a main distributor road for the region. For residents exiting Veluwse Poort to travel to almost any destination, one of two roads will be accessed. For the road segments Klinkenbergerweg North, Klinkenbergerweg South, N224 North and N224 South, passing traffic originating from Veluwse Poort will be examined.

It is assumed that approximately half of the trips made are regional, and the other half are local (in Ede). For regional trips; to access larger cities in the area such as Utrecht, Amersfoort and Arnhem, the most optimal route leads along the N224 for zones 2 and 3. For zone 1 however, it is more logical to take the route of the Klinkenbergerweg. For smaller towns closer in the region, Bennekom, Wageningen and Renkum, the most logical route for all zones is the Klinkenbergerweg. For the regional directions Utrecht, Amersfoort and Arnhem, equal proportions of regional traffic are assumed.

For local trips, almost all traffic will flow to the Klinkenbergerweg, as it is the main distributor road connecting to other parts of Ede. For people of zone 3, travelling to the center, west or north side of Ede, the Arnhemseweg is a logical route. Eventually, this road connects to the Klinkenbergerweg North. The district connects to the Klinkenbergerweg in the middle of Ede. For local traffic from zone 1 and 2,

approximately half of traffic will flow north, and half or traffic will flow south. A direct route to the west is restricted for being a one-way road to the east.

In Table 8.3, the above mentioned considerations are presented in percentages. To each road segment, per zone and per traffic type (local vs regional), a percentage is given that indicates to what extend the traffic flow coming from the particular zone is travelling over that specific road segment. With the total trips calculated based on the households in BAG current in zone 1, 2 and 3 (Table 8.2), the proportion of in- and outgoing traffic from district Veluwse Poort is determined.

Table 8.3: Division of in- and outgoing traffic by road per zone for district Veluwse Poort

Outgoing trips 100%	**Local 50%**		**Regional 50%**	
Zone 1	KW North	50%	KW North	33%
32% of BAG current trips	KW South	50%	KW South	33%
	N224 North	0%	N224 North	0%
	N224 South	0%	N224 South	33%
Zone 2	KW North	50%	KW North	0%
28% of BAG current trips	KW South	50%	KW South	33%
	N224 North	0%	N224 North	33%
	N224 South	0%	N224 South	33%
Zone 3	KW North	75%	KW North	0%
40% of BAG current trips	KW South	25%	KW South	33%
	N224 North	0%	N224 North	33%
	N224 South	0%	N224 South	33%

KW North = 0,50 * 0,32 * 0,50 + 0,50 * 0,32 * 0,33 + 0,50 * 0,28 * 0,50 + 0,50 * 0,40 * 0,75 = 0,35 or 35%

KW South = 0,50 * 0,32 * 0,50 + 0,50 * 0,32 * 0,33 + 0,5 * 0,28 * 0,50 + 0,50 * 0,28 * 0,33 + 0,50 * 0,40 * 0,25 + 0,50 * 0,40 * 0,33 = 0,37 or 37%

N224 North = 0,50 * 0,28 * 0,33 + 0,50 * 0,40 * 0,33 = 0,11 or 11%

N224 South = 0,50 * 0,32* 0,33 + 0,50 * 0,28 * 0,33 + 0,50 * 0,40 * 0,33 = 0,17 or 17%

With the above calculations it is determined what percentage of total trips generated in district Veluwse Poort is passing over what road segment, which can be seen in Figure 8.3. The used percentages are of course highly speculative, but deliberately considered. And the main point of the exercise is estimating the impact of the use of different data sources on surrounding traffic intensities, not determining the exact route choice for residents of Veluwse Poort.

With the values in Table 8.3, traffic intensities for each road segment have been calculated. These can be examined in Table 8.4. The division of the trips generated per zone for the open data and the Ede data have are based on the division of BAG current trips. These percentages are not applied to BAG future trips, for which its own zonal estimations are made.

Table 8.4: Traffic intensities per average working day per road segment on both directions with different data uses

	Open data	**Data Ede**	**BAG current**	**BAG future**
KW North	260	1.400	1.700	2.150
KW South	275	1.480	1.960	2.250
N244 North	80	440	580	720
N244 South	125	680	900	1.020

The differences between conventional open data and BAG data are obvious. For estimating trip generation for recent developments, using conventional open data may drastically affect trip estimations, and therefore traffic intensities.

When comparing estimations from BAG current with data Ede, significant traffic intensity differences are present on all road segments. Especially for KW South, traffic estimations are largely underestimated based on CBS data. For the N244, the differences may not lead to traffic problems, as these regional roads have a high capacity. When comparing BAG future data with data Ede, the increase of traffic on KW North of more than 50% is evident. This is mainly caused by the developments in zone 3, for which compared with other zones a relatively large amount of traffic passes over KW-North. BAG data enables developing zones at a low aggregation level. If district level would be used as aggregation level, the corresponding traffic intensities would have been significantly different, and likely bottlenecks would not be identified.

The traffic intensities are useful for imagining the actual impact of the development Veluwse Poort on surrounding traffic conditions. An example; approximately 1/3 of trips made are home-work trips. When comparing traffic conditions on KW North; 470 working trips per day based on data Ede, 720 working trips per day based on BAG future, half of which are made during morning-peak (to work), others during the evening (to home). During a morning peak of two hours, 120 cars per hour will pass KW North based on data Ede, and 180 cars based on BAG future. This is comparable for KW-South, 125 (data Ede) and 190 (BAG future). The result is an increase of 125 cars per hour accessing the Klinkenbergerweg, for which traffic lights at the Eikenlaan need to be passed (Figure 8.3 and Figure 8.4). This increase of traffic could lead to these traffic lights being a massive bottleneck. Thus, by using BAG data potential bottlenecks in current infrastructure can be identified, and preventive measures can be developed.

Figure 8.4: Traffic lights at the Eikenlaan accessing Klinkenbergerweg (Google, 2021)

Chapter 9. Conclusion

The research goal of this research project was to increase the possibilities of trip generation estimates based on open data in the Netherlands by researching the potential of BAG to be used as a data source for trip generation modelling. To achieve the goal, the following research questions have been formulated:

1. What trip generation factors and activities need to be discerned in BAG in order to enable trip generation estimates?
2. To what extent can BAG provide for trip generation factors and activities required for trip generation modelling?
3. How does BAG improve trip generation outcomes based on open data and how does this impact transport modelling outcomes?

Research question 1 is answered through a systematic and comprehensive analysis of motives, activities, trip rates, available open data and linkages between databases. It was found that for shopping activities, no open data was available while being one of the most important motives in a trip generation model. Also for work trips, no open data is available. Furthermore, other promising open data sources were identified that could complement the abilities of BAG, or fill up missing links. An example is the availability of DSA data that can supply for sport activities which is of significant value for the estimation of trip generation for the motive sport / hobby. And by combining BAG and DSA data, the GFA for activities such as the gym can be retrieved from BAG. A complete oversight of open data that is suitable for estimating trip generation has been developed. This practical handout (Chapter 4, Table 4.5) might prove useful for modelers. From RQ1 it can be concluded that in order to improve the potential trip generation estimations based on open data, BAG should provide data to estimate shopping trips and work / business trips.

To answer the second research question, analysis has been carried out to estimate to what extent BAG is able to provide for shopping activity locations, and it has been evaluated to what extent BAG is able to predict trip generation factors at the household side. A set of rules has been developed based on which it is possibly to identify clustered shopping activities in a city, by executing a DBSCAN, and examine the number of VBOs in the cluster, the shape of the cluster and the surface area of the VBOs within the cluster. The city center, shopping centers, wholesale activities and home decor boulevards could be successfully identified. Only separate supermarkets could not be identified based on BAG. OSM data has the potential to be linked to BAG and fill in this gap. By distinguishing shopping activities, and providing the surface area for each activity, BAG enables shopping trip generation estimations at the activity side based on open data.

To estimate to what extend BAG is able to predict trip generation factors at the household side, two factors have been analyzed: car ownership and the number of residents. Based on VBO *surface area* and *residence type*, BAG is well-abled to predict car ownership levels and the number of residents at zonal levels in municipalities, if properly used. For different types of cities, on average the estimations may turn out higher or lower. It is suggested to keep these differences in mind, and to perhaps account for these differences. Furthermore, it is advised to not rely on BAG predictions only, but to use BAG to subdivide trip generation factors from CBS Districts data into smaller zones. By comparing the number of households present in CBS and BAG it is possible to compensate for the outdated aspects of CBS data. Only for newly developed neighborhoods, BAG-only predictions should be used. Based on the results it can be concluded that BAG could provide for the most important trip generation factors relevant for trip generation modelling, car ownership and residents.

To answer the third research question, a case study has been carried out in which trip generation estimations of conventional open data, the best data available, and BAG data have been compared. Conventional open data cannot provide for the households that have been constructed in the past four years, which leads to a massive underestimation of trips, especially surrounding large housing developments. Even with the most recent CBS data available, trips in the case study underestimated by 24%, compared to current BAG data. However, one of the major advantages of BAG is that data of future housing development is present. By comparing future BAG trip estimations with current CBS data estimations and applying these to surrounding roads, potential future bottlenecks were identified. BAG enables anticipating on future transportation demand, because VBO *living* objects for which construction permits have been granted are registered in BAG approximately four years before realization.

All in all, it can be concluded that BAG tremendously increases the possibilities of estimating trip generation based on open data in the Netherlands. Five key characteristics of BAG were identified that make that BAG can increase the possibilities and the quality of trip generation estimates based on open data; providing for shopping activities, linkage to other open data sources and combining data attributes, predicting future transport demand and estimating residents and car ownership at lower aggregation levels or for new housing developments. For work trips, it was found that BAG has a large potential to fill in the void of the open availability of employment data. All these aspects open the door for institutions and organizations that do not have the means to heavily invest in data purchasing and development, to be able to estimate trip generation based on open data with improved reliability. Eventually, this can lead to better transportation decision-making and improved policies.

Chapter 10. Discussion & Recommendations

For all research that is carried out, it is of great importance to reflect upon assumptions that have been made, on the limitations of the research, and to recommend how future research could contribute to the knowledge base created by the research. In this chapter, several shortcomings and limitations of the research project will be discussed, together with future recommendations based on the conclusion and discussion points that are addressed.

This research explores the possibilities of BAG to predict trip generation factors and to discern activities. Based on these activities and factors, trip rates can be applied that can be developed based on travel surveys, counts, or other data sources, which determines how many trips a BAG object generates. This is an indirect way of estimating trip generation, that enables building on existing knowledge and the predictions in BAG include additional uncertainty in the process. In theory, it would be best if BAG can be used to estimate trip generation directly. Future research could focus on direct estimations of trip generation with BAG. To do so, reliable trip generation data is required, for which perhaps GPS data, mobile network data or reliable trip generation model outcomes could be used.

It is questionable whether all relevant activities for trip generation have been included in the research. The activities analyzed in the research, derived from the CROW, are incomplete. For activities that are included in the CROW, it could be analyzed how these are registered in BAG. However, activities not provided by the CROW, but present in BAG, could affect trip estimates. Future research could expand the number of activities that is considered.

In this research, it has not been considered in-depth to what extend what trip generation factors at the household-side are best used for specific activities. Carrying out such an analysis could lead to more substantiated conclusions for BAG and its potential to estimate the trip generation at the household side for specific activities. For example, household data can be obtained directly from BAG, while the number of residents in a zone needs to be estimated. If the number of residents is best used to predict for example leisure trips, but the number of residents is unsure, and the number of households is a second best alternative, then substantiated decisions can be made for using BAG. It is suggested that future research expands the literature studies in this research, and that for different activities it is examined what factors are best used. Then, substantiated considerations can be made about the use of factors for different activities, combined with the uncertainties in the prediction by BAG.

This research did not cover the possibilities of BAG to estimate trip generation for work and business activities, nor did it consider the possibilities of estimating freight trip generation. Work is the most important motive within trip generation, as it is producing the most trips. It is recommended that future research gives priority to analyzing the possibilities of BAG to estimate work trips. A starting point for doing so could be to estimate how BAG could provide job data. In Chapter 4 of this research, it is commented how BAG can be linked to NRM employment data.

It is also valuable to consider the practicalities of the results. All operations carried out during the research did take a lot of time and required a certain skillset for tasks like writing a script to operate the data, importing and updating data sources, and exporting data to a GIS environment. Although all data used is open, the time and effort it would take for institutions to produce a solid database for estimating trip generation would require an investment of many working hours. Therefore, future research should also focus on automatizing as many of the processes as possible. But first, it is valuable to extend the substantial analysis of BAG to research for activities and work and freight trips.

The case study developed and carried out in the research is rather simplistic and reveals only the potential of BAG in a district with many recent housing developments. It is expected that also in districts in which

no recent large housing developments have been taken place, BAG could be of added value. At the activity side, it is more of a challenge to estimate whether trip estimates are improved, as no conventional data of these activities are present.

In this research, no attention has been given towards the process of developing zones for trip generation models at the urban level. It has been concluded that the maximum zone size should be between Districts and PC6 level, based on zoning systems in current municipal studies. It could be possible to automatically create zones with BAG, based on the average HH surface area and residence type. In this research it was estimated that BAG attributes have a predictive ability towards factors important for trip generation. Develop a zoning algorithm that groups together BAG areas in which BAG properties are more or less the same, and making sure that zones comply with existing census data boundaries could be of great added value for estimating trip generation based on open data.

During the research it was found that BAG is not able to distinguish supermarkets from VBO *shop*. From OSM data, the location of each supermarket can be obtained. To discern for supermarkets in BAG, it is suggested to spatially join BAG VBO *shop* with OSM supermarkets and analyze the a number of nearest neighbors of OSM supermarkets, from nearest to farthest. If one of these limited nearest neighbors has a VBO *surface area* larger than 1.000 m^2, it is likely the OSM supermarket has been linked to the corresponding supermarket in BAG.

All in all, a lot can be improved, but a lot of new useful insights have been gathered as well. A broad approach has been used for the research, in which several aspects of the potential of BAG have been touched upon. Improvement is definitely possible, but the results that have been gathered in this research are relevant and significant.

References

Axhausen, K., & Rieser-Schüssler, N. (2013). *Self-tracing and Reporting: State-of-the-Art in the Capture of Revealed Behaviour*. https://doi.org/10.4337/9781781003152.00012

Brouwer, J., Kampen, H., & Wilmink, R. (2014). *Check op kencijfers*. which is an indication of the income of the owner, and therefore of the number of cars he might possesses.

Caiati, V., Bedogni, L., Bononi, L., Ferrero, F., Fiore, M., & Vesco, A. (2016). Estimating urban mobility with open data: A case study in Bologna. *2016 IEEE International Smart Cities Conference (ISC2)*, 1–8. https://doi.org/10.1109/ISC2.2016.7580765

CBS. (2020). *Organisation*. https://www.cbs.nl/en-gb/over-ons/organisation

CBS. (2021). *Kerncijfers per postcode*. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

Chang, J. S., Jung, D., Kim, J., & Kang, T. (2014). Comparative analysis of trip generation models: Results using home-based work trips in the Seoul metropolitan area. *Transportation Letters*, *6*(2), 78–88. https://doi.org/10.1179/1942787514Y.0000000011

Cooley, K., De Gruyter, C., & Delbosc, A. (2016). A best practice evaluation of traffic impact assessment guidelines in Australia and New Zealand. *Australasian Transport Research Forum*. http://atrf.info/papers/2016/index.aspx

CROW. (2018). *Toekomstbestendig parkeren. Van parkeerkencijfers naar parkeernormen.* https://www.crow.nl/publicaties/toekomstbestendig-parkeren

Currans, K. M. (2017). Issues in Trip Generation Methods for Transportation Impact Estimation of Land Use Development: A Review and Discussion of the State-of-the-art Approaches. *Journal of Planning Literature*, *32*(4), 335–345. https://doi.org/10.1177/0885412217706505

Daszykowski, M., & Walczak, B. (2009). Density-Based Clustering Methods. *Comprehensive Chemometrics*, *2*, 635–654. https://doi.org/10.1016/B978-044452701-1.00067-3

Digitale Overheid. (2020). *Stelsel van basisregistraties*. https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/basisregistraties-en-stelselafspraken/stelsel-van-basisregistraties/stelselplaat/

Ding, C. (1998). The GIS-Based Human-Interactive TAZ Design Algorithm: Examining the Impacts of Data Aggregation on Transportation-Planning Analysis. *Environment and Planning B: Planning and Design*, *25*(4), 601–616. https://doi.org/10.1068/b250601

DUO. (2021a). *Databestanden*.

DUO. (2021b). *Gegevens kinderopvanglocaties LKR*.

Egu, O., & Bonnel, P. (2020). How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? An empirical investigation in Lyon. *Transportation Research Part A: Policy and Practice*, *138*, 267–282. https://doi.org/10.1016/j.tra.2020.05.021

GeoParaat. (2021). *BAG GeoPackage*. https://geoparaat.baasgeo.com/bag/

Goudappel Coffeng. (2011). *Opstellen multimodaal verkeersmodel Regio Twente*. https://dloket.enschede.nl

Goudappel Coffeng. (2018a). *Technische Rapportage Verkeersmodel MRDH 2.0*. https://mrdh.nl/project/verkeersmodel

Goudappel Coffeng. (2018b). *Verkeersmodel Regio Utrecht VRU3.4*. https://www.utrecht.nl/fileadmin/uploads/documenten/bestuur-en-

organisatie/publicaties/onderzoek-en-cijfers/verkeerscijfers

Gruyter, C. De. (2019). *Multimodal Trip Generation from Land Use Developments : International Synthesis and Future Directions*. *2673*(3), 136–152. https://doi.org/10.1177/0361198119833967

IBIS. (2019). *IBIS Bedrijventerreinen*. https://data.overheid.nl/en/dataset/ibis-bedrijventerreinen

ITE. (2021). *INSTITUTE OF TRANSPORTATION ENGINEERS COMMON TRIP GENERATION RATES*. www.ITE.org

Jeon, J.-H., Kho, S.-Y., Park, J. J., & Kim, D.-K. (2012). Effects of spatial aggregation level on an urban transportation planning model. *KSCE Journal of Civil Engineering*, *16*(5), 835–844. https://doi.org/10.1007/s12205-012-1400-4

Kadaster. (2020). *Over BAG*. https://www.kadaster.nl/zakelijk/registraties/basisregistraties/bag/over-bag

Kadaster. (2021a). *BAG Producten*. https://www.kadaster.nl/zakelijk/registraties/basisregistraties/bag/bag-producten

Kadaster. (2021b). *Praktijkhandleiding BAG*. https://imbag.github.io/praktijkhandleiding/artikelen/wordt-bij-het-gebruiksdoel-het-feitelijk-gebruik-of-het-vergund-gebruik-opgenomen

KiM. (2019). *Mobiliteitsbeeld 2019*. https://www.kimnet.nl/publicaties

Locatus. (2020). *Meubelzaken*. https://www.retailinsiders.nl/branches/woninginrichting/meubelzaken/

Miller, J. S., Hoel, L. A., Goswami, A. K., & Ulmer, J. M. (2006). Borrowing residential trip generation rates. *Journal of Transportation Engineering*, *132*(2), 105–113. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(105)

Milne, A., & Abley, S. (2009). *Comparisons of NZ and UK Trips and Parking Rates*. Land Transport NZ Research Report.

Moeckel, R., Huntsinger, L., & Donnelly, R. (2015). Microscopic trip generation: Adding fidelity to trip-based travel demand models. *CUPUM 2015 - 14th International Conference on Computers in Urban Planning and Urban Management*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019168655&partnerID=40&md5=94cbbdcfee6febaa0082a2a55d580066

Mulier instituut. (2021). *Database Sportaanbod*.

Oranjewoud. (2009). *Verkeersmodel Harderwijk*. https://pilot.ruimtelijkeplannen.nl/documents

Ortúzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport. In *Modelling Transport* (4th ed.). Wiley. https://doi.org/10.1002/9781119993308

OSM. (2020). *OpenStreetMap*. https://www.openstreetmap.org/about

Rasouli, S., & Timmermans, H. (2014). Activity-based models of travel demand: Promises, progress and prospects. *International Journal of Urban Sciences*, *18*(1), 31–60. https://doi.org/10.1080/12265934.2013.835118

Roorda, M. J., Páez, A., Morency, C., Mercado, R., & Farber, S. (2010). Trip generation of vulnerable populations in three Canadian cities: a spatial ordered probit approach. *Transportation*, *37*(3), 525–548. https://doi.org/10.1007/s11116-010-9263-3

Royal Haskoning. (2015). *Verkeersmodel Hilversum*.

Royal Haskoning. (2018). *Actualisatie verkeersmodel Beter Bereikbaar Wageningen*. https://beterbereikbaarwageningen.gelderland.nl/bestanden/Documenten

Royal Haskoning. (2020). *Technische Rapportage Verkeersmodel gemeente Woerden*.

https://www.woerden.nl/bekendmakingen/verkeersvisie-2030

Saadi, I., Mustafa, A., Teller, J., & Cools, M. (2017). A bi-level Random Forest based approach for estimating O-D matrices: Preliminary results from the Belgium National Household Travel Survey. *Transportation Research Procedia*, *25*, 2566–2573. https://doi.org/10.1016/j.trpro.2017.05.301

Schoemakers, A., & Geurs, K. T. (2008). *Vroeger voorspelden we de toekomst beter*. https://www.cvs-congres.nl/cvspdfdocs/cvs08_83.pdf

Shi, F., & Zhu, L. (2019). *Analysis of trip generation rates in residential commuting based on mobile phone signaling data*. *5*, 201–220.

Thomas, T., Geurs, K. T., Koolwaaij, J., & Bijlsma, M. (2018). Automatic Trip Detection with the Dutch Mobile Mobility Panel: Towards Reliable Multiple-Week Trip Registration for Large Samples. *Journal of Urban Technology*, *25*(2), 143–161. https://doi.org/10.1080/10630732.2018.1471874

TNO. (2007). *Handbook of transport modelling in Europe: learning from best practice*. https://trimis.ec.europa.eu/

Zenina, N., & Borisov, A. (2013). Regression Analysis for Transport Trip Generation Evaluation. *Information Technology and Management Science*, *16*. https://doi.org/10.2478/itms-2013-0014

Appendix A: BAG data

Table A.1: BAG_VBO Attributes

BAG_VBO		
Attribute	**Type**	**Description**
BAG VBO	Point	Smallest unit of use located within one or more buildings and suitable for residential, commercial, or recreational purposes that is accessed via its own lockable entrance from the public road, a yard, or a shared traffic area, and can be the subject of property law legal transactions, and is functionally independent.
Identification	Character	Unique identifier of the VBO
Address identification	Character	Unique identifier of the address of the VBO
Building identification	Character	Unique identifier of the building in which the VBO is present
Surface area	Integer	Surface area of the VBO
Status code	Integer	VBO status: 0. Unrealized 1. Shaped 2. Out-of-use 3. In use 4. In use (unmeasured) 5. Withdrawn
Meeting	Logical: TRUE / FALSE	VBO for meeting people for art, culture, religion, communication, childcare, catering on the spot and watching sports
Jail	Logical: TRUE / FALSE	VBO for a coercive stay of people
Healthcare	Logical: TRUE / FALSE	VBO for medical examination, nursing, care or treatment
Industry	Logical: TRUE / FALSE	VBO for the commercial processing or storage of materials and goods, or for agricultural purposes
Office	Logical: TRUE / FALSE	VBO for administration
Lodgings	Logical: TRUE / FALSE	VBO for providing recreational or temporary accommodation to people
Education	Logical: TRUE / FALSE	VBO for teaching
Other	Logical: TRUE / FALSE	VBO for functions other than mentioned in which the stay of people plays a subordinate role
Sport	Logical: TRUE / FALSE	VBO for practising sports
Shop	Logical: TRUE / FALSE	VBO for trading materials, goods or services
Living	Logical: TRUE / FALSE	VBO for living

Table A.2: BAG_Building attributes

BAG_Building

Attribute	Type	Description
BAG Building	Polygon	A building
Identification	Character	Unique identifier of the building
Status code	Integer: 0 - 7	Status of the building: 4. Construction started 5. Construction permit granted 6. Unrealized building 7. Building out-of-use 8. Building demolished 9. Building in use 10. Building in use (unmeasured) 11. Demolishing permit granted
Residence type	Integer: 0 - 5	Type of residence: 0. No residence 1. Detached 2. Semi-detached 3. End-of-terrace 4. Terraced 5. Multi-family
Nr. Meeting	Integer	The number of VBOs with function Meeting present in the building
Nr. Jail	Integer	The number of VBOs with function Jail present in the building
Nr. Healthcare	Integer	The number of VBOs with function Healthcare present in the building
Nr. Industry	Integer	The number of VBOs with function Office present in the building
Nr. Office	Integer	The number of VBOs with function Lodging present in the building
Nr. Lodging	Integer	The number of VBOs with function Education present in the building
Nr. Education	Integer	The number of VBOs with function Other present in the building
Nr. Other	Integer	The number of VBOs with function Sport present in the building
Nr. Sport	Integer	The number of VBOs with function Shop present in the building
Nr. Shop	Integer	The number of VBOs with function present in the building
Nr. Living	Integer	The number of VBOs with function present in the building

Table A.3: BAG_Address attributes

BAG_Address

Attribute	Type	Description
BAG Address	Point	An address
Identification	Character	Unique identifier of the address
Zip code	Character	The zip code in which the address is present, at PC6 level

Appendix B: Operations abundant BAG data

To strip the raw BAG data of abundant records, attributes with the following values have been removed:

- BAG_VBO *Other* = TRUE for records with one function
- BAG_VBO *Status code* = 0, 1, 2, 5, NA
- BAG_Building *Status code* = 0, 1, 2, 3, 4, 7, NA
- BAG_VBO *Surface area* = 1, 9999, 99999, 999998, 999999, 888888 (these values area used as dummy values for a surface area that is unknown)
- BAG_VBO *Surface area* < 14 AND BAG_VBO *Living* = TRUE
- BAG_VBO *Surface area* > 2.700 AND BAG_VBO *Living* = TRUE

Appendix C: Open data sources

DSA

The Database Sport Accommodations (DSA) is an open database in the Netherlands containing information of around 28.000 sport facilities. Maintaining the database is sponsored by the national government and carried out by the Mulier institute. The data is used for monitoring sports facilities, and for research on themes such as sport and well-being. The use of the database for non-commercial and commercial purposes is permitted, provided that reference is made to the database. In addition, it is permitted to further develop, transform and build products based on the data, provided that these products are available under the same conditions.

The DSA can be requested at the webpage of the Mulier institute (Mulier instituut, 2021). At 06/2021, the database was available at the following webpage:

https://www.mulierinstituut.nl/producten-diensten/dataverzameling/database-sportaanbod/

The following attributes are present:

Export_DSA_2021/05/18

Attribute	Type	Description
ID	Character	Unique identified of the accommodation
Accommodation name	Character	Name of the accommodation
X	Integer	X coordinate of the accommodation
Y	Integer	Y coordinate of the accommodation
Address	Character	Street name and accommodation number
Postal code	Character	PC6 postal code
City / town name	Character	Name of the city or town
Municipality	Character	Name of the municipality
Number of accommodations	Integer	
Number of indoor accommodations	Integer	
Number of outdoor accommodations	Integer	
Number of artificial turf pitches	Integer	
Number of outdoor grass pitches	Integer	
Number of outdoor semi-water-based pitches	Integer	
Number of outdoor water pitches	Integer	
Number of sporting halls	Integer	
Number of indoor sporting pitches	Integer	
Number of indoor gymnastic pitches	Integer	
Sport	Character	The type of sport practiced at the accommodation, such as football, tennis, fitness, skiing

The database export of 18/05/2021 contains a total of 27.196 records.

Education Data

The educational institute in the Netherlands DUO publishes annual open figures of educational institutions in the Netherlands. The following educational institutions are included here:

- Primary school
- Secondary school
- ROC
- College

- University

Of each education accommodation, among others, the following data is available:

- Address
- PC6 postal code
- The number of students x
- The number of employees

The education data can be accessed at the webpage of DUO (DUO, 2021a). At 06/2021, the datasets of each educational institute was available at the following webpage:

https://duo.nl/open_onderwijsdata/databestanden/

Childcare data

The education institute in the Netherlands DUO publishes data being updated twice a week of locations of childcare accommodations in the Netherlands with detailed information. The data is maintained in the national childcare register (LKR). Of each accommodation, among others, the following data is available:

- Address
- PC6 postal code
- Number of childcare place

The data from the LKR can be accessed at the webpage of Overheid (DUO, 2021b). At 06/2021, the data of the LKR was available at the following webpage:

https://data.overheid.nl/dataset/gegevens-kinderopvanglocaties-lrk

BRT

The BRT contains topographical files that are made available at different scale levels. The BRT is a detailed and national covering geographical database describing the physical environment of the Netherlands. Among others, roads, railroads, heights, buildings and area functions are included. Prerequisites of objects included in the database is that the object is long-lasting and defining for the environment. Among others, the following buildings are included in the BRT data:

- Church
- Synagogue
- Mosque
- Abbey
- Other religious building
- Stadium
- Gas station
- Hospital
- Town hall
- Town office

Of each building, among other, the following attributes are known:

- Raw outline of the building
- Function

NRM employment data

NRM data is the only dataset described not containing any open data. The data is included because it is the only source of employment data that could be retrieved. The NRM employment data that is used originates from 2016 and is used in the NRM transportation model, a model which most of the municipalities mentioned in Table 4 use to determine external transportation demand for the study area.

At PC4 level, the number of employees for the following sectors are available:

- Industry
- Retail
- Services
- Government
- Agriculture
- Other

IBIS data

IBIS (Integral Business Area Information System) data is open geographical data containing information of business areas in the Netherlands, provided in a shapefile. For each record, among others, the work location type, environmental classifications and the total surface areas are available.

The IBIS data can be accessed at the webpage of Overheid ((IBIS, 2019). At 06/2021, the data was available at the following webpage:

https://data.overheid.nl/en/dataset/ibis-bedrijventerreinen

Appendix D: CBS data

CBS PC6 2016

Data form the Central Bureau of Statistics (CBS) at PC6 level. The following attributes are being used in the research:

CBS PC6 2016

Attribute	Type	Description
PC6	Character	PC6 postal code
GEM_HH_GR	Number	Average household size
AANTAL_HH	Integer	Number of households
INWONER	Integer	Number of residents
WONING	Integer	Number of residents

The CBS PC6 2016 data can be requested at the webpage of the CBS (CBS, 2021). At 06/2021, the data was available at the following webpage:

https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

CBS Districts 2018 / 2020

Data from the Central Bureau of Statistics (CBS) at District level. The following attributes are being used in the research:

CBS Districts

Attribute	Type	Description	2018	2020
GWB CODE	Integer	Unique identifier of the district	✓	✓
Gm_naam	Character	Name om the municipality	✓	✓
A_inw	Integer	Number of residents	✓	✓
A_hh	Integer	Number of households	✓	✓
A_woning	Integer	Number of residences	✓	✓
G_ink_pi	Character	Average income per earner	✓	X
G_pau_hh	Number	Average car ownership per household	✓	X
Ste_oad	Number	Surrounding address density	✓	✓

The CBS District 2018 and 2020 data can be requested at the webpage of the CBS (CBS, 2020). At 06/2021, the data was available at the following webpages:

https://www.cbs.nl/nl-nl/maatwerk/2018/30/kerncijfers-wijken-en-buurten-2018

https://www.cbs.nl/nl-nl/maatwerk/2020/29/kerncijfers-wijken-en-buurten-2020

CBS Linking Table

Data from the Central Bureau of Statistics that enables linking PC6 data and district data. The following attributes have been used within the research:

CBS Linking table

Attribute	Type	Description
PC6	Character	PC6 postal code
Huisnummer	Integer	House number
BUURT2018	Integer	Unique identifier of the district
WIJK2018	Integer	Unique identifier of the neighborhood
GEMEENTE2018	Integer	Unique identifier of the municipality

Appendix E: DBSCAN results

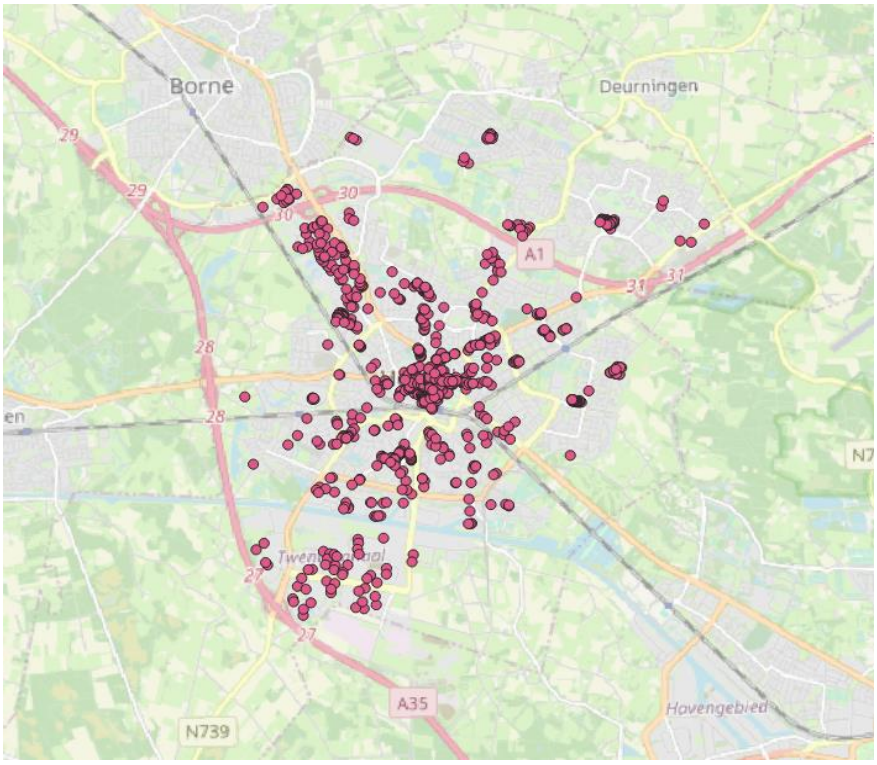


Figure E1: VBOs *shop* in the city of Hengelo

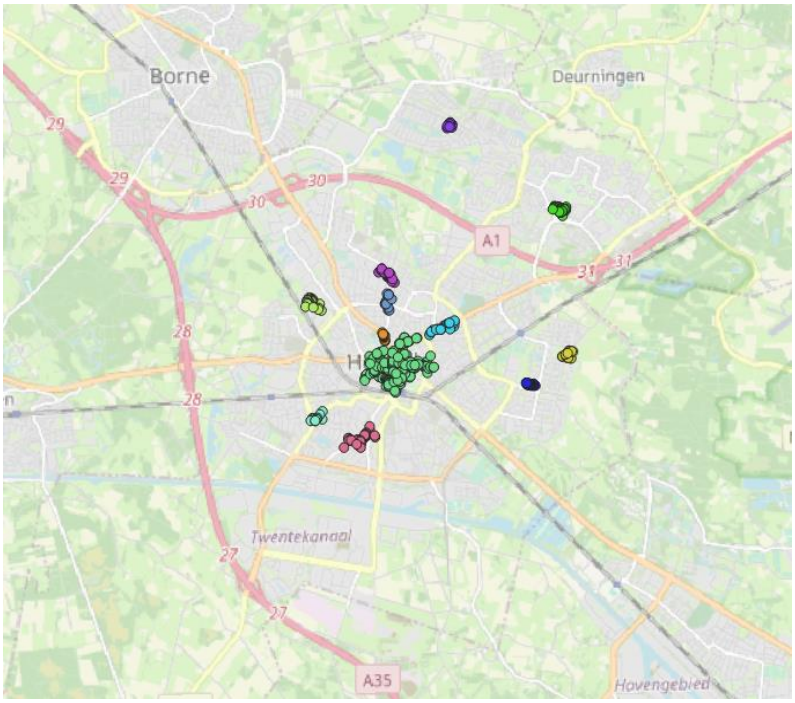


Figure E2: DBSCAN (*eps* = 120, *minPoints* = 11) VBOs *shop* in the city of Hengelo

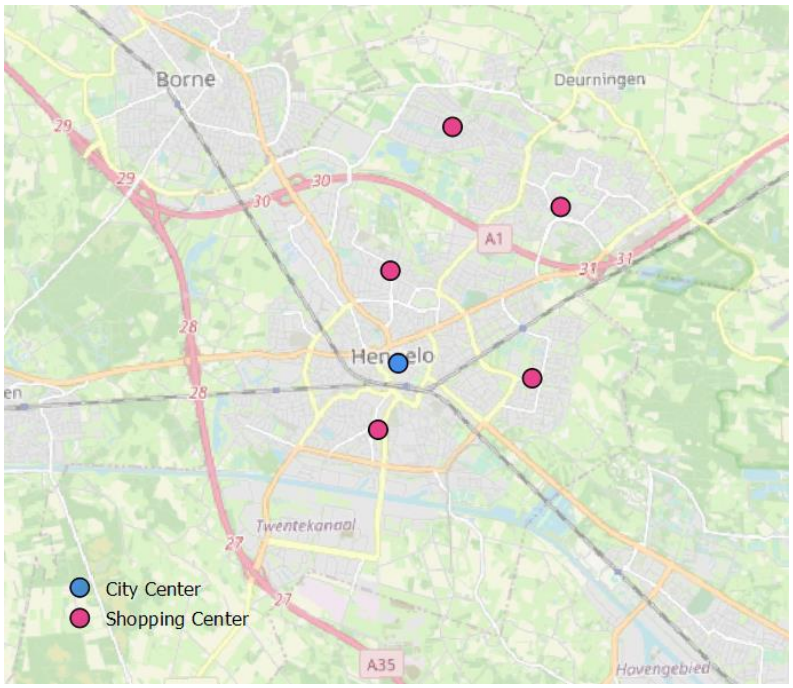


Figure E3: Shopping centers and city center in the city of Hengelo

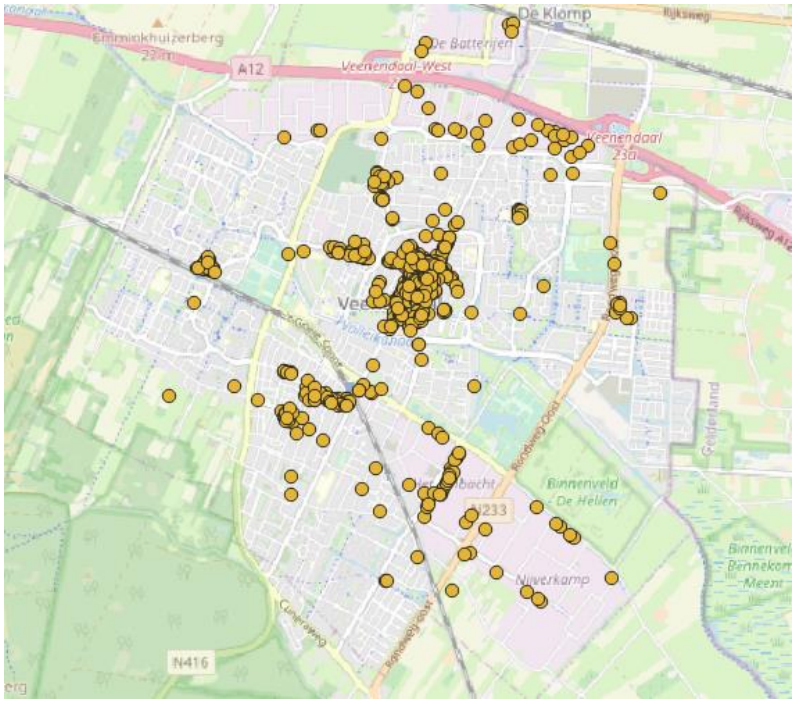


Figure E4: VBOs *shop* in the town of Veenendaal

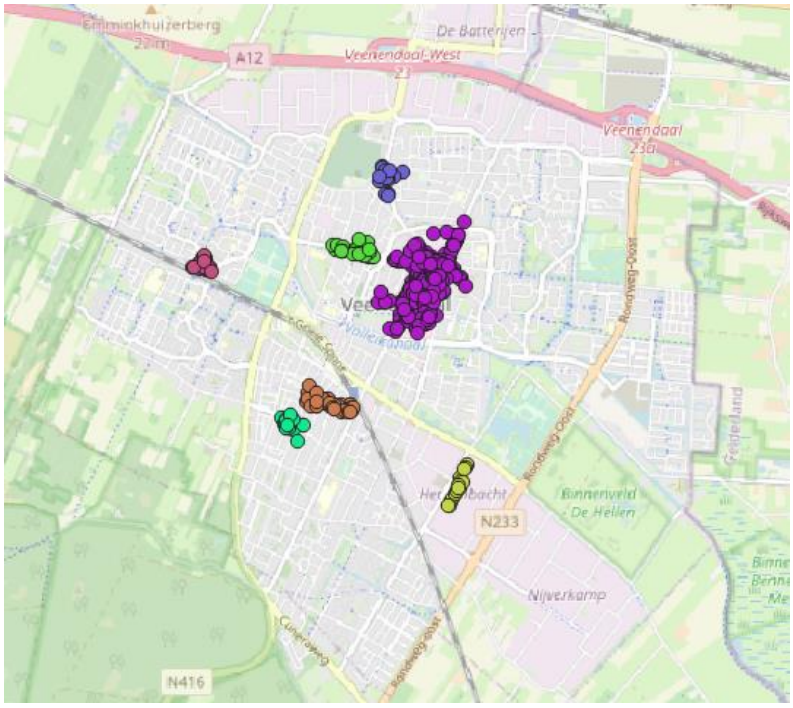


Figure E5: DBSCAN (*eps* = 120, *minPoints* = 11) VBOs *shop* in the city of Veenendaal

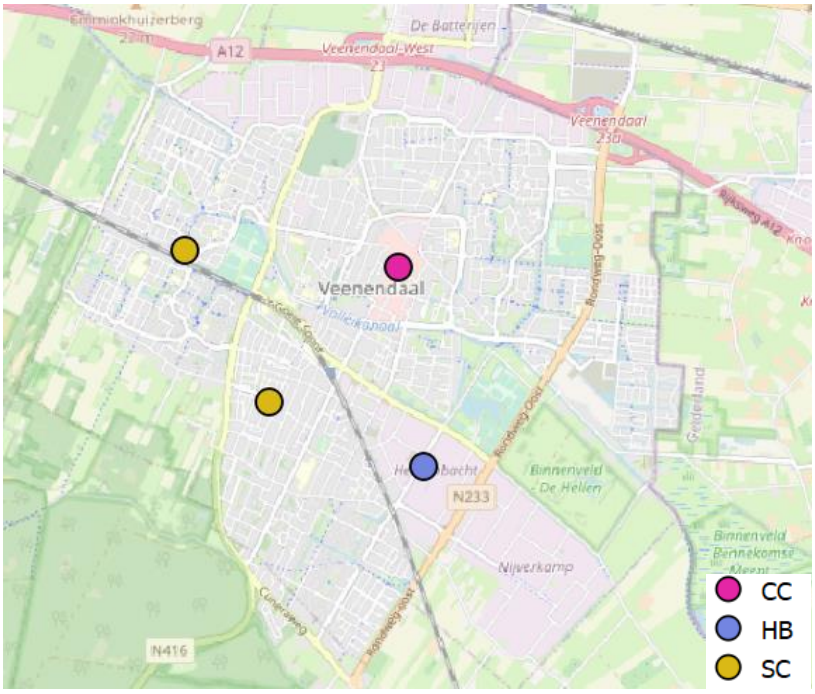


Figure E6: Clustered shopping activities in the city of Veenendaal