



## Master Thesis

submitted within the UNIGIS MSc programme  
Interfaculty Department of Geoinformatics – Z\_GIS  
University of Salzburg

# Identification and Classification of Retail Agglomerations

Submitted by

**Pascal Ramon Vorhaus**

Subscriber identification 105177, UNIGIS MSc Year 2018

A thesis submitted in partial fulfillment of the requirements of the degree of  
Master of Science (Geographical Information Science & Systems) – MSc (GIS)

Advisor

**Prof. Dr. Josef Strobl**

Duisburg, 8 March 2020

## **Science pledge**

I hereby confirm that this thesis was written independently by myself without the use of any sources beyond those cited, and all passages and ideas taken from other sources are cited accordingly.

Duisburg, 8 March 2020    Pascal Ramon Vorhaus

## Abstract

Comprehensive knowledge of retail agglomerations, their extents and composition provides the basis for an effective understanding and use of these spaces. As no nationwide data sets of retail agglomerations exists for Germany, this work aims at creating transparency on retail agglomerations on a national level, and proposes a consistent approach to a systematic identification and classification which can be updated in the future.

This work presents a comprehensive framework which helps identify, describe and classify retail agglomerations based on a set of retail points. The methodology consists of three main parts. Firstly, the study focuses on data acquisition and preparation. 236.944 center-relevant retail locations are sourced from the OpenStreetMap database and enriched with classifying data from research and retail-domain knowledge. Secondly, the methodology includes cluster identification. The well-established and widely implemented clustering algorithm DBSCAN, or density-based spatial clustering of applications with noise, has been chosen for the detection of retail agglomerations. DBSCAN has proved to be the most suitable alternative, as the algorithm takes the density of the points into consideration and performs better than other algorithms in cases where clusters are of arbitrary shape or size. Thirdly, the work puts forward the cluster characterization and classification. The clusters are described in detail and further classified by means of a rule-based classification logic. This logic is oriented towards a typical center classification. QGIS and the QGIS processing modeler were used to automate the process.

The results of this work show that identifying, characterizing and classifying retail agglomerations is possible over a large area and that the process can result in meaningful insights for individual centers, municipalities and regional studies. The analysis of the retail locations available from OpenStreetMap indicates that the data set is almost comprehensive, whereas missing retail locations are expected to be added within the next years. The validation of the identified agglomerations against external boundaries of retail agglomerations indicates that the proposed methodology provides accurate results.

The research concludes that the proposed analytical framework enhances the studies on retail agglomerations in several ways. Firstly, it identifies and classifies central supply areas at a country-wide scale, by fusing publicly available data and domain knowledge. Secondly, it defines a method meant to estimate clustering and classifying parameters, which can subsequently be applied across the study area. Thirdly, it integrates the process to such an extent that regular updates and investigations in other geographies become feasible. Moreover, the final data set offers, for the first time, a comprehensive overview on central supply areas in Germany. The resulting data helps to answer structural questions on how retail agglomerations are related to one another, how the parts of the system concentrate and share functions, and how these locations can best be described.

# Table of Contents

1 Introduction.....	7
1.1 Motivation.....	7
1.2 Problem statement.....	7
1.3 Objectives.....	8
1.4 Review and synthesis of literature.....	9
1.4.1 System of central places and spatial planning in Germany.....	9
1.4.2 Central retail agglomerations.....	11
1.4.3 Center formation of retail locations.....	13
1.4.4 Retail locations.....	15
1.4.5 Clustering of central retail agglomerations from retail points.....	16
1.5 Structure of this work.....	17
2 Methods.....	18
2.1 Research design.....	18
2.2 Study area and scale of analysis.....	19
2.3 Data.....	19
2.3.1 OpenStreetMap.....	20
2.3.2 Types of cities and municipalities.....	21
2.4 Cluster identification.....	22
2.4.1 Requirements.....	23
2.4.2 Comparison and selection of relevant cluster algorithm.....	24
2.4.3 DBSCAN description and mechanism.....	25
2.4.4 Tuning of clustering parameters.....	27
2.4.5 Cluster Detection.....	28
2.5 Cluster classification.....	29
2.5.1 Requirements.....	29
2.5.2 Comparison and selection of classification methods.....	30
2.5.3 Decision tree.....	31
2.5.4 Tuning of the decision tree brake points.....	32
2.5.5 Cluster characterization and classification.....	32
3 Results.....	33
3.1 Input variables and domain knowledge.....	33
3.1.1 Point of interest selection, classification and analysis.....	33
3.1.2 Parameterization of DBSCAN clustering.....	37
3.1.3 Characterization, decision tree and rules for the classification.....	39
3.2 Outcome of the clustering and classification.....	44
3.2.1 Overall country-wide results.....	44
3.2.2 Regional and city results.....	47
3.2.3 Results for exemplary individual clusters.....	50
3.3 Process automation.....	56
4 Discussion.....	58
4.1 Interpretation and implications.....	58
4.2 Validation.....	60
4.2.1 Quality assessment of the OpenStreetMap data.....	60
4.2.2 Validation of the cluster identification and classification.....	62
4.3 Reflecting on the research method.....	70
5 Conclusion.....	72
6 List of References.....	75
7 Appendix.....	81

## Table of Figures

Figure 1: Basic elements of a work-sharing, hierarchically tiered center model adapted representation from.....	12
Figure 2: Research design.....	19
Figure 3: Examples illustrating the diversity of shape and size of central retail agglomerations.....	24
Figure 4: Illustration of the function of DBSCAN.....	27
Figure 5: Exemplary decision tree for predicting if a retail cluster is the main-center in a large city .....	31
Figure 6: Location patterns of center relevant locations by their frequency of demand.....	35
Figure 7: Dispersed versus concentrated distribution of bakery and clothing retail.....	35
Figure 8: Center-relevant locations in Düsseldorf by attributes.....	36
Figure 9: Municipalities by type of centrality and population density.....	37
Figure 10: K-distance graph for the third nearest neighbor, sorted by nearest to farthest, for 90 % of the closest locations.....	38
Figure 11: Number of clusters detected for MinPts=3 and increasing values of esp.....	39
Figure 12: Simplified decision tree.....	42
Figure 13: Distribution of centers and other retail locations across Germany.....	44
Figure 14: Diversity of the 5 largest main-center at the same scale of 1:20.000.....	45
Figure 15: Distribution of central retail agglomerations and groups of stores in the Ruhr-Area.....	48
Figure 16: Main-center of Düsseldorf.....	51
Figure 17: Sub-center Rüttenscheid in Essen.....	52
Figure 18: Minor center Buchholz in Duisburg.....	53
Figure 19: Group of stores in Mülheim an der Ruhr Baakendorf / Duisburger Straße.....	54
Figure 20: Isolated location in Duisburg Großenbaum.....	55
Figure 21: Selected cities with single- or poly center setup, by share of low and medium frequency goods.....	59

## Index of Tables

Table 1: Characteristics of inner city centers.....	12
Table 2: Differentiation of inner-city hierarchical systems according to city size.....	13
Table 3: Factors influencing sales volume at a certain location.....	14
Table 4: Frequency of demand.....	16
Table 5: Overview of data sources used.....	20
Table 6: Attributes of the OpenStreetMap raw data.....	21
Table 7: Cities differentiated by centrality and size.....	22
Table 8: Requirements for choosing the clustering algorithm.....	23
Table 9: Comparison of DBSCAN with alternative clustering techniques. Adapted and extended...	24
Table 10: DBSCAN parameters.....	26
Table 11: Considerations for selecting parameters (esp) and (MinPts).....	28
Table 12: Requirements for classifying the clusters.....	29
Table 13: Comparison of approaches for classifying clusters of retail locations .....	30
Table 14: Added attributes to the OpenStreetMap data.....	34
Table 15: Number of clusters obtained by applying different parameters for (esp).....	39
Table 16: Enriched cluster characterizing data.....	39
Table 17: Center defining criteria, data and information used in the classification is highlighted in gray.....	41
Table 18: Sequenced rules for cluster classification.....	43
Table 19: Number of central retail agglomerations and minor sites by type of center and city hierarchy.....	44
Table 20: QGIS processing plugin and processing models.....	56
Table 21: Monthly OpenStreetMap contributor statistics for Germany.....	61
Table 22: Completeness of OpenStreetMap data in Germany, features 2018-2019.....	62
Table 23: Exemplary limitations.....	69
Table 24: Selected aspects for improving the results.....	73

# 1 Introduction

## 1.1 Motivation

The increasing availability of spatial data provides extensive opportunities for urban research; however, in most cases, the raw data is not sufficient to answer spatial questions directly (Miller and Goodchild, 2015). For the most part, as the majority of the public data is less integrated, incomplete and not particularly extensive, data cleaning and domain knowledge are required to parameterize the analysis. Simultaneously, the computer and GI-Science have advanced in providing methods to address specific challenges for a broad range of research areas. Transferring, adapting and recombining these methods into a new field, in this case urban planning and retail agglomerations, is one of the central challenges and contributions to the development of the field (Xia, Zou and Su, 2018).

Numerous approaches have been developed to build meaningful clusters based on sets of retail and service locations (Yang et al., 2018; Xia, Zou and Su, 2018). However, the research conducted on country-wide data is limited (Pavlis, Dolega and Singleton, 2018; Mackaness and Chaudhry, 2011). In addition, no research has been done on the specific situation of and regulations in Germany. Current approaches to defining central retail agglomerations are based on a comprehensive collection of data, site visits and surveys of the local population. However, they have been limited to individual municipalities. (Lichtenberger, 1963; Acocella, 2019).

Since no nationwide data set of central retail agglomerations is available, the results will—probably for the first time—shed light upon this topic and enable a thorough analysis across municipal borders. In this study, policy-makers and city planners, both at the local and the regional level, will find starting points for evaluating the needs of the population with regard to goods for daily use, or adapt to changes in demographic distribution. Another use case would involve companies evaluating, expanding or optimizing their store footprint.

This work follows two topics which are of interest for me. More specifically, this is the OpenStreetMap project that aims at describing the physical world with spatial data at an ever-increasing rate and level of detail (Anderson, Sarkar and Palen, 2019; Jonietz and Zipf, 2016; Touya et al., 2017). Using this rich data source to draw concrete conclusions, derive additional and helpful data or to answer strategic questions is one key motivation of this paper. Secondly, I am interested in expanding my expertise regarding the spatial analysis toolbox. In this context, it is crucial to apply spatial analysis to relevant, real world challenges, investigate the space of clustering techniques and learn about the structure of central retail places, as well as the application and automation of GIS-workflows.

## 1.2 Problem statement

A comprehensive classification of retail spaces, their extents and composition is the first step to understanding the relationship between the use of retail spaces and changing consumer behaviors (Pavlis, Dolega and Singleton, 2018; Lichtenberger, 1963). Transparency about where people shop and what defines these places helps to build a common understanding, sustain access to essential goods and develop attractive cities (Dolega et al., 2019). In addition, a comprehensive data set can act as an input for further analysis, enabling the acquisition of knowledge. City and regional plan-

ners can evaluate the supply level, while shop owners might focus on optimizing their store footprint. Learning from this data allows assumptions about the best practices and success factors when designing and developing attractive retail locations or networks of retail locations. Overall, comprehensive data would help to focus limited resources and plan for the future or set the ground for strategy development. Furthermore, recognizing temporal changes in this system allows for the identification of arising challenges and opportunities, thus supporting interventions or business opportunities.

Central retail agglomerations are undergoing constant changes (Dolega et al., 2019). The consistent observation of research and industry experts is that the level of change introduced to the system has increased over the years (Brown, 1994) and will not slow down in the future (Dearden and Wilson, 2011; Schiller, 2001). Central forces influencing the system include, among others: the aging population, increasing urbanization, consolidation of retail chains, competition from online retail and an increasing ecological awareness of the population, which leads to changes in behaviors; for example, the footfall is decreasing and trends towards larger store formats or concentration of some retail formats in the most central places are emerging (Dolega et al., 2019; BBSR, 2017). The changes overlap and the current and potential impact on the overall system is unclear, as it varies from location to location. What remains true is that today's central retail agglomerations are challenged and large numbers of centers are declining (Acocella, 2019). Simultaneously, the change in retail agglomerations is tracked inconsistently at a national scale, and public research is only available for small extents. Most large cities only perform thorough analyses of their central retail agglomerations on an irregular basis (Acocella, 2019). Moreover, this analysis becomes a challenge when done automatically, over a long time and at a larger scale (Pavlis, Dolega and Singleton, 2018), as no comprehensive data on retail outlet locations, suitable algorithms and workflows is available for Germany.

Location decisions are expensive and have long-term effects (Hickson, 1986). As a consequence, it is important for the stakeholders involved in retail agglomerations to learn about changes early on, thus being able to develop timely strategies to preserve the investment, steer potential new development, divest or solve local challenges. In this context, a lack of understanding and transparency might lead to unwanted effects or late and uncoordinated reactions. Municipalities might choose strategies to outdo the competition from neighboring municipalities and attract additional purchasing power. In large areas of rural Germany, access to local supply is already a challenge (Kulke, 2020b); without constant monitoring, this process can lead to unwanted effects, and without a clear center-favoring strategy, new offerings might appear at unwanted locations, potentially cannibalizing existing centers.

### **1.3 Objectives**

To create national wide transparency on the central retail agglomerations and to describe and understand the current situation in Germany, this work proposes a consistent approach for a systematic identification and classification which can be easily updated in the future. To achieve this aim, five sub-objectives have been defined: (1) develop a methodology using clustering algorithms to identify retail agglomerations and classify them by their type, (2) parameterize the analysis based on external research and observations to derive meaningful results, (3) validate the approach and results with official, commercial or self-sourced data and evaluate the quality of the outcome, (4) integrate the GIS-workflow in such a way that regular updates become feasible, and (5) describe the derived



insights from the data to learn about retail agglomerations and gain an initial perspective into the structure and composition of retail spaces in Germany.

In this context, three key challenges have to be solved:

- Finding the optimum clustering and classification methodology to account for the diverse size, shape, noise locations and composition of retail areas over a large extent
- Parameterizing and validating the tools and results based on retail domain knowledge and observations
- Automating the process to a high degree to enable future updates

The final output includes a data set referring to the central retail agglomerations, groups of stores, isolated locations and all the center-relevant retail locations which have been identified, characterized and classified. The second output is a standalone QGIS plugin automating the process of data cleaning, cluster identification, characterization and classification.

## 1.4 Review and synthesis of literature

The study of the system of central locations has a long tradition in the German-speaking countries (Lösch, 1940; Christaller, 1933; Carol, 1960). The principles of a hierarchical structure for the function of cities and inner city centers found their way into national, regional and municipal spatial planning paradigms (Greiving, Flex and ARL, 2016; Deutschland, 2020). **Central retail agglomerations** are spatially delimitable areas of a municipality which, due to existing retail uses have a supply function beyond the immediate vicinity and are often supplemented by various services and gastronomic offers (Bunzel and Difu, 2009). These areas are defined by the municipalities as main, sub- or minor retail agglomerations in spatial and functional terms. Municipalities are required to concentrate relevant retail functions and to develop and protect these areas (Heinritz, Klein and Popp, 2003). The centrality of a center is defined by the type and number of retail locations (Kulke, 2017). Center-relevant retail locations within these centers can be classified in various ways (Assortment, frequency of demand, local supply, market areas) (Daniels, 1993). The traditional method of identifying central retail agglomerations would be through detailed on-site surveys. Broadening the extent to a whole country such as Germany requires spatial techniques, e.g. clustering methods.

### 1.4.1 System of central places and spatial planning in Germany

Most theoretical approaches explain the spatial distribution and the formation of central places by the size of the market areas of provider-based services (e.g. supermarkets or hairdresser) (Lösch, 1940; Christaller, 1933). One theory to explain the regular occurrence of retail agglomerations and the formation of centers is the theory of central locations (Christaller, 1933). The theory assumes that every good offered at a location supplies a market area in the vicinity. When the distance to the supply location increases, the quantity demanded by the consumers decreases, because these, in addition to the price of the product, have to bear the higher transport cost as well. The following applies to central locations in the hierarchical system:

- The rank of the center results from the highest ranking good offered there
- Higher centers have all goods of the lower ranking centers
- Centers of the same hierarchical level provide a similar supply of goods

This system is also called the supply or market principle. Criticism of this approach focuses on its simplicity and the unrealistic assumption of homogeneity (Kulke, 2017). The concept of central places is crucial to the German Spatial Planning Act (Raumordnungsgesetz (ROG), which requires that settlement activity is concentrated in a system of efficient central places within a decentralized settlement structure (§ 2 Section 2 Nr. 2 ROG) (Greiving, Flex and ARL, 2016; Deutschland, 2020).

The ROG defines three fundamental concepts on the supply of the population with central goods (§ 2 Section 2 Nr. 3 ROG) (Heineberg, 2017). First, the provision of fundamental services and infrastructure, in particular the accessibility of basic services and facilities for all population groups, must be ensured appropriately to guarantee equal opportunities in the sub-regions; this also applies in sparsely populated regions. Second, social infrastructure is to be concentrated primarily in central locations; the accessibility and sustainability criteria of the central-location concept are to be flexibly geared to regional requirements. And third, the spatial prerequisites for maintaining inner cities and local centers as central retail agglomerations must be created. Other aspects guide the protection of critical infrastructure, a good and sustainable reach of all areas or transfer of traffic from road to rail and water.

The states detail the central framework law by developing land development plans, with which they designate, develop and promote their area in accordance with the basic concepts of the ROG. The municipalities are classified in a hierarchically and functionally structured model of order. Centers have ascending catchment areas (supply-, interrelationship or supplementary areas), an increasing range of offered goods and services and an increasingly dense infrastructure (Greiving, Flex and ARL, 2016; Kulke, 2017):

- **Sub-center** (basic or small centers) serve to particularly cover the basic supply of short-term or daily needs and parts of the medium-term needs. This includes the availability of a primary school, community rooms for cultural events, a supermarket, a bakery and doctors, among others. This type of municipalities have a rather small retail agglomeration and a population of less than 10,000 and up to 25,000 people.
- **Intermediate** centers serve to cover fundamental and medium-term or higher periodic needs. This includes primary and secondary schools, vocational schools, institutions of art and culture, theaters, hospitals, and retail of clothes and shoes, among others. Depending on their location within the federal state and the distance to the next regional center, this category typically corresponds to a city of 25,000 to around 500,000 inhabitants (e.g. Oberhausen).
- **Regional-centers** form the highest level within the center hierarchy. They serve to meet the basic, long-term and the higher and specialized periodic needs. These include, among others, institutions of higher education, institutions of art and culture such as Operas, concert halls or museums, medical care in hospitals with specialized departments, retail supply of periodic demands of all types. The category typically includes cities with 500,000 to several million inhabitants (e.g. Berlin).

As far as centers of one certain hierarchical level are concerned, it must be noted that each has typical features and a specific distribution of public services of different quality levels. Some services are only permitted if they correspond to the rank of the municipality. What central location is allocated to the respective supply or demand level of a city is defined in the regional planning and development plans of the 16 federal states. Due to the heterogeneous distribution of the population, it

is a political goal to maintain equal and efficient access to central services and to the historically grown centers for the population across Germany (NRW, 2020; Niedersachsen, 2020; Bayern, 2020).

### 1.4.2 Central retail agglomerations

A further development of the system of central places theory is the theory of market networks by Lösch. This theory assumes that the market area size is product specific and that the suppliers cannot achieve extra profits because of perfect competition (Lösch, 1940). Further to be considered are different population densities in an area. If the density is lower, the market areas must be larger to achieve the required minimum turnover. The application of the central-local theory to the inner structure of cities finally led to the realization of a hierarchical system of suburban centers (Carol, 1960). These inner city centers are called central retail agglomerations (Zentrale Versorgungsbereiche). **Central retail agglomerations** are spatially delimitable areas of a municipality which, due to existing retail uses, fulfill a supply function beyond their immediate vicinity and are often supplemented by various services and gastronomic offers (BVerwG, judgement of 11.10.2007, ref. 4 C 7.07)

In this context, municipalities are required to adapt their local land-use plans to comply with the goals of the regional and national ones (§ 1 Section 4 BauGB). Furthermore, municipalities are instructed to ensure the preservation and development of central retail agglomerations (§ 1 Section 6 Nr. 4 BauGB). Five principles of the building code (BauGB) guide the decision-making process to retain proper access to central services for the population:

1. **Congruence requirement:** Compliance with the principles of central places, according to which the range of goods and sales areas should correspond to the supply mandate and the area of interdependence
2. **Concentration requirement:** appropriate and sustainable bundling of services of general interest in central locations
3. **The need for integration:** securing and developing trade functions, especially in inner cities and town centers, in alignment with the urban development policy
4. **Coordination requirement:** Regional planning assessment of large-scale retail trade projects in a regional context and regional planning procedures
5. **Prohibition of impairments:** In case of deviation from the first three principles, projects have to be evaluated and planned in alignment with neighboring municipalities

To define the central retail agglomerations, the regional development program of the federal states provides the municipalities with guidelines of varying depth. The common development goal of all plans is focused on ensuring a sustainable and fair supply of goods for the population, including those that do not have the benefit of a privately used car (Bunzel and Difu, 2009). As an example, this paper will discuss the regional development program of the State of North Rhine-Westphalia (§ Section 24 a LEPro NRW) (NRW, 2020). Central retail agglomerations are defined by the municipalities as main, sub- or minor supply centers in spatial and functional terms. New retail outlets with center-relevant product ranges may only be allowed in cities of a respective centrality or in a city within a central retail agglomeration (§ 11 (3) BauNVO) (Junker and Kühn, 2006).

This work follows the classification of central retail agglomerations, as defined in the latest LEPro, as particularly worthy of protection, specified and detailed both in expert reports and in retail- and center concepts developed by cities (Bunzel and Difu, 2009; Kulke, 2020a; Acocella, 2019). One of the most important elements of this classification is the distinction of the areas by their supply function for the local community (Heineberg, 2017).

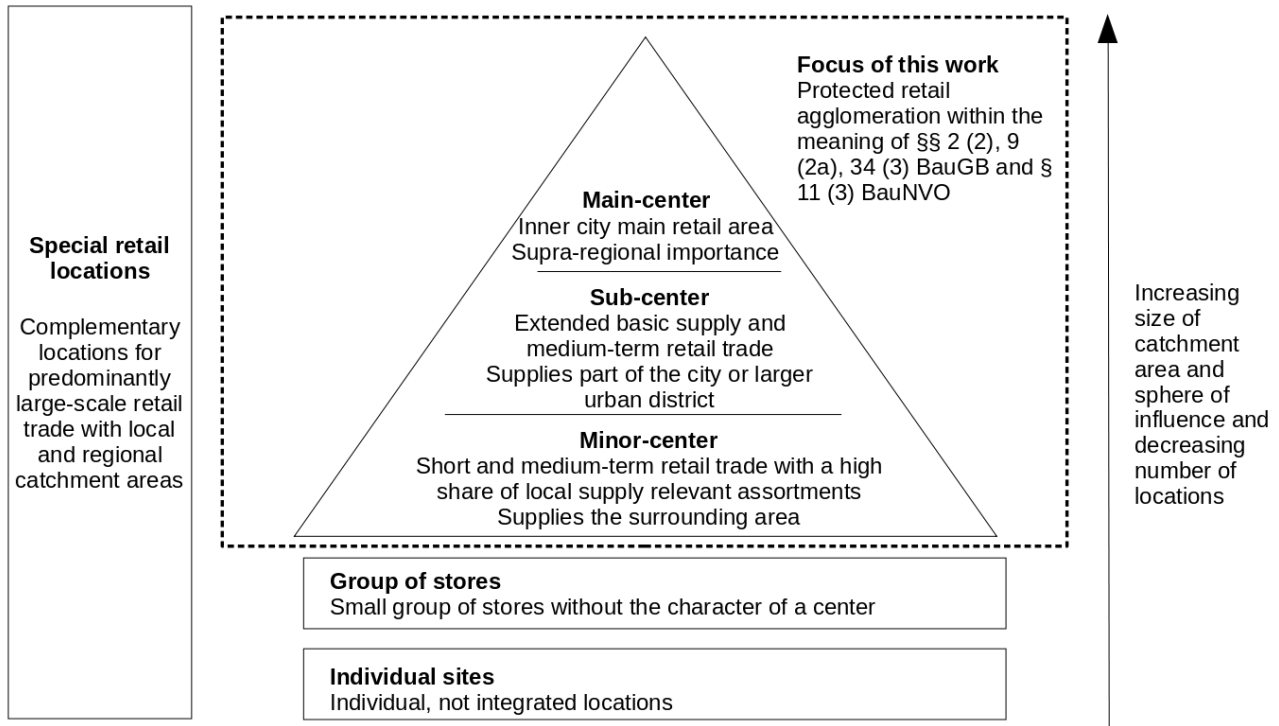


Figure 1: Basic elements of a work-sharing, hierarchically tiered center model adapted representation from (Bunzel and Difu, 2009)

Table 1: Characteristics of inner city centers, adapted and supplemented from (Carol, 1960; Kulke, 2017)

Center	Catchment area	Characterization and typical retail and service offering
<b>Main center (City center)</b>	Supra-regional importance for the city and the surrounding area	Inner city main business area with special emphasis on medium and long-term retail trade (including department stores and specialist shops), high-level business-oriented services (e.g. banks), high-quality public/social institutions (e.g. city administration, medical specialists), high-quality personal services (e.g. hotels and theaters).
<b>Sub-center (Nebenzentrum)</b>	Part of the city or larger urban district	Extended basic supply and medium-term retail trade (e.g. supermarkets, drugstores and specialist shops for books or clothing), individual business-oriented services (e.g. specialist lawyers, tax consultants), social facilities (e.g. medical specialists), personal services (e.g. restaurants and fitness studios). However, the focus is on relevant assortments of the local supply.

<b>Minor center (Nahversorgungszentrum)</b>	Surrounding area, urban district or surrounding settlement areas	Short and medium-term retail trade (e.g. supermarkets, bakeries, pharmacies and certain medium-term consumer goods), social institutions (e.g. general practitioners), simple personal services (e.g. hairdresser and post offices). With a high share of relevant assortments of the local supply.
<b>Group of stores (Nahversorgungsstandort)</b>	Surrounding building blocks	Small group of stores without the character of a center and low footfall frequency. Location often includes at least one grocery store and simple personal services (e.g. snack bar, hairdressers, postal services)
<b>Individual sites (Streulage)</b>	Local area	Areas that can no longer be delimited as shopping locations with individual, unintegrated locations, often grocery stores or personal services such as hairdresser and general practitioners.

The hierarchy shown above can be observed with all its characteristics in cities of a higher rank (regional centers). Cities usually comprise several delimited zones of different size and supply function (Heineberg, 2017). However, not all categories can be observed when it comes to intermediate centers and sub-centers. This is illustrated in the table below:

Table 2: Differentiation of inner-city hierarchical systems according to city size, by (Kulke, 2020b)

	<b>Large sized cities</b>	<b>Medium-sized cities</b>	<b>Small towns</b>
<b>Number, diversity and nomenclature of central retail agglomerations</b>	<i>Main center</i>	-	-
	Sub-center	<i>Main center</i>	<i>Main center</i>
	Minor center	Minor center	-
	Group of stores	Group of stores	Group of stores
	Individual sites	Individual sites	Individual sites

Beside the size, composition and location of these centers, a number of additional criteria have emerged from practice and case law on how to delimit central retail agglomerations (BVerwG, judgement of 11.10.2007, ref. 4 C 7.07) (Bunzel and Difu, 2009).

- Integrated location (surrounded by residential buildings)
- Density of the retail and service stocking
- Urban development qualities (architecture, street space design)
- Structural, natural, traffic barriers (water bodies, roads or railway lines)
- Urban discontinuities (street space design, building structure)
- Supply function of the center beyond the immediate vicinity

### 1.4.3 Center formation of retail locations

Theories of spatial agglomeration of retail locations explain the factors that influence the concentration of service locations at certain locations. The approaches recognize the interaction of supply and

demand, and that competing services often settle close to each other (Kulke, 2017). One explanation was developed by R. K. Nelson. The fundamental idea is that there are three factors influencing the sales volume at a certain location: the own attraction, the shared attraction and the foreign attraction.

*Table 3: Factors influencing sales volume at a certain location (Nelson, 1958; Kulke, 2020b, 2017; Krider and Putler, 2013; Lösch, 1940)*

<b>Own attraction</b>	<p>Describes the number of customers and the sales volume that a provider can achieve due to its attractiveness. Decisive factors are the type, size and variety of the offer, as well as activities that increase attractiveness (advertising and promotional activities). Providers that generate most of their sales from generative business are, for example, department stores or car dealers. They can generate sufficient customer flows at individual locations without neighbors. They serve as magnet operations for other businesses (stock exchange for stock traders, port for transport service providers, department store for retail).</p>
<b>Shared attraction</b>	<p>Describes the number of customers and the volume of sales that a supplier can achieve by being located in the vicinity of other suppliers (specialty shop for leather gloves, umbrellas). Companies with a large share of turnover from shared attraction seek locations near other suitable locations (central magnet).</p>
	<p><b>Compatibility advantages or coupling advantages</b></p> <p>Coupling advantages are evident in the geographical proximity of suppliers at one location, such as bank branches, hairdressers, general practitioners, pharmacies, supermarkets and beverage stores or DIY stores, garden centers, pet supplies and trailer rentals. The decisive element is to address the same target group, consisting of people who combine several errands in different types of businesses during one visit.</p>
<b>Foreign attraction</b>	<p><b>Comparative advantages or cumulative advantages</b></p> <p>Companies with comparable offers are concentrated in the immediate vicinity, and customers can compare offers. This is particularly evident for similar, highly specialized suppliers with large market areas. Comparative advantages can be seen in the geographical proximity of suppliers such as jewelers and galleries. In individual or dispersed locations, they are generally unable to attract the necessary clientele.</p>
<b>Foreign attraction</b>	<p>Describes the number of customers and the sales volume that a provider achieves through external frequency providers. Customers visit the location for another reason and use the service offer on the side. Companies with a large share of sales resulting from external attractiveness are hardly able to choose locations other than those of external frequency providers. External attractiveness can be seen in magazine sellers at traffic hubs, dentists, real estate agents, restaurants in retail centers or hotels at attractive holiday destinations.</p>

Further approaches which explain spatial concentrations are the nearest-center binding (Hotelling, 1928) and the ground rent model (Heineberg, 2017). The former describes the locations at which two similar suppliers with the same products and the same prices function settle (Hotelling, 1928).

Since the customers have to make the effort to travel to the nearest location in addition to paying the product price, they will mainly seek out the nearest location. Clear orientations of the population towards the nearest center can be observed for simple goods (supermarket or bakery), towards a sub-center for medium-term goods (clothing or shoes), and towards the city center for medium to long-term goods (optician or jeweler). This ultimately leads to the concentration of locations.

By contrast, the ground rent model is an explanatory approach to the spatial distribution of centers within the urban space. The model establishes a relationship between the turnover obtained at a location after covering the costs and the companies competing for the area (Heineberg, 2017). The highest values are found in the city center, as this is where the highest sales are to be attained. As the distance to the city center increases, the costs for the land decrease (Kulke, 2017). Different users compete for the central areas and can pay different land prices depending on the product offered. Less solvent locations are pushed to the periphery. Overall, a land value surface can be constructed for large areas, whereby higher land values usually mark the locations of central places (Kulke, 2017).

#### **1.4.4 Retail locations**

At the lowest level, the individual stores with their location and specific type of offer have to be recognized. Most retail locations considered to be relevant for a center require the direct interaction between the supplier and the customer, thus classifying as provider-based services. This is the most common form for simple and consumer-oriented services such as supermarkets, as well as for higher-ranking services like optician or jeweler (Kulke, 2017). The classification of retail trade locations into subgroups with common characteristics and, consequently, similar location requirements is helpful to structure the heterogeneous landscape of services. Within the retail-related services, the differentiation by frequency of the required good is most central (Heinritz, 1979).

Different types of shops have different catchment areas. While goods for daily use are cheap and frequently in demand, the willingness to travel long distances is limited (Daniels, 1993; Heinritz, 1979). Therefore, service providers have small market areas and many locations (baker, supermarket). For less frequently needed, high-quality and expensive products and services, customers are willing to travel longer distances (Daniels, 1993; Heinritz, 1979). Consequently, these service providers have larger market areas with fewer locations. Depending on the size of the market areas, the goods can be put in an order – the greater the range, the greater the centrality. The result is a hierarchical organization of the supply locations. Centrality is also often used as a synonym for the lifetime of a product or the short-, medium- and long-term demand.

Table 4: Frequency of demand, adapted from (Daniels, 1993; Kulke, 2020; Bunzel and Difu, 2009)

Frequency of demand	Description
<b>High</b>	All goods with short-term procurement rhythm. Essentially food, short-term consumer goods, which includes detergents, household paper goods, perfumery and over-the-counter pharmaceutical goods, newspapers and magazines, as well as cut flowers and pet food.
<b>Medium</b>	All goods with medium to long-term procurement cycles. These are, for example, textiles, shoes and toys.
<b>Low</b>	All goods with long-term procurement cycles. Products of long-term demand are, for example, furniture, garden and DIY products, home textiles or consumer electronics.

One element which is worth considering when discussing central retail agglomerations is the center-relevant assortment of shops. This term is anchored in state development plans and defines a basic set of assortments of relevance for centers and for the local supply (NRW, 2020). The municipalities are given the opportunity to adapt the list to their local requirement. The proposals range from detailed, multi-page product range subdivisions (e.g. the "Freiburger" and "Kölner-Liste" about "non-center-relevant" parts of the DIY store product range) to simple subdivisions of the entire retail product range based on proposals by the Hauptgemeinschaft des Deutschen Einzelhandels or retail center experts. In their retail and city development plans, most cities follow the suggestions of experts and consultants (Bunzel and Difu, 2009; Acocella, 2019, 2018; Orzessek-Kruppa, 2016). Typical assortments relevant to centers are books, clothing, shoes, household goods and toys. Additionally, relevant for local supply are the assortments food and beverages as well as health and body care products. Another related aspect is the inner city compatibility of retail and service offerings. The criterion "inner-city compatibility" is currently not defined conclusively; the same applies to the method of measurement and the definition of an "incompatibility threshold". Assortments like large electrical appliances, bicycles, garden products or furniture are, however, not relevant or compatible with centers (Bunzel and Difu, 2009).

Finally, some locations draw higher footfall than others. These play the role of magnet locations for shops situated close by (Brown, 1994). More specifically, within small centers, the role of magnet locations is fulfilled by supermarkets, drugstores and pharmacies, while supermarkets, department stores and stores with large sales areas, such as some electronics and cloth retailers, represent magnet locations within larger centers and city centers (Brown, 1994). To comprehensively evaluate the individual locations and centers, additional describing information can be helpful. Retail and city center studies include data such as sales area and the sales area by assortment (Bunzel and Difu, 2009).

#### 1.4.5 Clustering of central retail agglomerations from retail points

Clustering in the context of spatial analysis describes an unsupervised process that groups a set of elements based on their similarity, taking into account attributes such as their location. The distinction results in homogeneous subgroups and differentiates these from dissimilar groups (Gan, Ma and Wu, 2007). In this context, retail locations can be aggregated by cluster algorithms to identify and delimit most central locations. These techniques are part of the larger exploratory data mining



tools and help to reveal covert patterns from large data sets. When attempting to answer spatial questions, researchers only had access to limited data in the past; today, however, the opposite is true – there is an abundance of data available (Miller and Goodchild, 2015). Spatial clustering is one of the central techniques for aggregating and reducing the size of data, which can help in deriving underlying insights from large data sets.

As far as central retail agglomerations are concerned, the traditional method to identify them would be through detailed on-site surveys (Acocella, 2018, 2019). This data collection can be detailed and profound for small areas, but inherently limited when capturing information over larger extents. The volunteered geographic information project OpenStreetMap made a large data set of store locations available, exceeding the possibilities of individual surveys. Based on this data and with the knowledge that central retail agglomerations form clusters, studies can be carried out more thoroughly. As such, the identification and description of urban functional zones has become the topic of a number of recent studies (Pavlis, Dolega and Singleton, 2018; Mackaness and Chaudhry, 2011; Xia, Zou and Su, 2018; Yang et al., 2018). Even so, limited work has been done to explore retail centers over large extents and based on volunteered geographic data (Pavlis, Dolega and Singleton, 2018). A large share of authors studying similar topics recently choose geographies in china as study area (Xia, Zou and Su, 2018; Yang et al., 2018). The rapidly growing urban population makes the need to understand functional compositions most relevant for those geographies, especially when it comes to managing the fast growth and answering to the supply demands of the population.

## **1.5 Structure of this work**

This work is structured in seven chapters. Chapter one introduces the topic and the scientific background, defines the essential concepts, the foundation of this work, and introduces the problem statement, objectives and the scientific question to be answered. In the second chapter, the research design, methodology, the data used and the study area are described. The third chapter presents the results which were found to answer the research question. In chapter four, the results are interpreted, validated and the research method is critically discussed. The fifth chapter focuses on the conclusion and outlook, as well as the answers to the scientific questions. Chapter six lists the references and chapter seven, the appendix, includes some accompanying materials.

## 2 Methods

In this section a framework is proposed to identify, describe and classify central retail agglomerations from a set of retail points. The methodology consists of three main (one to three) and two supplementary parts (four to five): 1. Data acquisition and preparation, 2. Cluster identification and 3. Cluster characterization and classification, 4. Interpretation and implications, and 5. Validation of the results. Data from OpenStreetMap and other adjacent data sets are brought together and used to identify and describe central retail agglomerations in Germany. QGIS was chosen as the main GIS tool to automate the workflow.

### 2.1 Research design

Point location data is used to define and describe the extent of retail and service agglomerations in Germany. The locations are defined by their position (latitude and longitude), their retail and service category (supermarket, shoe shop, optician) and the function which the location provides to the population in its vicinity. The foundational data used in this work is sourced from OpenStreetMap. Data sets enriching the retail location information are obtained from different sources, such as domain knowledge from retail industry experts, the Federal Agency for Cartography and Geodesy Germany (BKG), the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR) and inherent information derived from the analysis of the raw data.

The key methodology to be used is cluster analysis. This type of analysis is counted among the unsupervised machine learning methods and is aimed at grouping objects based on their spatial distribution (Gan, Ma and Wu, 2007). Cluster identification is based on the density and location of retail points. Research shows that, for the identification of retail agglomerations, the DBSCAN (Ester et al., 1996) algorithm, with other assisting spatial analyses, leads to good results (Pavlis, Dolega and Singleton, 2018). This methodology seems viable for identifying retail spaces, as the extent of the agglomerations is defined by spatial discontinuity (Dearden and Wilson, 2011). The chosen cluster algorithm is also flexible enough to account for the diverse differences which retail clusters can show. In a subsequent step, the clusters are characterized and further classified by their supply function, using a transparent decision tree approach (Breiman, 1998). Separating the identification and the classification in two steps gives more flexibility when doing the individual analysis. Parameters in this process can easily be adapted to other geographies and their specific circumstances. The first three steps are then automated by developing processing models in QGIS. The fourth and fifth step comprise the interpretation and implications of the results, as well as a detailed validation of the underlying data and results, and a discussion of the method. This will be detailed in the chapter devoted to the results and discussion.

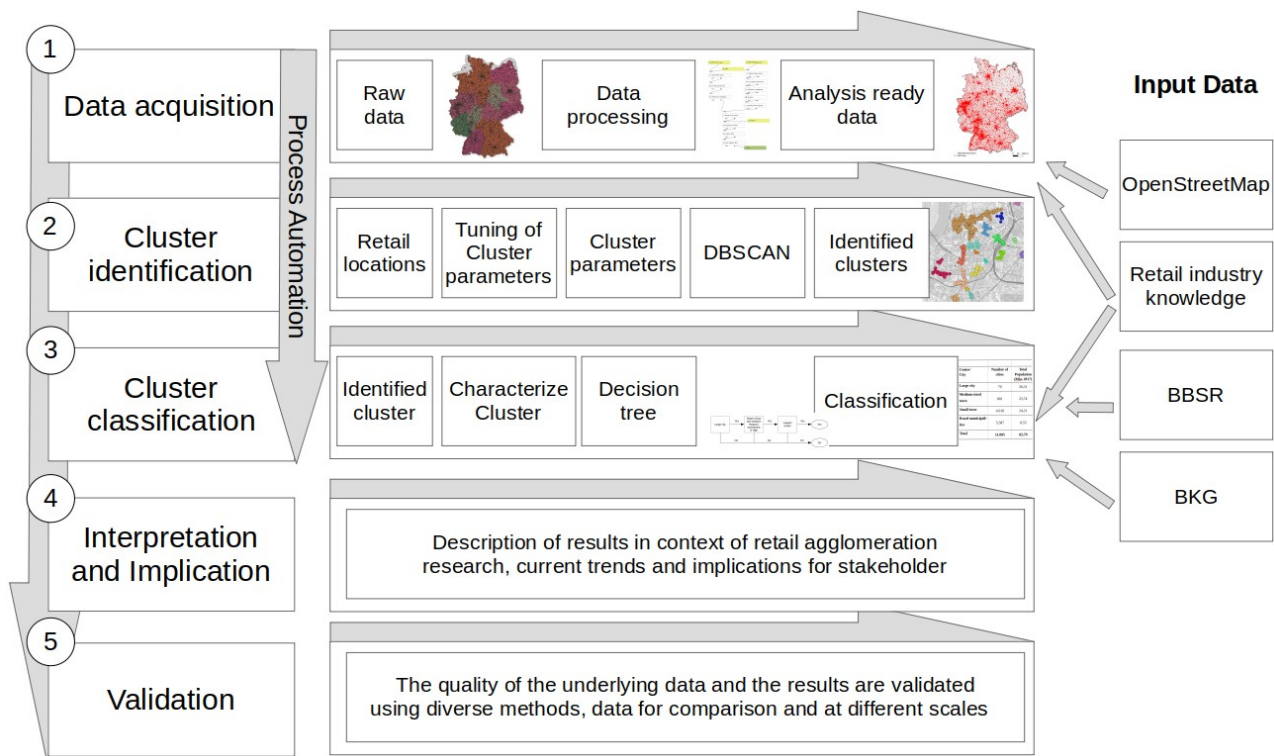


Figure 2: Research design

## 2.2 Study area and scale of analysis

The study area comprises the entire surface of Germany with a total area of 357.578,17 sq Km and a population of 83.019.213 as of 31th December 2018 (StaBu, 2020). A country-wide scale has been chosen, as there is little documented research on identifying and classifying central retail agglomerations in an automated way. Furthermore, no such research has been conducted for Germany. The subsequent possibilities for further analysis which the resulting data enables represent another argument in favor of the scale chosen. Since a replication of the analysis for other European countries is planned, a European-centric CRS has been chosen. In line with the best practice for the intended analysis, all used data is projected to “CRS EPSG:3035 - ETRS89 / LAEA Europe – Projected” prior to any analysis. The cluster analysis is performed on the highest available detail, which is the location of the center-relevant retail locations. The validation and analysis of the data is done at varying scales to the level (e.g. individual clusters to regional and nationwide scales).

## 2.3 Data

Points of interest data from OpenStreetMap, classified based on expert knowledge, are used for the central analysis alongside additional supporting data sourced from the Federal Agency for Cartography and Geodesy Germany (BKG) and The Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR). To concentrate on the conceptual design, the parameterization of the spatial analysis and the national scale of the analysis, only few publicly available data sources are used. This leads to a number of compromises in terms of comprehensiveness of the data and the level of detail per data point.

Table 5: Overview of data sources used

Providing institution	Data	Projection / Format / Scale	Vintage and Source
<p><b><u>OpenStreetMap</u></b>  OpenStreetMap is a global database of geographic information that can be observed on the ground. The data is contributed by private individuals and companies.</p>	<p>Clustering  Data preparation  Points of Interest  Country Border  Mapnik background map</p>	<p>Unprojected,  WGS84  (EPWG:4326)    Shapefile    Point location</p>	<p>January 1<sup>st</sup> at 21:59 o'clock  (OpenStreetMap contributors, 2020)</p>
<p><b><u>Federal Agency for Cartography and Geodesy Germany (BKG)</u></b>  BKG is a technical agency under the Federal Ministry of the Interior, Building and Community. The agency provides geodetic reference systems and basic information for the territory of the Federal Republic of Germany.</p>	<p>Data preparation  Administrative areas  Names of regions  Population  Area of municipality</p>	<p>UTM, zone 32,  Ellipsoid  GRS80, Date  ETRS89    Shapefile    1:250 000</p>	<p>December 31<sup>st</sup>  2017  (BKG, 2020)</p>
<p><b><u>The Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR)</u></b>  BBSR is a research institution under the portfolio of the Federal Ministry of the Interior, Building and Community. It provides the Federal Government with sectoral scientific consultation in the political fields of spatial planning, urban development, housing and building.</p>	<p>Ongoing urban monitoring - spatial delimitations    Types of cities and municipalities in Germany</p>	<p>None    Tabulated    Municipality</p>	<p>December 31<sup>st</sup>  2017, published  2019  (BBSR, 2020)</p>

Apart from the actual analysis, all data is also used for visualization and statistical assessments. Other adjacent data sources meant for visualization are mentioned at the appropriate place. The validation data is described in the respective chapter.

### 2.3.1 OpenStreetMap

The point-of-interest-related data for the clustering analysis is sourced from the OpenStreetMap database (OpenStreetMap contributors, 2020). OpenStreetMap is a collaborative project of volunteers whose aim is to create a global editable database of spatial information (Mooney and Corcoran, 2014). The data generated by the project, rather than the map itself, is considered to be the primary output (Mocnik, Mobasheri and Zipf, 2018; Mooney and Corcoran, 2014). The creation and growth of OpenStreetMap has been motivated by restrictions on use or availability of digital map information across much of the world, and the advent of inexpensive portable satellite navigation devices (Goodchild, 2007).

Retail and service locations make up a large part of the data found in OpenStreetMap. Globally, the database holds more than 4 million shops alone (OpenStreetMap contributors, 2020); the number of available center-relevant shops is roughly expected to be around 250.000 elements in Germany. Processes involving the sourcing, cleaning and preparation of the data will be limited to a minimum to achieve a maximum degree of automation and reproducibility.

The raw data from OpenStreetMap is stored in a specific format. Extracting and mapping the data to a usable and easily accessible format is a complex and time-consuming task. For instance, the tag Shop alone delivers 1.125 characteristics. Extracts of the OpenStreetMap data are therefore sourced from the free Geofabrik GmbH download service (Geofabrik GmbH, 2020). This product contains a series of shape files of which the once containing points of interests were used. The data set includes the location information of retail, services, health, public and other amenities, including a detailed description of their specific type, such as supermarket or police station. Other possibly available data such as the name, opening hours or wheelchair accessibility are not used, as these attributes are not available for all locations.

*Table 6: Attributes of the OpenStreetMap raw data*

<b>Attribute</b>	<b>Description</b>	<b>Example</b>
<b>osm_id</b>	OpenStreetMap object Id, not necessarily unique	1000012684
<b>code</b>	4-digit code describing the feature class (type)	2513
<b>type</b>	Class describing the type of the feature	florist
<b>name</b>	Name of the feature	Sander & Sanders

Although the data is already extracted, formatted and unified, additional processing steps have to be made to prepare the data for the planned analysis. Retail and service locations are usually represented as points and point locations are required for the cluster algorithms to identify retail agglomerations. However, numerous features have been collected as outlines of buildings or sites and have to be transformed. Additional steps include projecting the data and clipping it to cover the territory of Germany, as well as removing duplicate locations with identical or nearly identical locations. The processing is limited to a minimum, with few dependencies on additional data (e.g. the country boundary), is meant to be reproducible on most computers and open to data for different geographies.

Other sources of points of interest data have been considered but not further investigated, because they are either proprietary, accessible under strict license agreements or associated with high costs. Another limiting factor is that most commercial data sets are bound to certain geographies, specific types of retail and services or focus on chained retail outlets. Therefore, OpenStreetMap was the only choice.

### **2.3.2 Types of cities and municipalities**

The BBSR typification of cities and municipalities distinguishes between large cities, medium-sized cities, small towns and rural municipalities (BBSR, 2020). The city and municipality types were first developed in 2003 as part of a study on urban redevelopment and have been updated regularly ever since. The approach focuses on the function and significance of cities in relation to their size and importance for their surrounding (Heineberg, 2017). The criteria are the size of the municipality (population) and its function as a central town. If a municipality within an association of municipalities or the unitary municipality itself has at least 5,000 inhabitants or at least a basic central function, then it is referred to as a "town". If one of these conditions does not apply, then it is a rural municipality.

Table 7: Cities differentiated by centrality and size (BBSR, 2020)

<b>Large city</b>	Municipality of a municipal association or unitary municipality with at least 100,000 inhabitants; these cities usually have a super central function or at least a medium central function. The group of large cities can be divided into 15 large cities with at least 500,000 inhabitants and smaller cities with less than 500,000 inhabitants.
<b>Medium-sized city</b>	Municipality with 20,000 to less than 100,000 inhabitants; these towns mostly have a medium central function. The group of medium-sized towns can be divided into large medium-sized towns with at least 50,000 inhabitants and small medium-sized towns with less than 50,000 inhabitants.
<b>Small town</b>	Municipality with 5,000 to under 20,000 inhabitants or at least a basic central function. The group of small towns can be divided into larger small towns with at least 10,000 inhabitants and small towns with less than 10,000 inhabitants.
<b>Rural municipalities</b>	All municipalities falling under the threshold of a small town and without a central function.

The classification of municipalities by their centrality in accordance to the ROP law is in the responsibility of the 16 states of Germany (Deutschland, 2020). This leads to regional differences in definition and nomenclature of the center hierarchy, which is also established at different time intervals. Seeing as the classification is not centrally available, this results in significant challenges when attempting to work with the data. As an alternative, data from the BBSR is used. The latter follows a coherent approach, resulting in a consistent data set, and ensures comparability across regions (BBSR, 2020). As such, the city and municipality defined by type data serve analytical and comparative purposes, as not all phenomena and trends can be represented. However, the data is suitable for an approximation of the hierarchy of municipalities or the search for explanations for certain observations. It is particularly useful for small-scale municipal analyses - both for describing the current situation (e.g. unemployment rates, purchasing power or real tax revenue) and for measuring development (e.g. population development or net migration). Therefore, the data can be seen as an appropriate alternative to the official classification.

## 2.4 Cluster identification

After preparing the data, the retail locations are used to identify potential central retail agglomerations. In order to reliably detect these agglomerations, a cluster algorithm that meets the numerous requirements, the specific use case and the national scale has to be found. Central to successful clustering is the tuning of the parameterization of the algorithm, because this significantly influences the number of clusters to be detected and the final quality of the results.

## 2.4.1 Requirements

Many clustering algorithms exist that are potentially suitable for defining the extent of central retail agglomerations. Therefore, one central decision when clustering spatial data is to choose a clustering algorithm that suits the requirements of the specific use case. As such, before choosing an algorithm, one must define and discuss the requirements for detecting central retail agglomerations.

*Table 8: Requirements for choosing the clustering algorithm*

<b>Description</b>	<b>Actual characteristics of clusters to be considered</b>
<b>Cluster from point geometries</b>	The retail locations to be used come in point format.
<b>Varying form</b>	Retail centers differ in their form depending on the local geography. For example, they can be compact or chained.
<b>Varying size</b>	Clusters comprise an arbitrary number of locations and can differ in size depending on their function.
<b>Varying density</b>	Depending on the type of the center, the density of locations can range from a few meters to larger distances.
<b>High density vs. Outliers</b>	Central retail agglomerations can be defined as dense agglomerations of locations which need to be detected, and numerous locations which are positioned at individual sites need to be identified as outliers.
<b>Exploratory and unsupervised for large extents</b>	The number of clusters is unknown before the clustering and the process has to be done on a large area with a large number of locations, returning results in a meaningful time frame on standard retail hardware.
<b>Few input parameters and self-adjusting algorithm</b>	Results are usually sensitive to the chosen Input parameters required for most of the cluster algorithms. As the diversity of the locations is high, fixed parameters for a large area should be limited and the algorithm should be capable of self-adjusting to the regional circumstances.
<b>Proven and commonly implemented algorithm</b>	The analysis shall be implemented in a workflow of a common GIS that can be easily applied to different data sets.

The diversity of central retail locations is high as the exemplary visualization of some clusters illustrate. The maps below show center-relevant retail locations at a scale of 1:10.000 (OpenStreetMap contributors, 2020).

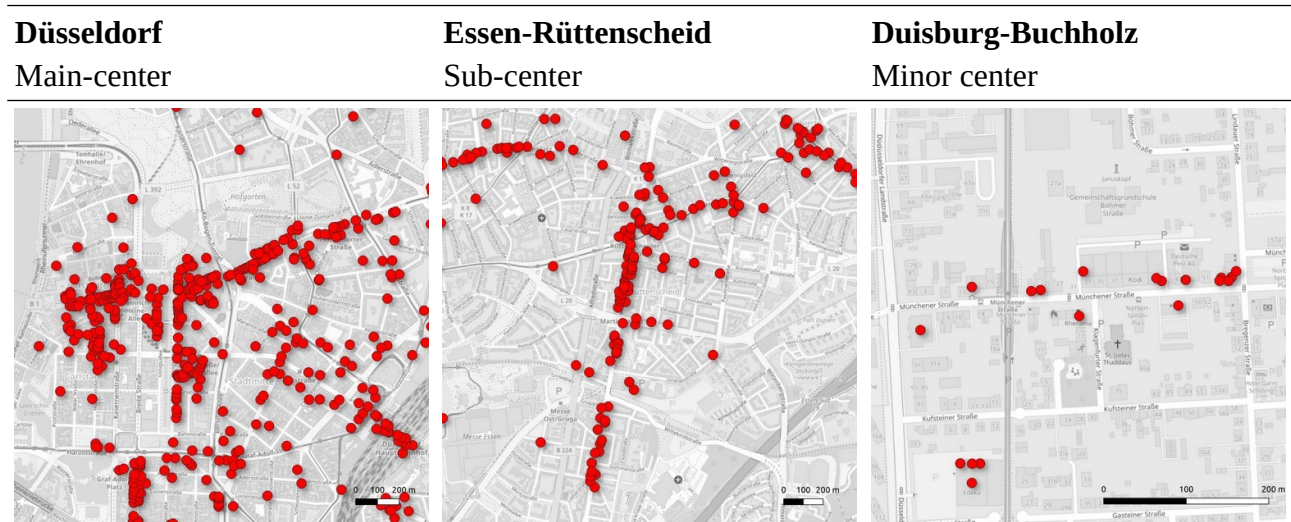


Figure 3: Examples illustrating the diversity of shape and size of central retail agglomerations

## 2.4.2 Comparison and selection of relevant cluster algorithm

A broad range of cluster algorithms have been applied in the context of identifying central places. Algorithms range from the traditional K-means clustering (Macqueen, 1967) to the widely used DBSCAN algorithm (Ester et al., 1996) or further developments such as HDBSCAN (Campello, Moulavi and Sander, 2013) or SNN (Ertöz, Steinbach and Kumar, 2003). The following table compares the above outlined requirements with a selection of clustering algorithms and gives reasons for the selection of DBSCAN in this work. The comparative summary of available clusters techniques indicates DBSCAN as the most appropriate algorithm.

Table 9: Comparison of DBSCAN with alternative clustering techniques. Adapted and extended (Devkota et al., 2019)

Cluster algorithm	Comment
<b>DBSCAN (Ester et al., 1996)</b>	<ul style="list-style-type: none"> <li>• Works unsupervised, as the number of clusters is not known before clustering</li> <li>• Clusters can be of arbitrary size and shape</li> <li>• Detect clusters over the entire study area</li> <li>• Handles outliers as noise</li> <li>• Requires little domain knowledge to determine the input parameters</li> <li>• Good performance when demarcating clusters</li> <li>• The parameters can be set based on domain knowledge or in case the data is well understood.</li> <li>• Stable results over multiple runs</li> </ul>
<b>HDBSCAN (Campello, Moulavi and Sander, 2013)</b>	<ul style="list-style-type: none"> <li>• Varying density leads to high fragmentation and clusters of low density</li> </ul>



<b>K-means (Macqueen, 1967) and K-medoids (Schubert and Rousseeuw, 2019)</b>	<ul style="list-style-type: none"> <li>• Requires knowledge about the number of clusters to be generated</li> <li>• Does not handle outliers as noise</li> <li>• Result is not stable over several runs</li> <li>• Assumes the form of the underlying clusters is globular</li> </ul>
<b>Spatial Point Processing methods such as Local Moran (Anselin, 2010) and Getic-ord Gi (Getis and Ord, 2010)</b>	<ul style="list-style-type: none"> <li>• Does not outperform generic clustering algorithms like DBSCAN in delineating aggregated data and shaping generated clusters</li> </ul>
<b>Self-Organizing Maps (Dehuri et al., 2006)</b>	<ul style="list-style-type: none"> <li>• Performs worse than DBSCAN for clusters of arbitrary shape and size</li> </ul>
<b>Mean-Shift Algorithm (Yizong Cheng, 1995)</b>	<ul style="list-style-type: none"> <li>• Does not handle outliers as noise</li> </ul>
<b>Kernel Density Estimation (Rosenblatt, 1956)</b>	<ul style="list-style-type: none"> <li>• Cannot draw a clear distinction between different clusters</li> </ul>
<b>Affinity Propagation (Dueck, 2009)</b>	<ul style="list-style-type: none"> <li>• Assumes the form of the underlying clusters is globular</li> </ul>
<b>Spectral clustering (Ng, Jordan and Weiss, 2001)</b>	<ul style="list-style-type: none"> <li>• Requires knowledge about the number of clusters to be generated</li> </ul>
<b>SNN (Yang et al., 2018; Ertöz, Steinbach and Kumar, 2003)</b>	<ul style="list-style-type: none"> <li>• Does not scale to a country wide extent when run on standard computers</li> </ul>
<b>Delaunay triangulation (ASCDT) (Xia, Zou and Su, 2018)</b>	<ul style="list-style-type: none"> <li>• Does not handle outliers as noise</li> </ul>

DBSCAN has proved to be the most suitable alternative. The algorithm takes the density of the points into consideration and finds the number of clusters explorativ. DBSCAN performs better than other algorithms in cases where clusters are of arbitrary shape or size (Dehuri et al., 2006). The same is true for the ability to differentiate clusters from each other (Wang et al., 2019). Furthermore, DBSCAN can handle outliers as noise. For parameterization only two parameters are required. In comparison to most of the mentioned algorithms, DBSCAN is conveniently built into most of the common GI-Software and performs well even on large areas and sets of data. Finally, other research showed that good results could be achieved in comparable settings (Pavlis, Dolega and Singleton, 2018).

### 2.4.3 DBSCAN description and mechanism

The density-based spatial clustering of applications with noise (DBSCAN) is a clustering algorithm used for class identification in spatial databases (Ester et al., 1996; Schubert et al., 2017). The primary idea behind DBSCAN calculation is to find areas of high density that are isolated from each other by areas of low density. DBSCAN clusters dense point agglomerations based on the point locations and a minimum number of points within a fixed distance. The points located outside of the identified clusters are marked as noise. The algorithm is widely implemented and commonly used

for exploitative data mining and unsupervised machine learning (Kriegel, Schubert and Zimek, 2017). Three central reasons for its popularity are:

1. The algorithm requires only two input parameters, which limits the domain knowledge needed
2. It discovers clusters of arbitrary shape and size
3. The algorithm is one of the fastest clustering methods (Kriegel, Schubert and Zimek, 2017; Ester et al., 1996).

The mechanics of the algorithm can be summarized in four steps (Schubert et al., 2017):

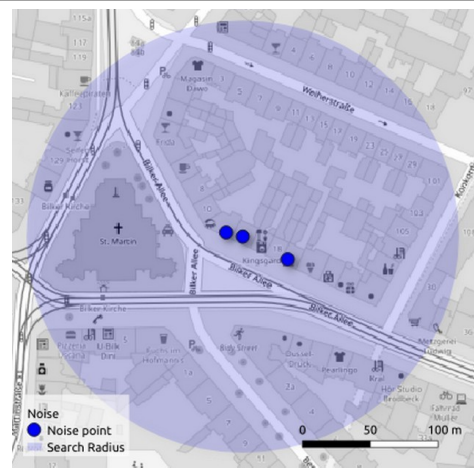
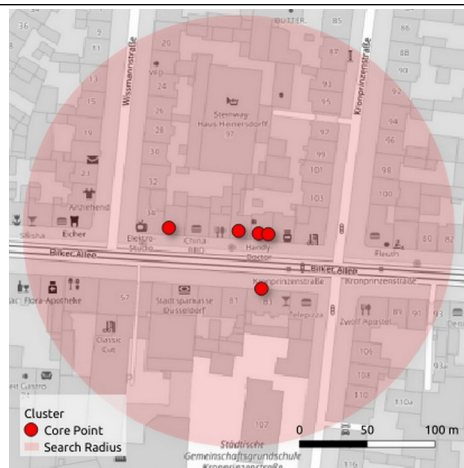
1. Identify the points within the defined radius ( $eps$ ) ; define the core points with more than the ones defined as minimum ( $minPts$ ) in the radius.
2. Ignoring all non-core points, identify the connected core points to form the clusters.
3. All non-core points are matched to the next cluster if they are located within the defined distance.
4. All remaining points are defined as noise.

Beside the point locations to be clustered, the DBSCAN algorithm requires two parameters ( $eps$ ) and ( $minPts$ ) .

Table 10: DBSCAN parameters

Parameter	( $eps$ )	( $minPts$ )
<b>Description</b>	Defines the maximum cut off distance between points to be recognized as part of a cluster.	Defines the minimum number of points to form a cluster.

**Illustration**



**Parameter value**

( $eps = 100 \text{ Meter}$ )

( $minPts = 4$ )

The following diagram illustrates the function of the DBSCAN algorithm for the parameters ( $eps = 100 \text{ Meter}$ ) and ( $minPts = 4$ ) . The red dots are core points, as a minimum of 3 points

can be found for each of them within 100 meters. These points are interconnected and therefore form the initial cluster. The yellow dots are border points, as they are located within the defined distance but reachable from at least one core point. Therefore, they are considered part of the cluster. The blue dot is a noise point, as it is not by any means connected to a cluster.

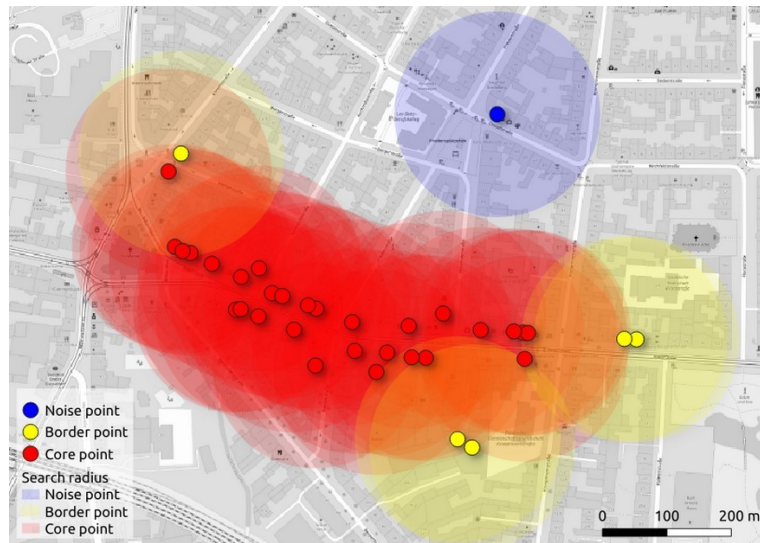


Figure 4: Illustration of the function of DBSCAN

In the context of this work, some challenges resulting from the nature of the DBSCAN algorithm have to be recognized and discussed. DBSCAN is deterministic for the core and noise points. However, it is also reliant on the order of the underlying data and therefore not deterministic for border points that could be connected to more than one cluster. To connect these points to the cluster, the algorithm follows the order of the data. However, such cases are not expected to often occur and should thus have a limited impact on the results (Schubert et al., 2017). Especially challenging is the definition of the parameter ( $\epsilon$ ) if the domain of the data is not well understood and the sensitivity to this parameter is relatively high. In case of highly varying densities and without the ability to define a maximum search distance that applies to all potential clusters, it is not recommended to use DBSCAN (Schubert et al., 2017).

#### 2.4.4 Tuning of clustering parameters

Apart from the location information, all the clustering algorithms discussed above require input parameters to guide the construction of the clusters (Dehuri et al., 2006). These parameters significantly influence the number of clusters to be detected and the final quality.

Two overall goals of the parameter tuning can be defined (Schubert et al., 2017). The first goal is to find good and robust parameters that can be derived from domain knowledge in the best case or test and compare against validation data to tune based on observations and test results. The second goal is to tune the parameters in such a way that more clusters are identified in case of doubt. This is desirable because, after the first step, we want to make sure that most places are detected regardless of whether they are included in the next step – the classification of the clusters–, described as central retail agglomerations or not. Tuning for over-detection is also preferable, as we can assume that the underlying data for the clustering is incomplete. Tuning the parameters for the algorithm which will be used is central to all data mining tools (Dehuri et al., 2006). Defining good parameters for DB-

SCAN requires an understanding of how they are used and knowledge about the underlying data. The developers of the DBSCAN algorithm advise that predominantly domain knowledge be recognized when setting the parameters (Ester et al., 1996). The two parameters, (*esp*) and (*MinPts*) determine the outcome of the DBSCAN clustering. Clusters are identified at locations where the defined minimum density threshold is surpassed.

Table 11: Considerations for selecting parameters (*esp*) and (*MinPts*) (Schubert et al., 2017; Ester et al., 1996; Campello, Moulavi and Sander, 2013; Sander et al., 1998)

Parameter	Considerations
( <i>esp</i> ) Search distance or radius	<ul style="list-style-type: none"> <li>• Choosing a small value leads to large parts of the data not being clustered and recognized as outliers.</li> <li>• Choosing large value leads to large numbers of points being in joined clusters.</li> <li>• A slight preference for a smaller value should be considered.</li> <li>• Calibrating the parameter on the observed distances in the data can be done using the results of a nearest neighbor analysis and a k-distance graph. The value for (<i>esp</i>) can be found by plotting the distance to the (<math>K = MinPts - 1</math>) nearest neighbors ordered from the smallest to the largest value. Optimum values of (<i>esp</i>) are to be found where this plot shows an "elbow".</li> </ul>
( <i>MinPts</i> ) Minimum points within the search radius	<ul style="list-style-type: none"> <li>• Choosing a larger number of points leads to fewer yet larger clusters and more robust results.</li> <li>• Choosing a smaller number of points leads to more sub-clusters, and clusters might include noise.</li> <li>• The (<i>MinPts</i>) can be derived from the dimensions of the data as <math>MinPts \geq dimensions + 1</math>, as such, a minimum of 3 should be selected.</li> <li>• Larger values are preferable, as more distinct clusters will be identified and noise points will be more easily detected.</li> </ul>

## 2.4.5 Cluster Detection

The center-relevant point locations, including the parameter defined in the parameter tuning process, are provided to the DBSCAN algorithm to calculate the respective clusters and identify noise points. Once the clusters are defined, the points identified as belonging to one cluster are extracted and outlined using a concave hull generation algorithm. Outlining the clusters supports their visualization and characterization. In other studies, convex hulls were used to outline the location data of a cluster (Zhou, Xu and Kimmons, 2015). This, however, leads to overlapping polygons and larger enclosed areas without cluster points, which is not desirable. For this work, an additional small buffer is added to the convex hull polygons to receive meaningful outlines. In some cases, the resulting hulls still include empty areas, not used by the store locations. For later analysis and visualization, the clusters are described as individual points and polygonal aggregates.

## 2.5 Cluster classification

After detecting potential central retail agglomerations, the clusters are further classified by a rule-based classification logic. This logic is oriented on the typical center classification, on data derived from the clusters themselves and additional domain knowledge about the type of urban centers. This process aims at reaching five goals:

1. Classify the clustering results by applying domain knowledge and additional external data
2. Clean the cluster results to separate small group of stores from actual retail agglomerations
3. Design a transparent process that can easily be adjusted to calibrate the classification based on domain knowledge and is adaptable to other geographies
4. Derive additional and more detailed information from the clusters
5. Prepare results for interpretation and validation

To reach this, a two-step approach is proposed. Firstly, it is necessary to characterize the clusters by describing and enriching data. Secondly, the classification of the clusters following a decision tree logic based on domain knowledge.

### 2.5.1 Requirements

The classification is performed on the clusters described as polygons and the noise points defined as individual points. To do the classification, obtain relevant results and prepare the analysis for future updates or adaptation to other geographies, some prerequisites must be fulfilled. These are described in the following table:

*Table 12: Requirements for classifying the clusters*

<b>Requirement</b>	<b>Description</b>
<b>Cluster-characterizing data</b>	To do a data-driven classification, cluster-describing data has to be calculated and enriched based on the data set.
<b>Rules for classification</b>	The rules are defined based on domain knowledge, an in-depth analysis of the clustering results and calibration against validation data.
<b>Automatable</b>	Characterization and classification have to work in a supervised way and be implementable in a programmable GIS-workflow.
<b>Handle hierarchical and categorical data</b>	A main center of a municipality can be the top city center of Berlin and the town center of a small town.
<b>Transparent and adaptable algorithm</b>	Implementation in a transparent workflow, to incorporate domain knowledge and enable adaptability to circumstances or geographies.
<b>No learning data set available</b>	No representative training data set is available for the classification algorithms to learn how to classify the clusters.
<b>Types of clusters is predefined</b>	The number and a definition of the resulting clusters is predefined.
<b>Prepared for future runs and potential time series evaluations</b>	To observe the evolution and change of the retail functions and central places in Germany or other places.

## 2.5.2 Comparison and selection of classification methods

Previous research proposed a very diverse set of methodologies to classify retail spaces based on variate information (Mackaness and Chaudhry, 2011; Pavlis, Dolega and Singleton, 2018). These methodologies range from simple approaches, such as testing if a certain characteristic can be observed, to testing if a certain range of different variables is true to evaluate unknown locations against the characteristics of known locations. Other potential methods can be found in the context of machine learning and data mining techniques.

Classification methods use the attributes of features to group them in delimitable clusters (Breiman, 1998). Supervised classification approaches are considered in the evaluation of potential methods. These methods use pre-existing knowledge on the classes an object might fall in. The classification of central retail agglomerations is of high complexity, as numerous parameters would have to be considered for a comprehensive picture, and few of these parameters are known or available in this context. Furthermore, larger numbers of parameters make the results difficult to explain, visualize and impossible to calculate quickly on a retail computer. This might explain the predominance of simple and heuristic approaches applied and proposed by other researchers (Mackaness and Chaudhry, 2011).

*Table 13: Comparison of approaches for classifying clusters of retail locations (Mackaness and Chaudhry, 2011; Pavlis, Dolega and Singleton, 2018; Xia, Zou and Su, 2018)*

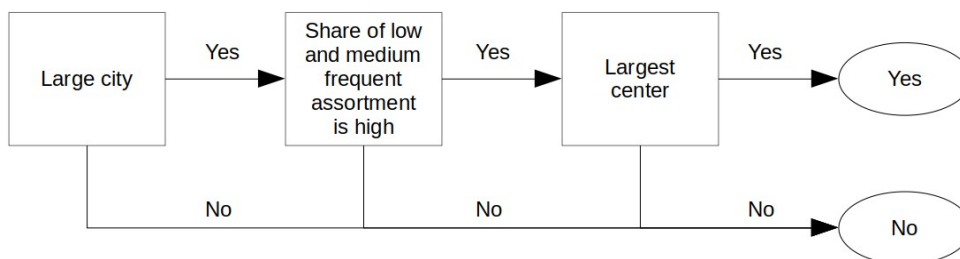
<b>Cluster algorithm</b>	<b>Comment</b>
<b>Decision tree or boolean logic (Breiman, 1998)</b>	<ul style="list-style-type: none"> <li>• Using sharp definitions and thresholds for distinguishing between two types</li> <li>• Significance to results potentially challenging</li> <li>• Easy to implement and to inform with domain knowledge</li> <li>• Able to handle numerical and categorical data</li> </ul>
<b>Fuzzy Logic (Ladner, Petry and Cobb, 2003)</b>	<ul style="list-style-type: none"> <li>• Using normalized data and recognizing the continuum retail areas of the same type show</li> <li>• A large share of omission and commission errors</li> </ul>
<b>Bayesian Inference (Berger, 1993)</b>	<ul style="list-style-type: none"> <li>• Using normalized and weighted criteria</li> <li>• Returns a probability concerning the extent to which a given feature may belong to a certain type of center</li> <li>• Does not require sharp thresholds</li> <li>• Requires training data, thus involving an intense collection effort</li> </ul>
<b>K-means (Macqueen, 1967) and partitioning around the medoids (Schubert and Rousseeuw, 2019)</b>	<ul style="list-style-type: none"> <li>• Results can be largely affected by extreme values in the data</li> <li>• Number of clusters to form has to be known beforehand</li> <li>• Not necessarily stable over multiple runs</li> <li>• Domain knowledge and constraints cannot be easily integrated</li> </ul>

The decision tree or boolean logic fulfills most of the requirements. The other, more sophisticated methods promise superior results; however, they also require additional data (data sets for training or more comprehensive characterization of the clusters), which is not necessarily available within a reasonable amount of time. Therefore, a rule-based filtering and sorting logic based on a decision tree is proposed that can be derived from analyzing the data and domain knowledge.

### 2.5.3 Decision tree

Decision trees are ordered, directional trees used to represent decision rules (Breiman, 1998). They are a simple representation of examples for classifications (Quinlan, 1983). Decision trees can be represented graphically as a tree diagram illustrating hierarchical and consecutive decisions that split the data in accordance to defined parameters (Liu, Xia and Yu, 2000). The technique can be counted among the supervised machine learning techniques. Decision trees are widely applied to automatically classify data based on rules derived from training data or from empirical domain knowledge (Lin et al., 2009; Machanavajjhala et al., 2009). Advantages of this technique include the easy construction and application, as small tasks does not even require specialized software. Calculation is usually fast and the results can be easily explained and interpreted as trees (Breiman, 1998). Accuracy is usually a par with other techniques for simple data sets. Other advantages are the modest requirements and the ability to handle numerical as well as categorical data. Decision trees consist of three elements (Breiman, 1998):

- Nodes: Decision question for a certain attribute or attribute combination
- Edges: The result of the decision and the connection to the next node or final leaf
- Leaf nodes: The final node that describes the outcome of the classification and represents the class label



*Figure 5: Exemplary decision tree for predicting if a retail cluster is the main-center in a large city*

Two main types of decision trees can be distinguished (Breiman, 1998). One is the classification type tree, built by means of binary recursive partitioning of the data. In this process, the data is split into partitions, sequentially at each branch. The resulting classification is categorical or discrete. The other common type is the regression type tree, which is most suitable for continuous types of data. In this case, the resulting variable can have continuous values.

Recursive partitioning is used to build decision trees, an algorithm design paradigm from the field of Computer Science (also known as divide and conquer) targeted at solving problems by repeat-

edly dividing a problem into smaller ones. The process ends when the problem is small enough to be solved or, in this context, when the data within the subsets are sufficiently homogeneous.

Numerous challenges can be described when using decision trees. These include the following (Lin et al., 2009; Liu, Xia and Yu, 2000): Decision trees can easily be overfilled by designing complex, poorly generalizing trees. Furthermore, large trees can be challenging to interpret. Results can become unstable as changes in the data might require a different tree structure. Bias might arise due to the tendency to split elements into multiple levels. Other disadvantages apply to optimal decision tree learning, applicable when more comprehensive data and a learning data set are available.

#### **2.5.4 Tuning of the decision tree brake points**

Some best practices when creating decision trees have to be followed to counter the disadvantages of the logic (Breiman, 1998). Before classification, it helps to balance the data (Breiman, 1998). A strategy for defining the tree is to start at its root and try to split each branch based on the attribute that has the highest information gain. This helps in reducing the uncertainty and reaching the final classification faster. For instance, a quick classification can be reached by interlacing several small trees in a larger one. At each break point of the decision tree to be developed, a profound decision has to be made on how to split the data. As limited information on the clusters can be derived and parts of the underlying data is incomplete, future changes in the data can have a measurable impact on the results. Therefore, in the design and calibration of the decision tree, the number of break points should be kept as low as possible. In addition, braking along known and comprehensive data, such as the classification of municipalities by type, should be preferred over cluster self-describing data. The defined decision tree will be transferred into simple data-splitting rules (Machanavajjhala et al., 2009) which can be implemented as a data classifying workflow. The resulting classified clusters can then be used to validate the results against the limited validation data. The results, in turn, can help to adjust the rules and improve the logic of the decision tree.

#### **2.5.5 Cluster characterization and classification**

Prior to being able to classify the identified clusters, fundamental characterizing data is calculated and enriched to the clusters. This includes, for example, the number of locations per cluster or the location density. The process of data enrichment entails making covert information in the spatial data set overt and available for analysis, visualizations and interpretation (Neun, Weibel and Burghardt, 2004; Mustière and van Smaalen, 2007). For the purpose of this work, only a few attributes are calculated, as the aim is to keep the results explainable and transparent. Based on this data, the clusters are classified following a set of rules derived from a detailed decision tree (Machanavajjhala et al., 2009). The filtering rules are implemented as a workflow in QGIS.



## 3 Results

The results are presented in the context of the three key challenges to be solved by this work. As such, this chapter focuses on the results of the applied methodology for clustering and classification of central retail agglomerations that are diverse in size, shape, density, contain a large amount of noise and cover a large area. Then, the chapter presents the final input data, the results of the parameterization and the incorporated domain knowledge. Finally, in this chapter the Workflows that automate the best part of the process are presented. The central findings of this work include: the sourcing, cleaning, preparation and well tuning of parameters for diverse spatial algorithms based on data mining and domain knowledge; a comprehensive data set of identified central retail agglomerations, including smaller groups of stores and individual locations and an enriched, commonly classified set of these clusters; an QGIS processing plugin which automates the entire process and can handle the large scale of the study area. The processing plugin in as well as the resulting data is available for download from this link: <https://nx4521.your-storageshare.de/s/nQ5PkANDYb2rDFE>.

### 3.1 Input variables and domain knowledge

To run the algorithms and processes, the underlying data and parameters have to be carefully selected and understood. The previous chapters outlined the selection criteria for the input of the models. This chapter describes the results of the data selection, the results of the parameter tuning process and where domain knowledge or assumptions informed the parameterization.

#### 3.1.1 Point of interest selection, classification and analysis

In the following, emphasis will be placed on the content-related decisions made when selecting and classifying the points of interest extracted from the OpenStreetMap database. Generally, the selection of the relevant data and the appending of additional data is based on the attribute, which describes the various types of individual features (e.g. supermarket or bakery).

To select the relevant elements, the feature types provided by OpenStreetMap were evaluated by their typical assortment and center relevance using respective lists (Bunzel and Difu, 2009; Acocella, 2018, 2019). The detailed composition of the assortment can differ considerably between locations of the same type. In the final selection, all locations named as being of relevance for a center, all relevant for local supply and all explicitly named as not relevant for a center are taken into consideration for analysis and later evaluation. The locations not relevant for a center are filtered before the clustering. Additional classifying data is added to the locations based on a broad range of relevant domain knowledge related to retail and central retail agglomerations (OpenStreetMap contributors, 2020; Geofabrik GmbH, 2020; Nelson, 1958; Kulke, 2020b, 2017; Bunzel and Difu, 2009; Acocella, 2018, 2019; Orzessek-Kruppa, 2016; Heineberg, 2017). The following table describes the appended data. A complete list of all features recognized as center-relevant and classified by type can be found in the appendix chapter of this document, also including the number of features recognized in the analysis.

Table 14: Added attributes to the OpenStreetMap data (OpenStreetMap contributors, 2020; Kulke, 2020a; Nelson, 1958; Bunzel and Difu, 2009; Heineberg, 2017)

Attribute	Description	Characteristics
<b>Category</b>	The classification generally follows the tag classification as developed by the OpenStreetMap community. In addition, detailing categories are introduced for service, sport and catering.	Shop; Public Service; Public Infrastructure; Service; Tourism/ Recreation; Education; Healthcare; Sport; Catering
<b>Description</b>	General textual description of what this location provides or supplies.	Shop focused on selling vegetables and fruits.
<b>Center relevance</b>	Classification of assortments by their relevance for a central retail agglomeration was matched to the feature class as provided by the data source.	yes yes_ls (ls= local supply) no
<b>Magnet</b>	Categorizes the shops by their own attraction and establishes if they draw larger footfalls, thus serving as magnet operations for other businesses.	yes yes_ls (=local supply) no
<b>Frequency</b>	The attribute describes the frequency of demand which the shop attracts with its assortment. This roughly translates to the size of the catchment area of this type of shop.	low medium high

Locations currently marked as being vacant are excluded. Using vacant locations in the clustering process can be challenging, as vacancy might be the result of a failed or currently declining agglomeration (Pavlis, Dolega and Singleton, 2018), signaling changes in the area covered by the center. In total, 9.319 shops are marked as being vacant in Germany (OpenStreetMap contributors, 2020).

The complete data set includes 1,05 million locations within the area of interest. Thereof, center-relevant shops are 236.944 (22.4 %). The most dominant types of shops are Bakeries, with 36.634 shops (15.5 %), Supermarkets, with 34.312 (14.5 %) locations, and clothes stores, with 30.506 locations (12.9 %). A total of 159.246 (67.2 %) shops are center relevant for local supply. 62.588 (26.4 %) are classified as magnet locations, while 94.4 % of these are classified as local supply magnets. In terms of frequency, 70.4 % of the locations can be counted as high, 21.7 % as medium and 7.9 % as low.

The results of a nearest-neighbor analysis indicate strong clustering tendencies, with an expected mean distance of 755 meters, a mean distance of 132 meters between the 236.944 center-relevant locations, and the nearest neighbor index of 0,17.

The exemplary visualization of the spatial distribution of the locations by frequency of demand illustrates the widely spread and large count of stores supplying goods of frequent demand, and, in contrast, the few and highly concentrated locations of shops supplying goods of lower frequency demand.

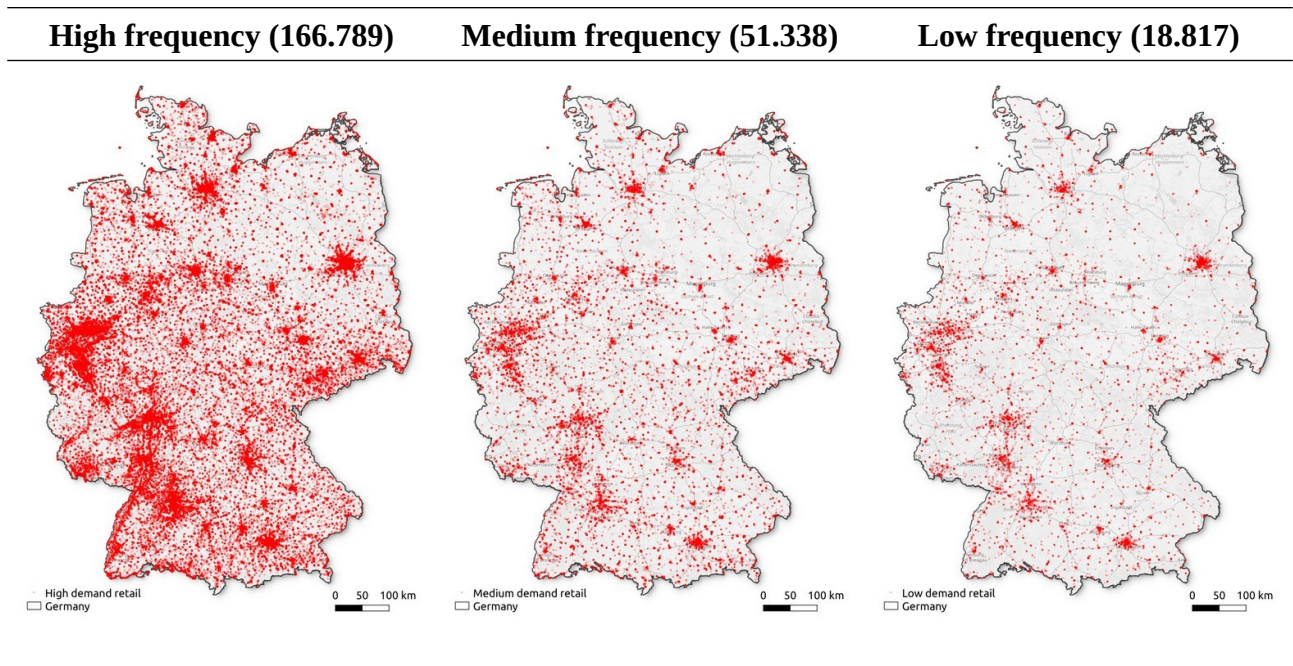


Figure 6: Location patterns of center relevant locations by their frequency of demand

The comparison of two of the most common features on the regional level highlights differences in clustering. While bakeries show a more dispersed distribution, clothing stores show a concatenated and clustered distribution.

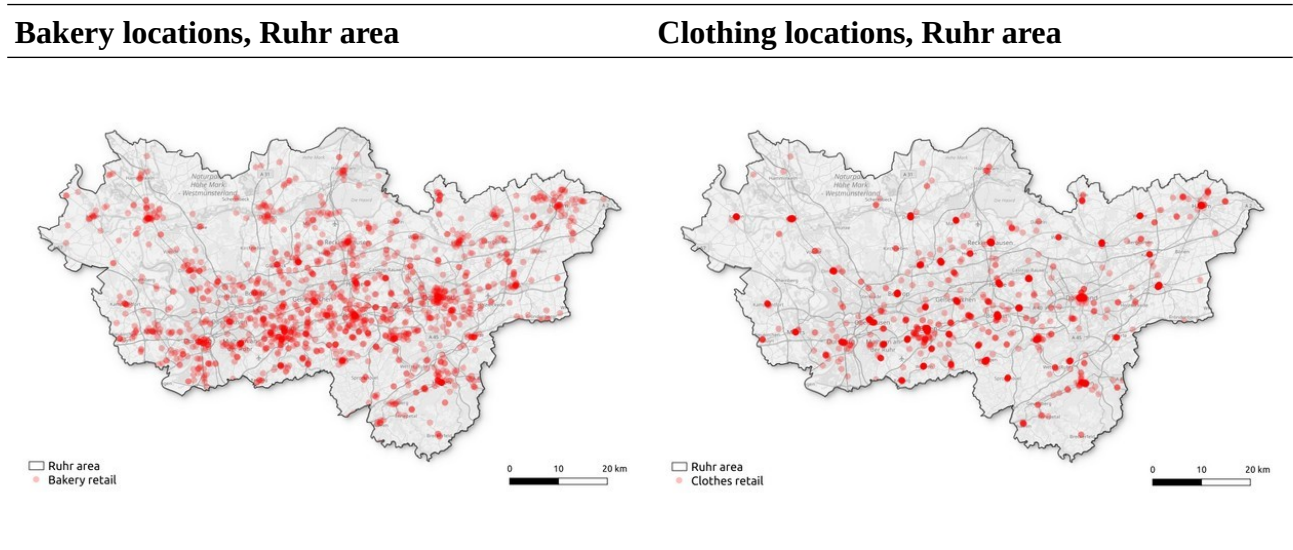


Figure 7: Dispersed versus concentrated distribution of bakery and clothing retail

At the sub-city level and the level of retail clusters, the individual center-relevant locations with their specific features have to be recognized. For the main center of Düsseldorf, the center relevant locations group close to each other and along the main shopping streets. Center-relevant locations of local supply can be found on the outskirts of the main center and show a looser distribution. The distribution of the main center magnets is relatively evenly distributed across the main center. Local supply magnets also spread around the border of the center.

**Center relevant locations and center relevant supply locations, Main center of Düsseldorf**

**Main center magnets and local supply magnets, main center of Düsseldorf**

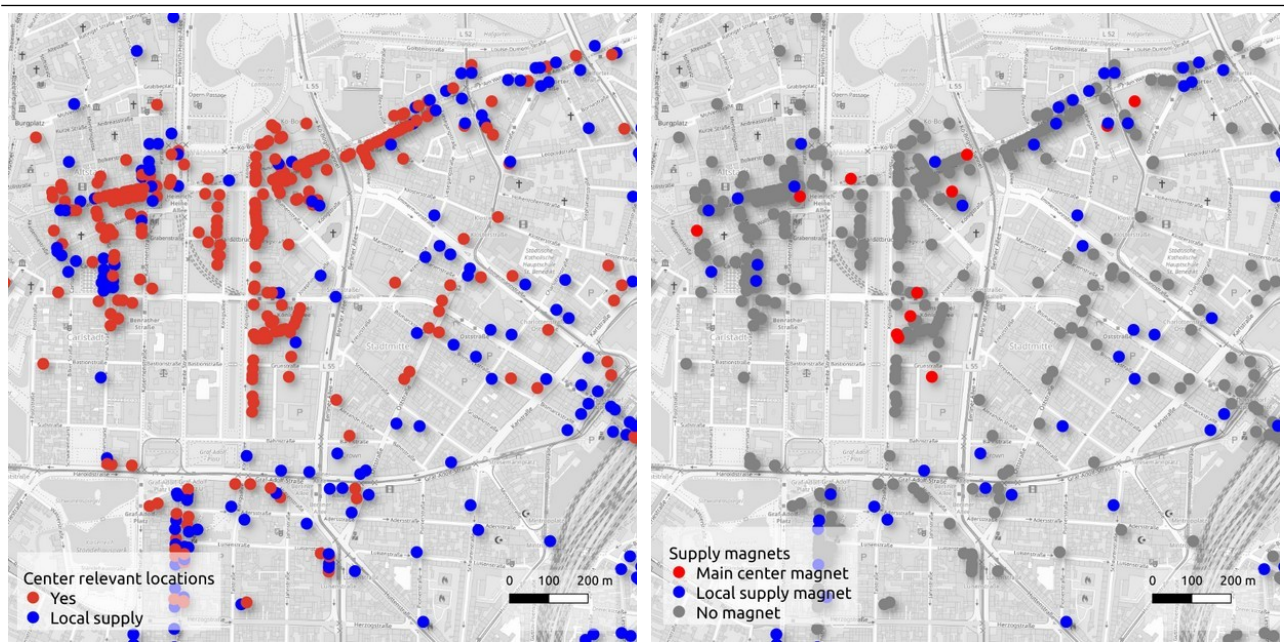


Figure 8: Center-relevant locations in Düsseldorf by attributes

The second data set included in this work is the classification of the municipalities by type of city and includes the total population. The typification of cities and municipalities includes the categories: large cities, medium-sized cities, small towns and rural municipalities. The typification represents the function and significance of a city in relation to its size and importance for the surrounding area.

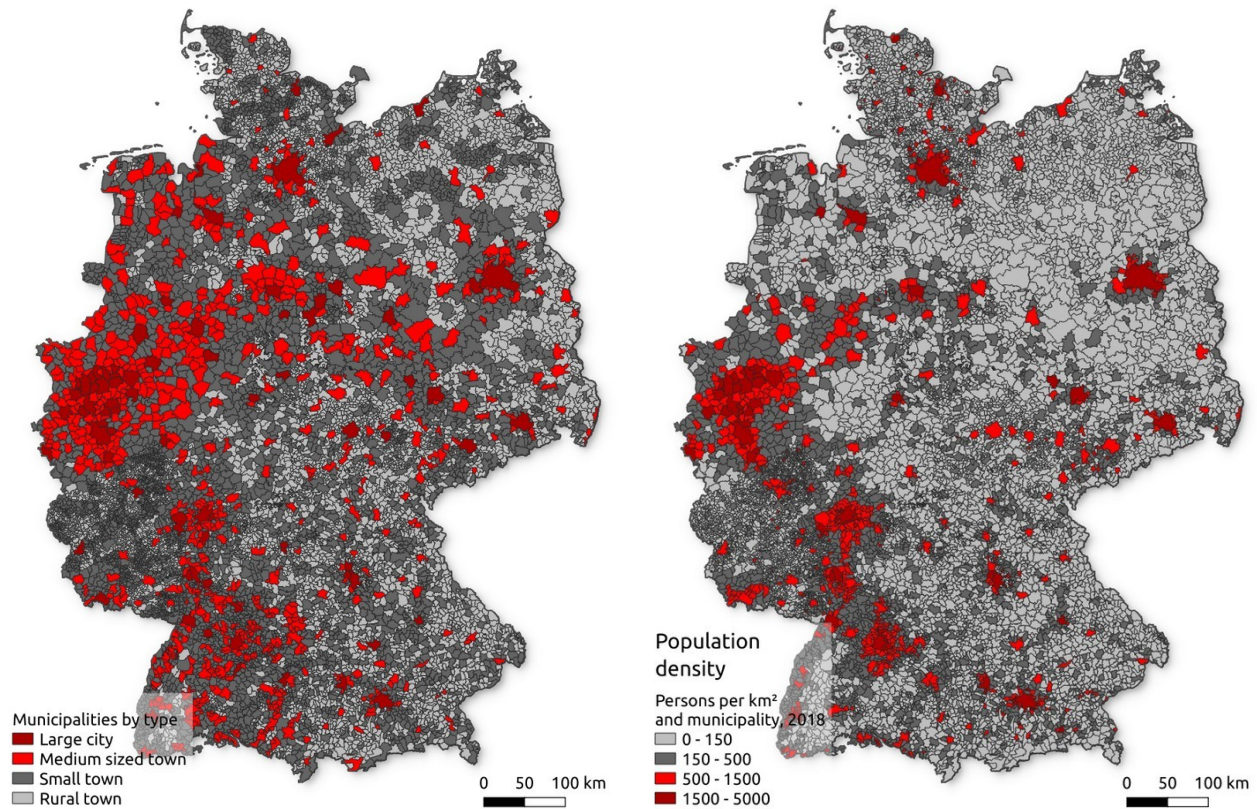


Figure 9: Municipalities by type of centrality and population density

### 3.1.2 Parameterization of DBSCAN clustering

For two-dimensional data, ( $minPts$ ) should be kept at the default value of  $MinPts=4$  (Ester et al., 1996). To counter missing location data and to over-detect clusters, the parameter was set to:

$MinPts=3$ . The ( $esp$ ) parameter is comparably more challenging to set; in the best case scenario, it is based on domain knowledge. Research for central retail agglomerations in the UK found optimum ( $esp$ ) values to be between 70 and 120 meters, depending on the individual situation in eight cities (Pavlis, Dolega and Singleton, 2018). No information is available on what minimum density is required in terms of the distance between stores to form a continuous retail agglomeration in Germany.

To define a global, ( $esp$ ) a data driven approximation is attempted. The parameter ( $esp$ ) depends on the distance functions observed within the data to be clustered (Ester et al., 1996). The observed mean distance between the 236.944 center-relevant locations is 132 meters. As initial tests showed a large share of noise points, the variable sought after is expected to be significantly below the observed mean distance. Even more detailed is the plot of a k-distance graph for all points, sorted from smallest to largest.  $K=3$  is used and ( $MinPts$ ) is set to be 3 (Sander et al., 1998). The graph is zoomed on 90 % of the closest mean distance locations. Some artifacts around the 0-Point are caused by different shops mapped to the same location. The graph shows a distinct ‘knee’ around 100 meters. Thus, it appears adequate to test values for ( $esp$ ) around this metric. As in

many other studies, the results were tested and compared with background knowledge of the study area. Based on said comparison, the absolute parameters were then selected (Chen, Arribas-Bel and Singleton, 2019; Hu et al., 2015).

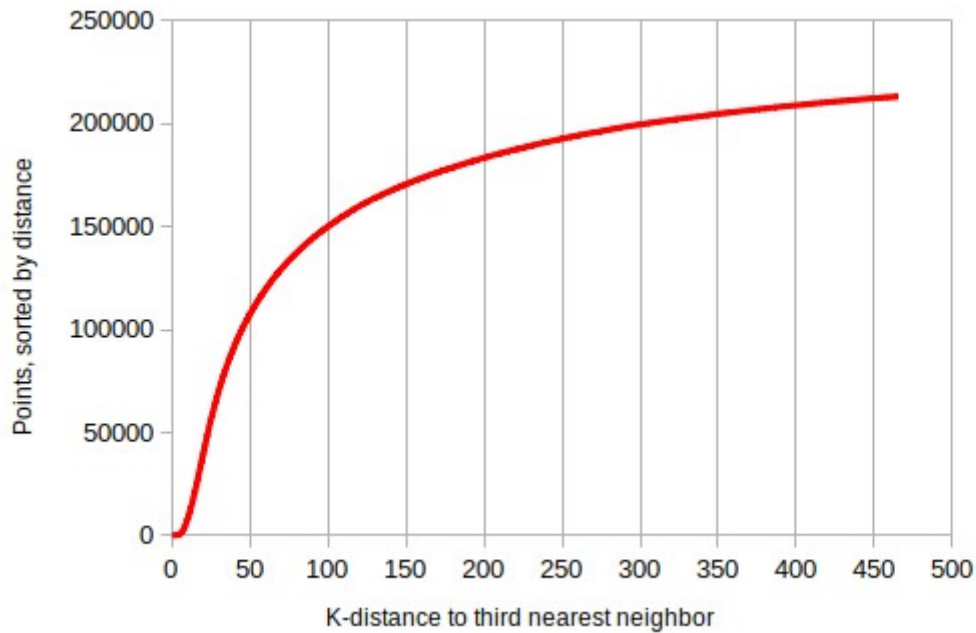


Figure 10: K-distance graph for the third nearest neighbor, sorted by nearest to farthest, for 90 % of the closest locations

The results of the test are summarized in the table and chart below and reveal that two contradicting forces influence the selection of ( $esp$ ). Firstly, the number of noise points would need to be minimized; secondly, the number of clusters detected would need to be optimized and set as large as possible. For the values tested, the sensitivity of ( $esp$ ) to the number of clusters is relatively low, between 95 and 120 meters, and varies between 16.081 and 16.318. This is in line with the general observation that the results are generally stable with varying choice of ( $MinPts$ ) (Schubert et al., 2017). This is different for the number of noise points detected, as they quickly decrease with increasing  $esp$ . The most stable result can be observed for values between 100 and 110 meters. In line with the best practice and the evaluations described above, the smaller value  $esp=100$  is selected (Ester et al., 1996).

Table 15: Number of clusters obtained by applying different parameters for (esp)

MinPts	Est (meter)	No of clusters	No of noise points
3	70	15.698	98.204
3	80	15.901	90.952
3	90	16.035	84.728
3	95	16.081	81.856
3	100	16.193	79.107
3	105	16.226	76.637
3	110	16.232	74.369
3	120	16.318	69.993
3	130	16.233	66.416

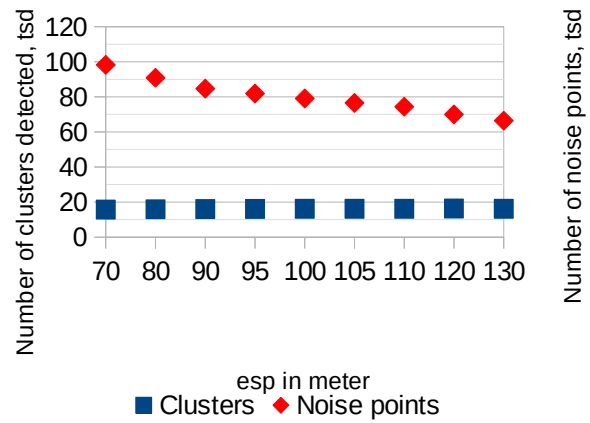


Figure 11: Number of clusters detected for MinPts=3 and increasing values of (esp)

### 3.1.3 Characterization, decision tree and rules for the classification

In this section, the results are threefold. Firstly, there is the detailed characterization of the clustering results; secondly, the decision tree used for classifying the clusters comprises extensive research of the underlying data and domain knowledge on central retail agglomerations; thirdly, the rules derived inform the workflow for classifying the clusters by their respective inner city center type.

#### Characterization of the clusters

The characterization of the clusters is solely based on the input data, the center-relevant retail locations, the extent of the derived cluster, and municipalities by their type of centrality. The respective data is enriched using spatial operations like spatial join, count in polygon or area calculations (Mustière and van Smaalen, 2007; Neun, Weibel and Burghardt, 2004).

Table 16: Enriched cluster characterizing data

Attribute	Unit	Description	Example output
City ID	ID	Official ID of the city.	11000000
City	Name	Name of the municipality that the centroid of the cluster is located in.	Berlin
Type of city	Type	Defines the type of municipality that the centroid of the center is located in.	Large city
Area	m <sup>2</sup>	Defines the area of the concave hull outlining each group of points identified as a cluster plus a 25-meter buffer.	1.477.076
Points total	Number	Counts the total number of points within each cluster polygon.	540
Local Supply	Number	Count of local supply stores defined as a magnet	47

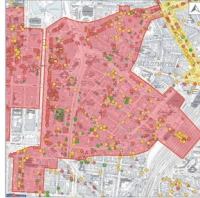


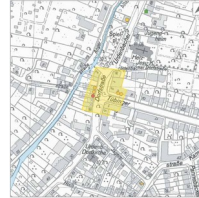
<b>Magnet</b>		within each polygon.	
<b>High frequency of demand</b>	Number	Count of stores supplying goods of high frequency demand.	222
<b>Medium frequency of demand</b>	Number	Count of stores supplying goods of medium frequency demand.	242
<b>Low frequency of demand</b>	Number	Count of stores supplying goods of low frequency demand.	76

## Definition of the centers

Most municipalities follow roughly similar definitions when classifying their central retail agglomerations. Therefore, general rules for describing center-bounding requirements can be defined. These requirements set the threshold for identifying a group of stores as a cluster of a certain type. The following overview is a collection of the center-defining criteria typically used by city planners and retail consultants in retail center reports written for municipalities in Germany (Orzessek-Kruppa, 2016; Acocella, 2019; Urban and Weidner, 2010; Acocella, 2018). Other center-relevant factors include the layout of the road network, the density of the structure built up, the availability of public transport infrastructure or estimations of the population demographics in the relevant catchment area. As limited information is known about the clusters and not all the information in the table below can be integrated into the classification algorithm, these other center-describing elements are left out. The table below summarizes the center definition for a large city.



Table 17: Center defining criteria, data and information used in the classification is highlighted in gray (Kulke, 2020a, 2017; Bunzel and Difu, 2009; Urban and Weidner, 2010; Acocella, 2018, 2019; Orzessek-Kruppa, 2016)

	Central retail agglomeration, character of a center			Other locations	
Center	Main center	Sub-center	Minor center	Group of stores	Individual sites
Center (German)	City Center	Nebenzentrum	Nachbarschafts-zentrum	Ladengruppen	Steulage
<b>Exemplary layout for locations in Düsseldorf</b>					Not specified
<b>Occurrence</b>	In large cities	In large cities This is the main center of medium-sized towns	In large and medium-sized cities. This is the main center of small towns and rural municipalities	In all cities	In all cities
<b>Catchment area</b>	Supra-regional importance, city and larger area	Part of the city or larger urban district	Urban district and surrounding settlement areas	Surrounding building blocks	Local area
<b>Footfall</b>	High	Medium	Medium to low	Low	No
<b>Focus of the supply offer</b>	High share of medium and long-term demand	Extended local supply, broad offer of other center-relevant assortments	Comprehensive local supply, high share of food, few other relevant assortments	Large variance of center mix	Store dependent
<b>Comprehensiveness of local supply</b>	Extended	Extended	Comprehensive	No	No
<b>Typical magnet</b>	Department store, mall	Pharmacy, supermarket, chemist	Chemists and pharmacies, one supermarket	-	-
<b>Number of stores</b>	The highest number of locations within the city	More than 30 to 50	At least 5 locations	Few, but varying	Single location
<b>Number of large locations</b>	More than 20	1-3, with focus on local supply	up to one	-	-
<b>Type of magnet locations</b>	Department stores and large cloth retail	Food retail, e.g. supermarkets and discounter	Smaller supermarkets and discounter	-	-
<b>Mix of locations</b>	All kinds of demand, department store, clothing retail, specialized retail	All kinds of demand. Supermarket, discounter and specialized retail	Supermarket, and other food retail like bakery, some specialized retail and pharmacy	Large variance	-

<b>Supplementary audience-oriented offers</b>	Diverse, simple to specialized services, gastronomy, tourist attractions, cultural institutions, public administration	Wide range of services (mainly retail-related, medical, financial and gastronomy related)	Primarily retail-related services, partly medical and financial, simple gastronomy	Some simple gastronomy or retail-related services	-
<b>Spatial structure</b>	No major spatial or functional gaps	No major spatial or functional gaps	No major spatial or functional gaps	Delimited from other centers	Single location
<b>Transportation</b>	Most centrally connected with roads, parking places and public transport	Well-connected, good parking situation and connection to public transport	Some parking and public transport	-	-

## Decision tree

Derived from the center-defining characteristics above, a decision tree is built. The structure of the tree is strategically selected to compensate for incomplete data. The advantage of such a condensed tree is, among others, the limited information required, the good comprehensibility and easy implementation.

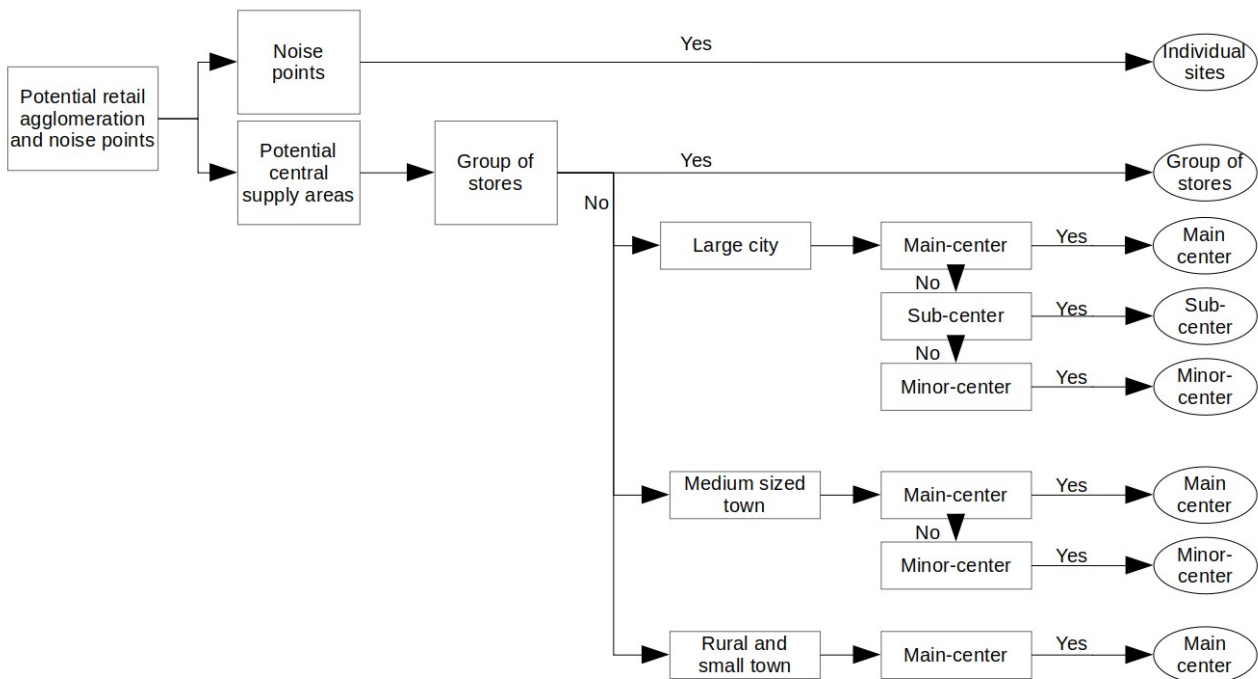


Figure 12: Simplified decision tree

## Rules for the classification

Rules for the classification are derived from the decision tree and the description of the centers, and formalized for implementation in a classifying workflow. However, the elaborate description of the centers gives little concrete information on what delimits a center. Moreover, the identified centers differ depending on the centrality of the municipality but also on other unobserved factors such as the availability of public transport, service locations or local cultural differences. As limited knowledge is available and especially as no clear breaking points are defined in literature (Heineberg, 2017), some assumptions had to be made to fill the gaps. The defined rules are therefore kept simple and establish the minimum criteria, which helps account for missing data but might also lead to the over-detection of certain types of centers. The assumptions integrated in this model have to be further challenged and refined based on additional observations or domain knowledge.

Table 18: Sequenced rules for cluster classification

Sequ ence	Rule	Minimum requirements	Classified as
1	<b>Individual sites</b>	All noise points as result from the DBSCAN algorithm	Individual sites
2	<b>Group of stores</b>	Clusters of 4 or fewer locations and no local supply magnet	Group of stores
3	<b>Main center</b>	The largest centers of each municipality with 5 or more locations and at least one local supply magnet, are classified as main center.	Main center
4	<b>Small towns and rural municipalities</b>	All remaining clusters	Group of stores
6	<b>Medium-sized cities</b>	Center of at least 5 locations, more than 1 local supply magnet and at least one location of low and medium frequency supply	Minor center
7		All remaining clusters within medium-sized towns	Group of stores
8		Center of 30 or more locations, with at least 1 local supply magnet and an extended range of locations of low and medium frequency supply, defined as $\geq 30\%$	Sub-center
9	<b>Large cities</b>	Center of at least 5 locations, more than 1 local supply magnet and at least one location of low and medium frequency supply	Minor center
10		All remaining clusters	Group of stores

## 3.2 Outcome of the clustering and classification

The results are presented at three different scales: firstly, for the scale of the whole study area, with summarized results at the country level; secondly, on the level of selected cities and regions, the distribution of the different functions of the clusters is shown; finally, single clusters with their individual store site locations are presented. Due to the number of sites and the large study area, the outcome of the clustering and classification is presented in small extracts, in a summarized form and with a focus on the main centers. In addition, the resulting data is available for a detailed view as GeoPackage from this link: <https://nx4521.your-storageshare.de/s/nQ5PkANDYb2rDFE>.

### 3.2.1 Overall country-wide results

A total of 5.062 central retail agglomerations and an additional 11.131 smaller groups of stores were identified and classified across Germany. The widest variety of centers can be observed within the 79 large cities, as these are the places where sub-centers can be observed. There 1.470 or 29 % of the centers were identified. Thereof, 79 are main centers, 192 are sub-centers and 1.199 are minor centers. Medium-sized cities host 1.611 or 31,8 % of the main and minor central retail agglomerations. By contrast, small towns host main centers (1.625, 32,1 %). The smallest number of 356 (7,0 %) centers can be found in rural municipalities.

Table 19: Number of central retail agglomerations and minor sites by type of center and city hierarchy

Center/ City	Number of cities	Total Population (Mio, 2018)	Central retail agglomerations			Other minor sites	
			Main center	Sub- center	Minor center	Group of stores	Individual sites
Large city	79	26,31	79	192	1.199	3.440	18.609
Medium-sized city	801	23,74	678	-	933	3.011	22.483
Small town	4.618	24,21	1.625	-	-	3.687	26.869
Rural municipality	5.587	8,53	356	-	-	993	11.138
<b>Total</b>	<b>11.085</b>	<b>82,79</b>	<b>2.738</b>	<b>192</b>	<b>2.132</b>	<b>11.131</b>	<b>79.099</b>

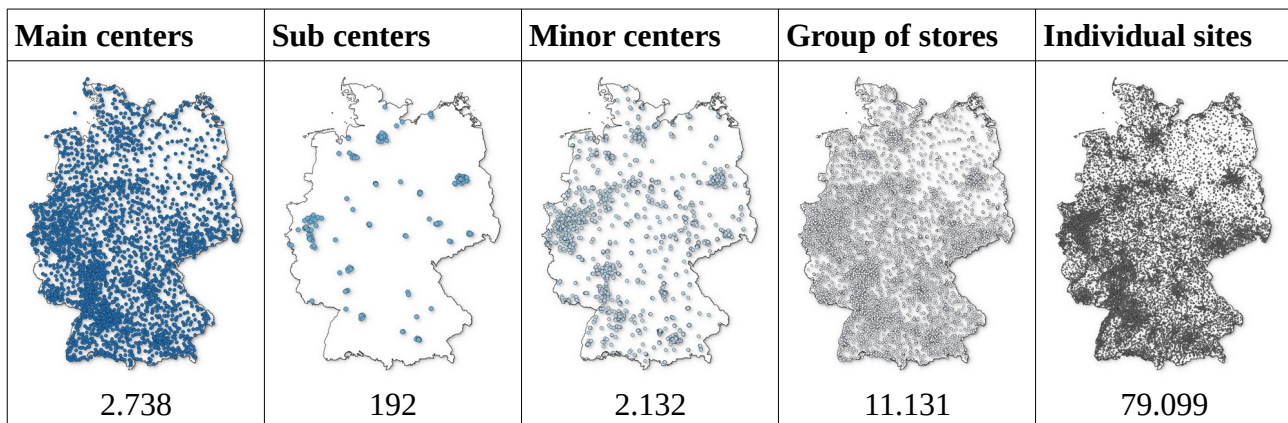


Figure 13: Distribution of centers and other retail locations across Germany

## Main center

Main centers can be found across all cities, irrespective of size. As this type describes the highest center of each municipality, the characteristics which depend on the type of city and each individual city significantly vary. The largest centers in terms of number of center-relevant locations can be found in Munich (922 locations), Cologne (807 locations), Hamburg (705 locations) and Berlin (540 locations). Zooming in to the five largest centers at a scale of 1:20.000 adds to the already observed diversity. What is worth noting is that Mannheim, with 521 locations in the category of main centers, is part of the top five centers; population-wise, it is the 22nd largest city in Germany.

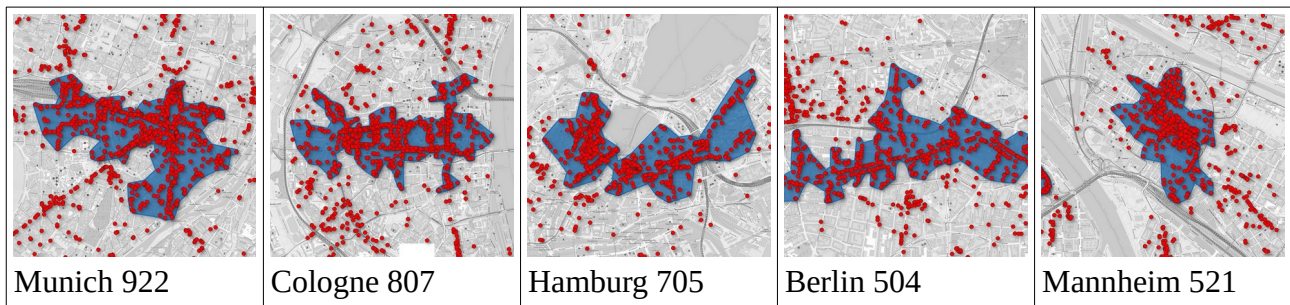


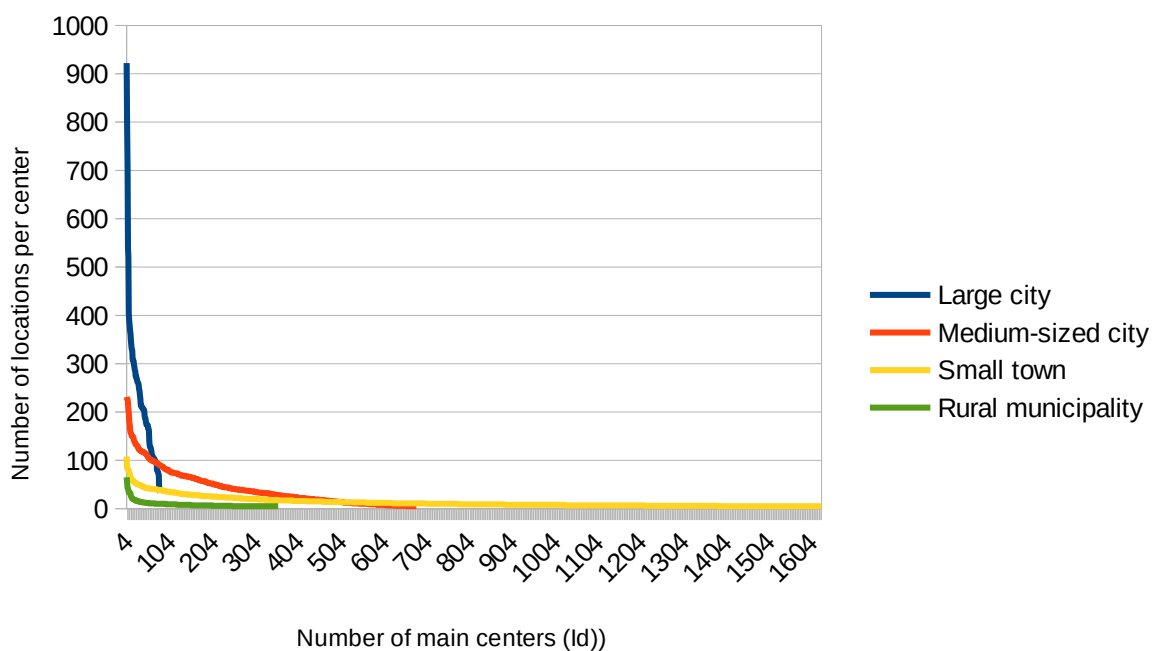
Figure 14: Diversity of the 5 largest main-center at the same scale of 1:20.000

Plotting the size of the main centers and the type of city shows a clear dependency. Similar results can be observed for the share of medium and low frequency demand and the type of city. However, the boundaries between the city types are not sharp, as are the size and composition of the main centers. The charts symbolize the wide variety of centers and the challenge to delimit them over a large scale.

---

### Main-center by sorted number of locations and type of city

---



In 77 municipalities, more than one main center is detected. This is because two centers show the same number of points and more than one local supply magnet. All the artifacts lie within small towns or rural municipalities and relatively small centers.

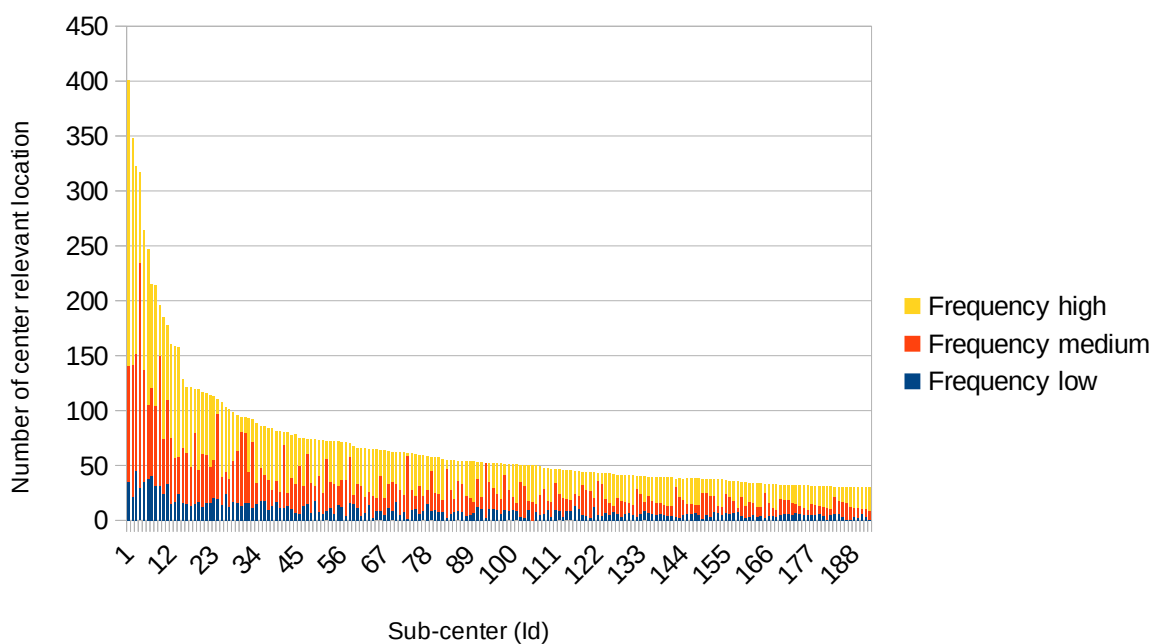
## Sub-center

Sub-centers are unique to the 79 large cities in Germany. However, the 192 sub-centers are distributed across 50 of the largest cities, of which 90 % count a population of 200.000 and inhabitants and more. The cities with the most sub-centers are Berlin (39), Hamburg (19), Cologne (11), Düsseldorf (11) and Bremen (10). The average size of the centers is 69 locations with a maximum of 401 in Berlin and a minimum of 30 in Berlin. The average share of assortments of low and medium demand is 50 %. Similar to the main centers, a broad diversity of sub-centers can be observed.

---

### Sub-center by sorted number of locations and locations by frequency of demand

---



---

## Minor center

Similar to the sub-centers, minor centers can be found in a subset of municipalities (not in small towns and rural municipalities). The total number of 2.132 is distributed across the large cities (1.199, 52,2 %) and the medium-sized cities (933, 43,8 %). With an average of 11 stores, minor centers are much smaller than sub-centers. Moreover, an average of 61,6 % of the locations supply products of frequent demand.

## Groups of stores

Groups of stores were identified in large numbers. The 11.131 locations are quite evenly distributed across the large, medium and small towns. Over 83 % have between three and five locations. Locations of local supply dominate in these clusters. 56 groups of stores count more than 20 locations and are probably misclassified.

## Individual sites

Even larger, with 79.099 locations, is the number of the individual sites identified. From the total number of center-relevant retail locations, 33,4 % were identified as individual sites, most of which can be found in small towns (34,0). The density of isolated locations is highest in the most central cities. The share of all locations differs across the study area, but no clear patterns of distribution emerge; in fact, vague tendencies can be described. Large cities and their surroundings tend to show lower shares of isolated locations. The same is true for the state North Rhine-Westphalia, showing a mostly continuous share below 40 %. The highest share of isolated locations is identified among beverage stores (57,6 %), kiosks, (50,2 %), convenience stores (50,1 %) and supermarkets (49,9 %). All locations have an assortment that addresses the high frequency demand and satisfies the local demand. On the other end, elements present least frequently in isolated locations are mobile phone shops (4,3 %), jeweler (6,5 %), optician (7,7 %), clothe stores (8,5 %) and chemists (10,0 %). These locations, supply the low and medium frequency demand, as well as a more central demand.

### 3.2.2 Regional and city results

At the scale of a larger region or city, the results are visualized under the aspects of the center hierarchy, what center focuses on providing the highest share of products of low and medium frequency demand, and the overall distribution within the regions. As one example of a larger, culturally connected region, the Ruhr-Area was selected; as one large city, Düsseldorf was selected. The regions were chosen based on the author's knowledge of the places.

#### Ruhr area

With a population of 5,1 million inhabitants, the Ruhr area is the largest continuous urban area in Germany. The region comprises 53, municipalities of which 24,5 % are large cities, 62,3 % medium-sized cities and 13,2 % small towns. 50 Main centers, 17 sub-center, 149 minor centers and 563 groups of stores are found in the area. The visualization shows the concentration of centers in the most densely populated municipalities located in the center of the region. Few municipalities in the north and northwest do not host a center that could be classified as a center. In these municipalities the supply of the population is ensured by stores in isolated locations. A comparison of the setup of the central retail agglomerations across municipalities reveals different center strategies. For instance, in Dortmund, a city with a population of 586.600 inhabitants, the central supply is highly focused on the city center, with a sub-center in the south and numerous minor centers across the urban area. By contrast, in Essen, a city with a population of 583.393 people, an even distribution of sub-centers around the main centers and some minor centers close to the main center can be observed. For better visualization, the central retail agglomerations and groups of stores are represented as points.

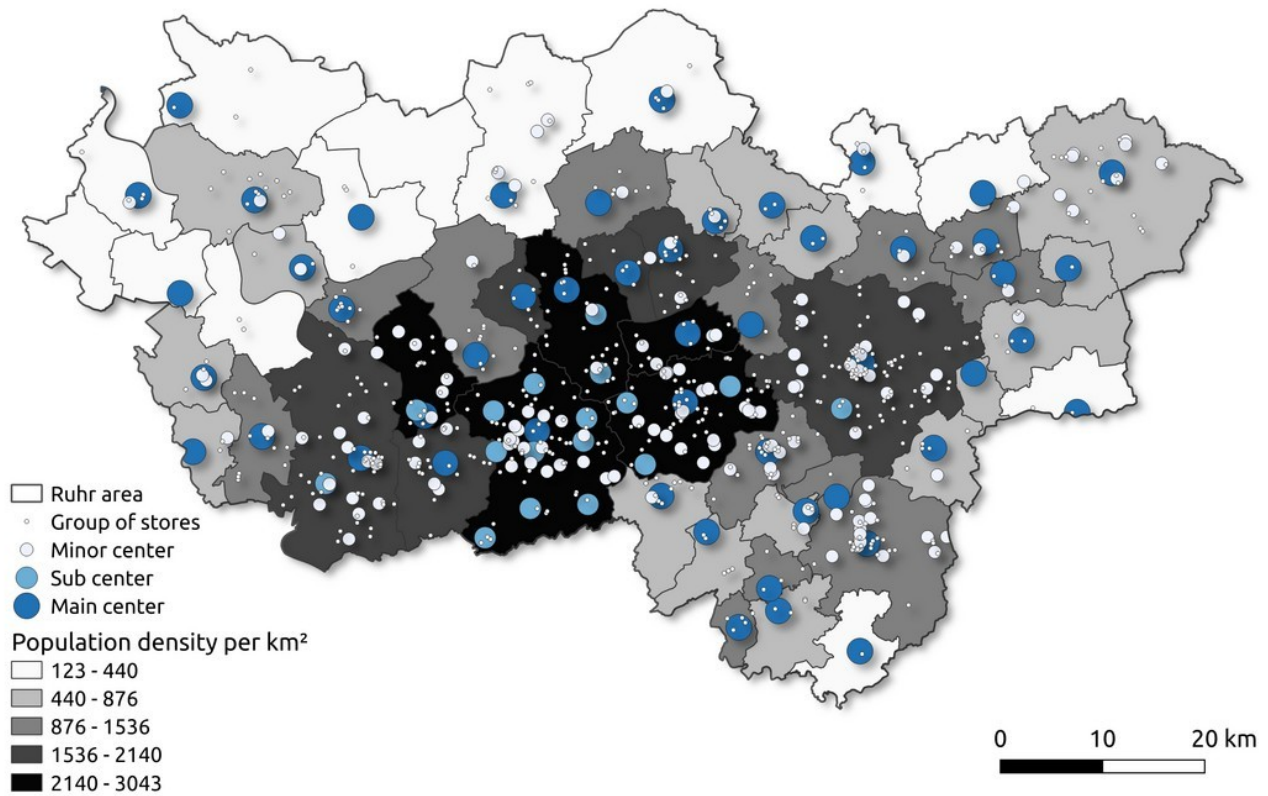


Figure 15: Distribution of central retail agglomerations and groups of stores in the Ruhr-Area

## Düsseldorf

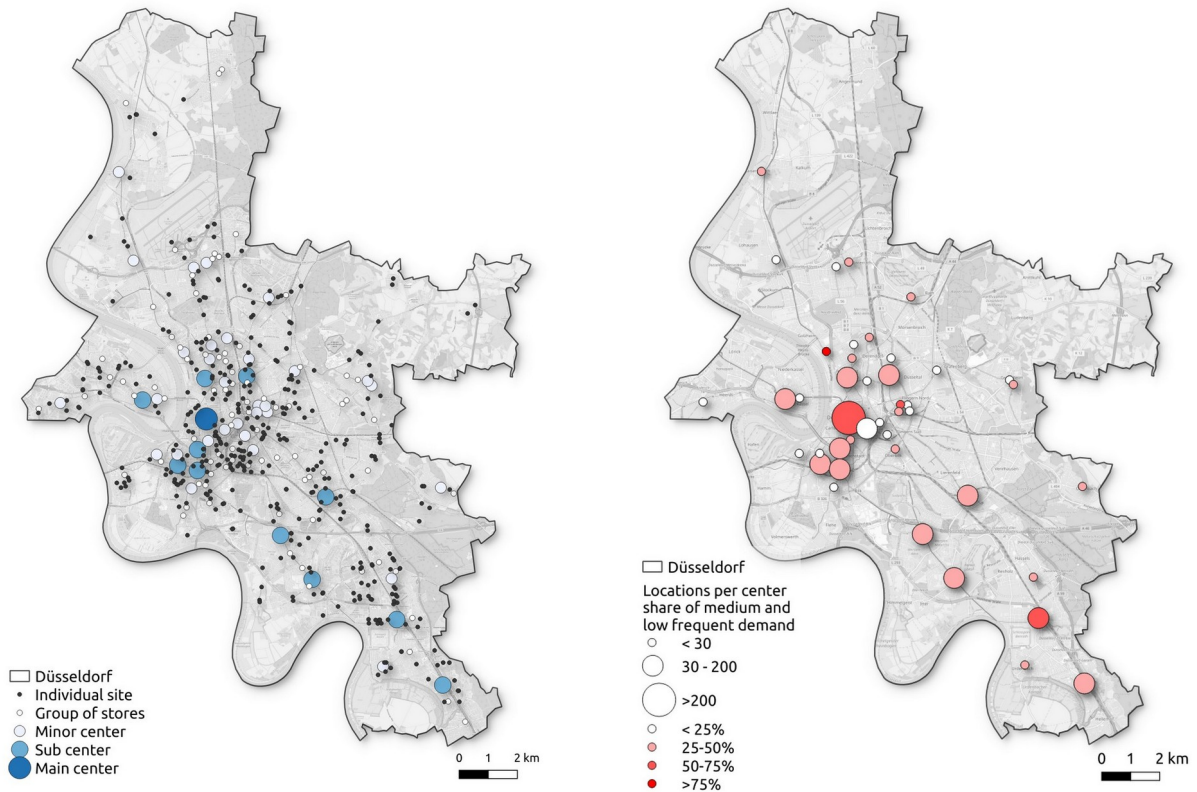
Düsseldorf is the capital city of North Rhine-Westphalia and has a population of 617.280 inhabitants. The city comprises 31 central retail agglomerations, 83 groups of stores and 348 individual sites. The summarized statistics for the centers show the function sharing between the centers. While stores that serve medium to long-term needs are concentrated in the main center, sub- and minor centers are more focused on the high frequency demand. The groups of stores and individual sites predominantly satisfy the high frequency demand. Relative to the whole study area, the share of isolated locations present in Düsseldorf is relatively low at 21.2 %. The most central retail agglomerations are concentrated in proximity to the main center.



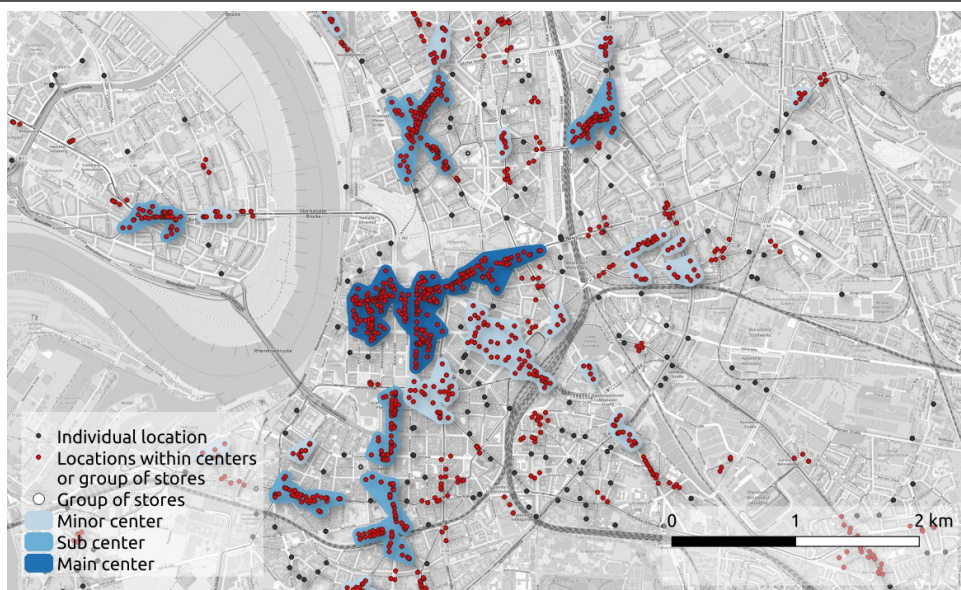
	Count	Total number of store locations	Average size	High frequency demand, %	Medium frequency demand, %	Low frequency demand, %
<b>Main-center</b>	1	269	269	26,8	58,7	14,5
<b>Sub-center</b>	11	562	51	60,0	25,6	14,4
<b>Minor center</b>	30	347	21,6	68,9	21,0	10,1
<b>Group of stores</b>	83	463	5,6	89,0	6,7	4,3
<b>Individual sites</b>	-	348	-	87,4	7,2	5,4

Center hierarchy

Center by size and share of medium and low frequency demand





**Extent of central retail agglomerations and groups of stores in central Düsseldorf**



### **3.2.3 Results for exemplary individual clusters**

At the smallest scale, individual clusters and the containing retail locations have to be recognized. Centers of all hierarchies from within the Ruhr area and Düsseldorf are shown exemplarily. The regions were chosen due to the author's knowledge of the places. For the centers shown, detailed statistics are calculated and presented alongside pictures of the sites. A map shows the extent of the center and the center-relevant locations by their frequency of demand.

## Main Center of Düsseldorf

Indicator	Unit	Result	On site images
Area	1.000 m <sup>2</sup>	434,3	
Locations	Number	269	
Density	Locations per 1.000m <sup>2</sup>	0,62	
Local supply magnet locations	Number	23	
Low frequency locations	Share	14,5	
Medium frequency locations	Share	58,7	
High frequency locations	Share	26,8	
Dominant features, top 5	Type, Share	Clothes (37,9 %), shoe shops (12,3 %), jeweler (7,4 %), bakery (4,5 %) and kiosk (4,1 %)	

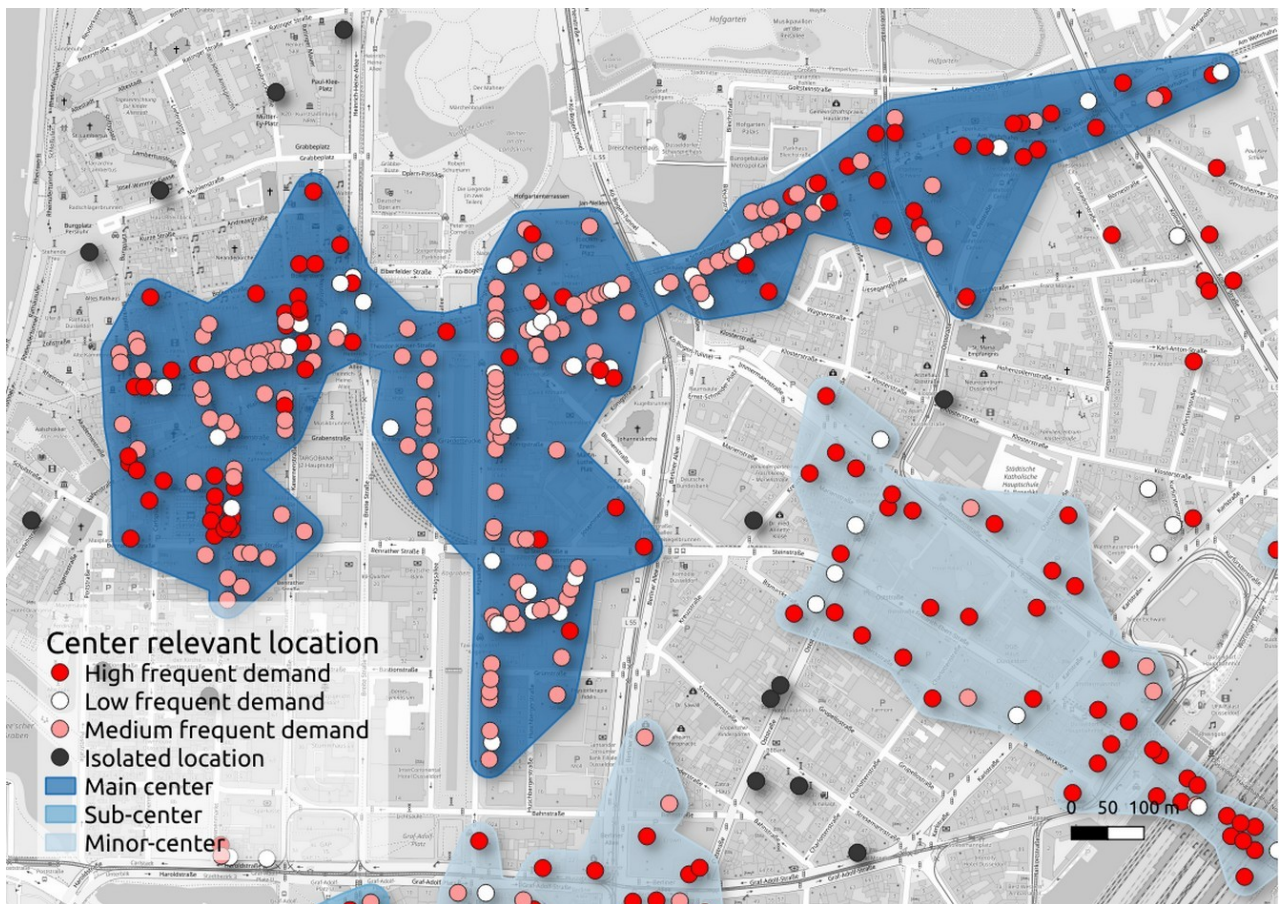




Figure 16: Main-center of Düsseldorf

## Sub-Center in Essen Rüttenscheid

Indicator	Unit	Result	On site images
Area	1.000 m <sup>2</sup>	291,4	
Locations	Number	93	
Density	Locations per 1.000m <sup>2</sup>	0,32	
Local supply magnet locations	Number	18	
Low frequency locations	Share	17,2	
Medium frequency locations	Share	30,1	
High frequency locations	Share	52,7	
Dominant features, top 5	Type, Share	Clothes (12,9 %), bakery (11,8 %), supermarket (8,6 %), pharmacy (7,5 %) and department store (7,5 %)	

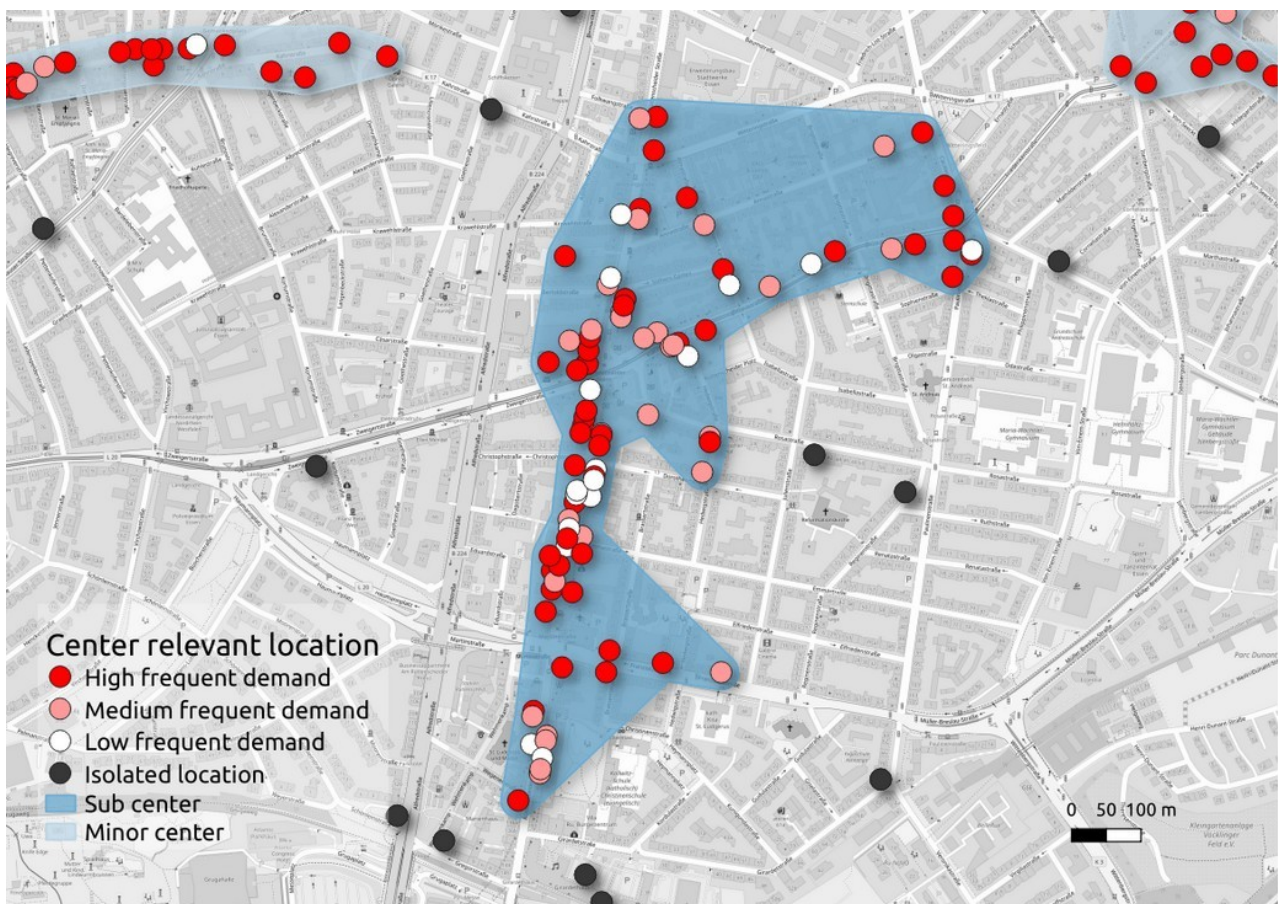


Figure 17: Sub-center Rüttenscheid in Essen

## Group of stores in Duisburg Buchholz



Indicator	Unit	Result	On site images
Area	1.000 m <sup>2</sup>	40,2	
Locations	Number	16	
Density	Locations per 1.000m <sup>2</sup>	0,4	
Local supply magnet locations	Number	5	
Low frequency locations	Share	18,75	
Medium frequency locations	Share	18,75	
High frequency locations	Share	62,5	
Dominant features, top 5	Type, Share	Optician (18,75 %) Supermarket (12,5 %), bakery (12,5 %), pharmacy (12,5 %) and shoe shop (12,5 %)	



Figure 18: Minor center Buchholz in Duisburg

## Group of stores in Mülheim an der Ruhr Baakendorf / Duisburger Straße





Indicator	Unit	Result	On site images
Area	1.000 m <sup>2</sup>	5,4	
Locations	Number	4	
Density	Locations per 1.000m <sup>2</sup>	0,74	
Local supply magnet locations	Number	3	
Low frequency locations	Share	0 %	
Medium frequency locations	Share	0 %	
High frequency locations	Share	100 %	
Dominant feature	Type, Share	Supermarkets (50 %)	



Figure 19: Group of stores in Mülheim an der Ruhr Baakendorf / Duisburger Straße

## Isolated location in Duisburg Großenbaum

Indicator	Unit	Result	On site images
Area	1.000 m <sup>2</sup>	-	
Locations	Number	2	
Density	Locations per 1.000m <sup>2</sup>	-	
Local supply magnet locations	Number	1	
Low frequency locations	Share	0	
Medium frequency locations	Share	0	
High frequency locations	Share	2	
Dominant features, top 5	Type, Share	1x Supermarket, 1x Bakery	

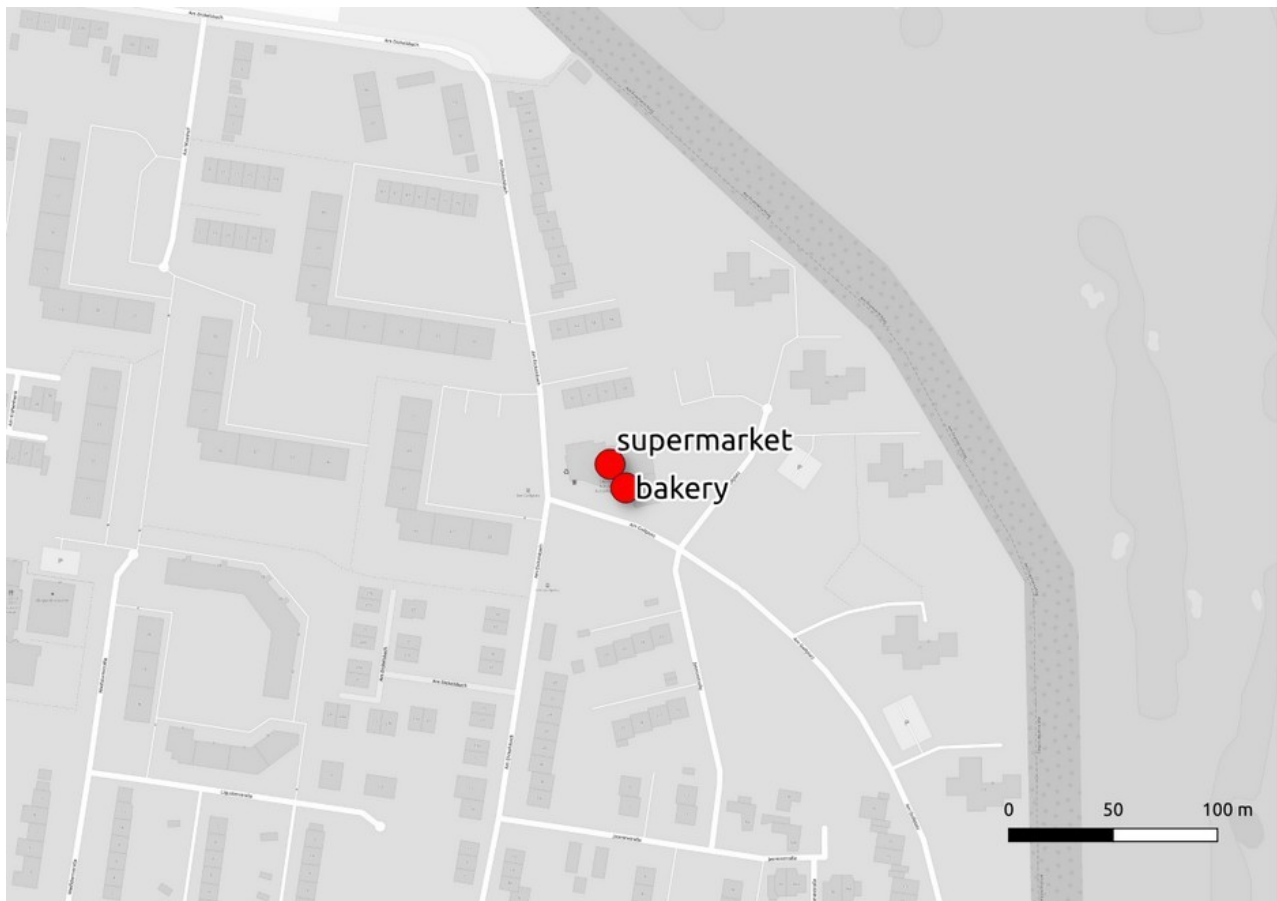


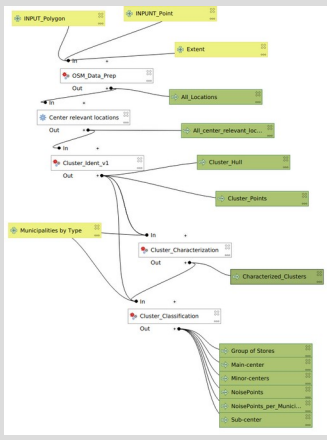
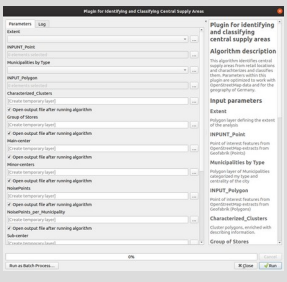
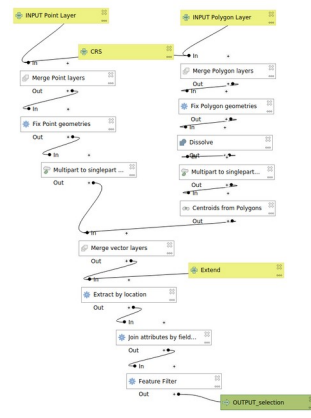
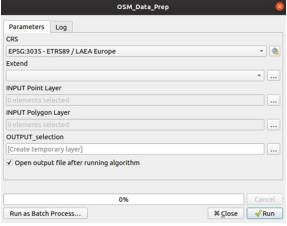
Figure 20: Isolated location in Duisburg Großenbaum

### 3.3 Process automation

Automating the process to a high degree was achieved by developing four processing models in QGIS 3.4. Each model solves an independent task. Apart from loading and initializing the processing models with the correct data and incorporating the required domain knowledge, the whole process is automated. The four processes are finally concatenated and combined into an independent processing plugin for QGIS. Due to the functional integration into the data processing framework of QGIS, the process can function as a stand-alone plugin or be integrated into any other spatial model. The documented plugin is available for download from this link:

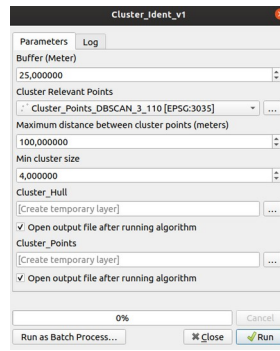
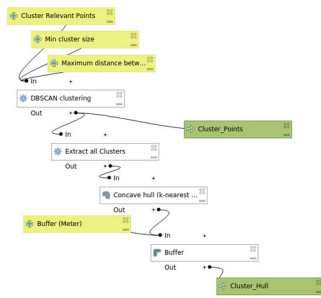
<https://nx4521.your-storageshare.de/s/nQ5PkANDYb2rDFE>

Table 20: QGIS processing plugin and processing models

Model and description	Processing Model	Model Initialization and documentation	Description and runtime
<b>QGIS processing plugin</b>			<p>All processing models presented below are concatenated and integrated into a QGIS processing plugin. The Plugin stands for its own and can be integrated into other processes. In order to run the plugin for a larger area, at least 16 Gigabyte of RAM are required. The plugin is documented and can be downloaded from this link: <a href="https://nx4521.your-storageshare.de/s/nQ5PkANDYb2rDFE">https://nx4521.your-storageshare.de/s/nQ5PkANDYb2rDFE</a></p> <p>1 hour and 20 minutes</p>
<b>Individual QGIS processing models</b>			<p>Outputs a cleaned and projected set of point locations that are necessary to form a center-relevant location-specifying data enriched in this step</p> <p>5 minutes</p>



## Cluster detection

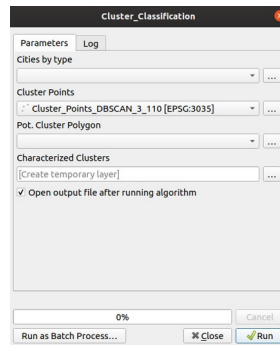
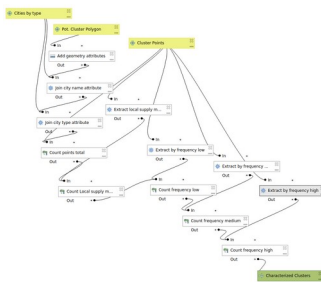


Outputs two data sets 1. Retail points and their association with a certain cluster or classification as noise. 2. Polygon enclosing each cluster.

The workflow is provisioned with the point locations from the first workflow, the parameters for the DBSCAN algorithm and the buffer distance

36 minutes

## Cluster characterization

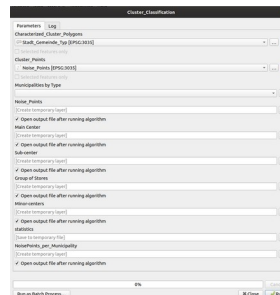
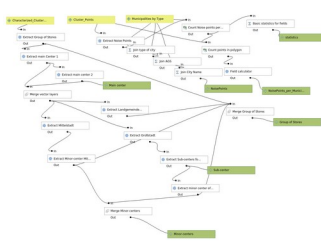


Outputs the cluster polygon data set enriched by location parameters, e.g. the type of the city and self-describing data such as the number of stores.

As input the workflow uses the polygons from the previous workflow, the initial point locations and the municipalities sorted by type.

19 minutes

## Cluster classification



Outputs two data sets. 1. Noise points 2. Cluster polygons classified by their inner city centrality. The workflow uses the enriched cluster polygons and the clustered point locations as input. The workflow is parameterized with classification rules based on a decision tree.

16 minutes

## 4 Discussion

The results have shown that clustering and the classification of central retail agglomerations is possible and can be integrated in an automated process for a large territory such as Germany. In the following discussion, a brief interpretation of the results is given and selected implications are highlighted. Then, the input data and results are validated at different levels of detail. The chapter closes with a critical reflection on the research method and some additional points questioning the results.

### 4.1 Interpretation and implications

The possible interpretations and implications of the results are manifold. Selected topics are outlined and described in context of the broader research on this topic. Other aspects for additional and deeper research are named for later reference or further research.

Overall, the retail center landscape shows a high fragmentation. Although there are a few large centers, even the largest center, with 922 locations in Berlin, comprises 0,4 % of the total retail locations in Germany. Moreover, 158 (3,1 %) of the central retail agglomerations sites count 100 or more center-relevant retail sites, and 90 % of sites have less than 44 locations. The fragmentation can be observed across all regions as well as centers are distribution across the study area.

The regional planning regulation aims to concentrate center-relevant retail offers in centers (NRW, 2020; Bayern, 2020). This concentration can be observed in the results, with 5.062 central retail agglomerations and 11.131 smaller groups of stores. Stores most often found in clusters are mobile phone shops (95,7 %), jewelers (93,5 %) and opticians (92,3 %). This proves the tendency of retail outlets with medium to low frequency of demand to cluster (Nelson, 1958; Kulke, 2017). Numerous stores and small groups of stores can be observed in isolated locations (88,7 % of all isolated locations are local supply stores). This shows the tendency of dispersion (Krider and Putler, 2013) and, at the same time, the goal of cities to supply the population with products of frequent demand close to their home. Most cities define an acceptable distance to the next supermarket for example to be less than 500 meters (Acocella, 2018; Bunzel and Difu, 2009). This ambition obviously results in large numbers of small groups of stores and single stores supplying goods of local demand in isolated locations.

Another perspective on the data would be the different center setup, cities developed over time. One observable distinction would be the single-center and the poly-center setup. Cities with a single center would have one large center with a relatively large share of retail outlets offering products of low and medium frequency. Poly-center cities would present multiple large centers with a high share of medium to low frequency products. Extensively discussed is the example of 'Neue Mitte Oberhausen' (Schulte, 2012; Heineberg, 2017; Heinritz, Klein and Popp, 2003). In the year 1996, just 3,5 km from the main center, Oberhausen opened the then-largest shopping center in Europe, with numerous entertainment and tourist attractions. The different layouts can be observed clearly by comparing the size of the locations and the share of low and medium frequency goods of the main and sub-centers. Observing the setup of centers, the spatial form and interconnection with supporting centers can help understand what makes a center perform well and give insights into lessons that should be derived from successful locations.

From the exemplary illustrations below, Cologne stands out due to its large center and few minor centers following the main roads. On the opposite end of the spectrum lies Berlin, with several large centers of various shapes and sizes, which reveals a diverse and multi-center setup.

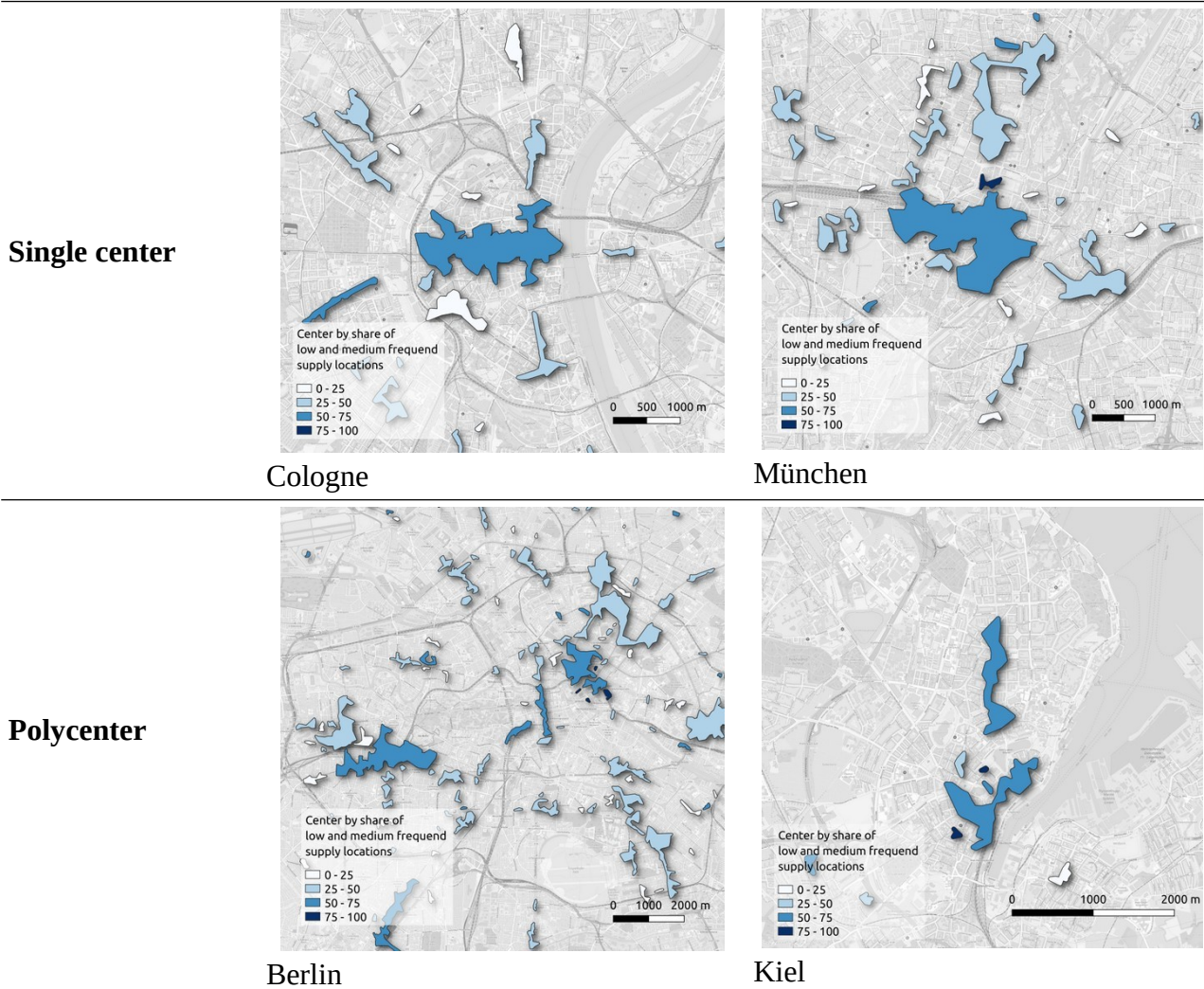


Figure 21: Selected cities with single- or poly center setup, by share of low and medium frequency goods

In the 1990s, a long-term study found that more than 10 % of center-relevant retail outside of centers is perceived as harmful to the performance of these centers (Vogels, Holl and Birk, 1998). This study and the 10 % threshold received critique (Kulke, 2017; Bunzel and Difu, 2009), as several contradicting examples and variations can be found. Assuming that the fundamental idea of a large share of center-relevant retail locations outside of centers is harmful, testing this hypothesis on a large data set can provide additional insights and possibly reveal cities that might face challenges due to a more decentralized and isolated retail footprint. In this context, the overall observed share of isolated locations, namely, 33,4 %, would have to be perceived as high. However, the two studies also present different assumptions which differentiate them. The share for the cities classified as large reaches 22,6 %, which is much lower. There are several large cities characterized by low shares of around 14 % (e.g., Bremen and Bremerhaven or Trier). At the higher end, there are cities such as Salzgitter and Duisburg with shares of noise above 40 %. Both cities are known for facing fundamental challenges such as high unemployment and economic and structural changes. Further research would have to investigate if the data drawn from OpenStreetMap leads to distorted conclu-

sions, if more rigorous city planning is required in some cities and if the centers in cities with less noise perform better in terms of, for example, attracting footfall.

When comparing the appearance of centers with the typification of the municipalities by their centrality, one exemplary contradiction stands out. For rural municipalities, no larger central retail agglomeration should be found (BBSR, 2020). However, this paper identified 31 municipalities with central retail agglomerations that host 15 or more retail locations. Some cases are special sites strongly influenced by seasonal tourism, such as the north sea islands Föhr (65 retail locations) and Helgoland (42 locations) or the winter sports and high-altitude health resort Reit im Winkel (33 locations). Other cases such as Lambrechtshagen or Hillesheim might be misclassified as rural municipalities.

## 4.2 Validation

236.944 center-relevant retail locations, 5.062 central retail agglomerations and 11.131 groups of stores were identified and described across Germany. Validating all locations individually is not viable, as validation data is not available nationwide; therefore, the underlying OpenStreetMap data is critically reviewed and several higher-level validations are performed alongside exemplary validations of individual locations in cases where the necessary data is available.

### 4.2.1 Quality assessment of the OpenStreetMap data

The following focuses on the need for as much transparency as possible, as well as the shortcomings and assumptions associated with the underlying data to be used in the analysis. The most crucial steps consist of understanding and assessing the quality of the point of interest data, as this helps establish whether the information is suitable for the analysis. Since OpenStreetMap data is neither comprehensive nor consistent (Mocnik, Mobasheri and Zipf, 2018), the results of this work have to be discussed and evaluated bearing the consequent constraints in mind. However, no formal and structured evaluation of the comprehensiveness will be performed. There is little research on how to evaluate the quality and completeness of point of interest data in OpenStreetMap (Touya et al., 2017) and on the quality of this data set in comparison to alternative sources (Zhang and Pfoser, 2019). Apart from some anecdotal observations, no extensive quality assessments on the retail location data has been performed for Germany.

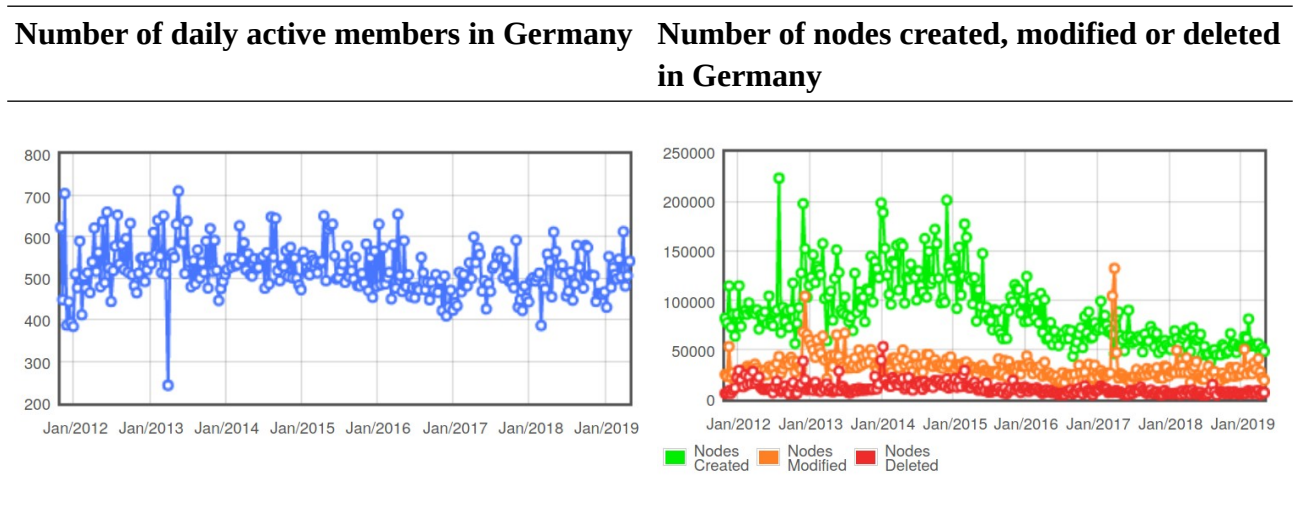
The OpenStreetMap project is considered as one of the largest collaborative projects and a prominent example of volunteered geographic information (Anderson, Sarkar and Palen, 2019). New data is collected following a few simple rules and characteristics (OpenStreetMap contributors, 2020).

**On the ground and verifiable** – requires that the information be mapped as it can be observed when physically examining the location. **Relevant** – requires the use of appropriate tags that make the data re-usable for others. **Legal** – requires the use of compatible licensed data or data from own surveys as source for mapping.

As of January 2020, 6 million users had opened an OpenStreetMap account, whereof an average of 45.000 users contributed with additional information each month in 2019 (Neis, 2020). Since 2012, Germany has usually seen around 500 daily active contributors, which makes it one of the largest active communities. At the same time, the number of newly created notes has decreased, as most of the easily traceable information from satellite imagery is collected and new contributions are more

time-consuming to collect. The number of nodes modified and deleted has stayed constant over the same period of time.

Table 21: Monthly OpenStreetMap contributor statistics for Germany (Neis, 2020)



OpenStreetMap data is widely used in private, scientific and commercial situations (Anderson, Sarkar and Palen, 2019). This has led to most large tech companies maintaining their own mapping teams to correct and complete the database (Anderson, Sarkar and Palen, 2019). One such example would be the American multinational technology company Amazon. Their OpenStreetMap wiki page lists around 450 employees who add and correct roads and access roads with data from their delivery vehicles.

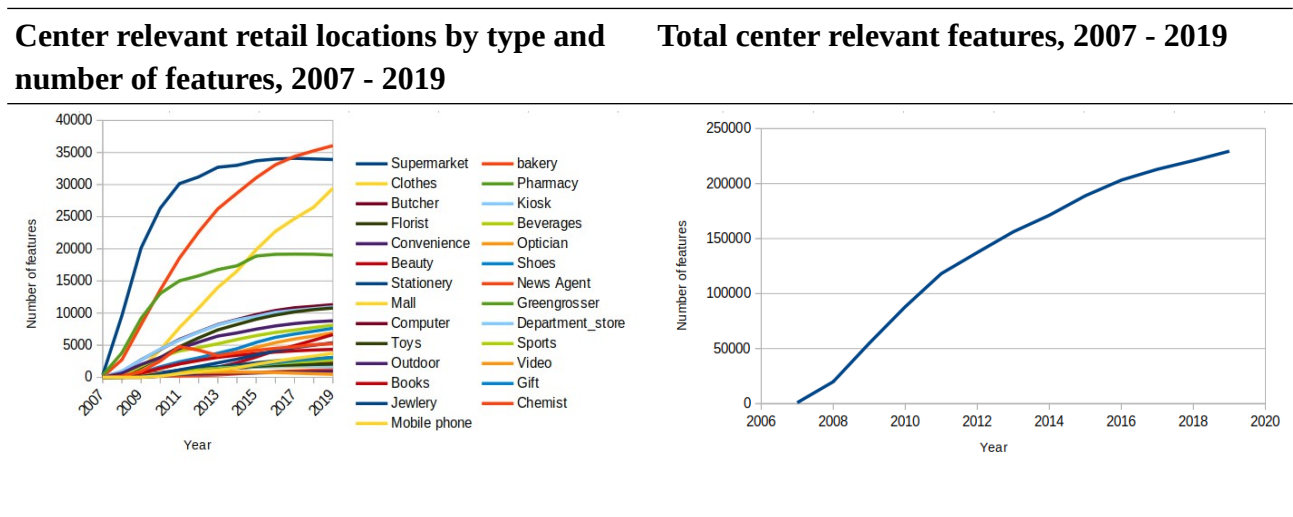
The mapping and tagging system of OpenStreetMap is open, but in some cases, it is not well-defined (Mocnik, Mobasheri and Zipf, 2018). Only in recent years there has been an increase in the formalization of the tagging system, manifesting itself in an increasingly detailed tagging system (OpenStreetMap contributors, 2020). These processes usually take a lot of time and effort, as there is no central structure and all changes are discussed, negotiated and agreed upon by the community. This approach allows great flexibility for large and fast data collection, adaption to local circumstances and, generally, new ways of approaching tasks. However, the unstructured character of the data gives rise to challenges when attempting to map the data to other sources, more commonly used GIS-formats, or when performing spatial analyses. That is one of the main reasons why prepared data was used for this work. One central disadvantage is that the available types of retail locations do not represent all available locations from OpenStreetMap. This is because the conveniently available data merely includes a given set of locations. Less relevant shops, such as pet food shops or cheese shops, are missing. In a later stage, a more comprehensive set of data could be included, which would require checking and consolidate around 10 thousand different tags used to describe shops globally.

The precision of the locations varies depending on the tools used to create the points and depending on the guidelines the mapper followed. For example, placing a node for a shop at the centroid of the building is a frequent approach. However, some mappers place the node close to the entrance. An inconsistency that cannot be corrected easily. In cases where buildings (polygons) were classified as relevant locations, the centroid was calculated and this location taken into consideration.

The comprehensive set of locations is known for a limited number of retail and service outlets. For the first quarter of 2019, the ABDA (Federal Union of German Associations of Pharmacists association) reported 19.268 pharmacies in Germany. By contrast, on January 1st 2020, OpenStreetMap had 19.370 locations in their database. Similar results can be found for florists, in which case the association estimates a number of 10 to 12 thousand locations (2018), while OpenStreetMap counts around 11.031 locations.

A more experimental way of assessing the comprehensiveness of the points of interest is by analyzing the number of locations in the database over time. The assumption is that the more comprehensive the data gets, the fewer new entries will be added. As the tests performed on all elements considered in this work prove, the thesis holds true for Germany. By this measure, the most comprehensive data sets with few new entries or a shrinking count of locations are supermarkets, pharmacies and department stores. The strongest growth, with more than 10 % for the time-frame 2018-2019, can be seen in beauty shops, mobile phone shops and clothes shops. Although the growth curve is flattening out for most shops, the overall year-over-year growth between 2018 and 2019 still highlighted a plus of 8.560, or 3,7 %, locations.

Table 22: Completeness of OpenStreetMap data in Germany, features 2018-2019 (HeiGIT, 2020)



Finally, the comprehensiveness is reliant on the existence of a local mapper and his or her interest in collecting information on shops. As such, the comprehensiveness at a given place is also a representation of the perspective on the world and bias of the mapping community. Anecdotal examples from the authors' observations are the missing luxury shops on one of the most central shopping streets in Düsseldorf, the Königsallee (Kö). Another example is the completely unmapped local supply center along the Eppinghofer Straße in Mülheim an der Ruhr. More specifically, the latter is a center mainly visited by migrants, which comprises at least 15 shops, multiple bars, restaurants and a variety of personal services.

#### 4.2.2 Validation of the cluster identification and classification

The issue of missing data affecting the identification of potential clusters is relevant due to a number of reasons. One challenge is the detection of small clusters, because a single missing location might be enough to cause the observer to overlook a potential center. Missing locations in larger centers could lead to discontinued clusters, clusters slit in sub-clusters, cluster outlines that appear

more condensed than they truly are in the real world or distorted data when calculating statistics for the center.

Two criteria are proposed to evaluate the results of the DBSCAN cluster detection and detect degenerated results (Schubert et al., 2017). Firstly, the size of the largest clusters should be taken into consideration. For this analysis, this cluster would be the main center of Berlin, with 922 locations representing 0,4 % of the total number of recognized locations. The small value stays well below the suggested 20 to 50 % range, at which point smaller values for *esp* or hierarchical approaches like OPTICS or HDBSCAN should be considered. The small size of the largest cluster is also an indicator for the high fragmentation of the retail function in Germany.

The second indicator is the share of noise points. A desirable amount of noise is described to be in the range of 1 % and 30 % (Schubert et al., 2017). For the data set used and the variables selected, it was observed that the share exceeds the suggested value of 33,4 %. The tuning of the parameters of the DBSCAN algorithm showed little variation on changing the parameter values. This suggests that the underlying location data and its distribution shows distinct and delimited clusters. The data provided by OpenStreetMap is incomplete, which should be considered as a reason for the high shares. It can be expected that a more comprehensive data set would return more small store clusters and the already detected clusters would be more complete and well defined. To test this, the OpenStreetMap community was asked to evaluate cities with a nearly comprehensive coverage of all shops. For the suggested cities, Hamburg (18,5 %) and Bochum (25,9 %), the shares are significantly lower. The same is true for cities with a community recognized as large, such as Trier (14,8 %), Heidelberg (17,5 %) or Düsseldorf (19,1 %). The few noise points in these cities might also be the result of strict city planning, which focuses on concentrating center-relevant retail in defined areas.

Another aspect is related to the types of locations considered for the clustering. Some types of stores turned out to be less relevant for a central retail agglomeration than their typical assortment might have initially suggested. The most relevant examples are beverage stores with 57,6 % of their locations outside of central retail agglomerations, kiosks with 50,2 % and convenience stores with 50,1 %. When subtracting these from the results, the overall noise level reaches a value way below the threshold of 30 %. The decrease is not accompanied by much information loss, as the sites presented comprise 12,9% of all locations.

## Validation against municipal retail development concepts

Recent validation data on the extent and composition of central retail agglomerations is available for some municipalities which have published reports on their most central retail agglomerations. These reports are usually created every five years and rely on comprehensive data and extensive surveys collecting the size of retail stores and the detailed assortment of each shop. The results are also calibrated to the size and local circumstances of the municipality. Since these reports usually cover one municipality and are published in pdf format, a detailed comparison for one city or a larger region is impossible without extensive digitization of the center outlines and statistics. The results obtained by retail center experts and the calculations of this work are therefore put side to side.

For comparison, this paper uses data from the retail center reports of Düsseldorf (a city with known good OpenStreetMap coverage) (Orzessek-Kruppa, 2016) and the city of Duisburg (a city with less comprehensive OpenStreetMap coverage) (Acocella, 2019). Both reports were developed by retail

consultancies, in close cooperation with the cities. However, as all cities define the concept of “center” differently, some discrepancies can be expected. The objective of both reports is to build the foundation for municipalities to develop and regulate the retail function within the city. As such, the ambition of the results is fairly high and might set the course for public decision-making or legal disputes (Junker and Kühn, 2006; Bunzel and Difu, 2009). Certain changes might have taken place and already been included in the OpenStreetMap data set since the creation of the reports. The clusters highlighted in this work are expected to be smaller and potentially discontinued, because exclusively retail locations were considered, whereas the comprehensive reports also include services, tourist attractions or public administration and entertainment sites. Furthermore, the external data available for comparison covers the most central retail agglomerations (main, sub- and minor centers), whilst group of stores and individual sites are not covered at all. Overall, validation against external data is limited because of different approaches, foundational data, scales of analysis and adjustments to regional specifics. Therefore, some exemplary central retail agglomerations are described, presented and discussed opposite to each other.

## Düsseldorf

Most cities update their retail center plan every five years; so does the city of Düsseldorf. The most recent plan was published in 2016 and is currently being updated. The results for 2020 are expected to be published in the third quarter of the year. The report does not include any statistical information on the centers that could be compared to the results of this work. All maps for the validation of the Düsseldorf centers are taken from the 2016 retail center report of the city (Orzessek-Kruppa, 2016). Based on the author’s personal observations, it could be said that the retail locations from Düsseldorf are thoroughly represented on OpenStreetMap.

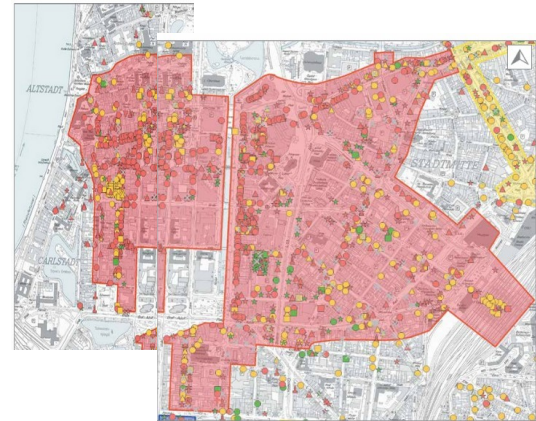
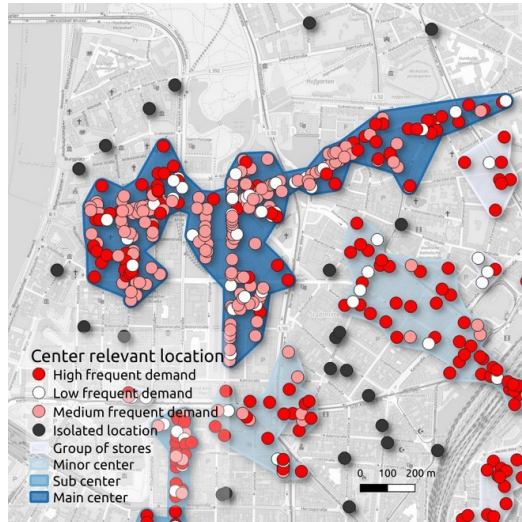


---

## Results

## Validation

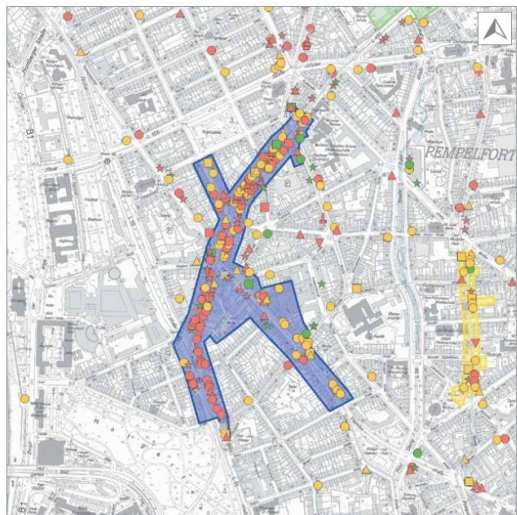
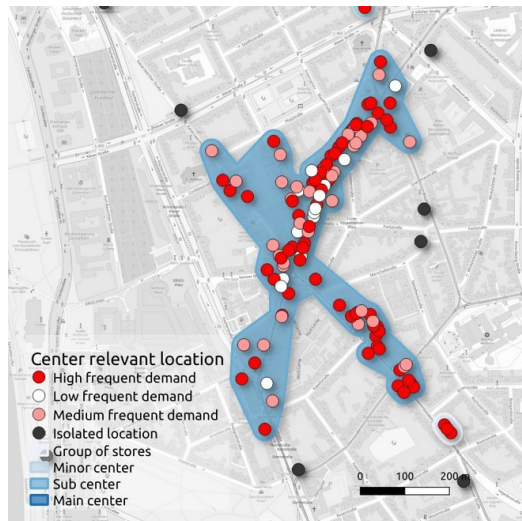
### Düsseldorf Main-center



When comparing the outlines, significant differences can be observed. While the results of this work are divided in smaller centers and compressed, the city results are more summarizing and reach across areas with a less dense population of retail outlets. One aspect that could not be recognized within the scope of this work is the political or strategical decision to divide the main center in an eastern and western part, although both parts feel connected and no major discontinuities exist. Similar to this aspect is the integration of the railway station area. An area that is disjunct from the main center and has a quite different profile from the main center.

---

### Düsseldorf Sub-center Nordstraße

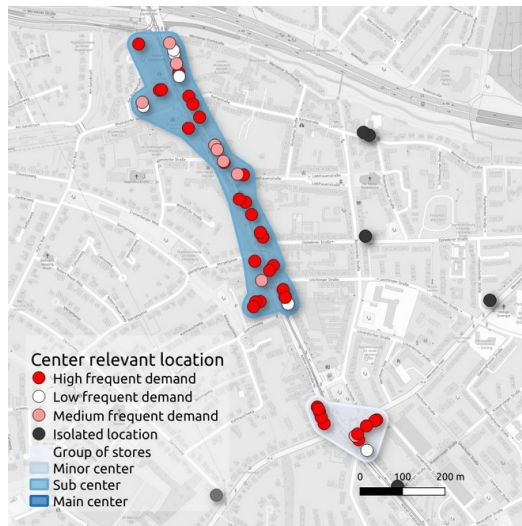


The central retail agglomeration 'Nordstraße' is by far the largest sub-center in Düsseldorf. Both outlines are similar in their extent. Some slight differences can be observed along the outskirts of the centers. In this case, the official results are better defined. For example, on the south-eastern arm of the center, a clearly delimited group of stores with three locations is officially integrated into the center.

---

**Düsseldorf**  
**Minor center**

**Kölner**  
**Landstraße**



In the results of this work, the central retail agglomerations along the Kölner Landstraße in Düsseldorf are classified as a sub-center, while the official classifications describe this center as a large district center. The outline for the most central part of the centers is similar. In the south, however, the official outline integrates a smaller group of stores into the center. A distance of about 250 meters separates the two centers.

## Duisburg

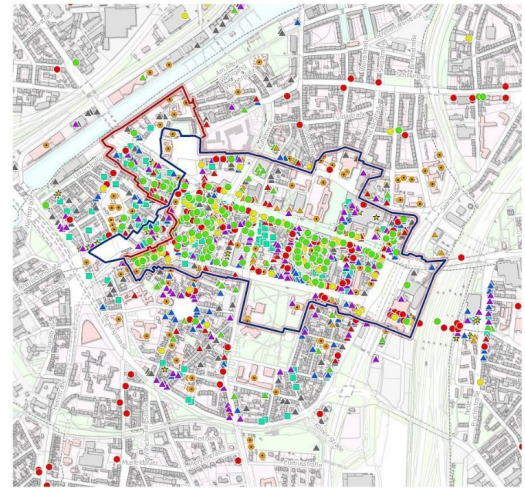
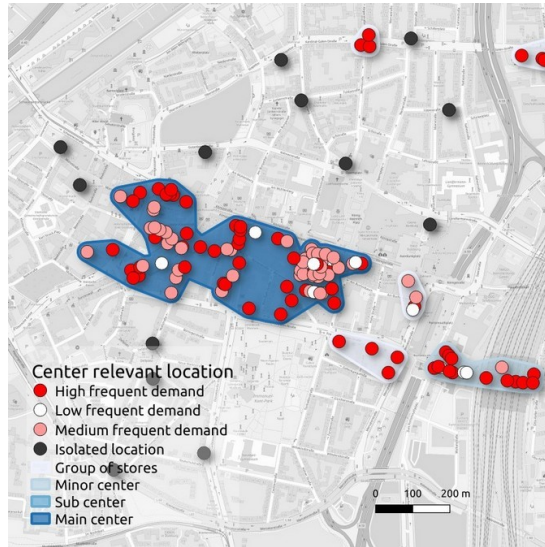
By contrast with Düsseldorf, the retail and center concept of the city of Duisburg was updated in 2019 (Acocella, 2019). The report is an update to the first detailed center definition from 2008. The results for Duisburg therefore show two extents for the central retail agglomerations: in red - the current extent; in blue - the future extent; in yellow - the extent from 2008. For the outline of the centers, services, hospitality and entertainment locations are recognized. Contrary to the mostly comprehensive data set of retail locations in Düsseldorf, in Duisburg the author has noted that large parts of the center-relevant retail locations are missing from OpenStreetMap.

---

## Results

## Validation

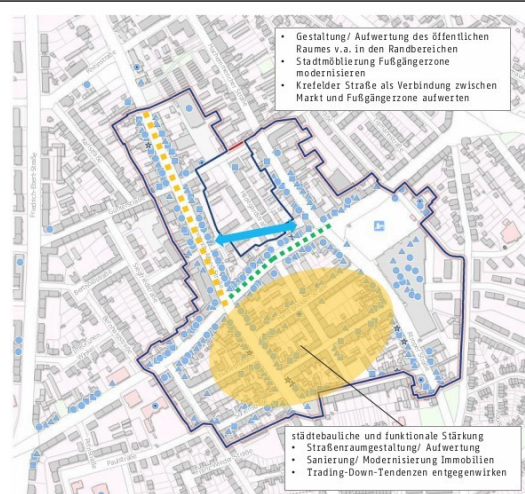
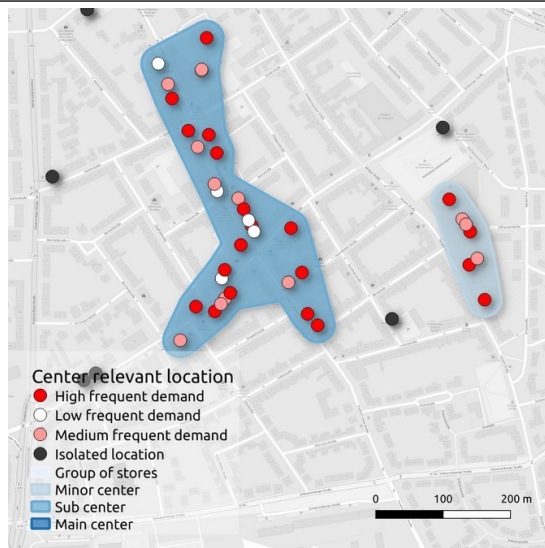
### Duisburg Main-center



The main center is correctly detected in Duisburg. While the location of the center itself is indeed well detected, the extent of the results of this work is compressed and focused on the most central area. The northern part of the shopping street and the areas close to the main station in the east are not included in the main center, but rather detected as smaller centers.

### Duisburg Sub-center

### Rheinhausen

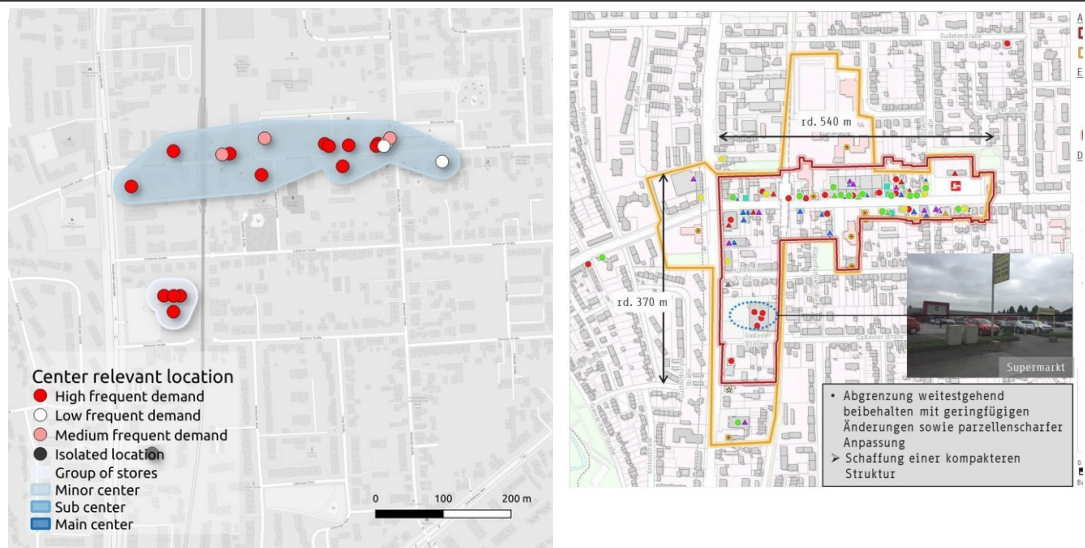


Moreover, the sub-center in Rheinhausen is identified with its correct center type. The outline is correct for the parts where center-relevant locations are available. As locations are missing, especially in the eastern part of the center, the detected center is focused on the main street in the west. Neither the shopping center in the south-east nor the surrounding isolated shopping locations are integrated.

---

**Duisburg**  
**Minor center**

**Buchholz**

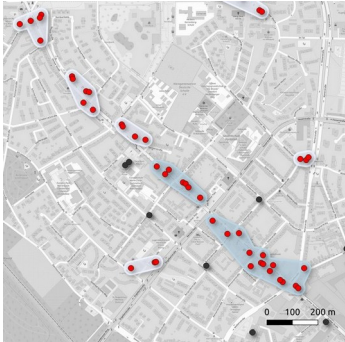
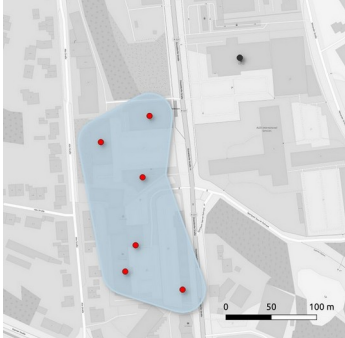
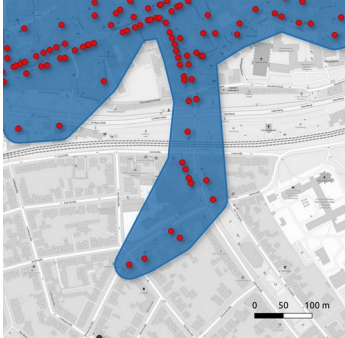



The minor center in Duisburg Buchholz is well identified and outlined in its current extent. The group of stores in the south is identified as a small group (including a supermarket and some smaller stores, e.g. a bakery), and not connected to the center. A distance of around 170 meters separates the two centers.

A comparison between the official outlines of central retail agglomerations and the results of this work highlighted a general conformity and several topics for discussion. The officially defined centers have a more generalized outline and follow the local infrastructure, e.g. roads, rivers or building outlines. Another general observation is that the results of this work are usually more compact, because no other center-defining facilities apart from retail locations were considered. The definition and outline of centers is often political or strategical. Although the definition of a center requires continuous retail locations in close proximity, cities often count smaller groups of stores as part of a larger center situated nearby. The results also show that, even in cases where the retail location data available on OpenStreetMap is incomplete, centers and their hierarchy within the city are correctly detected.

Finally, the results were inspected in depth, relying on local knowledge. This brings some additional limitations to the fore. However, not all alleged limitations necessarily represent real challenges; instead, they may require more details or a site visit for actual validation.

Table 23: Exemplary limitations

Limitation	Description	Exemplary
<b>Interrupted centers</b>	In some cases, the identified centers appear to be interrupted because of multiple smaller centers identified in the proximity.	 <p data-bbox="1222 622 1305 649"><b>Neuss</b></p>
<b>Isolated location in proximity to center</b>	In some rather isolated locations with large parking spaces, the maximum distance of 100 meters is somewhat low. This leads to some isolated locations being excluded from an identified center.	 <p data-bbox="1110 1037 1417 1108"><b>Mülheim an der Ruhr Saarn</b></p>
<b>Center detection across obstacle</b>	In rare cases, obstacles cut through the detected clusters. In the few observed cases, these obstacles are predominantly represented by small rivers or railways and motorways without direct access to a bridge or tunnel.	 <p data-bbox="1187 1478 1337 1505"><b>Paderborn</b></p>
<b>Identification of multiple main centers within one municipality</b>	In a few cases, multiple main centers have been detected. This is the case when two centers fit the definition of a center and have the same characteristics.	 <p data-bbox="1187 1874 1337 1901"><b>Rüdesheim</b></p>

To sum up, the validation revealed three main limiting factors influencing the quality of the results of this work 1. The comprehensiveness of the OpenStreetMap data 2. Missing integration of local infrastructural circumstances, e.g. building outlines or roads. 3. Sole concentration of this work on center-relevant retail locations and their continuous and dense distribution. The validation against the official outlines and classification of the central retail agglomerations of the cities of Düsseldorf and Duisburg showed that the comprehensiveness of the OpenStreetMap retail store locations has some influence on the outline of the centers and a limited influence on the correct detection of the center hierarchy within the city.

### 4.3 Reflecting on the research method

The research method applied was affected by numerous bounding criteria, which had a limiting impact on the available methodologies and subsequent the results. More specifically, no comprehensive retail location data is currently available, the area of interest was large and the available computer is equipped with limited processing power.

The national scale is very large for this kind of analysis, and, with the exception of the United Kingdom of Great Britain and Northern Ireland (Pavlis, Dolega and Singleton, 2018), no known research has been done at a national level on this topic. The scale is central to most of the decisions made in terms of choosing the appropriate methodology, parameters and algorithms, as the large area and number of locations automatically requires more computing power and processing time. However, for applications such as calculating the customer attraction to certain centers, the territory might prove to be too small, as cities close to the border, e.g. Nijmegen (NL), Strasbourg (FR) or Salzburg (AT), are relevant shopping destinations for the population living in the border region in Germany.

One implication of the large territory analyzed is that one set of variables was used across the whole study area. This was decided to keep the algorithm simple and processable on common computer hardware. Similar research showed that retail clusters are successfully discovered in both urban and rural locations even though the point density is higher in urban locations (Yang et al., 2018). Other researchers argued that it might be reasonable to adapt the parameters to local circumstances (Devkota et al., 2019; Pavlis, Dolega and Singleton, 2018; Yang et al., 2018). In comparison to the research works mentioned above, this paper focused on center-relevant retail locations rather than all retail and service locations which might be found within a center. In the case of Germany, this distinction results in mostly separated clusters of locations. Furthermore, this research showed that a central retail agglomeration can no longer be described as continuous if it exceeds a minimum density of around 100 meters. DBSCAN returned suggestive results and it was also one of the fastest algorithms tested.

While the cluster algorithm for the identification of the central retail agglomerations worked well at national scale, the classification of the clusters was challenging because the data used for calibration is limited and the centers show different signatures, even in cases when they are of the same type. The classification is based on data characterizing the centers, which, in turn, is based on the locations that the centers are comprised of. Apart from the varying composition and form, the characterizing information is further distorted by missing location data. This is most relevant for small central retail agglomerations, where even one missing location has a major impact on the statistics describing the composition of the place. As such, the classification results for the smallest centers have to remain indicative data and be continually questioned.

Another challenge stems from the fact that, even if the classification of main center, sub-center, minor center, group of stores and isolated locations is rather common, municipalities define the notion of “sub-center” differently. It might therefore be worthwhile to challenge the hierarchical approach to classifying centers which is traditionally applied in Germany (Christaller, 1933; Lösch, 1940). Alternative approaches might have considered the number of locations present within each center, as this is the most reliable attribute available for classification. Alternative methods not bound to the hierarchical system could include multi-variable classifications (e.g. k-means), or other machine learning methods.

Retail locations are attributed with additional values, such as the frequency of demand or if they function as a supply magnet for the surrounding locations. This attribution is solely based on the associated OpenStreetMap tag, as more reliable and comprehensive data is not available. Shops might be insufficiently described with a single tag like supermarket. This and adding additional attributes for the classification might have introduced systematical errors to the data-set.

QGIS was selected as the main tool to do the analysis and the modeling of the automation workflow. Other commercial GIS-tools were available and considered but eventually dropped. The choice of QGIS brought the advantage of pre-existing knowledge and the availability of the workflows past expired licenses, as well as faster performance in situations when limited processing power is available. Within QGIS and based on the processing framework, the workflow was designed using the ‘Processing Model Designer’. Apart from performing the spatial operations and algorithms step by step, the whole process could be automated in a way that enables updates and multiple runs in a matter of reasonable processing times. Since the plugin developed is published alongside this work, the results are traceable as well as reproducible. The processing framework and the ‘Processing Model Designer’ of QGIS work well and have seen constant development over the last years. Some details are, however, still missing. For instance, more comprehensive documentation is needed; the process can not be run on a sample of the data; and saving the results to a database requires improvement. Given the fact that QGIS is an open-source tool, the described challenges are already known to the developer community and a crowdfunding campaign have been launched to improve certain aspects of the software by early 2020.

The runtime of the models is relatively high, which is due to the large data set and limitations within QGIS. Further optimization of the runtime can be found in testing different processing providers within QGIS (e.g. the Grass tools), choosing different file formats and optimizing the sequencing of the processes. The automation of additional steps such as loading the data and a more flexible integration of the required domain knowledge would also be beneficial. Despite the fact that the runtime is 1 hours and 20 minutes, it should be noted that performing all the steps individually would take several hours. The update cycle of the data infrequent, what relativizes the runtime and the need for optimization.

## 5 Conclusion

The results of this work show that it is possible to understand the current situation of central retail agglomerations over a large area and that meaningful insights for individual centers, municipalities or regional studies can be derived. Traditional analyses comparable to this work were limited to municipal extents or chained retail outlets. The final data sets provide a comprehensive perspective on where central retail agglomerations are, how they could be fundamentally described and classified, and consequently where people shop for center relevant assortments in Germany.

The proposed analytical framework enhances the studies on central retail agglomeration in several ways. Firstly, it identifies and classifies central retail agglomerations at a country-wide scale, by fusing publicly available data and domain knowledge. Secondly, it defines a method meant to estimate clustering and classifying parameters, which can subsequently be applied across the study area. Thirdly, it integrates the process to such an extent that regular updates and investigations in other geographies become feasible. Moreover, the final data set offers, for the first time, a comprehensive overview on central retail agglomerations in Germany.

The novelty of this research and the results lie in the methodology identified, which is based on volunteered geographic information and take into account a large area. The DBSCAN clustering algorithm was shown to be the most suitable and performant choice for cluster detection. The parameterization was based on domain knowledge and data derived from the original information available. Characterizing and classifying the identified potential centers in the five predefined categories (main center, sub-center, minor center, group of stores and isolated locations) was possible – with conviction – for most of the locations and clusters. However, the fine gradient between the most central clusters remained challenging due to the location data used, the large variety of sizes, forms and compositions of centers, as well as missing data for calibration.

The analysis of the retail locations available on OpenStreetMap showed that the data set is mostly comprehensive and final blank spaces will probably be closed within next years. In addition, the results prove that VGI provides a high level of detail and that the sources are sufficiently thorough to enable the derivation of reliable and meaningful results for numerous purposes.

Every major city in Germany has its own center concept. Depending on the situation, these concepts differ to a large extent from one another, so that a 1:1 comparison is difficult. The results of the classification might therefore describe central retail agglomerations differently from municipal retail center plans or local knowledge. Similarly, two areas described as minor-centers might have different functions for the local community. The classification, although not comparable to municipal retail-center planning, provides a universal, transparent and easy to understand perspective on central retail agglomerations in Germany.

Finally, it was possible to combine the workflow comprising the data preparation with the clustering, characterization, classification and preparation of basic statistics for interpretation into a stand-alone Plugin for QGIS.

The resulting data holds numerous valuable information for additional in-depth analyses for a range of interest groups and use cases. Stakeholders who might benefit most from transforming the data into value are retail companies and shop owners, city planners and the local, regional and national authorities. Various use cases and suggestive analyses can be defined for all three groups. For instance, retail companies and shop owners could use the local retail center composition to tailor their



offering to the local circumstances or use the data to identify the most attractive centers for a market entry or expansion. City planners and local governments would typically carry out more detailed analyses for their own city, based on comprehensive data; this is because the results become part of the regulations and the ways in which cities decide on and control the retail function, protecting the central retail agglomerations. The data presented in this work might help them to learn best practices about efficient center design or center composition from other locations. Furthermore, investigations concerning the competing centers in the larger vicinity would also be relevant. The regional and national governments benefit from the results because they are based on the same approach across the whole country. The data enables these interest groups to analyze the supply needs of the population across large areas, optimize the regional and national spatial planning concepts or identify distressed centers.

The open location data source and the highly automated clustering and classification algorithm used enable the application of this approach to new geographies and the implementation of updates for time series analyses. The first such analyses show that similar positive results can be achieved with the same parameterization at least for culturally similar countries, as is the case of Austria and Germany.

The focus of this work lies on deriving high-level insights from unstructured data by describing the current state and distribution of central retail agglomerations in Germany. The results are meant to set the foundational knowledge to understand these areas, plan for a future state and make informed decisions when adapting to changes related to, for example, online retail, decreasing footfall or the emergence of new store concepts. The research conducted anticipates an increase in the speed at which retail locations are forced to adapt to changing circumstances (Dearden and Wilson, 2011; Brown, 1994; Dolega et al., 2019). In this context, taking the discussion a step further, from describing the current conditions to predictive and spatially aware actions for city planners or shop owners, will remain open for subsequent research.

There are multiple directions for moving this research forward and to answer arising and more detailed questions. The next potential steps can be grouped by three aspects. The first priority would be to improve the foundational data and to incorporate additional data. The second step is to improve the methodology, the domain knowledge and how both are integrated. The third and last step would be the expansion of the results, including more details to derive even more actionable results.

*Table 24: Selected aspects for improving the results*

<b>Improvement</b>	<b>Aspects</b>
<b>Foundational data and incorporating additional data</b>	<ul style="list-style-type: none"> <li>• Supplement the retail locations to gain a more comprehensive data set.</li> <li>• Expand the level of detail per location (e.g. affiliation to certain brand, size, assortment, situation within the center, opening hours, more differentiated classification of stores).</li> <li>• Expand the research to other center-relevant aspects such as services, tourism, entertainment and hospitality.</li> <li>• Incorporate spatial data on the urban land layout and infrastructure (e.g. roads, buildings, parking, public transport).</li> <li>• Include the demand perspective (e.g. demographic information such as total population or purchasing power, tourism or footfall).</li> </ul>
<b>Methodology</b>	<ul style="list-style-type: none"> <li>• Improve and expand the automation of the workflow by integrating the</li> </ul>

---

**and integration  
of domain  
knowledge**

sourcing of the relevant data.

- Adapt the algorithm to recognize impassable barriers.
- Re-calibrate the selection of center-relevant retail and the parameters based on the results.
- Develop a training data set for the classification, which could help to improve the classification results.
- Expand the algorithm to handle other center-relevant functions (e.g. services and hospitality).
- Eliminate parts of the center polygons not containing any locations.
- Increase the performance of the algorithm.

---

**Expansion of  
the results**

- Observe the agglomerations over time to expand the static perspective with a perspective on the temporal processes.
  - Describe in detail and distinguish between different types of centers (e.g. business density, mix of industries, share of chained retail, attractiveness of flanking retail-related services, shop vacancies, or the level of agglomeration).
  - Name the locations based on dominant features such as roads.
  - Estimate the level of supply to find under- and oversupplied areas by type of good.
  - Extend and fuse the analysis with similar centers (e.g. services, entertainment, tourism or hospitality).
  - Incorporate and recognize the center close to the border of Germany.
  - Check for the distribution within the center-relevant spots (e.g. Hot- and cold-spots).
  - Evaluate how the clusters are interrelated.
  - Calculate the demand for the centers (e.g. by using a gravity model).
  - Derive location-specific recommendations for action for the different stakeholders.
- 

Additional improvements could be made to other aspects, such as the elimination of bias resulting from the characteristics of the OpenStreetMap data. Incorporating the most beneficial of the optimizations outlined above would significantly improve the clustering, classification and informative value of the results. However, some limitations and edge cases will remain, which will have to be solved manually or by visiting the places themselves for validation.

This work showed that a current urban question can be transformed into a data question. It also highlighted how OpenStreetMap data can be sourced and effectively set into power, and how spatial thinking, methodology and technology can be used to derive meaningful classified data and insights that can help to inform decision-making processes.

## 6 List of References

Acocella, D., 2018. *Gutachten als Grundlage zur Fortschreibung des Einzelhandels- und Zentrenkonzeptes für die Stadt Freiburg 2018*.

Acocella, D., 2019. *Einzelhandels- und Zentrenkonzept der Stadt Duisburg 2019*.

Anderson, J., Sarkar, D. and Palen, L., 2019. Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), p.232.

Anselin, L., 2010. Local Indicators of Spatial Association-LISA. *Geographical Analysis*, 27(2), pp.93–115.

Bayern, 2020. *Landesentwicklungsprogramm Bayern (LEP) - nicht-amtliche Lesefassung Stand 01.01.2020: Landesentwicklung Bayern*. [online] Available at: <<https://www.landesentwicklung-bayern.de/instrumente/landesentwicklungsprogramm/landesentwicklungs-programm-bayern-stand-2018/>> [Accessed 10 Feb. 2020].

BBSR, 2017. *Online-Handel – Mögliche räumliche Auswirkungen auf Innenstädte, Stadtteil- und Ortszentren*. [online] Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Bundesamt für Bauwesen und Raumordnung (BBR). Available at: <<https://difu.de/publikationen/2017/online-handel-moegliche-raeumliche-auswirkungen-auf.html>> [Accessed 13 Feb. 2020].

BBSR, 2020. *Stadt- und Gemeindetyp*. [online] Available at: <[https://www.bbsr.bund.de/BBSR/DE/Raubeobachtung/Raumabgrenzungen/deutschland/gemeinden/StadtGemeindetyp/StadtGemeindetyp\\_node.html](https://www.bbsr.bund.de/BBSR/DE/Raubeobachtung/Raumabgrenzungen/deutschland/gemeinden/StadtGemeindetyp/StadtGemeindetyp_node.html)> [Accessed 14 Jan. 2020].

Berger, J.O., 1993. *Statistical decision theory and Bayesian analysis*. 2nd ed. Springer series in statistics. New York: Springer-Verlag.

BKG, 2020. *GeoBasis-DE - Verwaltungsgebiete 1:250 000 mit Einwohnerzahlen (Ebenen), Stand 31.12*. [online] Available at: <<https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/verwaltungsgebiete/verwaltungsgebiete-1-250-000-mit-einwohnerzahlen-ebenen-stand-31-12-vg250-ew-ebenen-31-12.html>> [Accessed 14 Jan. 2020].

Breiman, L. ed., 1998. *Classification and regression trees*. Repr ed. Boca Raton: Chapman & Hall [u.a.].

Brown, S., 1994. Retail Location at the Micro-Scale: Inventory and Prospect. *The Service Industries Journal*, 14(4), pp.542–576.

Bunzel, A. and Difu eds., 2009. *Erhaltung und Entwicklung zentraler Versorgungsbereiche*. Difu-Arbeitshilfe. Berlin: Deutsches Institut für Urbanistik.

Campello, R.J.G.B., Moulavi, D. and Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In: J. Pei, V.S. Tseng, L. Cao, H. Motoda and G. Xu, eds. *Advances in Knowledge Discovery and Data Mining*. [online] Berlin, Heidelberg: Springer Berlin Heidelberg, pp.160–172. Available at: <[http://link.springer.com/10.1007/978-3-642-37456-2\\_14](http://link.springer.com/10.1007/978-3-642-37456-2_14)> [Accessed 9 Feb. 2020].

- Carol, H., 1960. The Hierarchy of Central Functions within the City. *Annals of the Association of American Geographers*, 50(4), pp.419–438.
- Chen, M., Arribas-Bel, D. and Singleton, A., 2019. Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*, 21(1), pp.89–109.
- Christaller, W., 1933. *Die zentralen Orte in Süddeutschland: eine ökonomisch-geographische Untersuchung über die Gesetzmäßigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*. Sonderausg. der 2., unverändert. Aufl ed. WBG-Bibliothek. Darmstadt: Wiss. Buchges.
- Daniels, P.W., 1993. *Service industries in the world economy*. IBG studies in geography. Oxford, UK ; Cambridge, Mass: Blackwell.
- Dearden, J. and Wilson, A., 2011. A Framework for Exploring Urban Retail Discontinuities. 探索城市零售业不连续性的框架: Exploring Urban Retail Discontinuities. *Geographical Analysis*, 43(2), pp.172–187.
- Dehuri, S., Mohapatra, C., Ghosh, A. and Mall, R., 2006. A Comparative Study of Clustering Algorithms. *Information Technology Journal*, 5(3), pp.551–559.
- Deutschland, 2020. *ROG - nichtamtliches Inhaltsverzeichnis*. [online] Available at: <[https://www.gesetze-im-internet.de/rog\\_2008/](https://www.gesetze-im-internet.de/rog_2008/)> [Accessed 10 Feb. 2020].
- Devkota, B., Miyazaki, H., Witayangkurn, A. and Kim, S.M., 2019. Using Volunteered Geographic Information and Nighttime Light Remote Sensing Data to Identify Tourism Areas of Interest. *Sustainability*, 11(17), p.4718.
- Dolega, L., Reynolds, J., Singleton, A. and Pavlis, M., 2019. Beyond retail: New ways of classifying UK shopping and consumption spaces. *Environment and Planning B: Urban Analytics and City Science*. [online] Available at: <<http://journals.sagepub.com/doi/10.1177/2399808319840666>> [Accessed 13 Feb. 2020].
- Dueck, D., 2009. *Affinity Propagation: Clustering Data by Passing Messages*.
- Ertöz, L., Steinbach, M. and Kumar, V., 2003. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In: *Proceedings of the 2003 SIAM International Conference on Data Mining*. [online] Proceedings of the 2003 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics.pp.47–58. Available at: <<https://epubs.siam.org/doi/10.1137/1.9781611972733.5>> [Accessed 29 Dec. 2019].
- Ester, M., Kriegel, H.-P., Sander, J. and Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press.pp.226–231.
- Gan, G., Ma, C. and Wu, J., 2007. *Data clustering: theory, algorithms, and applications*. ASA-SIAM series on statistics and applied probability. Philadelphia, Pa. : Alexandria, Va: SIAM, Society for Industrial and Applied Mathematics ; American Statistical Association.
- Geofabrik GmbH, 2020. *OpenStreetMap Data Extracts*. [online] Available at: <<http://download.geofabrik.de/>> [Accessed 1 Jan. 2020].
- Getis, A. and Ord, J.K., 2010. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3), pp.189–206.

- Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211–221.
- Greiving, S., Flex, F. and ARL eds., 2016. *Neuaufstellung des Zentrale-Orte-Konzepts in Nordrhein-Westfalen*. Arbeitsberichte der ARL. Hannover: ARL, Akademie für Raumforschung und Landesplanung, Leibniz-Forum für Raumwissenschaften.
- HeiGIT, 2020. *Ohsome - OSM History Analyzer*. [online] Available at: <<https://ohsome.org/apps/dashboard/>> [Accessed 15 Jan. 2020].
- Heineberg, H., 2017. *Stadtgeographie*. 5., überarbeitete Auflage ed. utb. Paderborn: Ferdinand Schöningh.
- Heinritz, G., 1979. *Zentralität und zentrale Orte: eine Einführung*. Teubner-Studienbücher Geographie. Stuttgart: Teubner.
- Heinritz, G., Klein, K. and Popp, M., 2003. *Geographische Handelsforschung*. Studienbücher der Geographie. Berlin: Borntraeger.
- Hickson, D.J. ed., 1986. *Top decisions: strategic decision-making in organizations*. 1st ed. The Jossey-Bass management series. San Francisco: Jossey-Bass Publishers.
- Hotelling, H., 1928. Stability in Competition. In: A.C. Darnell, ed. *The Collected Economics Articles of Harold Hotelling*. [online] New York, NY: Springer New York. pp.50–63. Available at: <[http://link.springer.com/10.1007/978-1-4613-8905-7\\_4](http://link.springer.com/10.1007/978-1-4613-8905-7_4)> [Accessed 11 Jan. 2020].
- Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W. and Prasad, S., 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54, pp.240–254.
- Jonietz, D. and Zipf, A., 2016. Defining Fitness-for-Use for Crowdsourced Points of Interest (POI). *ISPRS International Journal of Geo-Information*, 5(9), p.149.
- Junker, R. and Kühn, G., 2006. *Nahversorgung in Grossstädten*. Difu-Beiträge zur Stadtforschung. Berlin: Deutsches Institut für Urbanistik.
- Krider, R.E. and Putler, D.S., 2013. Which Birds of a Feather Flock Together? Clustering and Avoidance Patterns of Similar Retail Outlets: Retail Clustering and Avoidance Patterns. *Geographical Analysis*, 45(2), pp.123–149.
- Kriegel, H.-P., Schubert, E. and Zimek, A., 2017. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, 52(2), pp.341–378.
- Kulke, E., 2017. *Wirtschaftsgeographie*. 6., aktualisierte Auflage ed. utb Geographie. Paderborn: Ferdinand Schöningh.
- Kulke, E., 2020a. Dynamik von Zentralsystemen. In: C. Neiberger and B. Hahn, eds. *Geographische Handelsforschung*. [online] Berlin, Heidelberg: Springer Berlin Heidelberg. pp.183–192. Available at: <[http://link.springer.com/10.1007/978-3-662-59080-5\\_16](http://link.springer.com/10.1007/978-3-662-59080-5_16)> [Accessed 17 Feb. 2020].

- Kulke, E., 2020b. Zentren und Zentrensysteme. In: C. Neiberger and B. Hahn, eds. *Geographische Handelsforschung*. [online] Berlin, Heidelberg: Springer Berlin Heidelberg, pp.171–181. Available at: <[http://link.springer.com/10.1007/978-3-662-59080-5\\_15](http://link.springer.com/10.1007/978-3-662-59080-5_15)> [Accessed 17 Feb. 2020].
- Ladner, R., Petry, F.E. and Cobb, M.A., 2003. Fuzzy Set Approaches to Spatial Data Mining of Association Rules. *Transactions in GIS*, 7(1), pp.123–138.
- Lichtenberger, E., 1963. Die Geschäftsstrassen Wiens : eine statistischphysiognomische Analyse. In: *Mitteilungen der Österreichischen Geographischen Gesellschaft: MÖGG ; Jahresband*. Wien, pp.463–504.
- Lin, P., Lei, Z., Chen, L., Yang, J. and Liu, F., 2009. Decision Tree Network Traffic Classifier Via Adaptive Hierarchical Clustering for Imperfect Training Dataset. In: *2009 5th International Conference on Wireless Communications, Networking and Mobile Computing*. [online] 2009 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM). Beijing, China: IEEE, pp.1–6. Available at: <<http://ieeexplore.ieee.org/document/5302133/>> [Accessed 13 Feb. 2020].
- Liu, B., Xia, Y. and Yu, P.S., 2000. Clustering through decision tree construction. In: *Proceedings of the ninth international conference on Information and knowledge management - CIKM '00*. [online] the ninth international conference. McLean, Virginia, United States: ACM Press, pp.20–29. Available at: <<http://portal.acm.org/citation.cfm?doid=354756.354775>> [Accessed 13 Feb. 2020].
- Lösch, A., 1940. *Die räumliche Ordnung der Wirtschaft: eine Untersuchung über Standort, Wirtschaftsgebiete und internationalen Handel*. Die Handelsblatt-Bibliothek 'Klassiker der Nationalökonomie'. Düsseldorf: Verl. Wirtschaft und Finanzen.
- Machanavajjhala, A., Gehrke, J., Moro, M.M., Marian, A., Schenkel, R., Theobald, M., Baumann, P., Brinkhoff, T., Hansson, J., Xiong, M., Zhang, E., Zhang, Y., Tan, P.-N., Domingo-Ferrer, J., Arasu, A., Domingo-Ferrer, J., Wada, K., Patel, C., Weng, C., Young-Lai, M., Sion, R., Tannen, V., Tannen, V., W. Eembley, D., Lalmas, M., Jensen, C.S., Snodgrass, R.T., Pehcevski, J., Larsen, B., He, B., Wang, X.-J., Zhang, L., Vechtomova, O., Jimenez-Peris, R., Patiño-Martínez, M., Fekete, A., Kemme, B., Wada, K., Pedone, F., Kemme, B., Jiménez-Peris, R., Patiño-Martínez, M., Jimenez-Peris, R., Patiño-Martínez, M., Despotovic, Z., Gokhale, A., Novák, V., Zhang, D., Du, Y., Weiss, M., Christophides, V., Janée, G., Plaisant, C., Etzion, O., Papadias, D., Tao, Y., Li, C., Gokhale, A., Shahabi, C., He, B., Zhang, Y., Joshi, J.B.D., Craswell, N., Papadopoulos, A.N., Corral, A., Nanopoulos, A., Theodoridis, Y. and Tung, A.K.H., 2009. Rule-based Classification. In: L. Liu and M.T. Özsu, eds. *Encyclopedia of Database Systems*. [online] Boston, MA: Springer US, pp.2459–2462. Available at: <[http://link.springer.com/10.1007/978-0-387-39940-9\\_559](http://link.springer.com/10.1007/978-0-387-39940-9_559)> [Accessed 25 Feb. 2020].
- Mackaness, W.A. and Chaudhry, O.Z., 2011. Automatic Classification of Retail Spaces from a Large Scale Topographic Database: Automatic Classification of Retail Spaces. *Transactions in GIS*, 15(3), pp.291–307.
- Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. pp.281–297.
- Miller, H.J. and Goodchild, M.F., 2015. Data-driven geography. *GeoJournal*, 80(4), pp.449–461.
- Mocnik, F.-B., Mobasher, A. and Zipf, A., 2018. Open source data mining infrastructure for exploring and analysing OpenStreetMap. *Open Geospatial Data, Software and Standards*, [online] 3(1). Available at: <<https://opengeospatialdata.springeropen.com/articles/10.1186/s40965-018-0047-6>> [Accessed 30 Dec. 2019].

Mooney, P. and Corcoran, P., 2014. Has OpenStreetMap a role in Digital Earth applications? *International Journal of Digital Earth*, 7(7), pp.534–553.

Mustière, S. and van Smaalen, J., 2007. Database Requirements for Generalisation and Multiple Representations. In: *Generalisation of Geographic Information*. [online] Elsevier. pp.113–136. Available at: <<https://linkinghub.elsevier.com/retrieve/pii/B9780080453743500089>> [Accessed 9 Feb. 2020].

Neis, P., 2020. *OSMstats*. [online] Available at: <<https://osmstats.neis-one.org/?item=countries&country=Germany&date=1-5-2019>> [Accessed 14 Jan. 2020].

Nelson, R.L., 1958. *The selection of retail locations*. FW Dodge Corporation.

Neun, M., Weibel, R. and Burghardt, D., 2004. Data Enrichment for Adaptive Generalisation. *The 7th ICA Workshop on Generalisation and Multiple Representation, Leicester*.

Ng, A.Y., Jordan, M.I. and Weiss, Y., 2001. On Spectral Clustering: Analysis and an algorithm. In: *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*. MIT Press. pp.849--856.

Niedersachsen, 2020. *Neubekanntmachung der LROP-Verordnung 2017 | Nds. Ministerium für Ernährung, Landwirtschaft und Verbraucherschutz*. [online] Available at: <[https://www.ml.niedersachsen.de/startseite/themen/raumordnung\\_landesplanung/landes\\_raumordnungsprogramm/neubekanntmachung-der-lrop-verordnung-2017-158596.html](https://www.ml.niedersachsen.de/startseite/themen/raumordnung_landesplanung/landes_raumordnungsprogramm/neubekanntmachung-der-lrop-verordnung-2017-158596.html)> [Accessed 10 Feb. 2020].

NRW, 2020. *Landesentwicklungsplan Nordrhein-Westfalen*. [online] Available at: <<https://www.regioplaner.de/planung-raum/raumordnung/landesentwicklungsplan-nordrhein-westfalen>> [Accessed 10 Feb. 2020].

OpenStreetMap contributors, 2020. *OpenStreetMap*. [online] Available at: <<https://openstreetmap.org>>.

Orzessek-Kruppa, R., 2016. *Düsseldorfer Rahmenplan Einzelhandel 2016*.

Pavlis, M., Dolega, L. and Singleton, A., 2018. A Modified DBSCAN Clustering Method to Estimate Retail Center Extent. *Geographical Analysis*, 50(2), pp.141–161.

Quinlan, J.R., 1983. Learning Efficient Classification Procedures and Their Application to Chess End Games. In: R.S. Michalski, J.G. Carbonell and T.M. Mitchell, eds. *Machine Learning*. [online] Berlin, Heidelberg: Springer Berlin Heidelberg. pp.463–482. Available at: <[http://link.springer.com/10.1007/978-3-662-12405-5\\_15](http://link.springer.com/10.1007/978-3-662-12405-5_15)> [Accessed 13 Feb. 2020].

Rosenblatt, M., 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3), pp.832–837.

Sander, J., Ester, M., Kriegel, H.-P. and Xu, X., 1998. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, 2(2), pp.169–194.

Schiller, R., 2001. *Dynamics of property location*. [online] Available at: <<https://ebookcentral.proquest.com/lib/uqac-ebooks/detail.action?docID=167613>> [Accessed 12 Feb. 2020].

- Schubert, E. and Rousseeuw, P.J., 2019. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *arXiv:1810.05691 [cs, stat]*, 11807, pp.171–187.
- Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X., 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*, 42(3), pp.1–21.
- Schulte, T.E., 2012. *Städtebauliche Auswirkungen von innerstädtischen Einkaufszentren*.
- StaBu, 2020. *Statistisches Bundesamt*. [online] Available at: <[https://www.destatis.de/EN/Home/\\_node.html](https://www.destatis.de/EN/Home/_node.html)> [Accessed 10 Feb. 2020].
- Touya, G., Antoniou, V., Olteanu-Raimond, A.-M. and Van Damme, M.-D., 2017. Assessing Crowdsourced POI Quality: Combining Methods Based on Reference Data, History, and Spatial Relations. *ISPRS International Journal of Geo-Information*, 6(3), p.80.
- Urban, H. and Weidner, M., 2010. *Zentrenkonzept München*.
- Vogels, P.-H., Holl, S. and Birk, H.-J., 1998. *Auswirkungen grossflächiger Einzelhandelsbetriebe*. Stadtforschung aktuell. Basel: Birkhäuser.
- Wang, T., Ren, C., Luo, Y. and Tian, J., 2019. NS-DBSCAN: A Density-Based Clustering Algorithm in Network Space. *ISPRS International Journal of Geo-Information*, 8(5), p.218.
- Xia, B., Zou, Z. and Su, W., 2018. POI Based Urban Commercial Centers Identification and Classification. *DEStech Transactions on Computer Science and Engineering*, [online] (iece). Available at: <<http://dpi-proceedings.com/index.php/dtcse/article/view/26606>> [Accessed 29 Dec. 2019].
- Yang, J., Cao, J., He, R. and Zhang, L., 2018. A unified clustering approach for identifying functional zones in suburban and urban areas. In: *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. [online] IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Honolulu, HI, USA: IEEE, pp.94–99. Available at: <<https://ieeexplore.ieee.org/document/8406847/>> [Accessed 27 Dec. 2019].
- Yizong Cheng, 1995. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8), pp.790–799.
- Zhang, L. and Pfoser, D., 2019. Using OpenStreetMap point-of-interest data to model urban change —A feasibility study. *PLOS ONE*, 14(2), p.e0212606.
- Zhou, X., Xu, C. and Kimmons, B., 2015. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers, Environment and Urban Systems*, 54, pp.144–153.



## 7 Appendix

### Resulting data and processing plugin

The resulting data as GeoPackage and the processing plugin for QGIS are available for download from this link: <https://nx4521.your-storageshare.de/s/nQ5PkANDYb2rDFE>

<b>Name</b>	<b>Description</b>
<b>Plugin_RetailAgglomeration_v1.model3</b>	QGIS processing plugin
	Municipalities by type
	All locations from OpenStreetMap
	All center relevant locations
	Cluster results (points and polygon)
<b>Data_RetailAgglomerations_Germany_2020.gpkg</b>	Noise point
	Group of store
	Minor center
	Sub-center
	Main center

## Center relevant and classified features considered for the analysis

The point locations are sourced from OpenStreetMap (OpenStreetMap contributors, 2020; Geofabrik GmbH, 2020). The classification is derived from multiple sources (Nelson, 1958; Kulke, 2017, 2020a; Bunzel and Difu, 2009; Acocella, 2018, 2019; Orzessek-Kruppa, 2016; Heineberg, 2017)

code	type	description	c_relev	magnet	freq_o_dem	no_of_feat
2502	bakery	Shop focused on selling bread	yes_ls	no	high	36.634
2501	supermarket	Supermarket – a large store with groceries and other items	yes_ls	yes_ls	high	34.312
2512	clothes	Shop focused on selling clothes or underwear	yes	no	medium	30.506
2101	pharmacy	A shop where a pharmacist sells medications	yes_ls	yes_ls	high	19.370
2516	butcher	Shop focused on selling meat	yes_ls	no	high	11.460
2503	kiosk	A small shop on the pavement that sells magazines, tobacco, newspapers, sweets and stamps	yes_ls	no	high	11.172
2513	florist	Shop focused on selling bouquets of flowers	yes_ls	no	high	11.031
2518	beverages	Shop focused on selling alcoholic and non-alcoholic beverages.	yes_ls	no	high	10.472
2511	convenience	A small local shop carrying a small subset of the items you would find in a supermarket	yes_ls	no	high	9.014
2517	shoe_shop	Shop focused on selling shoes	yes	no	medium	7.828
2519	optician	Shop focused on selling eyeglasses, contact lenses	yes	no	low	7.153
2529	beauty_shop	A non-hairdresser beauty shop, spa, nail salon	yes_ls	no	high	6.951
2520	jeweller	Jewelers shops	yes	no	low	5.559
2514	chemist	Shop focused on selling articles of personal hygiene, cosmetics, and household cleaning products	yes_ls	yes_ls	high	5.404
2515	bookshop	Shop focused on selling books	yes	no	high	4.446
2525	mobile_phone_shop	Shop focused on selling mobile phones and accessories	yes	no	low	3.880
2521	gift_shop	Shop focused on selling gifts, greeting cards, or tourist gifts	yes	no	medium	3.196
2523	stationery	Shop focused on selling office supplies	yes	no	high	3.097
2522	sports_shop	Shop focused on selling sporting goods	yes	no	medium	2.796
2546	computer_shop	Shop focused on selling computers, peripherals, software	yes	no	low	2.225
2528	greengrocer	Shop focused on selling vegetables and fruits	yes_ls	no	high	2.204
2526	toy_shop	Shop focused on selling toys	yes	no	medium	2.093
2504	mall	A shopping mall – multiple stores under one roof (also known as a shopping center)	yes	yes	medium	1.780
2505	department_store	A single large store – often multiple storeys high – selling a large variety of goods	yes	yes	medium	1.722
2527	newsagent	Shop focused on selling newspapers, cigarettes, other goods	yes_ls	no	high	1.222
2524	outdoor_shop	Shop focused on selling camping, walking, climbing, and other outdoor sports equipment	yes	no	medium	942
2530	video_shop	Shop focused on selling or renting out videos/DVDs	yes	no	medium	475
<b>Total</b>						<b>236.944</b>

## Center relevant assortment

The classification is derived from multiple sources (Nelson, 1958; Kulke, 2017, 2020a; Acocella, 2018, 2019; Bunzel and Difu, 2009; Heineberg, 2017)

Center relevant assortment	Center and local supply relevant	Not center relevant assortment
<ul style="list-style-type: none"> <li>• Opticians</li> <li>• Clothing / underwear</li> <li>• Bedding</li> <li>• Books</li> <li>• Computers (PC hardware and software)</li> <li>• Small electrical appliances</li> <li>• Photographic and optical products and accessories</li> <li>• Glass / porcelain / ceramics</li> <li>• Household/bed/table linen</li> <li>• Household contents</li> <li>• Home textiles / curtains</li> <li>• Haberdashery / Tailoring supplies / Handicrafts Yard goods for clothing and underwear</li> <li>• Medical and orthopedic equipment</li> <li>• Musical instruments and music supplies</li> <li>• Paper / office / stationery Artists' supplies</li> <li>• Perfumery</li> <li>• Shoes and leather goods</li> <li>• Toys</li> <li>• Sports equipment, sportswear / shoes and camping equipment</li> <li>• Telecommunications articles</li> <li>• Watches/ Jewellery</li> <li>• Consumer electronics incl. media</li> <li>• Weapons / hunting supplies / fishing</li> <li>• Home furnishings pictures / posters / picture frames / art objects</li> </ul>	<ul style="list-style-type: none"> <li>• Cut flowers</li> <li>• Drugstore, cosmetics</li> <li>• Food / luxury food</li> <li>• Pharmaceutical articles</li> <li>• Animal food</li> <li>• Newspapers / magazines</li> </ul>	<ul style="list-style-type: none"> <li>• Home improvement assortment in the narrower sense:               <ul style="list-style-type: none"> <li>◦ Bathroom- Sanitary equipment and accessories</li> <li>◦ Components. Building materials</li> <li>◦ Fittings. Hardware</li> <li>◦ Paints Lacquers</li> <li>◦ Tiles</li> <li>◦ Fireplaces (tiled stoves)</li> <li>◦ Installation material</li> <li>◦ Tools</li> </ul> </li> <li>• Boats and accessories</li> <li>• Large electrical appliances</li> <li>• Bicycles and accessories</li> <li>• Garden products</li> <li>• Lights / lamps</li> <li>• Musical instruments</li> <li>• Car accessories</li> <li>• Stroller</li> <li>• Furniture</li> <li>• Plants / seeds</li> <li>• Carpets, curtains floor coverings wallpapers</li> <li>• Zoological needs and live animals</li> </ul>