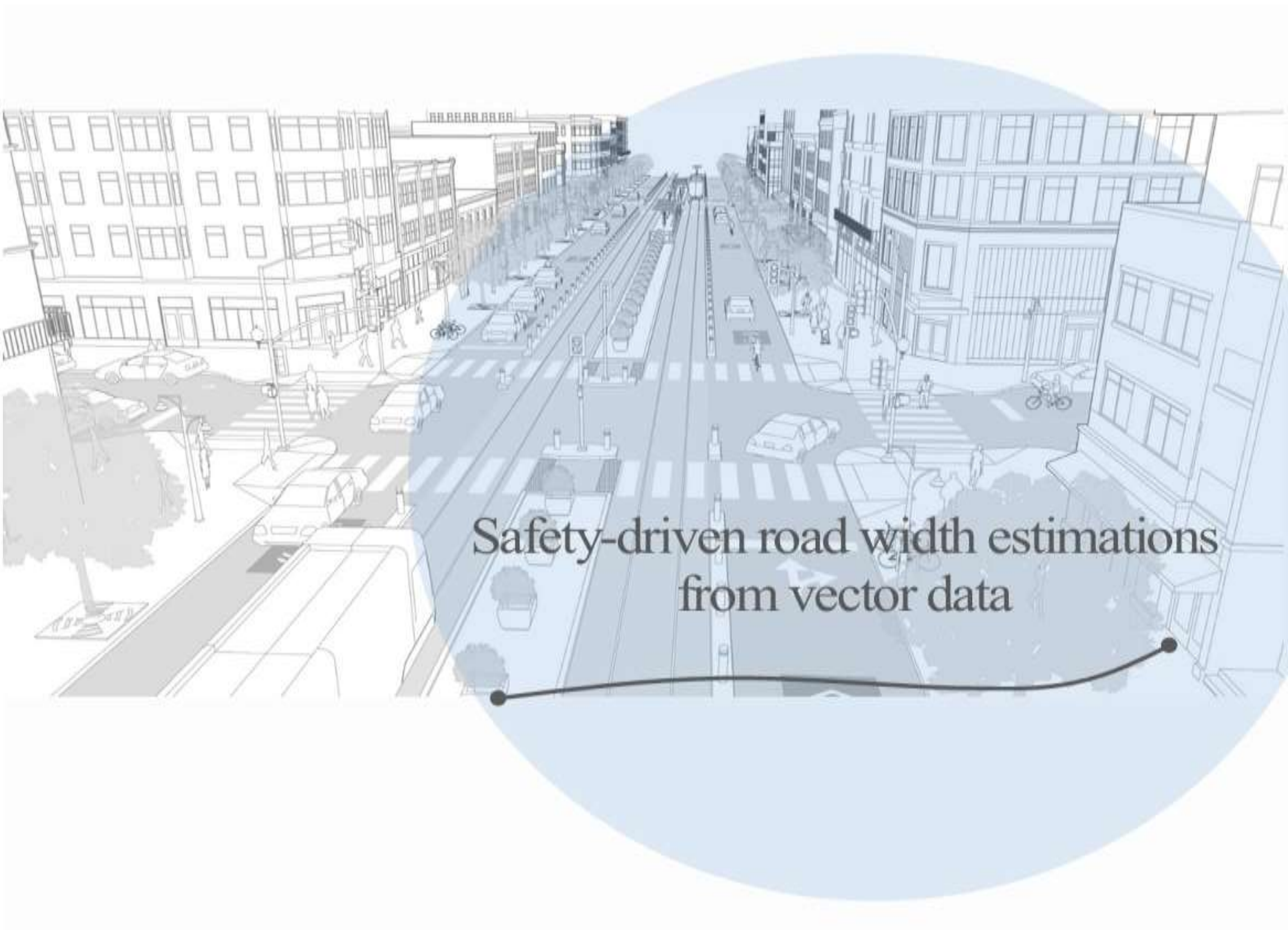
Delft University of Technology

Master of Science in Geomatics for the Built Environment

MSc Thesis by **Charalampos Chatzidiakos**

Supervisors: Anna Labetski, Dr. (Ken) Arroyo Ohori, Stelios Vitalis

Co-Reader: Dr. Ravi Peters



Safety-driven road width estimations from vector data

Delft, September 2021

Abstract

In recent decades, it is well known that transportation systems have a crucial role in the economic development and social prosperity of modern societies. While road width can be considered as one of the most important factors of the road environments, there is no clear definition of it. To better understand this, we need to reconsider the width of a road not as a single numerical value but as a complex concept that can be interpreted in different ways. Different road users may refer to different width values for the same road. Thus, while estimating road width seems to come with great benefits for different real-world cases, a fundamental rethinking of the purpose of this entire process is necessary.

Road safety management is an application whose overall process is strongly affected by the width of the road. While there are conflicting theories about the effect that road width can have on road safety, we may need to reconsider a few things before exploring their relation. In this thesis, we introduce a novel approach for estimating road width and linking it with roads in such a way that the overall process of road safety management application could benefit.

Moreover, even road width estimation is not a new topic most of the studies so far use LiDAR point cloud or satellite images as input. In our approach vector data coming from open-sources are used. Different inputs come with different limitations. The advantages, as well as the drawbacks of the different inputs, will be discussed.

Finally, one of the main objectives of this project is to develop a methodology that will be generally applicable. Vector data that can be found in different datasets would be used. Some additional deliverables will result from our efforts to overcome some of the challenges that have arisen in achieving this goal. A methodology that standardizes road vector data and a methodology that identifies the location and the type of the different intersections will be developed.

Acknowledgements

Before I go into more details, I would like to express my gratitude to my supervisors for their contributions to my Geomatics Graduation Project. The realization of this project would not have been possible without all these professionals to support and guide me during critical moments.

In particular, I would like to thank Anna Labetski for her help, guidance and patience throughout this research. She gave me important feedback and made me realize some aspects of my project that I couldn't see before. In addition, I would like to thank Dr. (Ken) Arroyo Ohori for his helpful advice. He helped me a lot with the directions he gave me for this project. Finally, I would like to thank Stelios Vitalis for his support. His technical advice and ideas helped me a lot in the implementation of this project.

Lastly, I would like to thank my friends and family who helped me during this long period. I worked remotely for most of this Thesis and daily contact with them was a fact. I would really like to thank them for their support and patience.

Contents

1 Introduction

In recent decades, it is well known that transportation systems have a crucial role in the economic development and social prosperity of modern societies. One of the main components of transportation systems is the road environment. In this thesis, I focus on that pillar of road networks and specifically, on one of the most crucial aspects of the road environment, the road width. Road width as an essential physical characteristic of roads affects many different applications of the modern world, that vary from road safety management, navigation systems, crime analysis, and many other processes. While estimating road width seems to come with great benefits for different real-world cases, a fundamental rethinking of the purpose of this entire process is necessary.

To better understand this, we need to reconsider the width of a road not as a simple numerical value but as a complex concept that can be interpreted in different ways. Different road users might have different needs regarding road width knowledge. For example, let's have a look at Figure 1. This is a road called Käskynhaltijantie, located in Helsinki (Finland). As it is obvious from the image, several width values can be linked to this road. Some notable changes in width occur along its geometry. Those changes are due to a change in the number of road lanes and the existence of a median strip at the end of the road. So here comes the question: *What would be a representative width value for this road?* Is it useful to combine all the individual measurements to obtain a single mean value, or could another approach benefit the road user more? This is what my thesis tries to answer. The motivation behind this research lies in the dependence of the road width on the application to be used. In order to know which width value is preferable to link to a road, we must first investigate the needs of the application for which the width will be used. For instance, if we need to provide a width value for large vehicle navigation purposes, then knowledge of the narrowest points of the road is essential. So as to ensure that the vehicles will fit on the road in every possible position. On the other hand, if we want to provide a width value to a car driver so that he has a broad understanding of road geometry, then, some statistical values of road width (mean, median, etc.) that result from the combination of various measurements may be sufficient. Thus, the reference to the width of the road as a single numerical value without associating it with the specific application can cause more confusion than benefits.

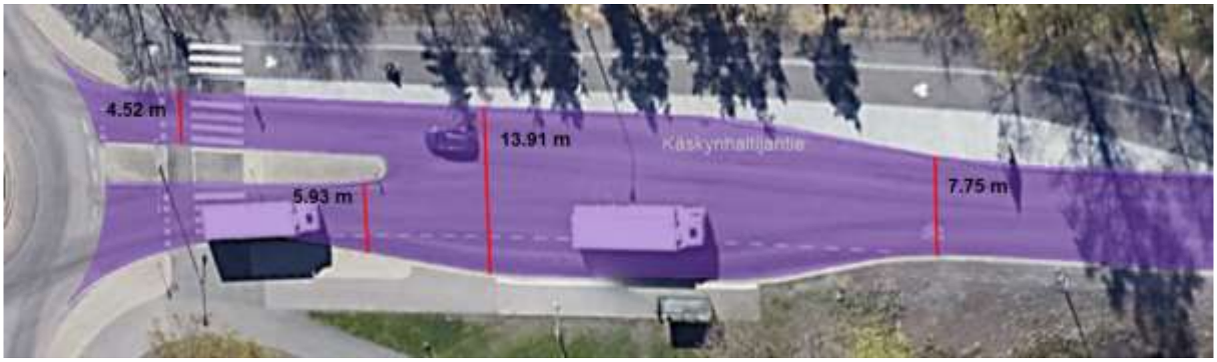


Figure 1: Road called Käskynhaltijantie located in Helsinki (Finland). Different width measurements are available for that road. The width interpretation of this road depends on the road user (application that needs width information). Source: Google earth V9.147.0.1

For the purposes of this Thesis, road safety management will be used to explore this application-dependent subsistence of road width. I will investigate the relation of road safety and road width. Based on my findings, I will develop an application-driven methodology to estimate and provide width in such a way that road safety could benefit.

Contradictory theories exist about the influence that road width can have on road safety. The conventional theory of roadway design is that wider, straighter, flatter, and more open is better from the standpoint of traffic safety [16]. On the other hand, other researchers ([15], [16], [38] et al.), found that wider roads are associated with statistically notable growth in total and fatal road accidents. While the correlation between these two phenomena seems to be a relatively straightforward process, we may need to reconsider a few things first.

To better understand this, let's have a look at Figure 2. This figure indicates a road showing several changes in its width. There are narrower and wider parts and different width values exist at different locations. Let’s now connect this case with road safety management application. Let us assume that 2 accidents occurred on that road at the points marked with x. By having a look at the image, we can easily derive the conclusion that 2 accidents occurred in the locations that the road becomes narrower. If we had associated this road only with a mean width value and based on that we were trying to correlate traffic accidents with road width, could we draw such a conclusion? The answer is much more complex than a simple no. Under some conditions, a mean value might be sufficient. For example, if we divide the original road based on these width changes and create some new smaller roads, then a mean value may be enough. In this thesis, various ways of dividing roads in a meaningful sense to benefit road safety application will be explored.

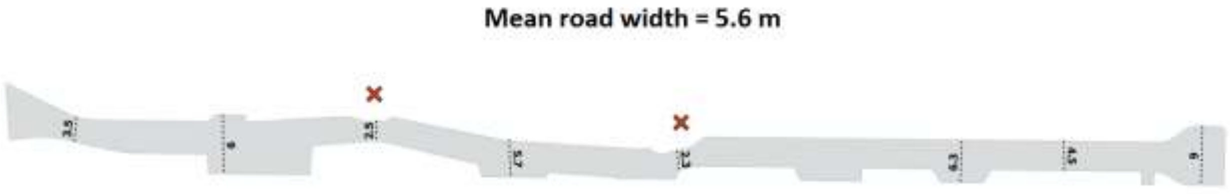


Figure 2: Road that shows some changes in its geometry. The mean width value for that road is 5.6m. If we simply use that value for road safety management purposes, we might trick the user that there is no correlation between width changes and accidents location. In practice there are some narrow spots that can be problematic. Thus, more investigation of what are the needs of real-world application in regards to road width is required.

Although road width estimation is not a new topic, to date there is no open-source GIS database, such as OpenStreetMap, that consists of road width information [48]. Moreover, most studies focus on using sensing images or LiDAR point clouds as input for width calculation. Different inputs can lead to different limitations. For example, access to high-resolution satellite imagery may be quite limited, or the complexity and size of LiDAR point clouds may require special analytical skills from the user.

This thesis introduces an approach that estimates road width by making use of vector road data from open sources such as OpenStreetMap. Vector data, can overcome some of the drawbacks that are related with other sources as explained before. Topographic maps all around the world use vector data to store information about topology of the roads. Moreover, online services that provide freely such information are available. Finally, road vector data usually have a rather simple structure compared to other inputs.

Hoffmans W. [25], created a script that calculates the width of the road sections using vector data. His methodology was developed to support the specific application of snow removal from roads in the Netherlands. Thus, it focuses on the automatic calculation of the width of the road sections as they are included in the Basic Large-Scale Topography of Netherlands (Basis Grootschalige-Topografie (BGT). Although Hoffman's work had great practical meaning for the particular application that was created, I explored the limitations of whether it is used for road safety management application. Moreover, I will expand the functionality of this methodology to be used with vector data from different sources.

One of the main goals of this research is to implement a generic methodology that can be used for different datasets. The main challenge is that there is not a unique way of modelling roads with vector data. Different datasets indicate different modelling strategies. Figure 3 illustrates a typical example of different modelling approaches of an intersection.

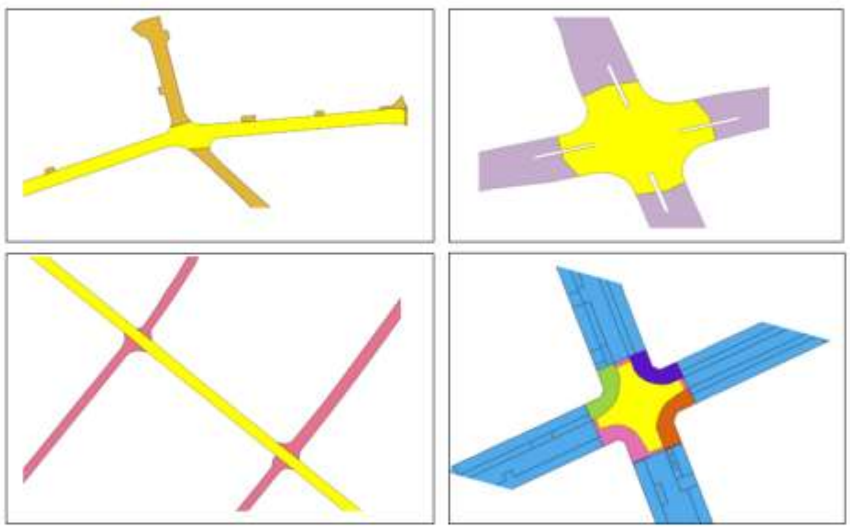


Figure 3: Cross intersection modelled in different datasets. Main road is prioritized by taking most area of intersection in the dataset of Helsinki (top left). Explicit intersection polygon exist to represent the intersection in the dataset of Toronto (top right). Sub-road are prioritized in the dataset of Poznan (bottom left), Five polygons are used to represent a cross intersection in the dataset of (Den Haag).

Developing an approach that standardizes the way that roads are modelled with vector data will be an extra deliverable of this thesis and it and will serve as an auxiliary feature to achieve one of my main goals. Moreover, I will investigate different modelling approaches and I will choose one that could be used as prototype for my standardization process. As an extra step, I will exploit the particularities of the selected modelling strategy, to identify the different intersection types. At later chapter, I will explore the complexity of intersections and define the reasons why they should be considered as special parts of the road network and why they should be treated differently.

1.1 Research Questions

The main question I want to answer in this thesis is: ***"How road width estimations can be derived from vector data to benefit road safety management application?"***

With this research I will expand the functionality of the existing methodology that works with road vector data. The main goals are: 1) Estimate and provide road width in such a way that can be used in practice for road safety management purposes and 2) Develop a methodology that works with road vector data from different sources.

Although road width can be considered as one of the most crucial aspects of the road environment, not enough emphasis has been placed on the different needs that different applications might have. I will implement a methodology that is driven by the needs of road safety management application. I will explore how width estimations could be used to benefit the overall process of this application. Moreover, the existing approach of Hoffmans W. [25] is developed to be used with a specific vector dataset as input (Basic Large-Scale Topography of Netherlands (Basis Grootschalige Topografie (BGT)). My goal is to extend the functionality of this approach so that it can be used with data from different sources. Working with data from different sources comes with the limitation of having different modelling approaches for road vector data. An approach that standardizes road vector data will be developed to allow the implementation of a generic methodology.

Therefore, some extra features will be added to the final output of this thesis. The standardization of road modelling with vector data and the identification of different types of intersections will be the main extra features that I will deliver with this research.

In order to answer the main research question multiple sub-questions need to be answered as well. The following sub-questions will be the building blocks of this thesis.

1.1.1 Road safety and road width

One of the main goals of this research is to implement a methodology that could benefit road safety management. In order to achieve this, the importance of road width for this application should be explored. The influence that road width can have on the safety of the various road users, should be investigated. My findings, will be used to guide my decisions at some crucial stages of my methodology and will be used to evaluate my results. In respect to the above the first sub-question is:

“How road width can affect the safety of different road users?“

1.1.2 Generic Methodology

The second sub-question is:

"How can road vector data be standardized in such a way as to benefit the development of a generic methodology for estimating road width?"

One of my main goals is to expand the functionality of the existing methodology in order to be used with different vector datasets. Datasets coming from different sources show crucial differences in road modelling. This, makes it hard to establish a generic methodology that can use input from different datasets. I will explore some different modelling strategies of road data. Therefore, I will implement a methodology that standardizes road modelling with vector data based on a specific existing modelling approach. The modelling approach that will be used as prototype, will be selected in such a way that the overall process of width estimation will benefit.

1.1.3 Units of measure

Throughout this research I explored different ways that road vector data could be combined in order to benefit my use case. Different ways to represent a road with vector data exist. Moreover, a road can be divided in various ways. For example, in the introduction a road faces some significant width changes was presented (Figure 1). While someone may think of this polygon as a single road, other possible approaches exist. The road could be divided based on those width changes into new smaller road parts with more representative mean width values. Other possible approaches were explored during this research. In § 2.2 the concept of units of measure is elaborated in more detail. Therefore, the third sub-question is:

"In what way original roads could be divided to benefit road safety management application?"

1.1.4 Impact of width estimation methodology to safety analysis

To result to a generic and robust safety-driven methodology for estimating road width, the analysis were conducted in various steps. Different features were implemented and tested with different datasets. I explored the relation of safety and width, I chose a way to standardize road vector data, I explored how to divide roads to address the needs of my use case etc. Finally, I result to a methodology that is expected to benefit the process of my selected use case, road safety management. My final sub-question is:

"How do the different aspects of the final width estimation methodology affect the process and result of a road safety analysis?"

1.2 Thesis Outline

The content of this research extends in 7 chapters. They structured as it follows:

Chapter 2 → introduces the reader to the relevant theoretical background of this thesis. Definitions and concepts that are used during this project are explained. Vector data as input data, modelling approaches of road vector data, intersections, ways of dividing and representing roads, clustering, and other notions are discussed in detail.

Chapter 3 → Scientific research related to this graduation project. Related work regarding road width estimations, intersections identification and road safety analysis are presented.

Chapter 4 → Analysis of road safety management application. The selected use case which will drive the methodology of this thesis is discussed in detail. The importance of road width and its influence on the safety of various road user groups is explored. Then, the width clustering approach as explained in the previous chapter is associated with road safety. The positive impacts that this specific unit of measure approach is expected to have on road safety are presented.

Chapter 5 → Methodology. The initial methodology that we build upon is discussed first. Then, the steps followed for standardization of road vector data based on Toronto modelling approach are explained. Next, the clustering approach that is developed in order to serve road safety application is described. Finally, a clustering validation approach that was implemented to evaluate the results of clustering is presented.

Chapter 6 → Results and Analysis. I present the results for the different features of this methodology. Finally, I analyze the results of each separate step by interpreting some the quantitative and qualitative results.

Chapter 7→ Road safety analysis. The purpose of this chapter is to explore whether certain features implemented for this thesis, could affect the process and the result of a road safety analysis.

Chapter 8 → Contains the summary of the most important conclusions and future work. My main contributions are summarized and I answer the main research question and the sub questions.

2 Theoretical Background

This chapter provides an overview of the relevant theoretical background related to the content of the following sections. In § 2.1 details about road vector data are discussed. Following is § 2.2. There, it is explained that different interpretations of roads with vector data exist. § 2.3 is about road intersections. In this thesis, road intersections are considered as special parts of road networks and they are treated differently in regards width estimation. Finally, § 2.4 elaborates the main idea of clustering which will be used later in this research.

2.1 Road vector data

As already mentioned for this research road vector data from open sources is used. Vector data is split into three types: point, line, and polygon data. Point data is most commonly used to represent nonadjacent objects and discrete data points, while the other 2 types of representations are commonly used to represent roads. In this chapter, the limitations, as well as the advantages of vector data as input data, will be discussed. In addition, areal and linear road representation types that are used in this thesis will be analyzed. Finally, the main challenge that arises when we process road data from different sources will be explored in detail.

2.1.1 Vector data as input data

The method that is developed in this thesis, uses different input for width estimation in regards to most of the studies so far. Most of the existing methodologies are using LiDAR point clouds or high resolution satellite imagery as input data. While data from sensing technologies (e.g. LiDAR pc, satellite imagery) come with some significant advantages, they are associated with some limitations that make them difficult to use as well. As it comes from the literature, sensing data can be used to estimate road width with a rather good accuracy. Many studies had used such input and result to a pretty good accuracy in regards width estimation (see § 3.1). In addition, they come with other advantages, such as low human dependency and the ability to integrate them with other data sources for complex data analysis. However, there are some drawbacks to these inputs.

First, sometimes, those kind of inputs require skilled data analysis techniques to be used. For example, for LiDAR point clouds, because of the large datasets and the complexity of the data being collected, a familiarity with such kind of input needed in order to analyze data. A second limitation is related to the availability-accessibility of those inputs. While most of the studies that work with aerial imagery are using high-resolution satellite images (30cm – 5m/pixel), there is quite limited free public access to such data. Most freely available information correspond to medium-resolution imagery (10 – 30m/pixel) [10].

Vector data coming from open sources will be used to overcome some of those limitations. Nowadays, topographic maps all around the world usually use vector data to store information about topology of the roads. While vector areal representations of roads are rarer than linear representations, there is some information available for free. There are many online services (geo-portals) that provide road polygons and centerlines for free. Some examples of such portals that also going to be used for this theses are: [24] [40] [44] [46]. Moreover, road vector data usually has a simple structure. They can be interpreted and used for data analysis even by users unfamiliar with this type of input. The methodology to be developed for this thesis will not require the user to be familiar with the scientific field of Geomatics. In addition, the data to be used is the less expensive and most easily accessible data. This is why open source vector data is chosen.

Important to consider is that vector data is a processed product that is created either based on data coming from sensing technologies (e.g. from aerial imagery or lidar) or by measurements in the field. Thus, it depends on the accuracy and the availability of them. For example, if high resolution satellites not be able to capture images over a work zone area, then vector data may also missing for that area. Moreover, working with open vector data is related to other limitations as well. The ease with which open data can be accessed by the public, indicates the risk that the data may be falsified (on purpose or not). In addition, vector data is created by humans and they does not correspond to 'raw' data collected in the field. Thus, the human dependency of such input is pretty high. Some particular essential challenges could arise from that human dependency of our input. In later section (§ 2.1.3), different modelling strategies of roads with vector data will be elaborated. Later in this thesis, the problems related with this inconsistent way of road modelling will be presented.

2.1.2 Areal and Linear representations

Road polygons

While a linear representation typically is sufficient for applications like navigation, traffic simulations or noise mapping, often an areal street model is needed in order to represent geometric details [5]. Overall, the areal representation of a road network can be viewed as a set of closed polygons. Each polygon represents a part of a road network. These polygons can represent road segments, road intersections or other parts of road network such as bike lanes, intersections and roundabouts. In general, there is no homogeneity in areal representations of road networks and most of the times polygons can vary in shape and size in the same dataset. In general, polygon shape can be described using some basic measures such as their area, eccentricity, Euler number, convexity, compactness, and aspect ratio [54]. Figure 4 illustrates an example of 3 different road polygons that can be found in the dataset of Toronto. Image A shows a long curving polygon, image B shows a short straight polygon and image C shows a pretty small polygon.

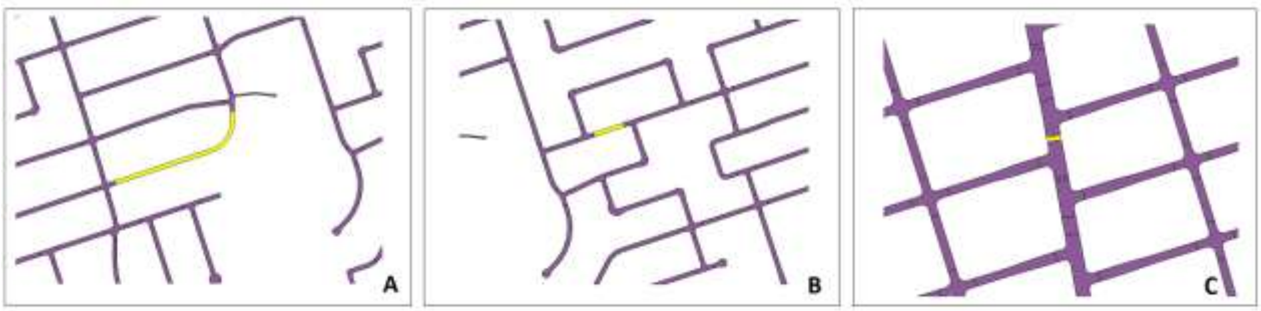


Figure 4: Examples of different road polygons, A) long-curving, B) short straight, C) Small polygon

Road centerlines

Linear road features, only represent the physical location of the street. As already explained, for the purposes of this research by referring to linear road representation we referring to road centerlines. While other linear representations of roads exist (i.e., lines that correspond to the lanes of the road) in this research we are focusing on road centerlines. Thus, when we refer to linear road representation we refer to road centerlines. Road centerlines are lines that represent the geographic center of roads on transportation networks and they are utilized in many different applications [4].Similarly to road polygons, road centerlines can vary in size and shape. Figure 5 shows 3 different examples of road centerlines that can be found in the dataset of Toronto.

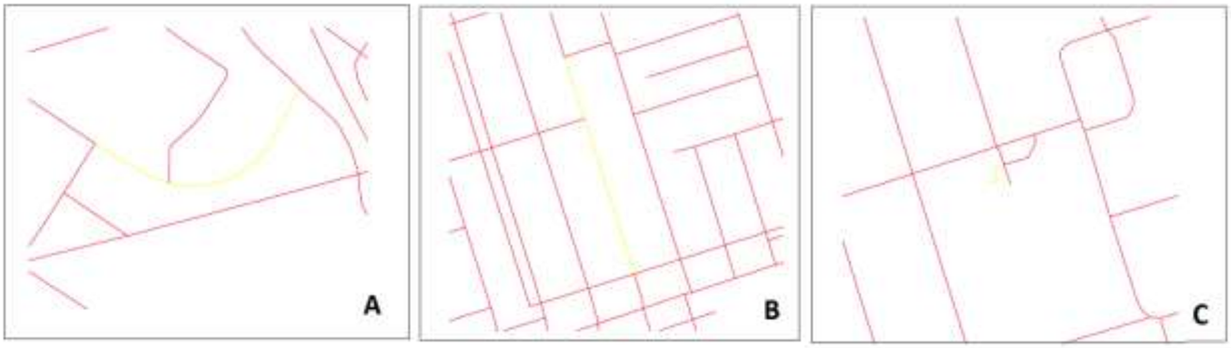


Figure 5: Different road centerlines, A) Curving, B) Straight, C) Small

Relation of road polygon and centerline in a dataset

The first thing that relates those 2 features in a dataset is associated with the relative position of them. In most cases, the road centerline is passing through the road polygon. Since the centerline represents the geographic center of the road and the road polygon represents the road geometry, it is reasonable that the centerline of a road passes through the center of the road polygon that corresponds to the same road (Figure 6, A). Although that is the most common case, it is possible to find a centerline that does not pass through the road polygon that represents the same road (Figure 6, B). This could mean that the centerline does not pass at all through the polygon or that it crosses the polygon at a certain location but not at all of its length. This might be due to noise in the data or due to the different sources that the data are coming from or for other reasons. Also possible but less common case is to find a centerline that is located inside the polygon but not in the center of it (Figure 6, C).

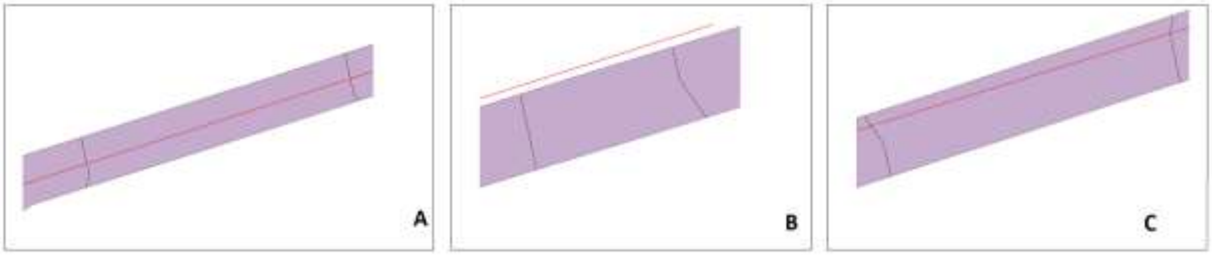


Figure 6: Different possibilities regarding the relative position of road centerline and road polygon, A) Centerline passes through the middle of the polygon, B) Centerline do not pass through polygon at all, C) Centerline passes through polygon but not from center

Another relation of those 2 features is about their length. In most dataset there is not 1-1 mapping between road polygons and road centerlines. Most of the times either a centerline crosses more than one polygon or a big polygon contains more than one centerlines. Figure 7 presents such an example. A road centerline that crosses 3 different road polygons in Figure 7, A. A big polygon that contains 5 road centerlines (Figure 7, B).

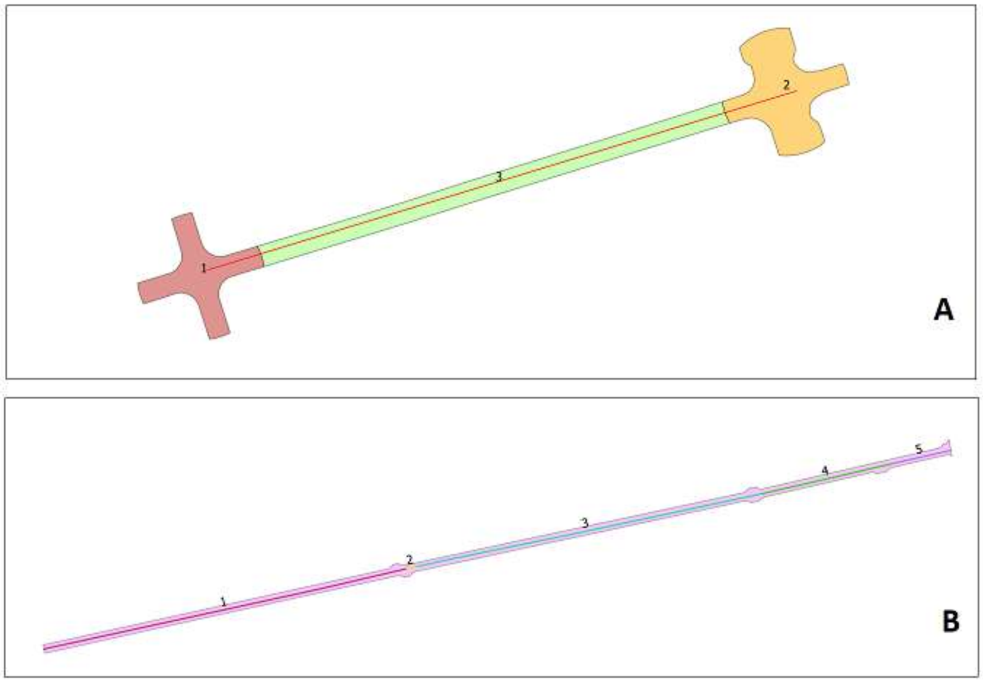


Figure 7: Relation between the size of road centerline and road polygon, A) Big centerline that crosses 3 different polygons, B) Big polygon that contains 5 centerlines, both centerlines and polygons are coming from the same sources

2.1.3 Modelling approaches

As already explain, human dependency of vector data is pretty high. Different people make decisions on how things should be structured. The main limitation of that is that different datasets coming with different modelling strategies. While one of the main goals of this thesis is to implement a generic methodology, not having a unique way to model roads makes this goal harder.

While differences can be found in linear representations, most of the times the same strategy is followed. The road centerline extends from one intersection to another. Figure 8 shows an example of 4 different centerlines in 4 different datasets.

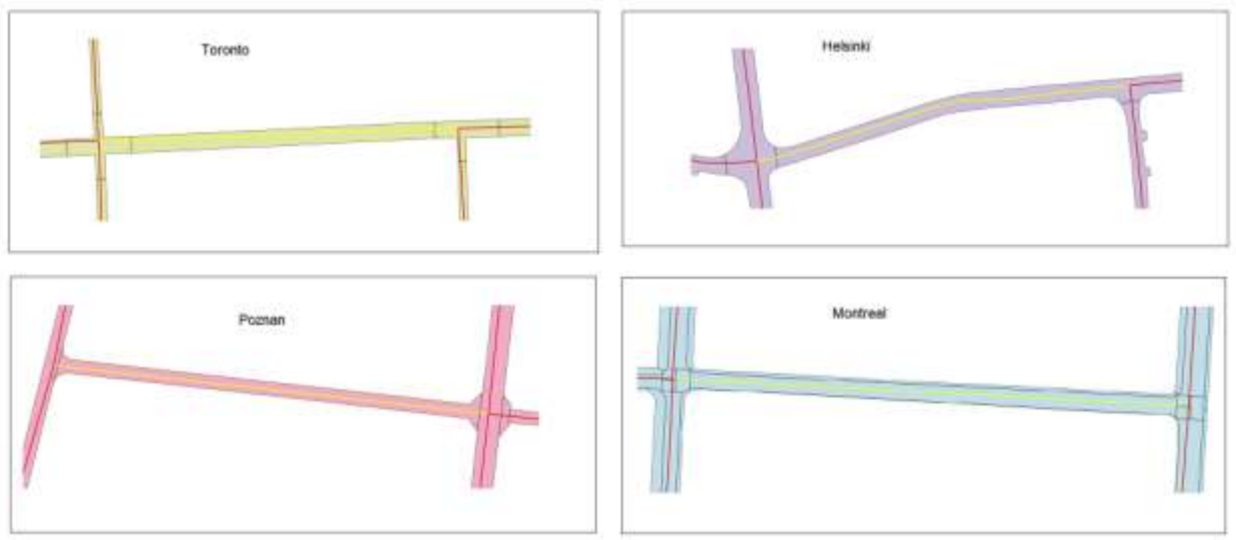


Figure 8: Centerlines usually, follow the same modelling strategy and they extend from one intersection to another. In this figure, centerlines into 4 different datasets are shown

Although road centerlines seem to follow the same modelling strategy, major differences can be found in areal representations of roads. The main differences are lying on how each modelling approach structures the places where 3 or more roads are met (intersections). Different datasets are following different strategies regarding intersection modelling.

Figure 9 illustrates a typical example of different modelling approach of a cross intersection that can be found in 4 different road vector datasets. Top left image shows a cross intersection in a polygon vector dataset of Helsinki. There is no unique polygon for representing the intersection and the main road is prioritized (road in yellow). The opposite strategy is followed in the road dataset of Poznan (bottom left) where the most of the space of road intersection is given to the 2 sub roads. In case of Toronto (Top right) the intersection is modelled by a unique polygon. Thus, the space is shared equally between the different road branches. Finally, the most complex representation can be considered the one that is shown in the bottom right image. This modelling approach of cross type intersection can be found in road vector dataset of Den Haag. Similarly to Toronto dataset, there is a polygon that corresponds to the cross intersection. In addition, to that polygon there are 4 more polygons that represent the different connections between road branches that intersect. Other approaches of modelling intersection can be found in other datasets as well.

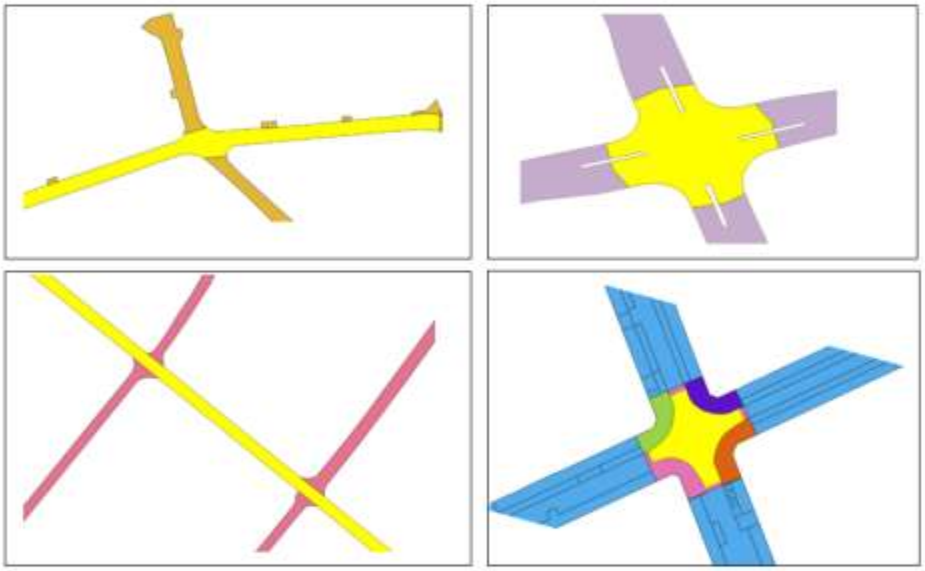


Figure 9: Different modelling strategies of a cross intersection

Apart intersection modelling other differences can be found in different datasets. In some cases other features of road networks such as parking spots are modelled explicitly (Figure 10 left, shows an example of parking spots modelled in dataset of Helsinki). Moreover, bike lanes or sidewalks could be modelled as well (Figure 10 right, sidewalks modelled in dataset of Montreal).

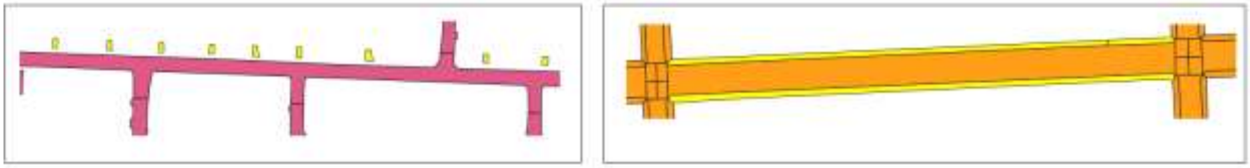


Figure 10: Parking spots modelled explicitly in dataset of Helsinki (left), sidewalks are modelled explicitly in dataset of Montreal (right)

2.1.3.4 Toronto modelling

In this section, the modelling strategy that can be found in Toronto dataset is further analyzed. This approach will be used as a prototype approach for our standardization methodology. The main reason for choosing this approach as prototype is that intersections are modelled explicitly. In § 2.3 the complexity of intersections is further examined. In general, intersections can be seen as one of the most complex parts of road networks. Thus, different treatment needed regarding width estimation in these parts as well. Having a unique road polygon at every intersection allows them to be handled better.

Toronto dataset follows explicit modelling of intersections with one polygon to represent each intersection. Different types of intersection polygons can be found in the dataset, Figure 11 displays the 6 main intersection types that can be found in the dataset of Toronto. As it is apparent from the figure, a unique and simple polygon is used to represent the different intersection types. While other approaches which follow explicit intersection modelling exist, Toronto follows a more simple and consistent way of intersection modelling.

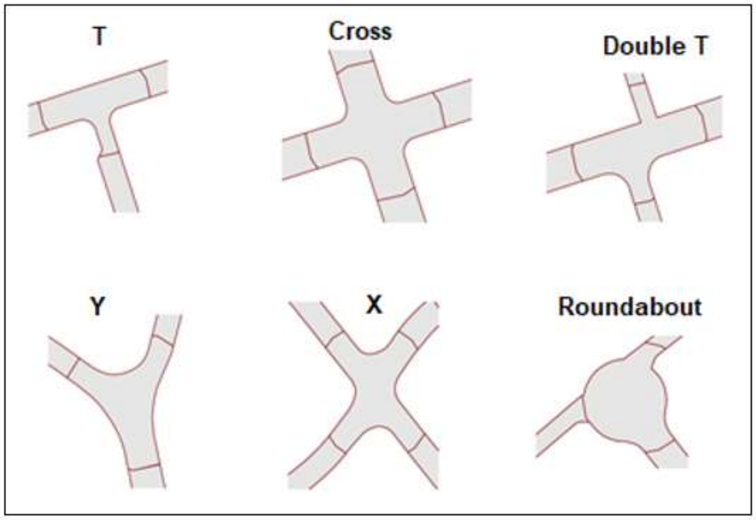


Figure 11 Six main interseciton polyons that can be found in Toronto dataset. A unique and simple polygon is used to represent each intersection

There is no explicit modelling of other features of the road network such as median strips or parking spaces. In general, the shape of the road polygons is very regular. Figure 12, illustrates the road polygons of a sample area of the city of Toronto.

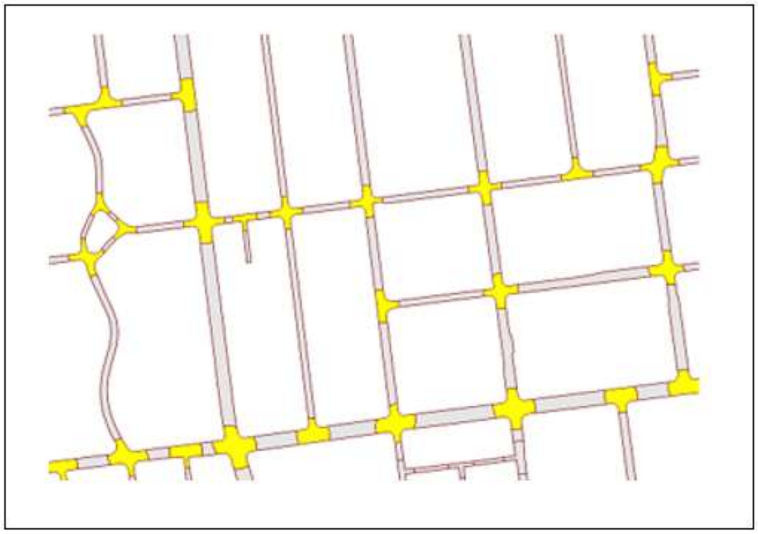


Figure 12 Sample area of Toronto. Explicit modelling of intersections is followed by this dataset

2.2 Unit of measure

The purpose of this section is to introduce the concept of units of measure. Units of measure can be translated like this: *What do we consider as a road with vector data?.* Unit of measure depends on 2 crucial aspects of vector data. First, as discussed in § 2.1, different representation types of roads with vector data exist. So, for example, with what representation type of a road should we link the width estimations?, Road polygon or road centerline? Moreover, road polygons and centerlines have an original structure. But there are many ways to divide a road. For example, should we estimate the width for each initial centerline as can be found in the original data set, or should we divide the initial centerlines into parts based on a specific attribute and estimate the width for the various parts?

Unit of measure is an essential concept of my research, It affects the input data, it's part of the standardization step, and ultimately it affects the computation and the storage of width estimations. Finally, since we are implementing a methodology that is driven by the needs of a specific application, choosing a specific unit of measure approach that fits the needs of the selected use case can have a positive impact on the overall methodology. In this section, some different approaches of units of measure are presented.

.

2.2.1 Approach 1 Initial road polygons (polygons-lines pairs)

Road polygons can differ in shape and size inside an areal road vector dataset. As already explained in § 2.1, although linear representation is typically is sufficient for some applications, often an areal representation is needed in order to represent geometric details. Thus, linking width estimations with road polygons might benefit some uses cases that make use of such representations. In order to estimate width per road polygon we need 1 to 1 mapping between road polygons and road centerlines. In most of the cases this mapping is not present in the dataset. The centerline-polygon pairs approach, is relying on **creating pairs of road polygons and road centerline parts**. Those pairs will be created after the segmentation the original road centerlines, as they downloaded from an online open-source GIS database (such as OSM). The main requirement of each pair is that each newly created centerline part will fit exactly one and only road polygon. Figure 13 illustrates that example. At the top of the image, the initial road centerline is shown with yellow color. It passes through 4 different polygons (1,2,3,4). Then at the bottom of Figure 13, the 4 different polygon-centerline pairs are shown (A,B,C,D). The initial centerline is cut to 4 parts based on the geometry of the polygons that it passes through.

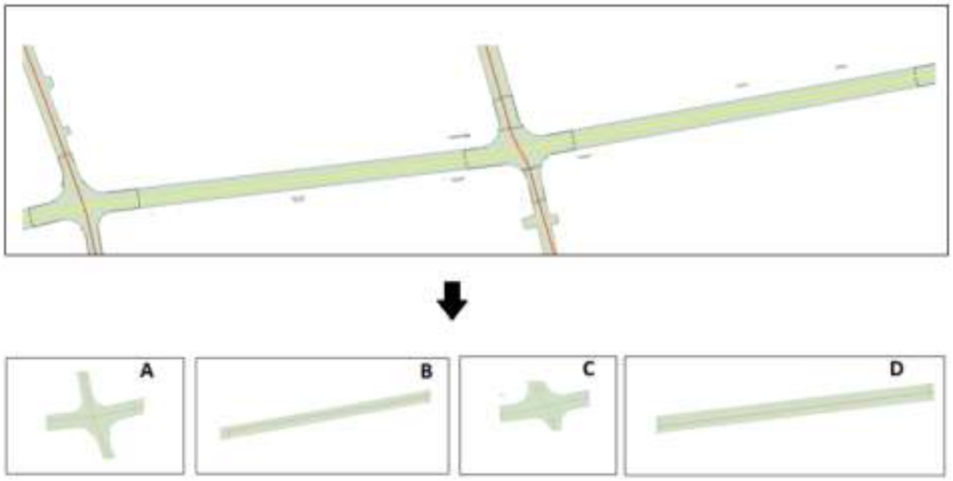


Figure 13: Centerline-Polygon pairs approach. The initial centerline (top), the polygon-centerline pairs (bottom)

With this approach it is possible to compute width statistics for the areal representations. Width will be estimated for each road polygon based on its geometry, and it will be linked with it.

2.2.2 Approach 2 Initial road centerlines

This approach is a bit different from the previous one. The width is still estimated using the polygon geometry but no pairs are created. The original road centerlines are not divided into parts, thus we do not need 1-1 mapping between centerlines and polygons. All the information will be stored to the initial road centerlines.

Figure 14 illustrates an example of this approach for width estimation. The centerline it passes through 3 different polygons. No pairs are created and the geometry of the 3 polygons is used to estimate road width at different parts of the centerline. After the computation of the measuring lines, the final width will be assigned only to the original centerline.

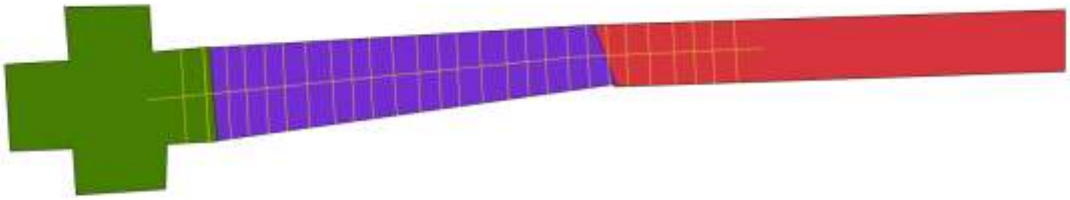


Figure 14: Example of approach 2, the initial centerline is not split to parts but is used as it is to directly estimate width based on areal representation. The final width estimation is assigned to the initial centerline only.

2.2.3 Approach 3 Width Clustering

With this approach we are not taking into account the initial structure of road centerlines or polygons as they can be found in the original dataset. The main purpose of this approach is to re-create the initial centerlines based on clusters of measuring lines with similar width. By using a user-defined threshold we will cluster successive measuring lines with similar width. All the width information will be computed and stored to the newly created centerlines.

Figure 15 illustrates an example of the clustering approach. Image A shows the a road polygon that is used for width calculation, the original centerline (red) and the measuring lines (black doted). After clustering successive measuring lines based on their length, we can conclude to 3 different clusters with different mean road width values, as shown in image B.

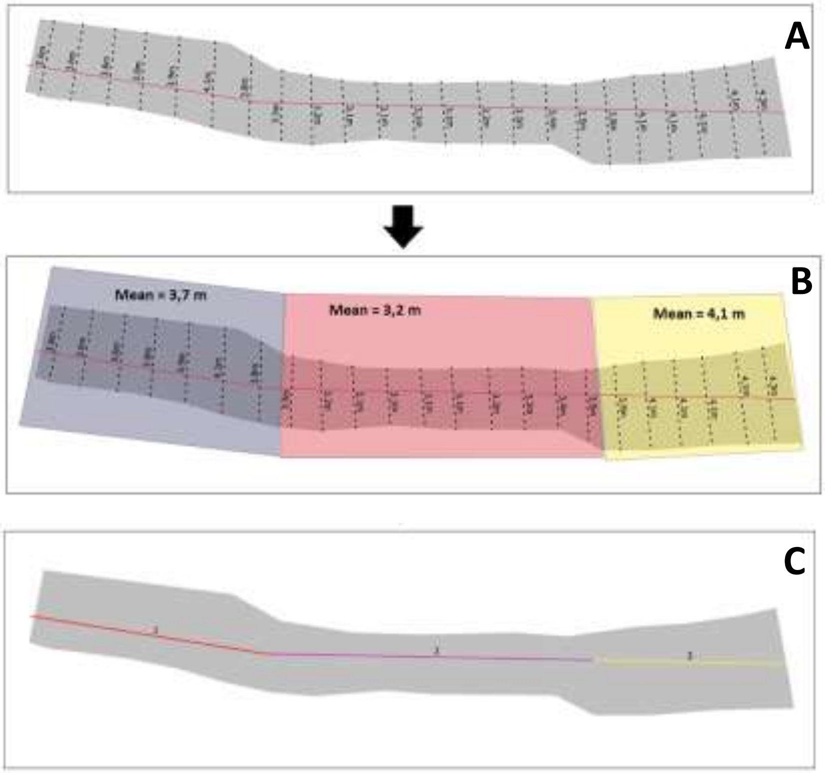


Figure 15: Overall pipeline of clustering appraoch. After computing width at certain locations of a centerline we create clusters based on width similarities. Finally, new centerlines are created based on clusters.

Using those clusters, we can divide the initial centerline to 3 new centerlines and assign to them the 3 different widths, as shown in image C. With this approach width estimation will be stored to the newly created centerlines

2.2.4 Approach 4 Equal intervals

Same as the previous approach, we do not take into account the original structure of road vector data. For this approach the original centerlines will be divided into equal parts based on length intervals. All the information will be linked to the newly-created centerlines.

Figure 16 illustrates an example of how width will be estimated with this approach. The initial centerline is sub-divided to 3 equal parts of 100m length. Then for each part the width will be calculated separately and it will be assigned to the linear representation (the new centerline that created after the division of the original one).

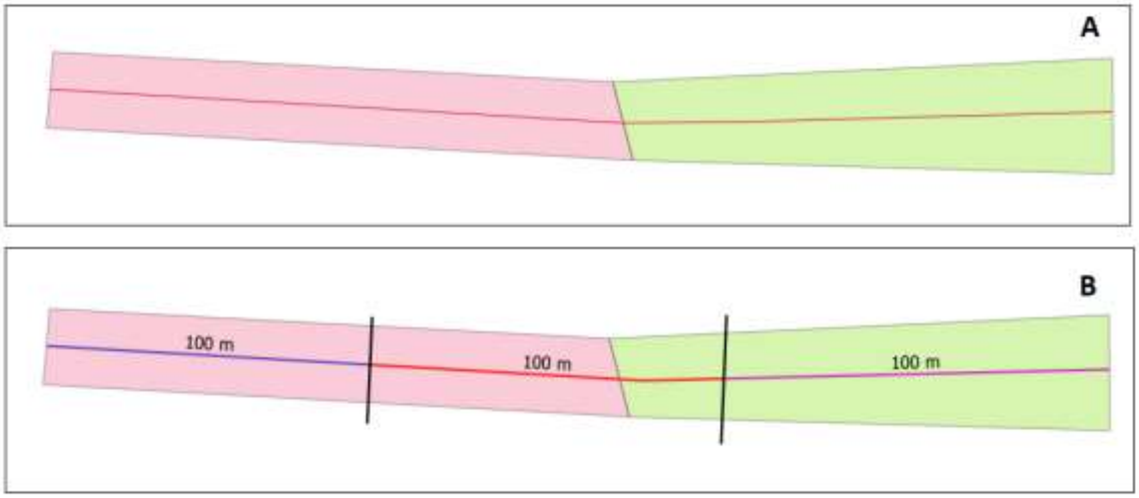


Figure 16: Example of approach 4, In this example, the initial centerline (top) has length of 300m at the bottom of the image we see that is subdivided into 3 new centerlines of 100m each. Finally, the width will be estimated for each one of those newly created centerlines

2.2.5 Summary

The characteristics of the 4 units of measure approaches are summarized in Table 1.

Approach	Use original road centerlines/polygons	Create new road centerlines/polygons based on attribute	Representation types to be linked with width estimation
1	✓	✗	Areal
2	✓	✗	Linear
3	✗	✓ (width)	Linear
4	✗	✓ (length)	Linear

Table 1: Summary of characteristics of 4 units of measure approaches

2.3 Intersections

The road intersections are defined as the areas obtained by the convergence in the same points of three or more road branches [45]. As stated by Quartieri et al. [45], intersections, including roundabouts, are seen as the most complex parts of road networks, because they can have many different configurations. In addition, they may have a rather complex structure and an unusual shape. Due to this peculiarity, intersections can be quite challenging regarding road width estimation. The main reason lies on the fact that intersections contain area that is shared between different roads. Thus, uncertainty arises as to how the area should be divided between the intersecting road branches. Figure 17 illustrates such an example. As aforementioned, different datasets are following different approaches to distribute intersection areas between road branches (see § 2.3.1). In a later chapter, we will explore in more depth how intersections can cause noise to the overall process of width estimation with vector data.

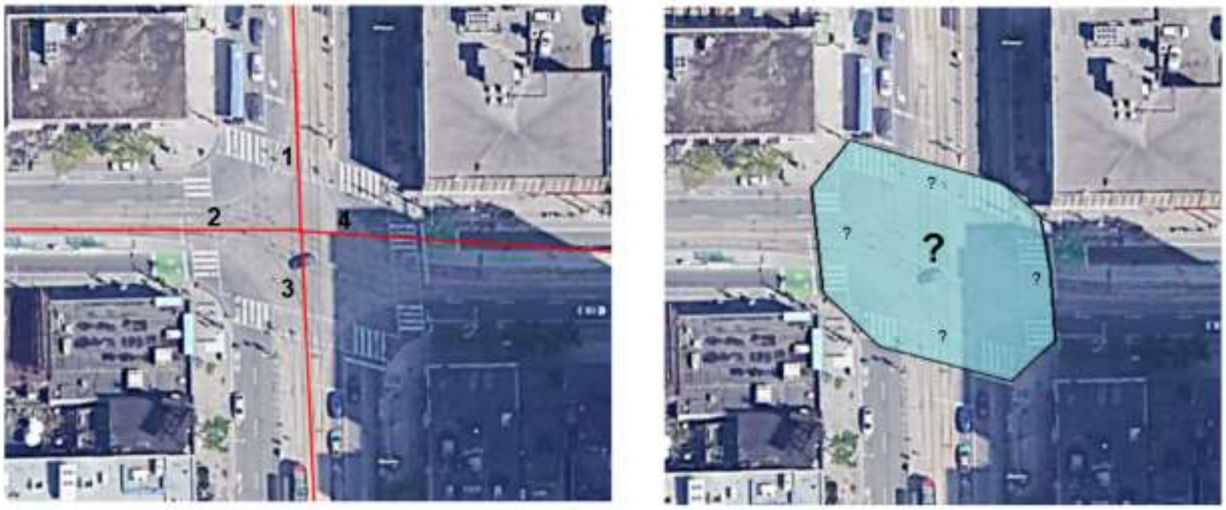


Figure 17 Cross intersection where 4 road branches met (left). Uncertainty arises as to how the area should be divided between the intersecting roads Source: Google earth V9.147.0.1

Apart their complexity there is another reason that led us to handle intersections differently. The main purpose of this research is to develop a methodology that benefits road safety management application. Intersections have been identified as crash "hot spots" for dangerous driving leading to crashes [56]. Moreover, according to Choi [8], intersections are particularly hazardous due to activities such as turning across traffic. Thus, identifying intersection types will be an extra deliverable of this thesis.

There are different ways to classify intersections. For example, Quartieri et al. [45] in their research for traffic noise impact of road intersections, divide them by how traffic flow is controlled. For the purposes of this Thesis, we are more interested in the topologically differences between them. Thus, we need to classify them by the amount of roads that connect and in what configuration. The main intersection types for this research are defined based on Toronto intersection modelling approach. Figure 11 illustrates the 6 main intersection types, while all the other intersection types that can be found among the different datasets will be considered as "another type" of intersection.

2.4 Clustering

In § 2.2 we refer to an approach called width clustering. Since, clustering will be used for the purposes of this research we need to elaborate more on some basic characteristics of it. Clustering is the process of grouping similar objects into different groups, or more precisely, the partitioning of a data set into subsets. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [33] .In this thesis, clustering techniques will be used in order to provide width estimations in such a way that road safety management could benefit. Thus, the main clustering algorithms, their advantages and disadvantages and finally, the different ways of evaluating clustering results will be explained in this section.

The main distinction that can be made to data clustering algorithms is between hierarchical and partitional algorithms. **Hierarchical algorithms** find successive clusters using previously established clusters, whereas **partitional algorithms** determine all clusters at time [33].

2.4.1 Hierarchical clustering

Hierarchical algorithms can be **agglomerative (bottom-up)** or **divisive (top-down).** Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters [33]. The main output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters [58]. So, just by looking at the Dendorgram you can tell how the cluster is formed [36]. Figure 18 illustrates an example of a dendrogram.

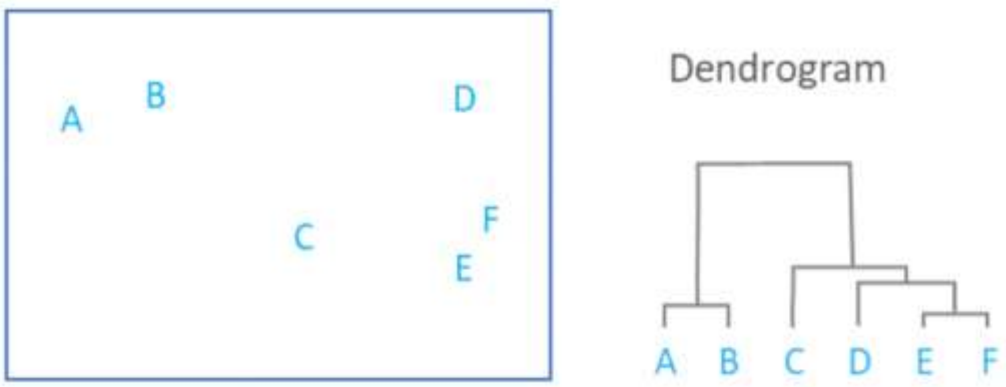


Figure 18: Six points in 2D space (left), their dendrogram formed after they merged to clusters hierarchically (right). The Distance between data points represents dissimilarities while Height of the blocks represents the distance between clusters. Source: Tim Bock (2018)

2.4.1.1 Agglomerative Hierarchical Clustering

It's a “bottom-up” method where every observation starts as a unique cluster and then pairs of clusters are merged hierarchically. The first step is to make N clusters for N observations. Then, we take the data that are closest and we merge them in one cluster (N-1 clusters exist now). Then, we repeat that process until we have left only one cluster.

As mentioned above we merge the observations that are closest. But how do we measure distance? There are many different distance metrics that can be used for hierarchical clustering such as Euclidean distance, Manhattan distance and others. The choice of distance metric should be made based on theoretical concerns from the domain of study.

After selecting a distance metric, for this clustering approach we need to decide the rules for clustering (from where distance is computed). These rules are often called linkage methods. Many linkage methods have been developed. Nagesh Singh Chauhan [36] explain some of the most popular:

- Complete-linkage: the distance between two clusters is defined as the longest distance between two points in each cluster
- Single-linkage: the distance between two clusters is defined as the shortest distance between two points in each cluster
- Average-linkage: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.
- Ward-linkage: This method reducing the sum of squared distances of each observation from the average observation in a cluster. This concept of distance matches the standard assumptions of how to compute differences between groups in statistics [58].

As with distance metrics, the choice of linkage criteria should be made based on theoretical considerations from the domain of application [58].

2.4.1.2 Divisive Hierarchical Clustering

It's a "top-down" clustering method where we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. So this clustering approach is exactly opposite to Agglomerative clustering [36].

2.4.1.3 Advantages and Drawbacks

Leung et al. [31] point out some of the main advantages of hierarchical clustering:

- It is not sensitive to initialization
- It is not required prior knowledge about the dataset
- Users do not need to define the initial number of clusters and
- It is robust in the presence of noise in the dataset

However, hierarchical clustering techniques generally suffer from some drawbacks [39]:

- They are computationally expensive. Hence, they are not suitable for very large data sets
- They may fail to separate overlapping clusters due to a lack of information about the global shape or size of the clusters

2.4.2 Partitional Clustering

Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. This algorithm typically determines all clusters at once. Most applications adopt one of two popular heuristic methods like k-mean algorithm k-medoids algorithm.

2.4.2.1 *K mean algorithm*

The most widely used partitional algorithm is the iterative K-means approach [18]. The K-means algorithm starts with K centroids (initial values for the centroids are randomly selected or derived from a priori information). Then, each pattern in the data set is assigned to the closest cluster (i.e. closest centroid). Finally, the centroids are recalculated according to the associated patterns. This process is repeated until convergence is achieved [39].

2.4.2.2 *K medoids algorithm*

The basic strategy of k-medoids algorithm is each cluster is represented by one of the objects located near the center of the cluster [33]. K-medoids algorithm starts with arbitrarily choose k objects as the initial medoids. Then repeat and assign each remaining object to the cluster with the nearest medoids. Finally, randomly choose a non-medoid object O1. Then compute the total cost (S) , of swapping Oj with O1. If S < 0 ,swap Oj with O1 to form the new set of k-medoids. We repeat this process until there are no changes [33].

2.4.2.3 *Advantages and Drawbacks*

In their paper [39] point out 2 main advantages for partitional algorithms:

- It is very easy to implement, and
- Its time complexity is O(Np) making it suitable for very large data sets.

However, according to Davies [11] those algorithms have the following drawbacks:

- The algorithm is data-dependent,
- It is a greedy algorithm that depends on the initial conditions, which may cause the algorithm to converge to suboptimal solutions, and
- The user needs to specify the number of clusters in advance

2.4.3 Other clustering techniques

2.4.3.1 *Spectral Clustering*

This new type of clustering algorithms called spectral clustering algorithms has been proposed by computer vision researchers and graph theorists [39].Spectral clustering is based on spectral graph theory where a graph representing the data (the graph is analogous to a matrix of the distance between the patterns in the data set) is searched by the spectral clustering algorithm for globally optimal cuts [39]. As pointed out by Forgy [18] the major advantage of this type of clustering is that it can generate arbitrary-shaped clusters. However, spectral clustering suffers from two major drawbacks:

- It is computationally expensive. Hence, they are not suitable for moderately large data sets
- It requires the user to specify a kernel width parameter which has a profound effect on the result of the spectral clustering algorithm. Choosing a good value for this parameter is usually difficult

2.4.3.2 Density Based Clustering

Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold [33] .The main advantages of this approach are [33]:

- Able to work with large datasets,
- Robust to noise, and is
- Able to identify clusters with different sizes and shapes

However Density-based clustering algorithms also face a few drawbacks. Some of the main drawbacks are:

- Does not deal very well with clusters of different densities [33]
- Does not work well in case of high dimensional data [1]

2.4.4 Cluster validation

The procedure of evaluating the results of a clustering algorithm is known under the term cluster validity [43]. In general, cluster validity can be classified into internal and external. When the output of a clustering algorithm is validated based on internal data (data that was clustered itself) then, this is known as internal validation. Internal methods set the highest score to algorithms that produce clusters with high degree of similarity within them and low similarity between other clusters [34]. One drawback of using internal criteria in cluster evaluation is that high scores on an internal measure do not necessarily result in effective information retrieval applications [34].

When the output of a clustering algorithm is validated based on external data (data that was not used for clustering) then, this is known as external validation. Those data are known as truth labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by expert humans. These types of evaluation methods measure how close the clustering is to the predetermined benchmark classes [43]. A drawback of this approach is that prior information of the dataset is required (in order to define the truth labels).

In this thesis, in order to validate the results of clustering algorithms external validation will be used. I will introduce an approach that examines the results of clustering and the clusters quality based on some weighted indicators. The resulting clusters are compared for their similarity with some ground truth labels.

3 Related Work

This chapter gives an overview of the previous research and work related to the content of this thesis. As aforementioned, in this research apart width estimation some other features are implemented as well (standardization of modelling, intersection identification etc.). In the end of each section, how our research differentiates itself from previous work in that field is discussed.

3.1 Road width estimation

3.1.1 Road width estimation from remote sensing images

In 2017 Xia et al. [63], in their paper present a novel approach for road width measurement from high-resolution satellite or aerial images. The methodology of this approach can be summarized in the following steps. First, they extracted linear features and road centerlines from given remote sensing images. Then, the distance overlap ratio and the difference of slope between line segments were computed to match parallel lines and cluster them by their width. The next step was to give each cluster a cluster label. Every cluster contains several pairs of parallel lines with similar width, so a cluster represents a width range. Finally, using the road centerlines and the clusters of parallel lines they assigned a cluster label to each pixel on the road centerline. That results in every road pixel being assigned a width range. This strategy was tested and produced quite good results regarding the road width estimation. Important to mention that this approach relies on having high-resolution satellite or aerial images and accurate extraction of road centerline to give proper results.

A year later a paper written by Luo et al. [32], also propose a learning-based approach for estimating road width from high resolution satellite or aerial images. The methodology of this approach is based on the road pixels intensity distribution to create a road width descriptor that describes the road pixel distribution on a local patch centered on the target pixel then the established features, along with the labels, are employed to train a CNN model to generate a probability map describing roads with different width [32]. In order to evaluate their approach the writers test data (images) taken from different satellite sensors which involved various roads of different type and width. Moreover they test their approach with reference data of various regions. The results showed that the learning based approach gives accurate and stable results in most cases regarding road width estimation.

3.1.2 Lane Width Estimation using LiDAR point clouds

Holgado-Barco et al. [26] at 2017 proposed a strategy to obtain road cross-section information using point cloud obtained with a 2D LiDAR. Among this information is the number of lanes, the width of the roadway, and the width of the shoulder and lanes. The main steps of this strategy are the road surface extraction from the initial point cloud, line fitting, and lane markings detection.

Although they proposed a method for extracting road width their approach is only suitable for 2D laser scanners acquiring data at high driving speed. Besides, their strategy requires the availability of raw measurements of captured points, such as timestamp and scan angle [48].

A few years later, another approach to derive lane width using dense LiDAR point cloud is introduced by R. Ravi et al. [48]. The point cloud was acquired from a geometrically-calibrated mobile mapping system. Moreover, this approach introduces a reporting mechanism for areas with ambiguous lane markings, narrow lanes, and/or wide lanes. The methodology that is used for lane width estimation can be described in the following steps. First, the road surface is extracted by the original point cloud. Next, lane markings are extracted by identifying high-intensity points. Then, the center lines of the derived lane markings are used for the final lane width estimation. The lane width is defined as the distance between the centerline points on either side of the lane along the direction normal to the lane direction.

For the evaluation of this strategy, 5 experiments were carried out and presented. The results of the experiments were quite promising. The accuracy of the estimated lane width was verified by comparing the values to manually measured lane width (ground truth) at different sample points. For most cases, estimates from the proposed algorithm differ from the manual measurements a few cm.

3.1.3 Road width Calculation using vector data

Hoffmans W. [25], created a script that calculates the width of the road sections as included in the basic topographic map of Netherlands (Basis Grootschalige Topografie (BGT)). The script uses areal and linear vector data of roads in order to estimate their width. The script is written in PostgeSQL and to be able to work with it, a number of things are required. Among others, you need the table with the existing BGT road surfaces. A short overview of the methodology can be found in Figure 19. The linear representation is cut into 10 equal parts and based on their intersection with the areal representation and after the necessary statistics are used an estimation of road width is provided. The method also takes into account outliers.

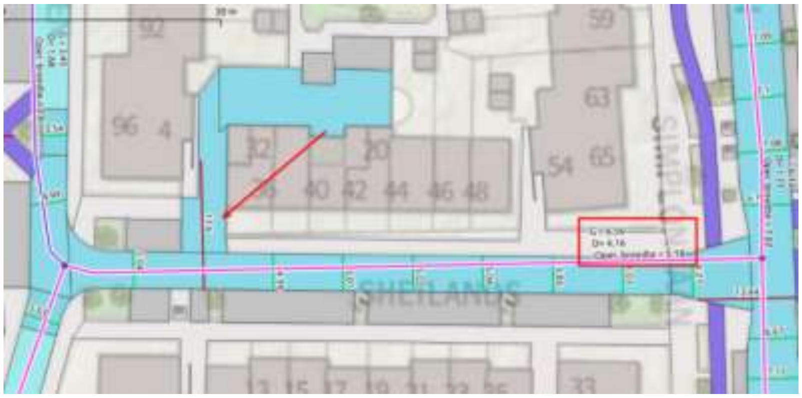


Figure 19 Methodology of ten lines road width calculations with vector data Source: Hoffmans W. [25]

3.1.4 Differentiation of our approach

After having a clear look on previous research regarding road width estimation we can now examine how the novelty of our approach will be ensured. First, a key difference can be detected at the different inputs. Most approaches use as input remote sensing images or LiDAR point clouds. In particular, 2 approaches using remote sensing images were examined. Both studies relying on the existence of high resolution satellite imagery. As pointed out by R. Ravi et al. [48] high resolution satellites may not be able to capture data over work zone areas at sufficiently frequent time intervals. Furthermore, traffic conditions might hinder lane width estimation from satellite imagery [48]. The main drawback of working with high resolution imagery lies on the availability of such data. To the best of our knowledge, there is quite limited free public access to such data. Most freely available information correspond to medium-resolution imagery (10 – 30m/pixel) [10].Two studies using LiDAR point clouds were examined. While the results of those studies in terms of accuracy for estimating width are quite promising, they both using dense LiDAR point clouds. The size and the complexity of dense point clouds requires from the user to be familiar with such an input in order to perform a complex data analysis. Moreover, approach of one study [26] is only suitable for 2D laser scanners acquiring data at high driving speed. This makes the availability of such an input even more difficult.

Finally, an approach that uses vector data was analyzed. Since this approach uses the same input with the one that we use in our methodology it seems to be the most similar approach with the one that we are about to develop. Indeed, the methodology that is used by Hoffmans W. [25] can be seen as our starting point. However, even if this approach has a great practical meaning for the specific application that is developed for (snow removal), we have identified some of its limitations. First, this approach work only with a specific vector dataset (BGT). Moreover, our approach focuses on the implementation of a width estimation methodology that is driven by the needs of a specific use case, road safety management. None of the previously analyzed studies, focus on considering the width of a road as an application-dependent concept.

3.2 Intersection identification for road safety purposes

Wijnands et al. [61], present a study to systematically analyze the design of all intersections in a large country (Australia), based on aerial imagery and deep learning. Their study was developed in order to benefit road safety, by identifying the safe intersection designs. Their methodology was conducted in several steps. The first step was to determine the location of the intersections. They did that by using OpenStreetMap (OSM). They used OSM nodes (geographic location), ways (list of sequential nodes) and relations (collection of nodes and/or ways). intersections were identified as nodes that appeared in at least two ways. This resulted in a large collection of three-way intersections, including T and Y types, four-way intersections, multiway intersections and roundabouts. Then, imagery based on satellite remote sensing was used for all identified intersection locations. As images were not available at all locations of their study, the sample size was decreased. Using deep AE high level features were extracted for each image. Finally, the images were clustered using t-SNE based on the extracted features.. Based on image clustering, they result to 'simplified' classes of intersections into 4 basic categories: i) roundabouts (O), ii) three-way intersections (T), iii) simple four-way intersections (X), and iv) complex intersections (#).

The results of all the separate steps of this study were quite promising. They manage to extract features including the intersection's type, size, shape, lane markings, and complexity in a large-scale. In addition, they obtained similar intersections through unsupervised clustering and finally, design features that have been linked empirically to extreme driving behaviors.

The approach that described in this section is quite different to our intersections identification approach. First, the input is different. Wijnands et al. [61], use vector data for identifying the location of intersections but not for identifying the type of intersections. For doing so they use satellite imagery. Moreover, the types of intersections of our approach are defined based on some prototype types as they can be found in Toronto dataset while Wijnands et al. [61] result to similar intersections through clustering of similar images. Finally, to relate intersections with road safety, deep learning were used (extracting important information of each intersection such as shape, size, etc.).

3.3 Road safety and Road width analysis

The relationship between road width and road safety has been explored by several researchers in different scientific fields. Plenty of studies that examine the relation of road width with road safety exist. In this section, a road safety analysis carried out for the Czech Republic will be analyzed in more detail. The study was performed by Ambros J. [3]. Ambros J. used three datasets: i) accident data that are reported by Czech Traffic Police, ii) road network data obtained from Road and Motorway Directorate of the Czech Republic (ŘSD), and traffic volume data come from National Traffic Census. Regarding accident data, he rejected the PDO accidents (property damage only) since he realized that they were biased. He worked only with serious injuries and fatalities data. He manipulated the road data as well. He excluded the intersections and he examined only the relation of road sections with accident data. Moreover, since data provided by RSD were very large, he choose to work only with 9.5m and 11.5m roads (only two-lane roads). Finally, he grouped data according to traffic conditions. In particular, he took into account the traffic speed and the traffic volume. Regarding traffic speed, he defined 2 main categories: i) rural roads (with speed limit 50 kph) and ii) urban roads (speed limit is 90 kph). Traffic volume conditions were set based on Czech national standard design volume limits. The limits set to 10 000 vpd (vehicles per day) on 9.5m roads and to 12 000 vpd on 11.5m roads.

After manipulating and choosing the data to improve the robustness of his analysis, he computed the accidents frequency (AF). In his research, AF is defined as the number of road accidents that occurred in a road divided by the length of that road. Finally, he generated some results. He derived the conclusion that the relation between road width and road safety differ significantly on 9.5m and 11.5m Czech rural roads (difference between AF means on 9.5m and 11.5m rural roads are almost twofold [3].

This road safety analysis is an example of how road data and accident data could be combined to examine the correlation between road width and safety. In this thesis, a similar analysis will be performed. The main difference regarding the procedure of the safety analysis is that in our analysis traffic data are not taken into account. Accident tendency in different areas is computed to normalize our data. Moreover, Ambros J. [3] is investigating the relation of 2 specific road categories (two-lane 9.5m and 11.5m roads) while we are using all the available road data. Other minor differences existing as well. The most essential aspect that should be mentioned is related to how width is estimated for the examined roads. In our case, we focusing more on investigating how different approaches of width estimation can affect the conclusion regarding safety while Ambros J. [3], focuses more on the actual relationship between Czech roads width and safety. He uses the fixed-width provided by RSD and he does not take into account how roads can be seen differently (different estimations, different unit of measure approaches, etc.).

4 Road safety management

Road width as an essential physical characteristic of roads affects many different applications of the modern world. In this chapter, the relationship between road width and road safety management application is presented. This application will be used to guide my decisions on some crucial points of my methodology. Moreover, it will be used to evaluate my results. Therefore, before diving into methodology a further investigation of why width is important for this particular use case is required. First, I introduce the application and its main tools, then I present my findings on its needs in terms of width knowledge. Finally, I link road safety needs to a unit of measure approach as described in § 2.2. The double positive impact that this approach is expected to have on road safety application is discussed in detail.

4.1 Road safety management and road width

The main purpose of road safety management is to correctly identify problems, risk factors, and priority treatments, and formulate strategy, set targets, and monitor the performance of a road network [60]. According to Capaldo et al. [7], the study of road safety in urban or rural area has shown that the number of accidents and their severity depends on: i) The characteristics of the road environment, ii) the traffic conditions, iii) the behavior of road users. Width of the road section can be considered as the main aspect of road environment both in urban and rural area, that affect the safety conditions [7]. In addition, in another study, Based on a matrix developed by Dr. William Haddon Jr. [54] , three different factors contribute to road accidents: a) Human Factors b) Vehicle Factors and c) Roadway/Environment Factors (Ahmed, Ishtiaque 2014). According to Ahmed, Ishtiaque [2] among the most prominent parameters of road environment that affect the road crash frequency as well as crash severity is road width. So far, plenty of studies relate road width with road accidents ([2], [3], [15], [16], [38]).

Main tool of road safety management application is the safety analysis. Aim of this analysis is the reduction of traffic accidents. This objective is pursued through the study of accidents that occurred under certain environmental and traffic conditions [7]. Accident studies allow the technician or the researchers to formulate instructions on possible actions to solve the technical failure of the traffic flows observed and to realize improvements to achieve a quantifiable reduction in crash frequency or severity [7]. In practice, a road safety analysis relates certain characteristics of road environment with traffic accidents and it can be realized in different levels. Three levels are proposed by Capaldo et al. [7]. Macro-level, Meso-level and Micro-level. While the first 2 levels are using aggregated accident data the micro level corresponds to the best degree of approximation possible based on the use of disaggregated data.

Finally, different road users can be found in the same road environment. Cyclists, pedestrians, and cars coexist and often share the same road area. It is reasonable to assume that changes in the road environment affect all different road users, and therefore it is necessary to consider the impact of road design on the safety of the different groups. For this thesis, I have defined 3 categories of road users: i) cyclists, ii) pedestrians and iii) motorized vehicles drivers. It is important to understand that each group has different needs and faces different hazards. The needs and the dangers of different road users can be contradictory. For instance, a temporary narrowing in a road might affect badly the safety of the cyclists but at the same time, it might not affect at all the car driver's safety. Thus, before moving into our methodology, we need to further analyze the relation of the 3 different road-user groups with road safety.

4.1.1 Road width – Cyclists Safety

As stated by Schramm et al. [51], the flow-on effects of lower speed choices by drivers increase the safety of cyclists. Moreover, in the same research they point out that road width has crucial effect on self-selected speed on road sections. Specifically, they found that in real road situations, an increase in lane width by 1m was predicted to result in an increase in speed by 15km/h. Reasonably turns out that a notable reduction/increase in road width that results in reduced/increased self-selected speed would result in a reduced/increased crash risk for cyclists. For this thesis, as a notable change in road width we define every change (width can change suddenly or in a more smooth way) more than 2m. Thus, identifying that changes of road width in a road network is considered important for cyclists.

In addition to that, research has shown ([17], [29]) that the total number of lanes also influences drivers self-selected travel speed. When the number of traffic lanes increases, driver travel speed increase. Thus, a change in the number of lanes is important for cyclists safety.

According to Schramm et al. [51], temporary narrowings is a source of concern for the cyclists. They point out that narrowings in road network can have negative effects on driver behavior, increasing risks. In addition to that, Gibbard et al. [21] mention that narrow points are especially dangerous for cyclists, since motorists are passing closer to cyclists, and attempting to overtake them prior to the narrow point.

Finally, Based on research of Ewing R et al. [16] ,on-street parking has negative effect on cyclists safety. One of the main causes of vehicle–bicycle incidents is “dooring” (when a vehicle occupant suddenly opening a door into the path of a cyclist).

4.2.2 Road width – Pedestrians Safety

According to WHO [62], road widening increases pedestrian injury risk. Wider lanes and roads, and higher design speed tend to increase motor vehicle traffic speed, which increases pedestrian risk. (WHO 2013).

In addition to that, WHO [62] points out that the number of driving lanes plays a role in pedestrians safety. With fewer driving lanes the safety of pedestrians increases.

Finally, in contrast to negative effects that on-street parking can have on bicycle safety, Ewing R. et al. [16], support that parking acts as a buffer between traffic and pedestrians and provides “friction” that slows vehicles. Thus provides pedestrians safety.

4.2.3 Road width – Motorized Vehicles Safety

While there are plenty of studies that relating road width with road safety, contradictory theories exist.

Main Theory

The conventional theory of roadway design is that wider, straighter, flatter, and more open is better from the standpoint of traffic safety [16]. This theory, relies on the fact that high-speed designs are expected to be more forgiving of driver error, and thus to result to reduced incidence of accidents. As mentioned by Zegeer and Council [64] high-speed design features such as wide shoulders and gentle curves improve highway safety, especially in rural areas. Furthermore, Ewing R. et al. [16] mention that highway system, which is designed for high vehicle speeds, normally shows lower crash rates than other roadway designs. In addition to that, Ahmed, Ishtiaque [2] claims that when the traffic volume is higher and the lane width is less, the probability for car crashes especially crashes like head-on or run-off the road, is greater. Based on their research, the probability for a car crash on a narrow lane i.e. 9 feet (2.75 meters) increases by more than thirty per cent (30%).

While this approach is accepted by many scientists, it fails to account some arguments that support the opposite theory. For example, the fact that highways seems to concentrate lower rates of crashes might be due to other factors. The degree of car conflict, the land-use context and the vehicle operating conditions are completely different in that areas.

Alternative Theory

There is a viewpoint of planners and urbanists that is 180 degrees counter to the conventional theory. Research has shown the road widenings occur at the expense of safety, even after controlling for traffic volumes [16]. Swift et al. [55] in their research found that a typical 11 m residential street (in Colorado) was associated with almost 4 times more accidents compared to a street with 7m width. Many other scientists have found similar results and have linked wider roads to more collisions. Noland and Oh [38], found that wider roads were related with statistically significant growth in the number of crashes in the state of Illinois. Other researchers ([30], [23]) support also this alternative theory. In general, the main argument of that approach is that the reason that links the road width with safety may be the speed [16]. It is known that drivers tend to behave more aggressive on wide streets in comparison with narrow streets.

Other important findings

An interesting conclusion comes from Ewing R et al. [16]. Based on their findings, less-“forgiving” design treatments such as road obstacles, median strips and street trees or other features close to the roadway seem to enhance a roadway’s safety performance when compared to more regular roadway designs.

4.2 Width clustering for road safety management

With respect to what is mentioned so far, I will propose an approach for deriving width estimations from vector data and linking them to vector road representation types in such a way that road safety management application could benefit. As it comes from the literature, the changes in the width of a road are usually indicating some special features of the road such as a change in the number of lanes, on-street parking existence, etc. By exploring the relation of road safety and road width it seems that some of those features are rather important for the safety of various road user groups. In addition, as a common point, it appears that plenty of studies, associating road width with the number of traffic accidents.

Moreover, in § 2.2 different approaches to what can be considered as road when using vector data were analyzed. Based on the relationship between road safety and road width, I believe that a specific unit of measure approach will potentially benefit road safety management. In particular, width clustering approach (as briefly described in § 2.2.3), is expected to have a double positive impact to road safety management application. The overall idea is to cluster the original road centerlines based on several width measurements. The goal is to create some new clustered road centerlines that will be more representative in terms of width. The newly created clusters will be formed by successive similar width measurements. Thus, there will be no major width changes on these "new" roads.

First, those new roads will allow will allow significant width changes to be detected for different clusters. Thus, we could provide the different groups of road users with the information they need. For example, imagine that we have a road where a temporary narrowing exists. While initially a single centerline would be used to represent this path, with the clustering approach it will be possible to split the initial centerline into further clustered centerlines based on width. This, will result in the temporary stenosis being represented as a separate cluster (which will obviously be associated with small width values). This information can benefit different road user groups that temporary narrowings seem to affect their safety.

Moreover, it is expected that the new clustered roads will be more representative in terms of width. This will help us to improve the process of correlating road accidents with the width of the road. Since no major width changes will occur on the "new" clustered roads, associating road accidents with width of roads will have more meaning. We will explore that in details in later chapter.

4.2.1 Clustering for identify important width changes

Based on the needs of the different road user groups as they described in § 4.1, I have identified some real world cases where clustering can be used to benefit road safety management. New clusters will be created to allow the identification of some important width changes of roads (i.e. new cluster when a road is facing a temporary narrowing). I have defined my view of how these real-world cases should be clustered based on width, in order to support our use case. This manual clustering will be used later as ground truth, so that I can use it to evaluate the different clustering approaches for each user group. Important to mention, that I do not expect to conclude into a single ‘correct’ clustering approach. Different parameters will result to different outcomes. Different road users are interested for different road features. Thus, different clustering approaches might suit to different user’s needs. The real-world case are described below:

Case 1: Different clusters needed when we face a notable change in road width. As already explained in § 4.1, a notable reduction/increase in road width that results in reduced/increased self-selected speed would result in a reduced/increased crash risk for cyclists. Figure 20 shows an example of a street that can be found in the city of Helsinki (Finland). The street shows a notable change in its width. Figure 20 left, shows the initial representation of the street with linear vector data (1 centerline in yellow color) while Figure 20 right, shows how it should be represented (3 different lines instead of 1) based on our ground truth. For this case, we define our ground truth based on the assumption that a notable change in width is a change of 2m or more. Important to mention, that for this case the change in width indicates a change in the drivable area of the road. Therefore, no change in the number of lanes or the existence of any obstacle or any other possible reason that may cause a change in width occurs in the examples of this case..

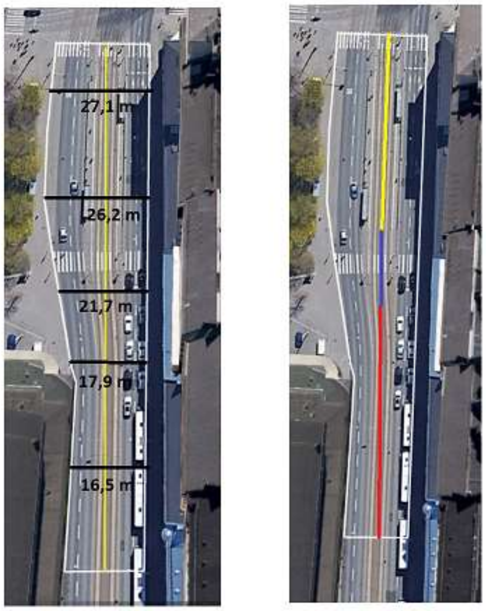


Figure 20 My view on how original road centerline (left) should be subdivided into 3 clusters based on the notable width changes that occur on that road (right). Source: Google earth V 6.2.2.6613

Case 2: Different clusters when on the street-parking exists. Based on research of Ewing R et al. [16], on-street parking has negative effect on bicycle safety. Moreover, in contrast to negative effects that on-street parking can have on bicycle safety, Ewing R et al. [16], support that parking acts as a buffer between traffic and pedestrians and provides “friction” that slows vehicles. Thus, having different clusters when on-street parking exists, could be rather useful for those road user groups. Figure 21 shows an example of a road that on the street-parking exists. Figure 21 left shows how this road is represented by a single line in the initial dataset. Figure 21 right shows how this road should be split into 3 different lines in order to indicates the existence of the parking spot. Figure 21 right corresponds to our ground truth for that case.

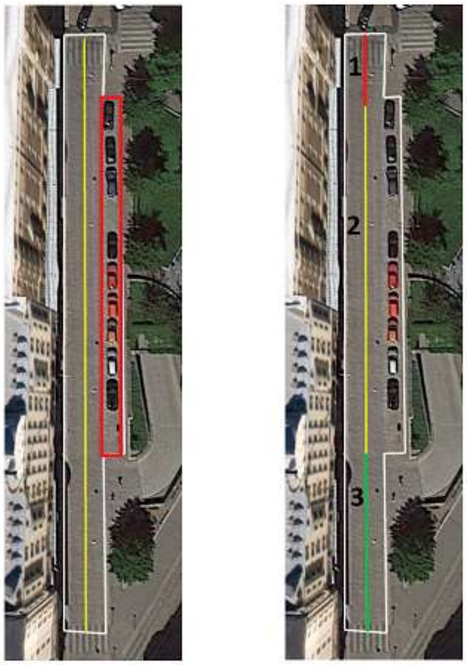


Figure 21 Road where on-street parkign exists. My view on how the single centerline (left) should be divided into 3 new lines. Source: Google earth V 6.2.2.6613

Case 3: Different clusters when a road lane is added/removed. This requirement also results from the literature study on road safety relation with cyclists and the 2 other road user groups. Number of lanes influence the selected speed of vehicles, thus the knowledge of the spot in road network where a lane is added/removed is important. Figure 22 left shows a an example of a linear road representation of a road that can be found in the dataset of Helsinki. In the initial dataset the road is represented by a single centerline. Figure 22 right shows how this road should be split and represented by 2 different lines based on our ground truth.

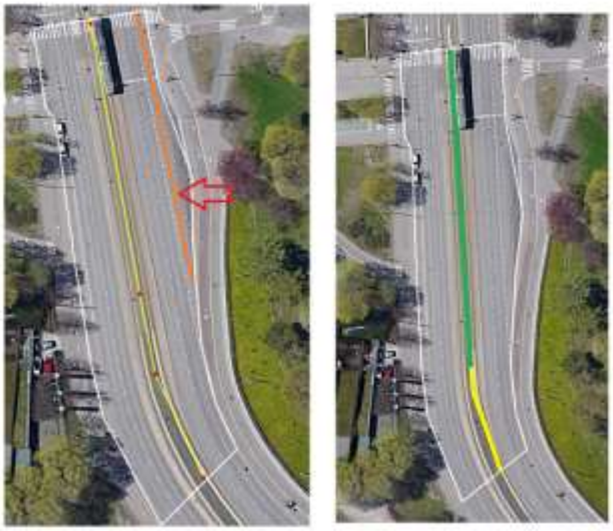


Figure 22 Road where change in number of lanes exists. Original road centerline (left), my view on how this centerline should be divided into 2 centerlines (right). Source: Google earth V 6.2.2.6613

Case 4: Different clusters when we face a narrow point in road network. Based on literature, narrow points are especially dangerous for cyclists, since motorists are passing closer to cyclists, and attempting to overtake them prior to the narrow point. Figure 23 shows a narrow point that can be found in a road. Figure 23 top shows how the road is represented in the initial dataset (single line). Figure 23 bottom shows how the road should be split to 3 different lines in order this narrow spot to be identified. For this case our ground truth is based on the assumption that a temporary change over 1m indicates a narrow point.

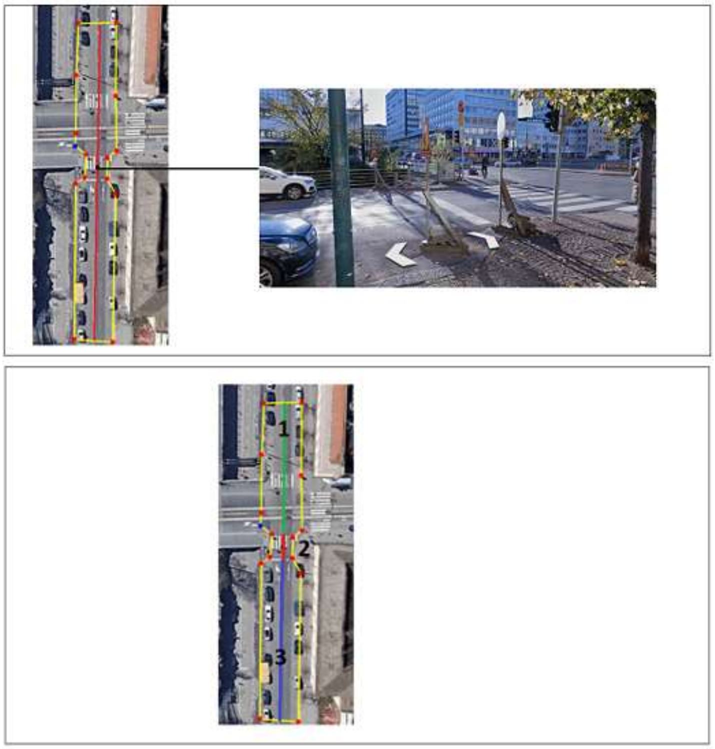


Figure 23 Case where a narrow spot exist in a road network. Original centerline (top left), ground truth of how the original centerline should be divided into 3 new clusters (bottom). Source: Google earth V 6.2.2.6613

4.2.2 Clustering for more representative roads

The second reason that width clustering approach will benefit road safety management application is based on the argument that clustered road centerlines will be more representative of the road geometry compared to the original road centerlines. Let’s explain this with an example. Figure 24 depicts a road as it can be found in the original Toronto dataset (no clustering is applied). As, we can see in the image this road faces some changes in its width along its geometry. If we apply the methodology of Hoffmans W. [25] for estimating road width (briefly explained in § 3.1), we will result to a mean value of 7.5 meters and to a median width value of 8.4 meters. Those 2 values are not indicating the width changes that occur among the road. Standard deviation around the mean is quite large (1.7 m) and partly indicates that there are different measurements along the road. But this is not enough.

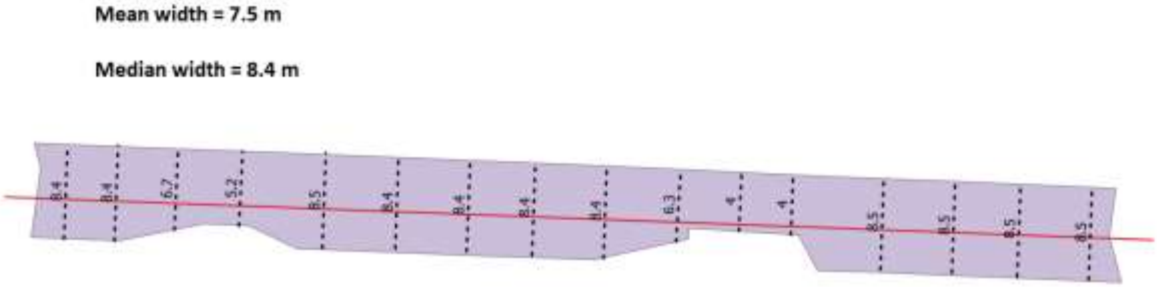


Figure 24 A mean and a median width values are assigned to that road after the combination of some measurements. Those values are not enough since they do not indicate the width changes that occur along the geometry of that road.

This will become even more obvious if we try to relate road width with road accidents. Let us assume, then, that this road is associated with a specific number of accidents (8). Suppose further that these 8 accidents occurred at the locations shown in Figure 25. It is apparent, that 7 out of 8 accidents have occurred at the points where the road becomes narrower. Let us finally assume that there are many roads like this in the original dataset. Therefore, if we had investigated the relationship between the width of un-clustered roads and the number of accidents, we could conclude that wider roads are quite dangerous and are associated with most accidents. Nevertheless, this is false. Clustering approach is expected to ensure that the width values of a clustered road are more representative of the overall structure of a road.

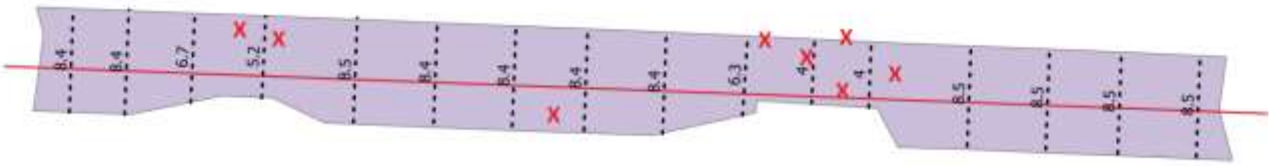


Figure 25 Road geometry is associated with traffic accidents. The most of the accidents occurred in the narrow parts of that road. By using only a mean or a median width value for this road this conclusion could not be derived.

5 Methodology

As aforementioned, the main goal of this thesis is to expand the scope of the existing methodology of Hoffmans (2018). I focused in implementing a more generic methodology that works with vector data from different sources. To do so, I developed an approach that standardizes road vector data based on an existing modelling approach. In addition to that, I exploit this modelling approach in order to identify different types of intersections. Moreover, I developed a methodology that is driven by the needs of road safety management application. The relation of road width with a selected use case is explained in chapter 4. Thus, clustering unit of measure approach is implemented in order to serve the needs of road safety management application (see § 4.2). This chapter explains how each step of my methodology is conducted. The schematic overview is shown in Figure 26.

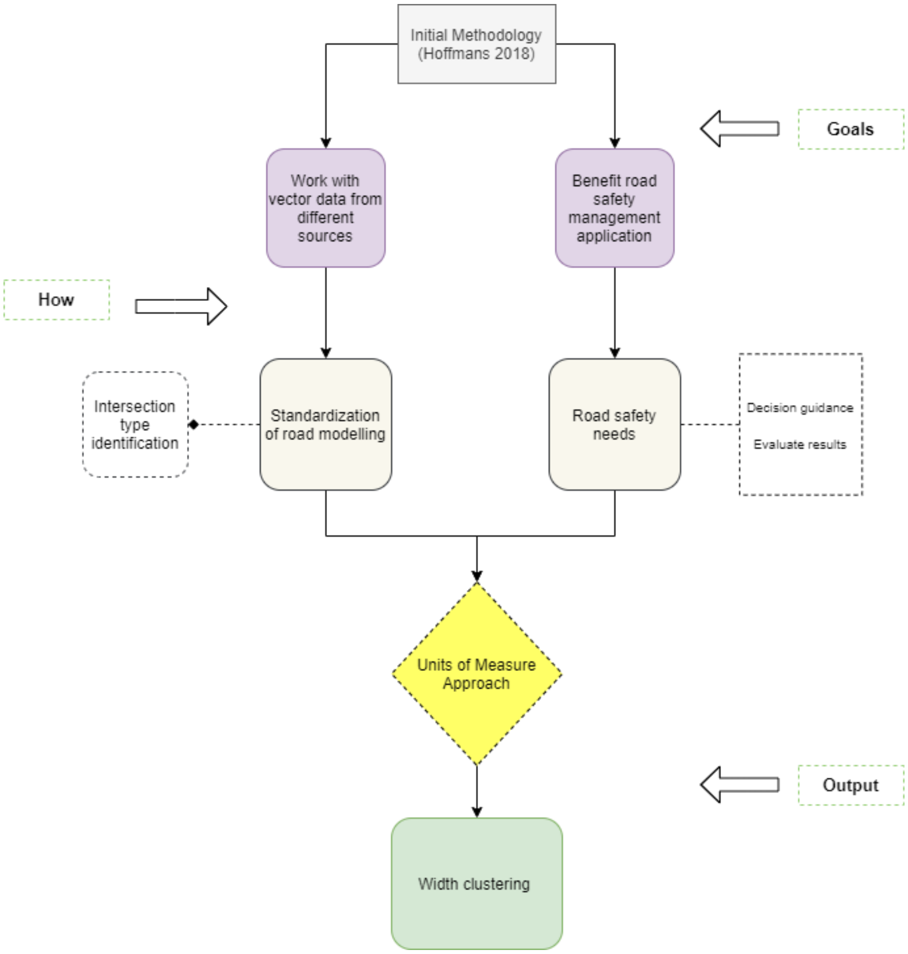

Initial Methodology (Hoffmans 2018)
Goals
Work with vector data from different sources
Benefit road safety management application
How
Intersection type identification
Standardization of road modelling
Road safety needs
Decision guidance
Evaluate results
Units of Measure Approach
Output
Width clustering

Figure 26: Flowchart of several steps of this research

5.1 Initial Methodology

Our initial methodology is based on work of Hoffmans W. [25]. The approach of Hoffmans which briefly explained in § 3.1.3 is the base for our approach since we both work with the same input. This approach focuses on using areal and linear road vector representations in order to estimate width. By areal representation we refer to road polygons and by linear representation we refer to road centerlines. For more details of those definitions you can refer to § 2.1. Initial methodology, exploits the main characteristics of vector data and by combining both representations of the same road a width value for each road is estimated. The main steps of how road width values are estimated for a road are explained below:

1) Road centerline and corresponding road polygons:

- For the initial road centerline that we are going to estimate width, we need to find the corresponding road polygon (it can be more than 1). By corresponding road polygon we mean the polygon that intersects with the road centerline (represents the same road but with different vector type).

2) Find the measuring lines: (Figure 27)

- (A+B) Road centerline is divided it into smaller parts every few meters. We 'cut' the initial centerline every x meters based on a **measuring interval**. The more the measuring lines is (the smaller the measuring interval is), the more detailed the measurement will be. On the other hand, smaller measuring interval leads to higher processing time (more measurements to manipulate).
- (C) We define 2 offset lines for each new part (one at the left and one at the right). We need to ensure that both offset lines are lying outside of the areal geometry of the pair.
- (D) We find the midpoints of the 2 offsets and we connect them. A new line, perpendicular to each small centerline part is created
- (E) Finally, we find the intersection between each perpendicular line and the areal geometry of the pair. A new line, the **measuring line** is created for each part.

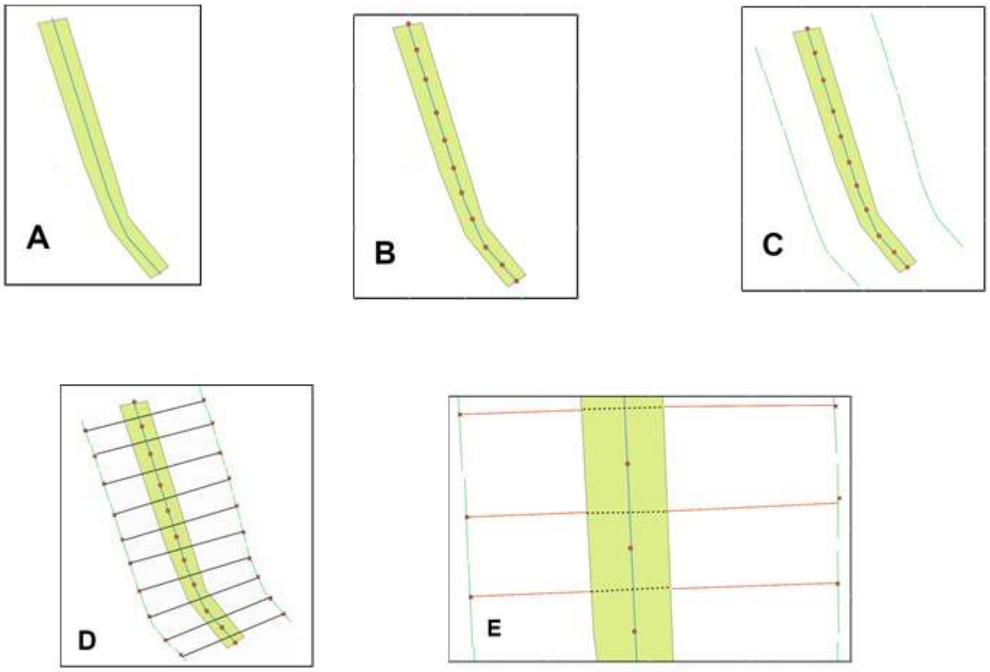


Figure 27 Steps for defining the measuring lines for a road. A) Input geometries, B) Cut lines to equal parts based on a user-defined measuring interval (10 in that case), C) Take 2 offsets in a big distance for each newly create centerline part, D) Connect the midpoint of those 2 offsets, a new perpendicular line is created for each part E) Find intersection of each perpendicular line and areal geometry, a new line, the measuring line is created (Black dotted line)

<u>3) Identify Outliers:</u>

- Find the mean length value of the measuring lines (length of a measuring line indicates a width measurement for the road)
- Find the standard deviation around that mean value
- Find the measuring lines that have length higher/lower of the mean length + 2 times the standard deviation. Mark those measuring lines as outliers. Other thresholds can be used to mark outliers. The higher the threshold will be the more outliers might be used to estimate road width.

<u>4) Estimate width values for each centerline:</u>

- Using the remaining measuring lines (after removing the outliers) to estimate width for each centerline. Width values are estimated. Those values is assigned to the linear geometry (road centerline). Different width values such as mean, median, max and min width are calculated.

5.2 Standardization of road vector data

The modelling approach that is used as our prototype is the Toronto modelling approach. We discussed in more details how Toronto chooses to model roads (especially intersections) with vector data in § 2.1.3. Now I will explain the steps that are followed in order to standardize road polygons based on this approach. The way that methodology is conducted is described below:

1) Dissolve all the road polygons:

All the separate areal geometries are dissolved into a single multi geometry.

2) Download graph data

The next step is to obtain the graph data (edges and nodes) of the specific area for which the standardization process will take place. After downloading the graph we keep only the nodes that correspond to an intersection. In order for a node to be considered an intersection node, at least 3 road branches must meet at it. Degree of a node indicates the number of roads that touch that node. Thus, we keep only the nodes of degree 3 or higher. Figure 28 illustrates the nodes of a graph of a specific area in Helsinki. Red nodes have a degree of 4, Green nodes degree of 3, and Blue nodes degree of 1. In that area, there are no nodes of degree 2. As you can see, nodes of degree 4 mostly correspond to cross/X intersections (4 roads are met at the same point) while nodes of degree 3 correspond at intersections that 3 roads are met. Finally, nodes of degree 1 mostly correspond at the endings of roads and do not represent an intersection. For our methodology, we keep only nodes of degree 3 or higher.

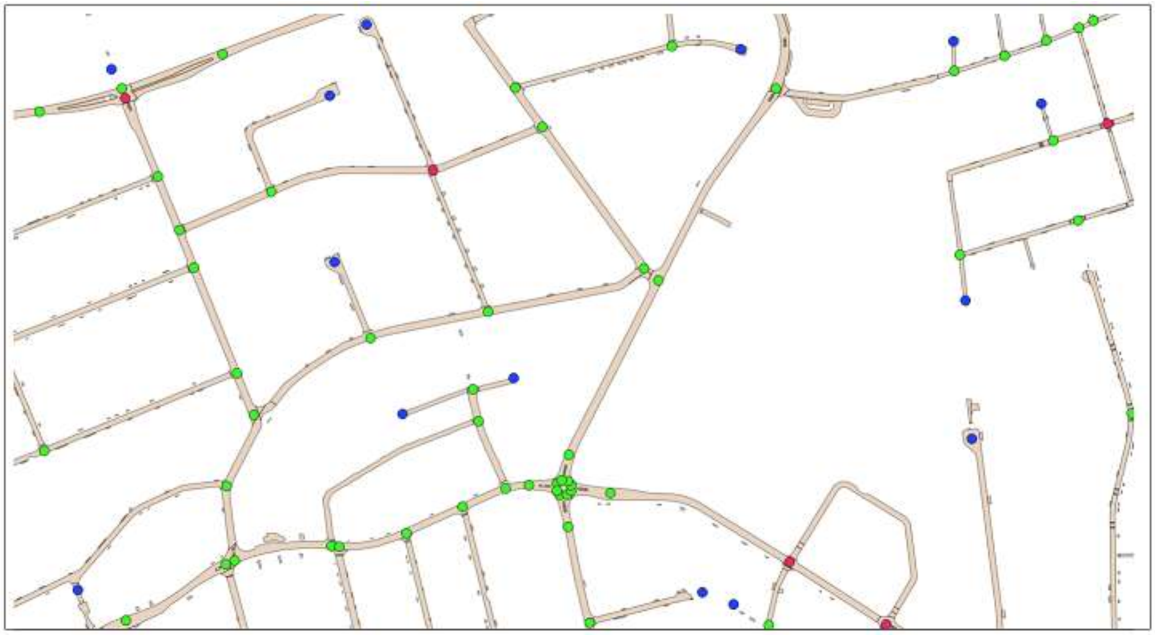


Figure 28 Nodes based on their degree

3) Remove motorway nodes

At this point, we need to remove the nodes that belong to motorways. A limitation of this process is that is not developed to work for motorways. Further details on the limitation of the standardization process will be discussed in a later chapter.

4) Remove roundabout nodes

Roundabouts can have a quite complex structure. Often they are modelled with many and large intersection polygons (Figure 29). We need to identify at this point, the nodes that are lying in a roundabout intersection in order later to re-create a new, more simple roundabout polygon. In order to identify the intersection nodes that lying to a roundabout polygon we use the properties of them. The nodes of the graph are assigned with an attribute called "junction". If they are roundabout points then "junction" is equal to roundabout.

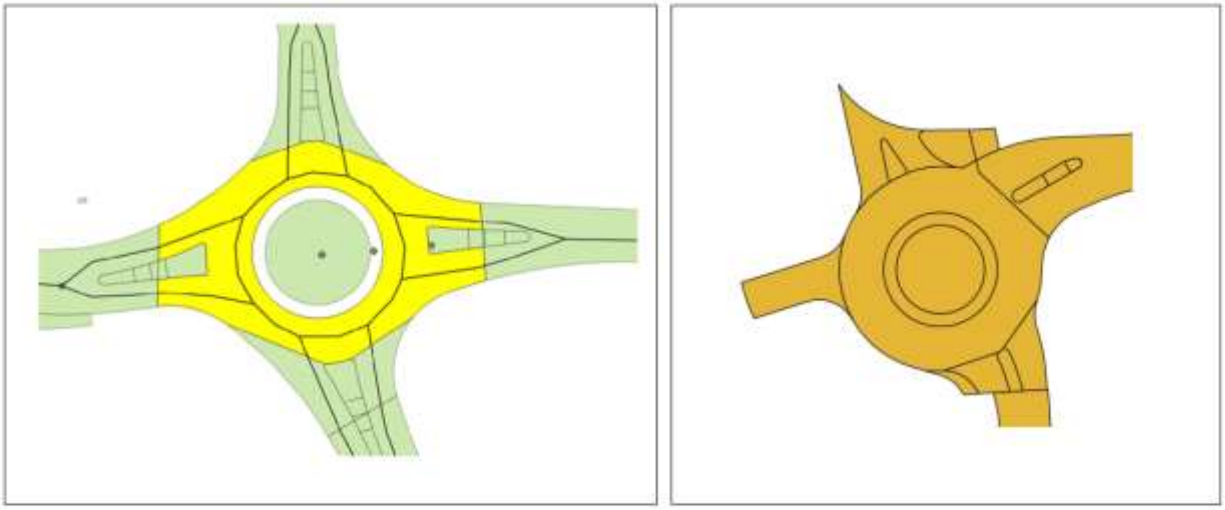


Figure 29 Roundabout polygons can have a rather complex structure. In many datasets more than 3-4 polygons are used to represent them. This will cause problems to our standardization methodology. We need to identify roundabouts and then create a new, simple roundabout polygon.

5) Create a small buffer around remaining nodes

In Figure 30 we can notice that some nodes are really close together. We create a small buffer to detect those nodes (when small buffers intersect).

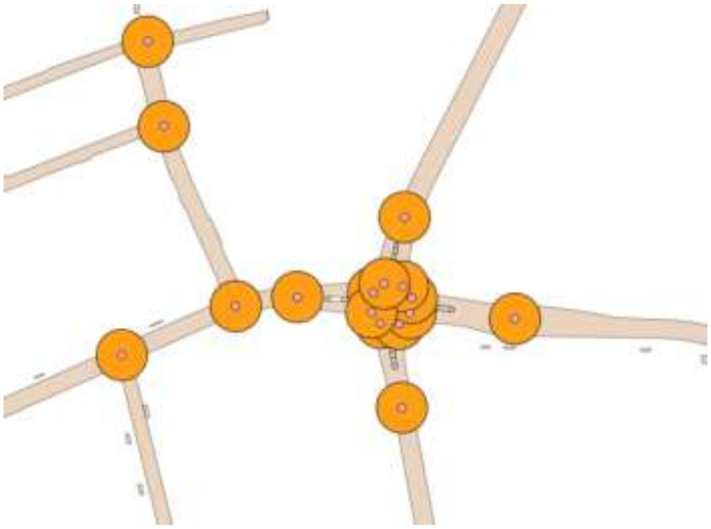


Figure 30

After detecting those nodes we create a union of all the buffers that intersect and we mark the centroid as a final intersection node (Figure 31).

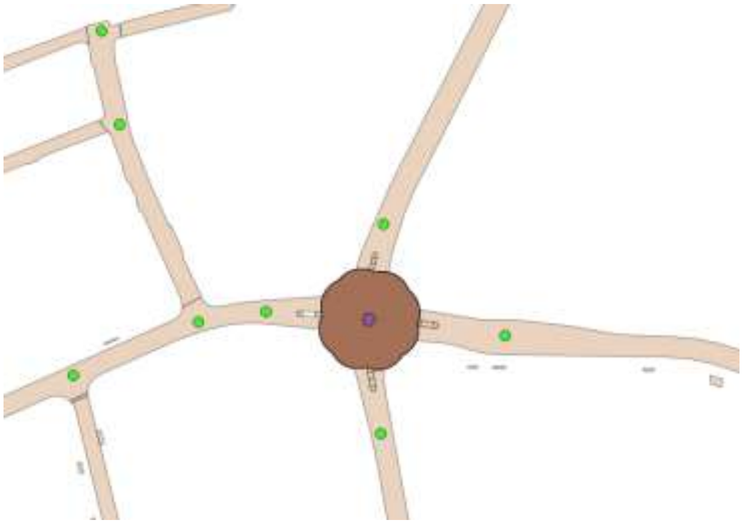


Figure 31

Those centroids and the other remaining nodes (all the nodes that their small buffer wasn't intersecting with buffer of other nodes) are the nodes that are going to be used for recreate the road polygons.

6) Create a buffer around intersection nodes

After merging the nodes that are too close together we create a buffer around those nodes (Figure 32). The size of this buffer is very important since it affects the final size of the new polygons. The larger the buffer, the larger the intersection polygons and the smaller the non-intersecting polygons.

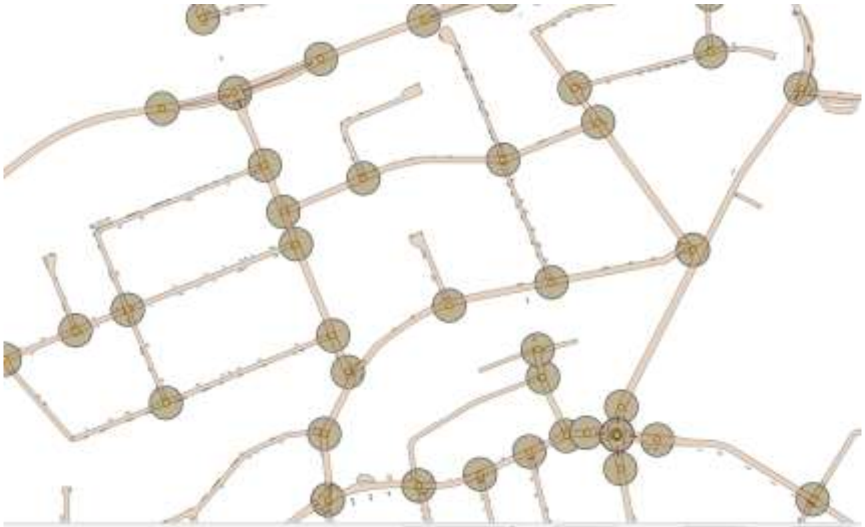


Figure 32

7) Find intersection between buffers and initial dissolved areal geometries

The next step is to find the intersection between the buffers of the previous step and the initial dissolved geometry that was created at step 1. By doing that we have create the new intersection polygons. Those intersection polygons are shown in Figure 33.

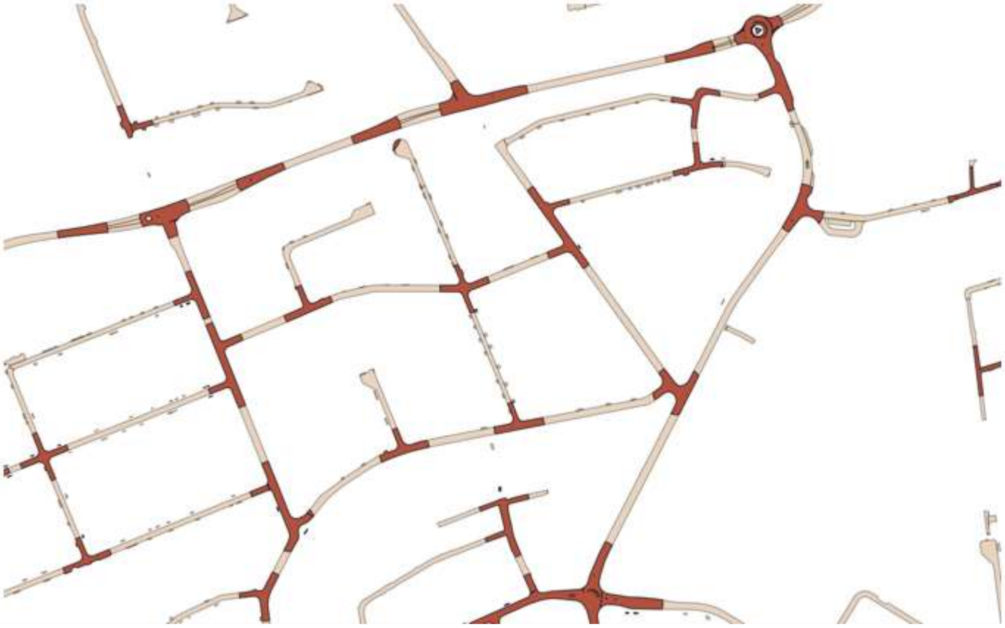


Figure 33 New intersection polygons created after computing the intersection between intersection node buffers and dissolved initial areal geometries

8) Find symmetric difference between buffers and initial dissolved areal geometries

The final step to create the new road polygons is to compute the symmetric difference between the newly created intersection polygons and the initial dissolved Multipolygon. By doing this we find the rest (non-intersection) polygons. Figure 34 shows the resulting new road polygons (left) in comparison with the old road polygons of the same dataset (right).

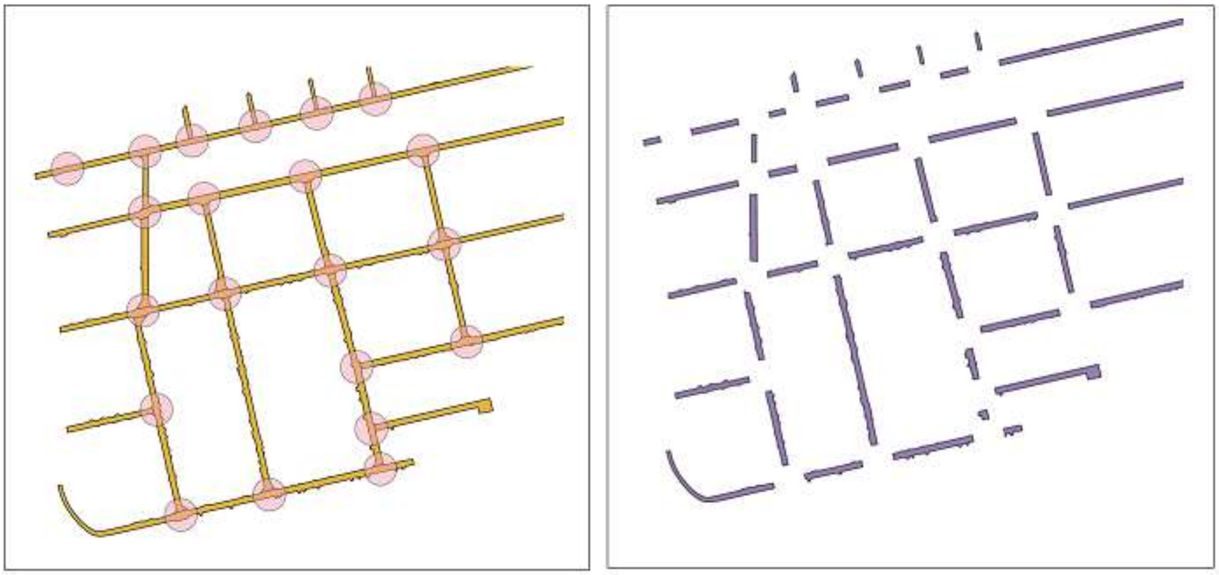


Figure 34:

9) Split intersection polygons:

The newly created intersection polygons need a further manipulation in order to follow our prototype modelling approach. We noticed that some of the new intersection polygons had a quite big shape. This, occurs when 2 or more intersections are close together. Thus, the buffers that were created in a previous step around their nodes are intersecting. This, had as a result, some 'new' intersection polygons that 2 or more different intersections are merged to 1 polygon (Figure 35).

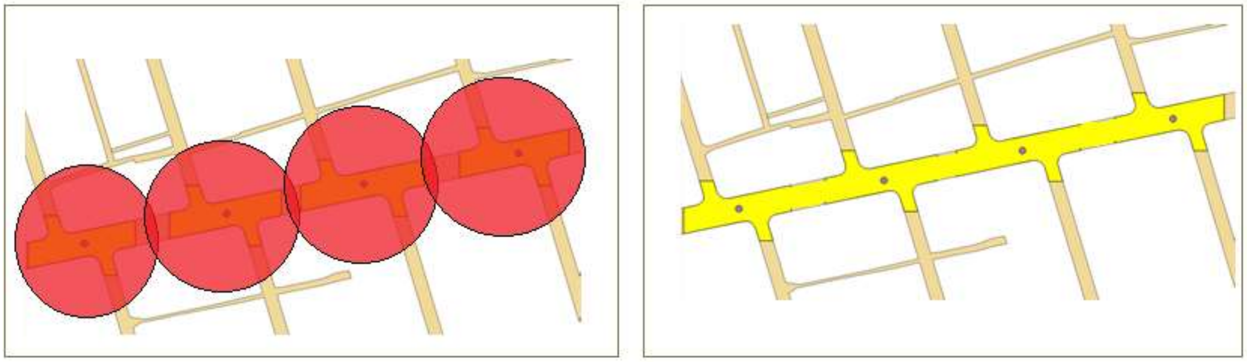


Figure 35: Extreme case of 4 intersection polygons that are close together and the buffers around their nodes are intersecting. This, has as a result to create a new big intersection polygon that contains all the 4 intersections

To solve this, we have to find the intersection polygons that contain more than 1 node. If the new intersection polygon contains **exactly 2 nodes,** then we check the distance between those 2 nodes. If distance is below a specific threshold (user-defined) then we do not split but we keep the intersection polygon as it is (it might be the case of an double T type of intersection polygon as it is described in section 2.2). If the distance is higher than the distance threshold, we need to split the polygon. We find the mid-point in the line between them, we define a perpendicular 'cutting' line and then split into **2 new polygons**, (Figure 36). If the distance is way bigger than the user-defined threshold then we create also a small road edge polygon in between the 2 new intersection polygons (Figure 37).

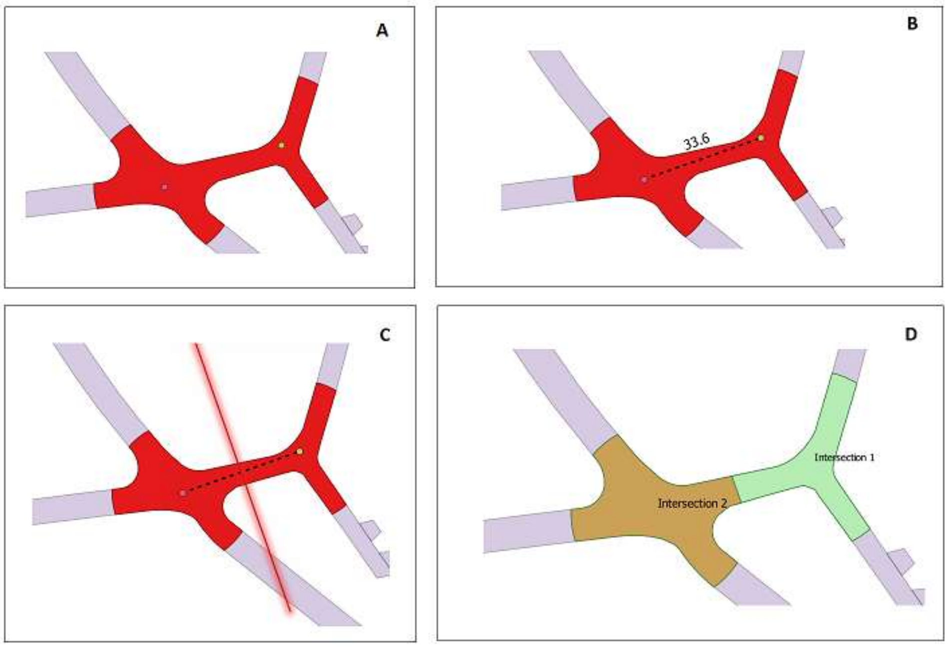


Figure 36: Process of splitting an intersection polygon that contains 2 nodes into 2 smaller intersection polygons

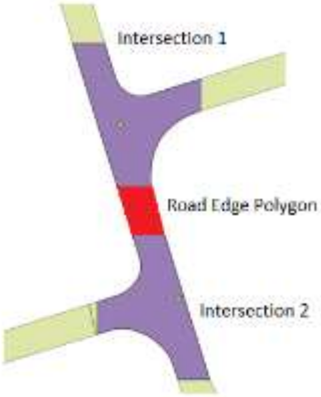


Figure 37: Case of distance between 2 nodes of an intersection polygon is too big and we create also a small road edge polygon in between

If the intersection polygon contains exactly 3 nodes, we need to check the angles between the 3 nodes (if they are performing a Y intersection we should not split). To check the angles, we take one of the 3 nodes and we create 2 lines based on that node as the start point and the 2 others as endpoints. Then, we check the angle created between these 2 lines, If they are parallel or near parallel then we need to split otherwise we don’t (Figure 38).

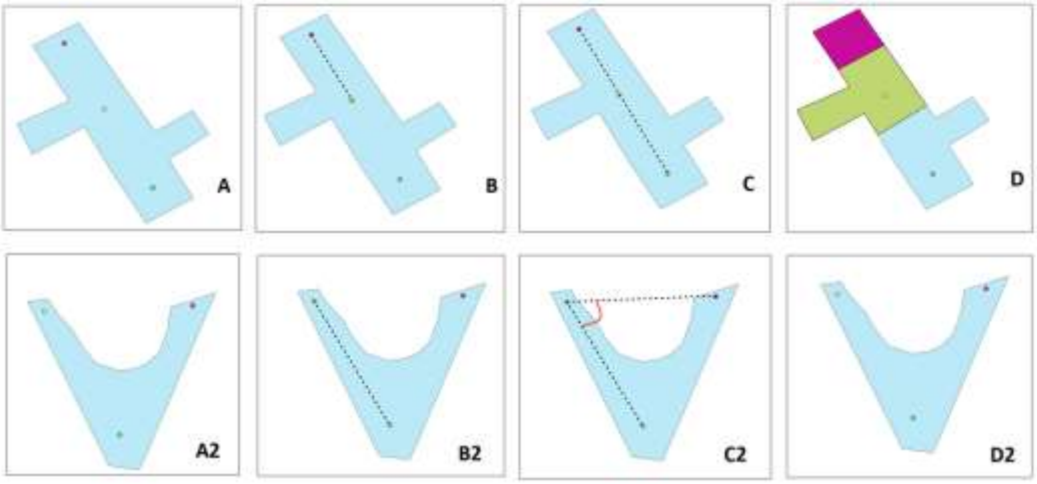


Figure 38: Two examples of polygons that contain exactly 3 nodes. In 1st case (Top) the polygon need to be splitted into 3 new polygons. We create 2 lines. One that has as start point the red node and as end point the yellow node (Top B) and one line that has as start point the red node and as end point the green node (Top C). By checking the angle between those 2 lines we find that they are near parallel thus, we do split the polygon. In the 2nd case (Bottom) the polygon should not be split. By checking the angles of the 2 lines (B2 and C2) we realize that the lines are not parallel or near parallel. Thus, we do not split this polygon

In case the 2 lines are almost parallel, we have to divide the polygon in such a way that 3 new intersection polygons are created. We define the shorter line from the 2 that were created to check the angle. Then, we simple split in the mid-point between the 2 nodes. Now, we have two new polygons. One of those 2 needs to be further split. We check the area of the 2 new polygons and for the bigger polygon we split again. Finally we result to **3 new polygons** Figure 39 shows this process.

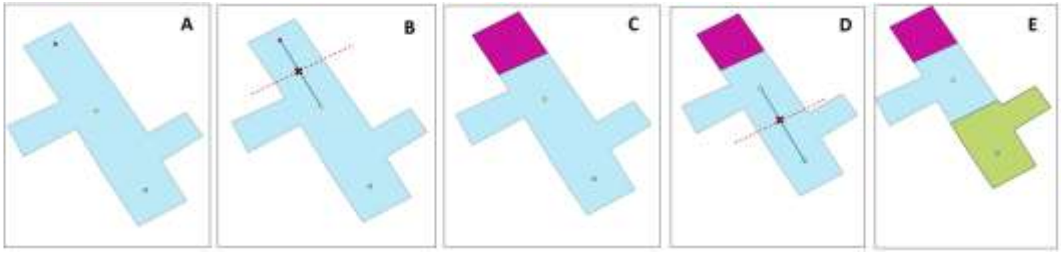


Figure 39: Process of splitting an intersection polygon with exactly 3 nodes

If the intersection polygon contains more than 4 nodes, it is case of a complex intersection. Thus, we keep it as it is.

<u>10) Merge small polygons with neighbors:</u>

From the overall standardization process, we noticed that some fairly small polygons are created. This might be due to the ‘not perfect’ intersection between the buffers and the initial road polygons. Since a perfect size for the buffers does not exists, some pretty small polygons might result from that geometrical operation. Thus, we need to merge those super small polygons with their neighboring polygons. To do so, we distinguish between really small polygons and regular size polygons based on a user-define area threshold. Finally, for each small polygon we find its regular size neighbor and we merge it.

11) Create a roundabout polygon:

Final step is to create a new simple roundabout polygon. We will use the roundabout nodes that was detected in step 4. We find the centroid of those points and we create a buffer around this centroid. Then we find the intersection of buffer with the initial road polygons. We combine all the separate geometries that was contribute the initial roundabout polygon and we simplify them. We assign an attribute that indicates that this polygon is a roundabout so we can use it later for the intersection identification.

5.3 Intersections identification

In this step of our methodology, some features of Toronto modelling strategy are exploited in order to identify the different intersection types. Toronto modelling as explained before, follows a strategy that intersections are modelled explicitly. A unique road polygon correspond to an intersection. First, I will discuss the reason why intersections should be identified and treated differently. Then, I will explain how I took advantage of the areal and linear geometries of intersections to identify their main types, as they are structured by the Toronto modelling approach.

5.3.1 Intersections will cause noise

As already explained, an intersection can have many different configurations. Different intersections might cause noise to our overall approach. Figure 40 illustrates 2 examples of how 2 different intersection polygons can cause noise if we simply apply our initial methodology (§ 5.1). Figure 40 left, indicates a road centerline (red) that passes through a cross intersection polygon. The dotted lines (black and red) are the final measuring lines as they result if we follow the steps of our initial methodology. The black dotted measuring lines are measuring the width of the road while the red dotted measuring lines are measuring the length of the perpendicular roads that meet at a cross intersection. Since the red dotted measuring lines are a big percentage of the total measuring lines of this polygon (3 out of 11) they are not marked as outliers. Thus, they influence the final width estimation for the road. Because, wrong measurements are included in the estimation of the average width, the final value that is calculated is 11.4 m, while a reasonable value for this road (based on the correct measurements) would be 7.8 meters. Something similar occurs in Figure 40 (right) but in that case, the road passes through a T intersection polygon.

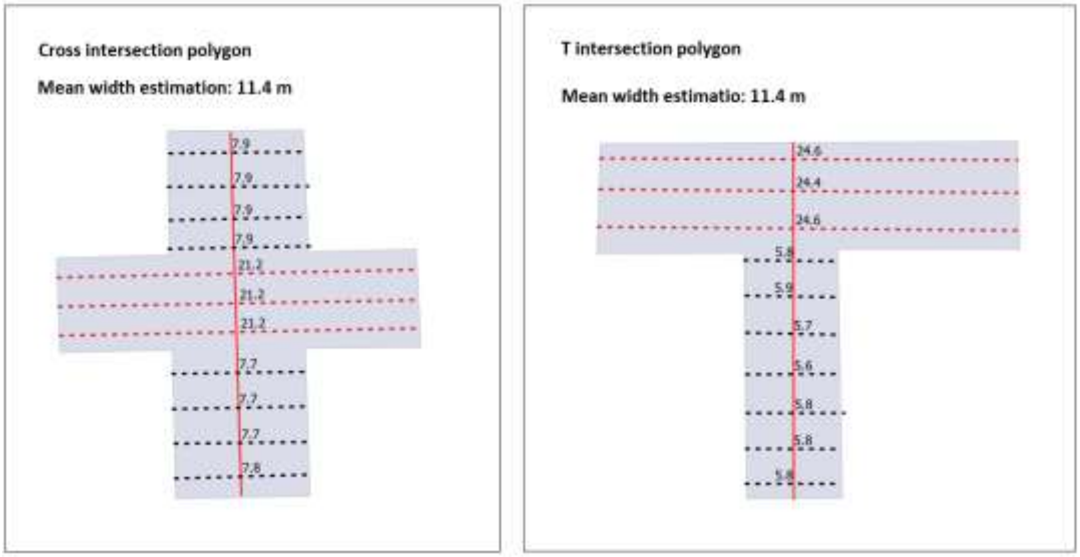


Figure 40: Intersection polygons might add noise to our approach if we simply rely to our initial methodology. 2 example illustrated. A cross and a T intersection polygons are resulting to wrong width estimations for the 2 road centerlines.

Figure 41 depicts an example of how intersections can have a big impact in the overall width estimation of a road. In this example, a road centerline (red) passes through 3 polygons. 2 of them are intersection polygons while the 2 other polygons correspond to road edge polygons. By applying the methodology as explained in chapter 6.1, we get a mean width value of 12.1 meters (Figure 41A). If we had excluded the intersections from the overall process the final value that we will get is 10.3 meters (Figure 41B). As is evident from the Figure 41C, where it depicts the mean width of the road if only the 'correct' measurements were used, the value that we get if we exclude the intersections from the process is way closer to the ‘correct’ value.

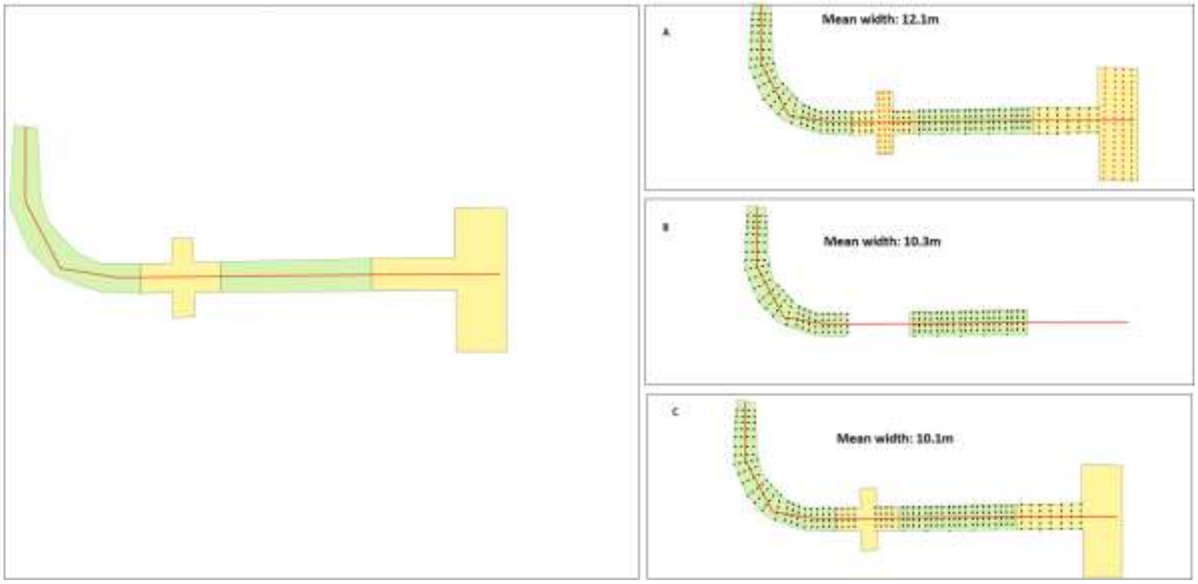


Figure 41:

In order to have a better look, on how the existence of 'wrong' measurements at intersection polygons can affect the final width estimations, we explored the results in width statistics for an area that can be found in the dataset of Toronto (Figure 42). We tested the results in width statistics with and without intersections and we compared those results with our ground truth. The total number of polygons is 153. The number of intersection polygons is 57 and road edge polygons are 97. Table 2 summarizes our findings.

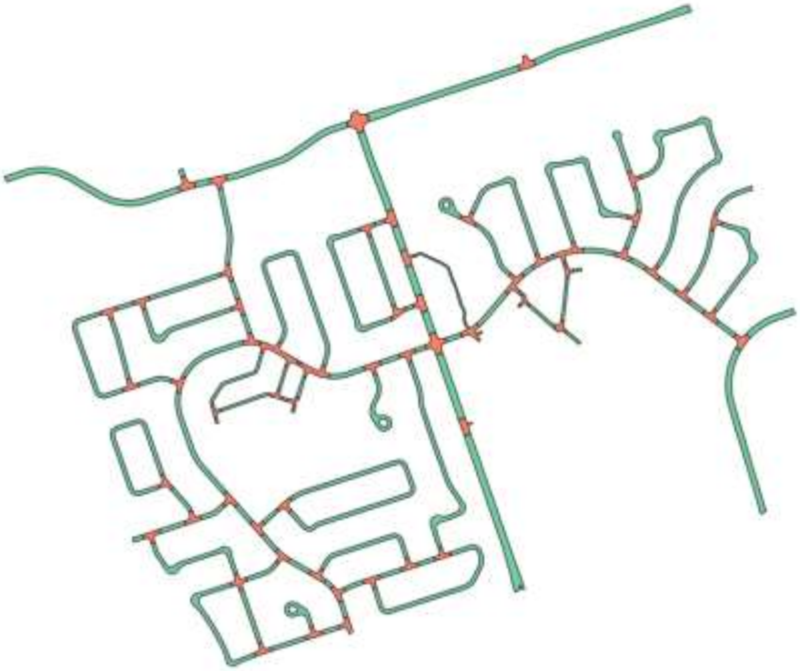


Figure 42: Tested area for impact of intersection polygons in width estimation process. Total number of polygons 153. 57 intersection polygons and 97 road edges polygons.

Measuring approach	Mean width (m)	Standard deviation from mean (m)	Median width (m)
Include intersection polygons	11.13	1.73	10.25
Exclude intersection polygons	9.34	1.03	9.23
Ground Truth	9.18	1.06	9.12

Table 2: Results of width estimation process with and without intersection polygons.

From the table above it is obvious that the existence of intersection polygons affect a lot the result of width estimation. The result when we exclude the intersection polygons and the 'wrong' measurements that they come with them are closer to our ground truth. Moreover, since the measuring lines are defined based on the length of the centerline (measuring interval that 'cuts' the line every x meters, see § 5.1), it turns out that shorter roads are affected more by the existence of intersection polygons than longer roads. For smaller roads, incorrect measurements will be a larger part of the total measurements. Table 3 contains data on how the width values of 150 roads found in Toronto are affected based on their length (50 roads for each length category). From this table, it can be argued that roads under 100 meters experience a quite larger change in their width compared to roads over 300 meters.

Road length category	Mean width with intersections	Mean width without intersections	Median width with intersections	Median width without intersections
Length <100m	16.43	11.2	12.65	11.16
Length 100-300m	14.21	11.89	11.97	11.15
Length > 300m	13.12	11.91	13.4	13.11

Table 3 Influence of intersection polygons based on road length

5.3.2 Identification of different intersection types

The types of intersections that are identified by our methodology are based on the types that can be found in Toronto dataset (see § 2.1.3). Note that, not all the different intersection types that can be found in Toronto dataset are identified. Specifically, T intersections, Cross intersections, X intersections and Roundabouts are identified. Regarding the main intersection types of Toronto modelling, there is still a limitation on identifying Y and double T intersections. Those 2 types are currently identified as another intersection type. Moreover, most complex intersections where more than 4 roads are meet are also identified as another intersection type.

<u>T intersections</u>

T intersections can be seen with 2 different ways based on the road we are examining. First case exists when the part of the T intersection polygon which correspond to the perpendicular road and is the part that lead to 'wrong' width measurements is located in the end of the road (Figure 43 left). Second case corresponds to cases where the wider part of T intersection is located in between the end and the start of the road (Figure 43 right). We will refer to the first case as main case and to the second case as case 2.

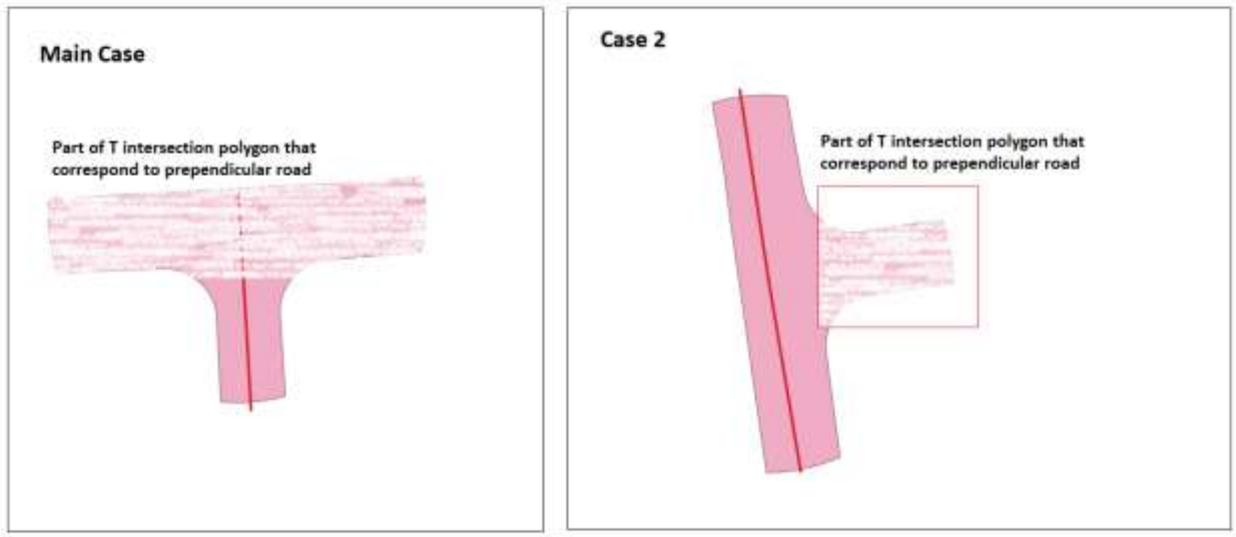

Main Case
Part of T intersection polygon that correspond to prependicular road
Case 2
Part of T intersection polygon that correspond to prependicular road

Figure 43 T intersection can be seen in 2 different ways based on the road that we examine each time. For the purposes of this research the case left we call it Main case and case at right we call it case 2

For identifying T intersection main case we compare mean length values of the measuring lines at different locations of the road. From Figure 44, it is obvious that the wrong measurements will be quite higher that the measurements that correspond to the actual road. The wrong measurements for the main case of T intersection are located in the end/start of the road. Thus, to mark an intersection as T intersection we compare the mean and median length of the measuring lines in the end/start of the road (3 measuring lines) with the mean and median length of the measuring lines in the rest part of the road. A user-defined threshold indicating the difference from which and above the polygon is denoted as the main T-case is required. For the purposes of our methodology we claimed that if the mean/median length value of measuring lines in the end/start of the road is 2 times higher than the mean/median length values in the rest road then it is a T intersection main case.

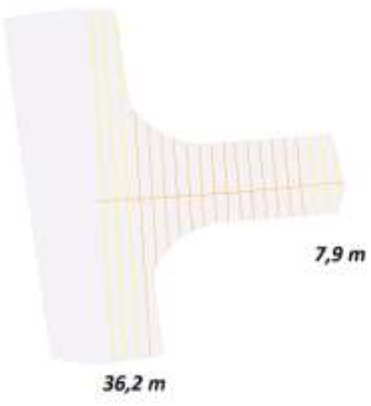


Figure 44 Main case of T intersection

For identifying **T intersection case 2** (Figure 45) we assume that the measuring lines in the middle of the road are longer than measuring lines at the end and at the start of the road centerline. Then in order to distinct between T and Cross intersection (in a cross intersection this would be the case as well), we check the distance between a few points in the middle of the centerline and the intersection points of the measuring line with the polygon. We assume that the distance to one edge of the polygon should be much higher than the distance to the other edge (blue lines are much higher than red lines). Thus, if the mean/median length of measuring lines in the middle of the road is 2 times higher than the mean/median length of measuring lines in the end and in the start of the road then we could possibly have a T intersection case 2. Thus, a second test is required. If the length of a 3 lines in the middle of the road centerline to one edge of the polygon is 2 times higher than the length of these lines to the other edge of the polygon, then we mark the polygon as a T intersection case 2.

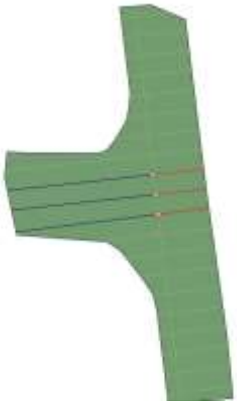


Figure 45 T inresection case 2

Cross and X intersections

A different strategy is followed for identifying Cross and X intersections. Methodology for identifying T intersections was based on the areal representation of the road. Now we use graph properties. Cross and X are road intersections that exactly 4 roads are met. The main difference between those 2 intersections is lying in the angle that the 4 roads are forming with each other. Figure 46 shows a typical example of those 2 intersections.

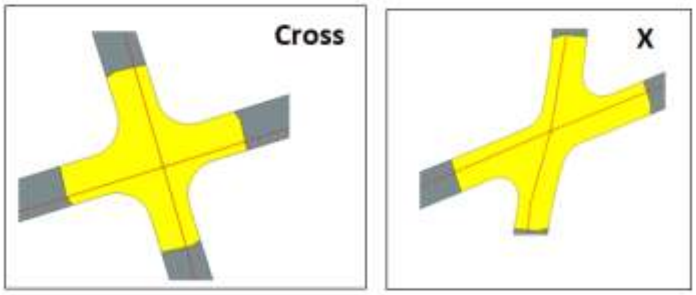


Figure 46 Typical example of Cross (left) and X (right) intersections

Thus, the first step is to identify all the polygons that exactly 4 roads are met. To do so, we exploit a property of the nodes and based on their degree we find all the polygons that contain nodes of degree 4. One more step is needed in order to distinguish between Cross and X. We need to perform an angle test between the 4 road that met. If the angle between the road is forming an angle of almost 90 degrees (85-95 degrees are acceptable) then it is a Cross intersection otherwise it is an X intersection.

Roundabout

Roundabout polygons are created from the scratch with our standardization methodology. In the final step of the aforementioned methodology we create a new, simple roundabout polygon replacing the complex roundabout polygons that could be found in the different datasets. Thus, when we create these new polygons we assign them an attribute that denotes that it is a roundabout.

Other complex Intersections

Every polygon that contains a node of degree higher than 4 (5 or more roads are met there) is simply identified as a complex intersection. Figure 47 shows some example of such possible intersections.

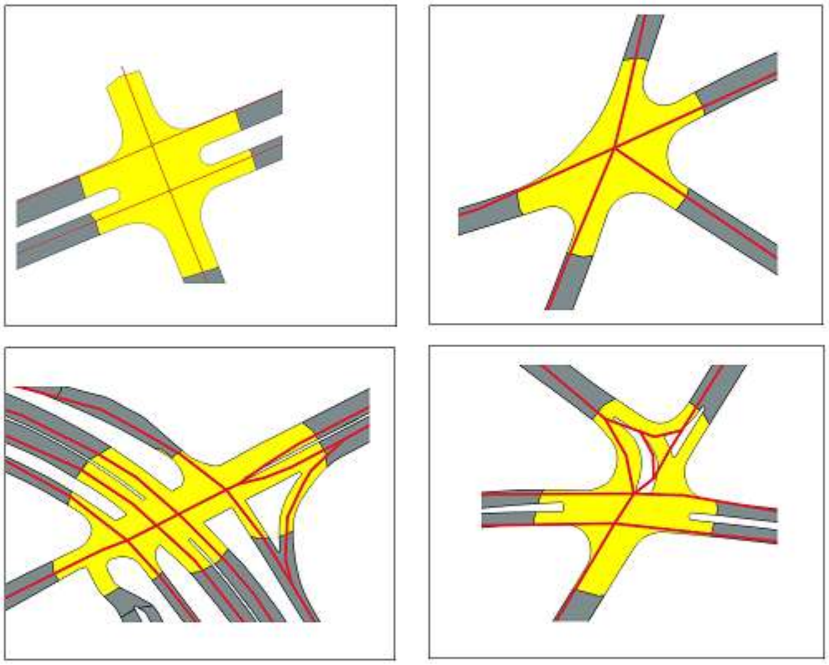


Figure 47 Complex intersections where more than 4 road brances are met at the same place.

5.4 Width Clustering

Driven by the needs of the selected use case the methodology that implemented creates new roads by clustering the measuring lines with similar width. In this section, I will explain how clustering is implemented, which clustering method is selected and what are the main parameters that influence the result of this application dependent clustering approach. Moreover, I will present you the clustering validation approach that is implemented in order to evaluate the result of the different clustering algorithms that will be tested in a later step.

5.4.1 Clustering methodology

Type of clustering

Among all the clustering techniques that are explored in § 2.4, the one that will be implemented for the purposes of this this is the agglomerative hierarchical clustering. In general, hierarchical and partitional clustering have key differences in running time, assumptions, input parameters and resultant clusters. The main point that drove me away from partitional clustering is that it requires more assumptions. In particular, I chose hierarchical clustering cause I do not need to define the number of clusters in advance. Since I do not know how many new roads should be formed an automatic way to define this is required. Moreover, hierarchical clustering is more robust to noise and it can handle high dimensional data in contrast with density based clustering techniques which faces difficulties with high dimensional data. Finally, the main drawback of hierarchical clustering is that it is computationally expensive. For our research the roads of a medium sized city seem to be clustered in accepted time.

Clustering of road centerlines based on width measurements

The steps of clustering methodology are explained below:

Find measuring lines of the road centerline by using methodology explained in § 5.1 (step 2). In this step a pre-defined measuring interval is used in order to cut the road centerlines into smaller parts (number of parts indicate the number of the measuring lines for the centerline)

Implement agglomerative hierarchical clustering for measuring lines of a centerline: Start with each measuring line as a separate cluster. Start randomly with a cluster and merge measuring lines that are similar. In order to check similarity between 2 measuring lines we need to define the distance metric (how do we check similarity). This explained in details in next paragraph. After a user-defined distance threshold stop merging measuring lines. Repeat the same process for another initial cluster that has not been merged yet. Figure 48 shows an initial road centerline sub-divided into 5 clusters.

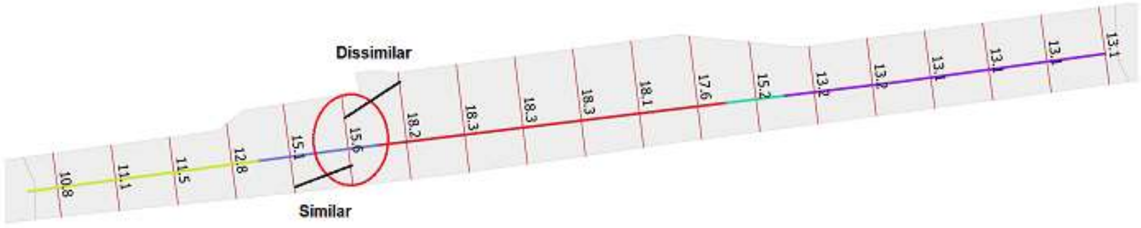


Figure 48 Initial road centerline that represents a road that width is not the same everywhere. After clustering is applied, 5 new roads are created based on similar width measurements.

Parameters that influence the result

Even if hierarchical clustering needs less input parameters than other approaches, it still needs a few things to be defined beforehand. The outcome of this approach in terms of the clusters created can be influenced by some parameters: i) the distance metric (how the distance between points is computed), ii) the linkage method (between which points of the cluster the distance is computed) and the iii) distance threshold (after which distance the clusters will stop merging).

Regarding **distance metric**, for the purposes of this thesis Euclidean distance between measuring lines will be used. The **Euclidean distance** between the measuring lines will be measured **in 2D + width dimension**. The first 2 dimensions correspond to x, y coordinates of the midpoint of a measuring line while the 3rd dimension will be the length of the measuring line (width of the road at a specific part). Figure 50 illustrates this distance metric. According to Tim Bock (2018), when there is no other theoretical justification, Euclidean distance should be preferred, as it is usually the appropriate measure of distance in the physical world.

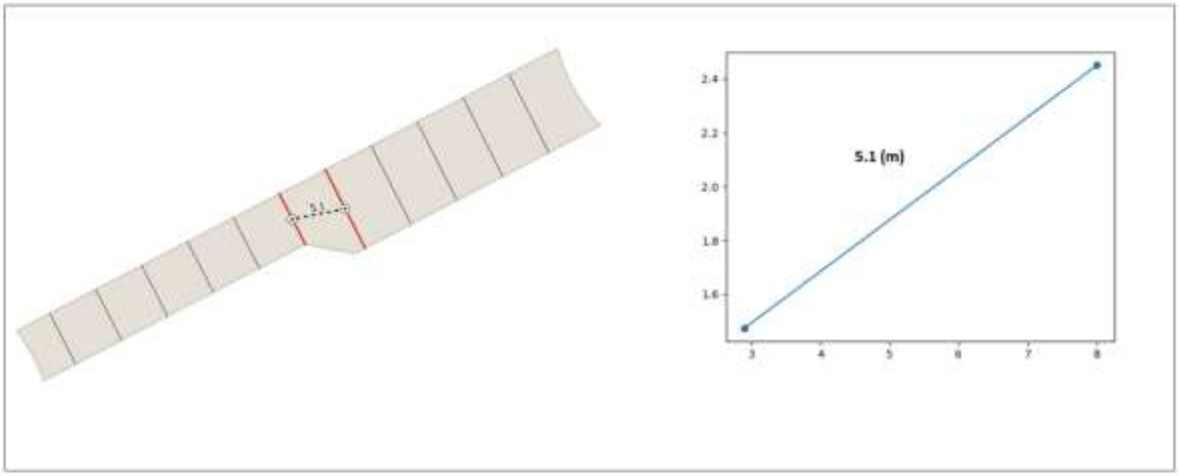


Figure 49 Euclidean distance in 2D space between the midpoints of two successive measuring lines

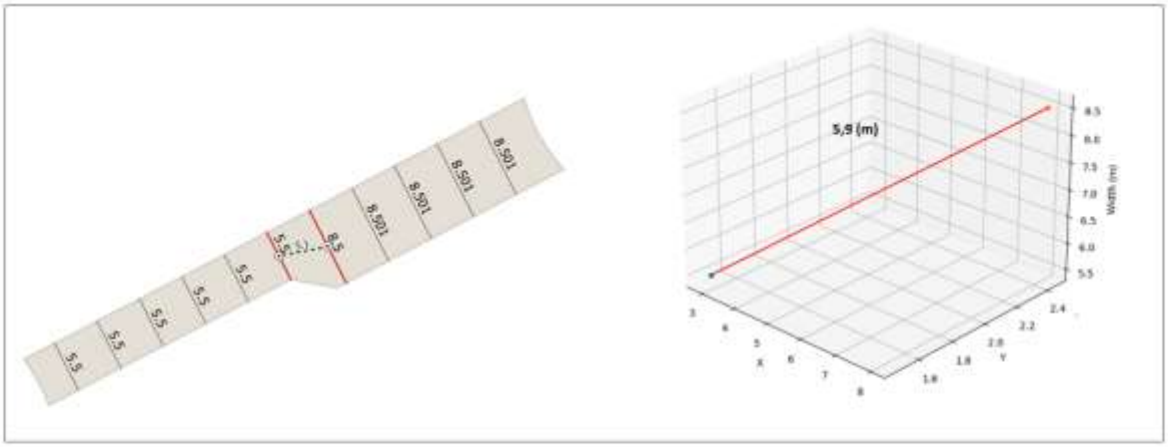


Figure 50 Adding a width as a 3rd dimension to the previous example, the Euclidean distance between the 2 midpoints is increasing. In image the distance was 5.1m. The 2 measuring lines have a difference in width around 3m. The new Euclidean distance that takes width into account is 5.9m

Regarding the **distance threshold**, different options will be examined. Since the distance metric is the Euclidean distance in 2D + width as 3rd dimension, the distance is computed based on the **physical distance** between 2 measuring lines **and the width difference** of them. Main goal is to merge in the same clusters, only successive measuring lines (in order to avoid overlapping of the new road centerlines). The physical distance depends on the interval that each initial centerline will be cut. For example if as interval the 5 meters are chosen, the distance between 2 successive measuring lines is 5 meters.

Thus, since the physical distance (x, y) between measuring lines will be fixed number, the parameter that actually changes the result of clustering is the third dimension (width). Regarding the **width difference** that is accepted in order 2 observations to be merged in the same cluster I will examine different options. In practice, distance threshold influences a lot the result of the clustering. The higher the threshold is, the less sensitive the clustering is to width changes. Figure 51 shows an example of the same centerline that is clustered differently by using different distance thresholds.

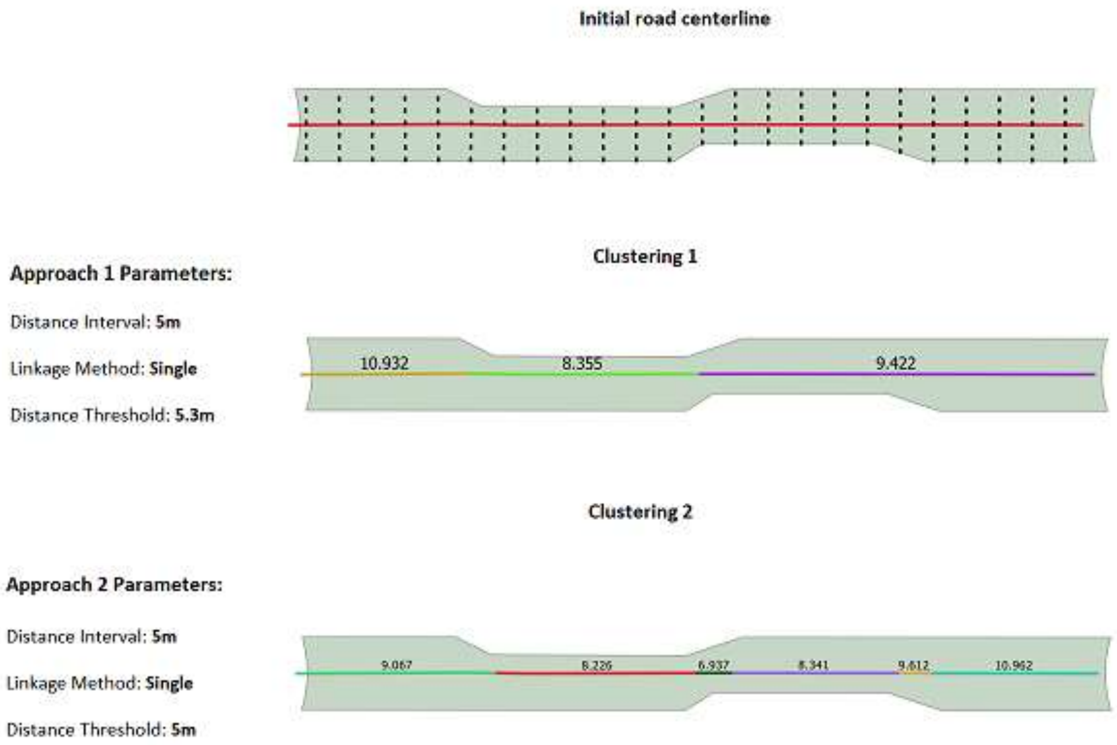
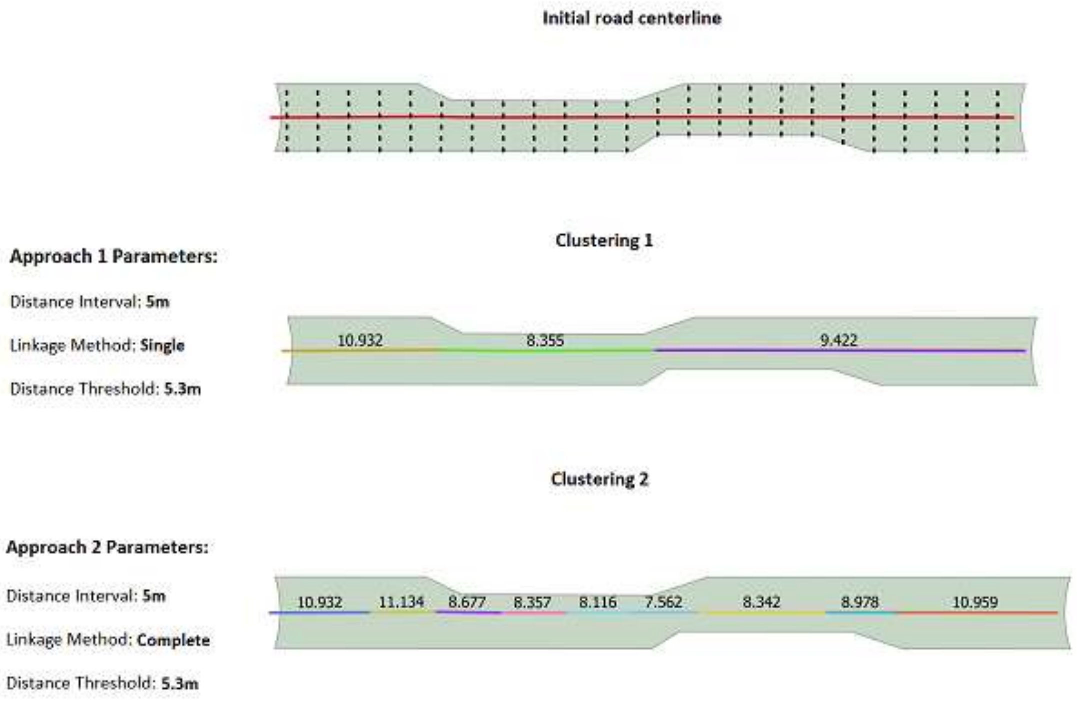


Figure 51 A small change in distance threshold (0.3m) gives us different clustering output. The clustering approach that uses the higher distance threshold results to 3 clusters while the approach that uses smaller threshold (5m) results to 6 clusters. The second approach is more sensitive to changes between the observations.

Regarding **Linkage methods**, they also play an important role in the whole process of hierarchical clustering. Linkage method indicates the rules for clustering (from where distance is computed). In chapter 2 the most well-known methods are discussed. In our methodology different linkage methods will be examined. Figure 52 shows an example of the same centerline that is clustered differently by using the 2 different linkage methods, single and complete. Single method considers as distance the shortest distance between two points in each cluster. That means that the distance computed between the most similar observations between 2 clusters. Complete method considers as distance the biggest distance between two points in each cluster. That means that the distance computed between the most unsimilar observations between 2 clusters. Thus, if all the other parameters are the same the complete method is going to be way more sensitive to differences between the observations.



Approach 1 Parameters:

Distance Interval: **5m**

Linkage Method: **Single**

Distance Threshold: **5.3m**

Approach 2 Parameters:

Distance Interval: **5m**

Linkage Method: **Complete**

Distance Threshold: **5.3m**

Figure 52 The different linkage methods result to different clustering outputs. Single method (approach 1 top) computes the distance between the most similar observation of each cluster. Thus, comparing it to the complete method (approach 2 bottom) which does the opposite we expect that single method is more conservative. The clustering approach 1 (single) results to way less clusters (3) comparing to approach 2 (complete) which results to 9 clusters.

Finally, there is another parameter that influences the result. It has to do with the distance interval that each initial road centerline will be cut (how many measuring lines and where). This parameter will change the number, the density and the position of our observations (measuring lines) so different intervals will be checked for their results. The higher the ‘measuring’ interval is the more sensitive the clustering is to changes between observations. Figure 53 shows an example of the same centerline that is clustered differently by using different ‘measuring’ intervals. Moreover, the choice of the ‘measuring’ interval influences the running time of the algorithm. Higher intervals needs less processing time (less observations).

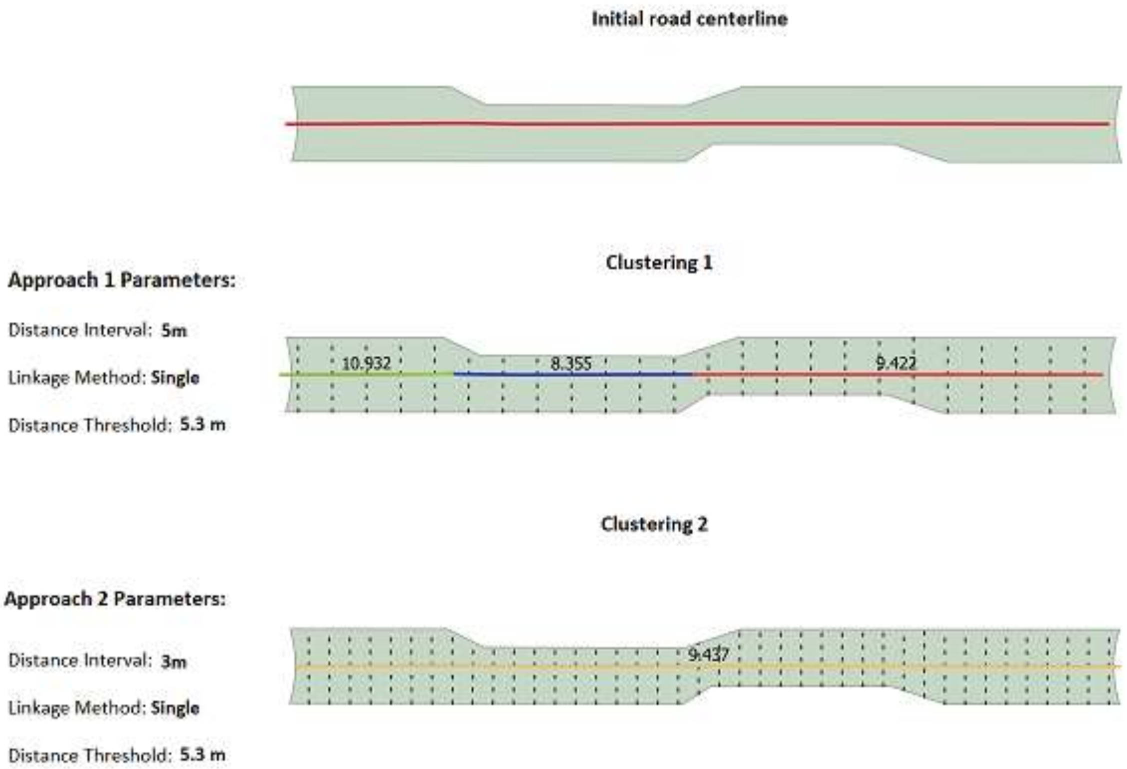


Figure 53 Same centerline is clustered by using different measuring intervals. The result is different

Intersections

As already explained, intersections are special parts of road networks. For this approach clustering will be performed only for road edges. Thus, we will re-create road polygons based on Toronto modelling (explicit intersection modelling). Then, intersection polygons will be identified and excluded from clustering process.

5.4.2 Clustering validation approach

In § 4.2.1, the specific needs of each road user group regarding road clustering was explored. Based on that needs some cases where clustering would benefit road safety applications were defined. Out of the 4 cases 2 will be used for evaluating clustering results. In § 2.4.1 the 2 types of validation techniques for evaluating clustering results where explained. In this approach, output of the clustering algorithm is validated based on external data (ground truth labels) thus, it can be seen as an external validation approach. Some pre-defined clustered centerlines will be used in order to compare with them the results of different clustering approaches.

The main steps of the validation approach are explained below:

i. Find polygons that correspond to roads of the 2 different real-world cases. Centerlines of those polygons will be used as ground truth labels. For case 1 I defined some polygons that show a notable change in their shape (width) and for case 2 I defined some polygons that on-street parking exists.

ii. Those polygons are further split into categories based on their characteristics (geometrical or other).
iii. Define ground truth for each polygon of each case based on some assumptions. The ground truth correspond to the desirable output of clustering for each polygon (based on the assumptions of each case). For example, for case 1 we made the assumption that a notable change in width is a change over 2m. Thus, for each polygon that shows such changes we defined the number and the geometry of clusters (centerlines) that are expected.
iv. Apply the different clustering algorithms and generate their results
v. Explore the relation between width changes that occur in a road and road length. As it reasonably turns out, I found that the longer the road is the more possible a width change to occur is. Specifically, we tested the 100 polygons (50 polygons from dataset of Helsinki and 50 polygons from dataset of Toronto) that show 0 or more changes in their width. Polygons are examined for relation of their length and the number of width changes occur. Then I associated their length and those changes. The results of this test is summarized in Figure 54 and Figure 55.

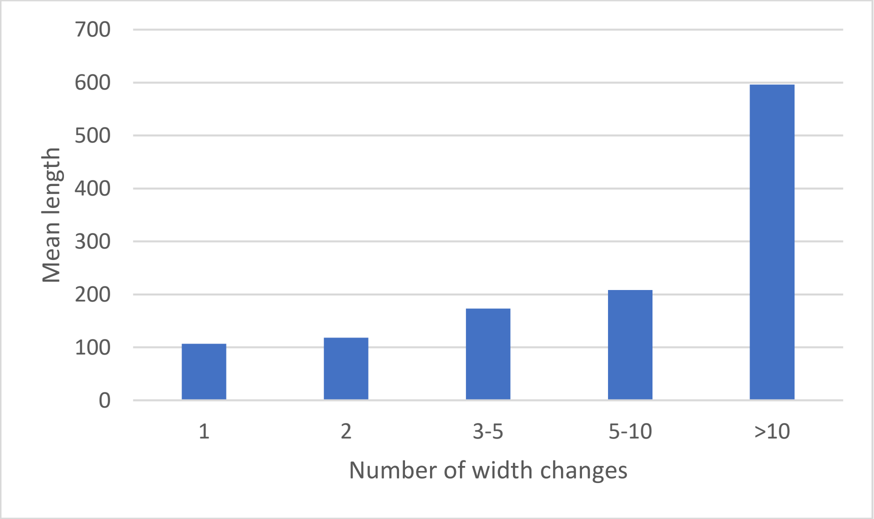


Figure 54 Relation of width changes that occur among roads and mean length of roads

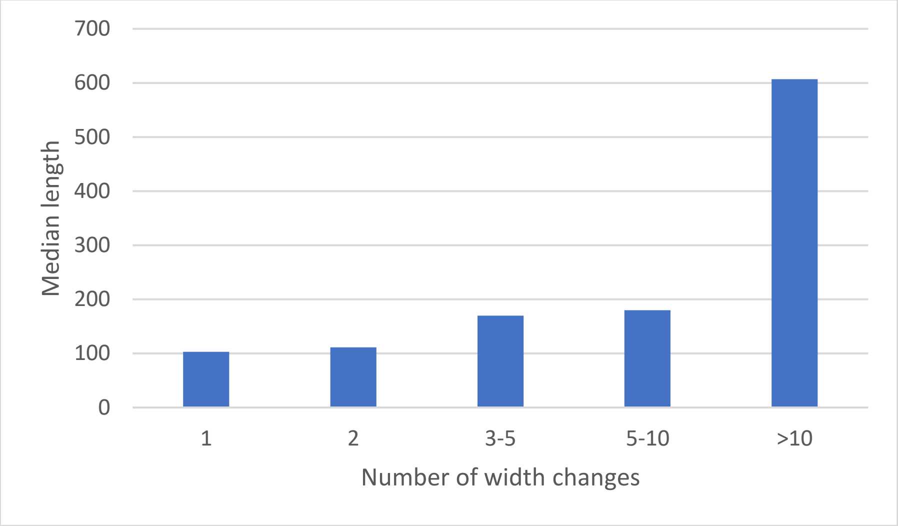


Figure 55 Relation of width changes that occur among roads and median length of roads

vi. Compared the results of the selected clustering approaches with the ground truth labels for each polygon. The clustering algorithms get a score based on some indicators for each polygon. In particular, to find out how close to our ground truth each approach is, I used 3 indicators. Those indicators can get a value between 0 and 1.Those numbers indicate how close the clusters that result from each approach are to the clusters that we defined for ground truth (Where 0 indicates no similarity at all and 1 indicates identical clusters). The first indicator correspond to the **number of clusters** (how many clusters we expect compared to how many clusters result from each clustering approach). The second indicator is related to the **geometry of the clusters** .The third indicator is based on **width statistics** of clusters. I compared some basic width statistics of our ground truth clusters with the basic statistics of clusters that results with the different approaches. Thus, each approach gets 3 values (one for each indicator) for each polygon. Finally, by weighting and aggregating the indicators into a composite index, each approach gets a score for each polygon. The score of a polygon could be from 0 (no similarity at all) to 1 (identical). In the next pages, we will go into more detail on: i) why each indicator is selected, ii) how each indicator gets the final value and how the indicators are weighted and aggregated into a composite index.

Why:

Indicator 1: This indicator compares the number of clusters between ground truth and the clustering approach. The number of clusters it can be consider as an important factor for evaluating process. This number denotes the number of the new roads that is created based on width changes. In some cases we might be interested only in the number of the roads that show changes in width. For example, think of the road de-icing application. In this case, we may need to know the exact number of the wider parts of a road network in order to 'clean' them so that trucks can pass. In practice, other characteristics of roads might be used too since it is a complex application, but the number of 'new' roads is a factor that partially examines the similarity of 2 approaches. Although it is important to consider the number of the new roads, this factor alone is not enough to indicate the similarity of the result of 2 approaches. In practice, it is possible for 2 approaches to result in the same number of clusters, but to have followed a completely different strategy and the resulting clusters to bear no resemblance at all. Figure 56 shows such an example. For this reason, the weight that will be assigned later to this indicator might be lower than the weight of other indicators that may better indicate the similarity of 2 approaches.

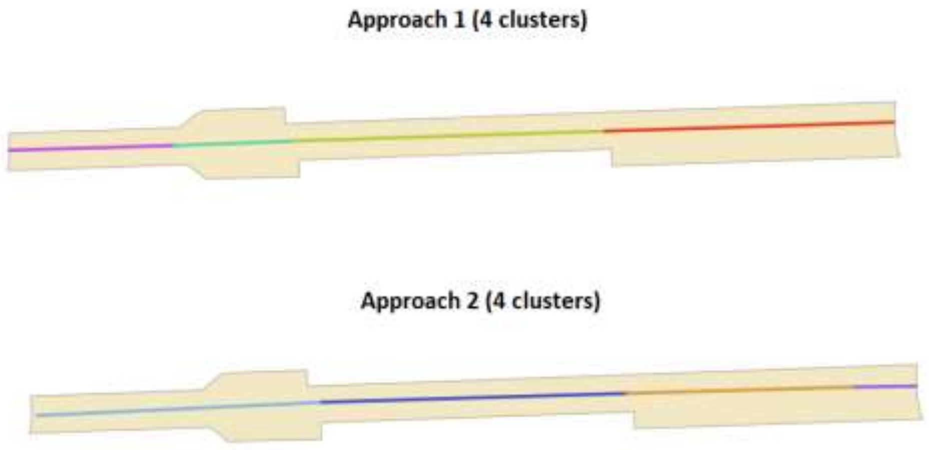


Figure 56 Both approaches are result to the same number of clusters (4). It is obvious that the overall result of clustering between the 2 approaches is different. Only one out of four clusters (green cluster for approach 1 and blue cluster for approach 2) seems to be identical between the 2 approaches.

Indicator 2: This indicator compares the geometry of the clusters of 2 approaches. The geometry can be consider as the most important factor for this evaluation process. The main reason for this is that if the clusters of the 2 approaches show similarity in their geometry, the strategy followed and the overall result of the 2 approaches will show at least some degree of similarity as well. Moreover, we manipulating roads. For most of the applications, the main requirement is to know the geometry of the road. For example, according to Claussen et al. [9] an absolute requirement of the navigation process is the knowledge of the geometry of the road network. Thus, the 'new' roads having similar geometry is quite important. In respect to the above, this indicator will be assigned the highest weight in comparison to other 2.

Indicator 3: This indicator compares the width mean and median values of the clusters of 2 approaches and uses a threshold in order to define whether the 2 clusters have similar width statistics or not. Important to mention that the threshold is not a static value but it is dependent to the length of the compared clusters. Our research, has shown that larger roads are more likely to show some width changes. Thus, the threshold becomes larger (more tolerance) when the compared clusters become longer. This indicator is important cause the main reason for perform the overall clustering approach is the width changes. So, if we have similar number of clusters with similar geometry but with completely different width statistics, it means that the result of clustering is not the desired. Although the width compliance between the clusters is important, as is the case with indicator 1, this indicator alone is not sufficient to show the similarity of the result of 2 approaches. It is possible for 2 approaches to result to clusters with similar width values, but to have followed a completely different strategy and the resulting clusters to bear no resemblance at all. Thus, this indicator will also be assigned with lower weight compared to indicator 2.

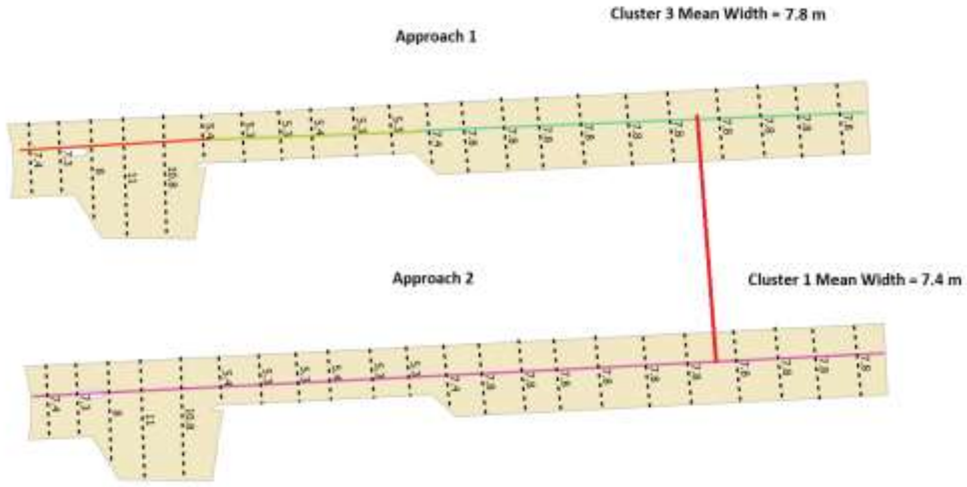


Figure 57 Although the 2 approaches result to different clusters, if we compare the mean width value of 1 and only cluster of approach 2 with the 3rd cluster (green) of approach 1 we will get similar results

<u>*Rules:*</u>

Indicator 1:

The value of indicator can be in the range of 0-1. The value is based on the ratio of the difference of cluster number. Figure 58 illustrates such an example

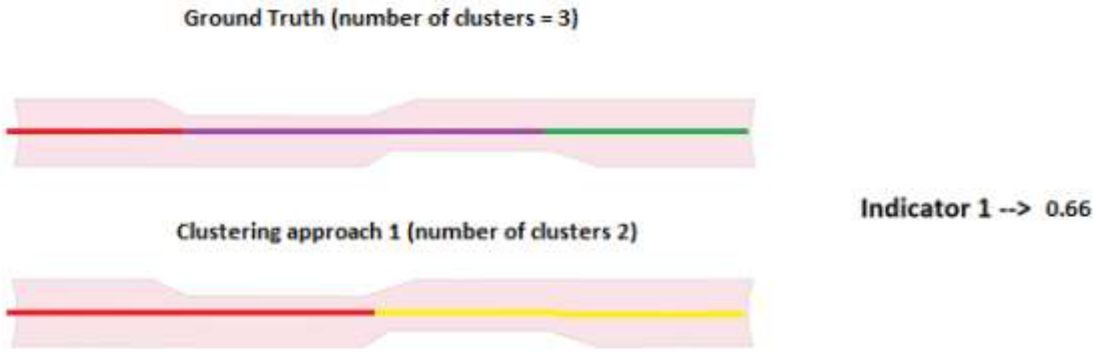
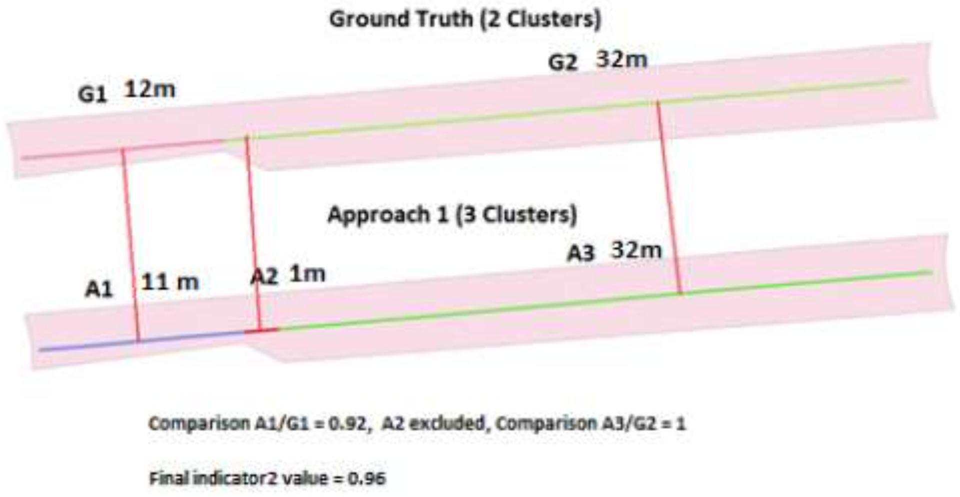

Ground Truth (number of clusters = 3)
Indicator 1 --> 0.66
Clustering approach 1 (number of clusters 2)

Figure 58 Tow clusters result from clustering approach while clusters of our ground truth are 3. Thus, indicator 1 gets the value of 0.66 (2/3)

Indicator 2

The purpose of this indicator is to compare the geometry of the clusters that result from clustering approach with the geometry of the clusters that result from our ground truth. The clustering approach results to a specific number of clusters. This number might be the same or not with the number of clusters that result from our ground truth. In order to define the value for this indicator we have to compare all clusters of the clustering approach with the 'correct' clusters of our ground truth. By 'correct' cluster we mean for each cluster of the clustering approach, the cluster of ground truth that has the biggest intersection with. Each cluster that we compare will get a value in range of 0-1 based on the percentage of the length difference between cluster of approach and cluster of ground truth. Then we combine all the scores and we define the final value (in range of 0-1 as well). Important to mention that from this comparison we exclude the clusters with pretty small length (less than 5m) since we assume that they do not influence a lot the geometrical result. Figure 59 shows one example of a case that ground truth and clustering approach result to different number of clusters and indicator 2 gets a pretty high value in scale of 1.



Comparison A1/G1 = 0.92, A2 excluded, Comparison A3/G2 = 1

Final indicator2 value = 0.96

Figure 59 In that case Ground truth results into 2 clusters (G1 and G2) while Clustering approach results into 3 clusters (A1, A2 and A3). For each one of the clusters of clustering approach we find the cluster of ground truth that has the biggest intersection with. For A1 is G1, for A2 and A3 is G2. Since A2 length is 1m (too low) it is excluded from the overall process. Thus, we have 2 comparisons. We compare A1 with G1 and A3 with G2 for their lengths. For fist comparison, there is a small difference, thus we get the value of 0.92 in scale of 1. For the second comparison we have identical result thus we get final value 1. By combining those 2 values we result to a final value for indicator of 0.96 ((result 1 + result 2) / number of comparisons) out of maximum 1. This, indicates a high degree of similarity between clusters of clustering approach and clusters of ground truth.

Indicator 3 (width statistics)

The purpose of this indicator is to compare the basic statistics of clusters width (mean and median values) between ground truth and clustering approaches. The comparison between the clusters is done similarly to indicator 2 (for each cluster of clustering approach we find the 'correct' cluster of ground truth to compare with). According to our research (step v) the longer a road is the more likely the changes in its width are. Thus, the longer the clusters are the higher the width threshold for comparing clusters is. We need to increase tolerance when we compare long clusters since changes in width are more likely to happen. The numbers for defining the threshold for the different length is derived based on our research (step v). Figure 60 illustrates an example of how indicator 3 gets its final score. In this example only mean width value of the clusters is compared for convenience of explanation. In practice median value also taken into consideration and the final value for each cluster is a combination of 2 values in range of 0-1.

For clusters with total length until **110m** → The value is based on the ratio of width difference. Final value in range of 0-1

For clusters with total length in between **110 m and 210 m** → The value is based on the ratio of width difference. A value of 0.05 is added to the final value (in range of 0-1). This value, indicates a small tolerance added due to the higher length of clusters compared to previous case.

For clusters with total length over **210m** → The value is based on the ratio of width difference. A value of 0.1 is added to the final value (in range of 0-1). This value, indicates the tolerance added due to the higher length of clusters compared to previous cases.

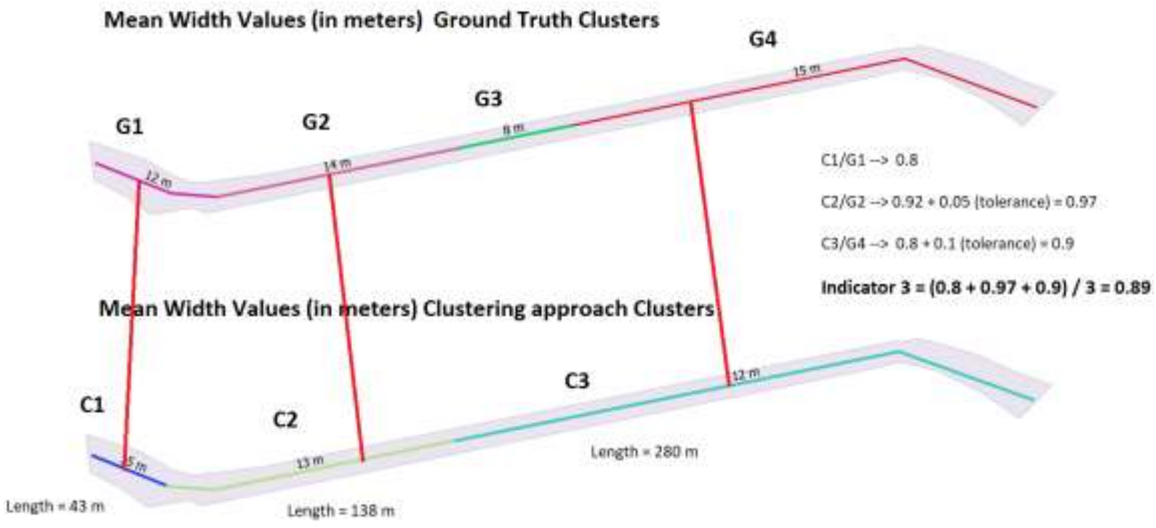

Mean Width Values (in meters) Ground Truth Clusters
G1
G2
G3
G4
12 m
14 m
8 m
15 m
C1/G1 --> 0.8
C2/G2 --> 0.92 + 0.05 (tolerance) = 0.97
C3/G4 --> 0.8 + 0.1 (tolerance) = 0.9
Indicator 3 = (0.8 + 0.97 + 0.9) / 3 = 0.89
Mean Width Values (in meters) Clustering approach Clusters
C1
C2
C3
5 m
13 m
12 m
Length = 43 m
Length = 138 m
Length = 280 m

Figure 60 Example of comparison width values between clusters of ground truth and clusters of clustering approach. For each cluster of approach 1 we find the corresponding cluster of ground truth to compare with (based on length of intersection). Then, based on width value (here we compare only mean value for ease of explanation) of the compared clusters we get a value in range of 0-1. Since length of the clusters is taken into consideration tolerance is added on that final value based on length of the clusters. For example C3 with length of 280m get a plus 0.1 tolerance value. The final value of indicator 3 for this example is formed by the combination of individual values of the different comparisons.

Weights and Aggregation

The weights of 3 indicators are revealed by the relative performance of a set of indicators, which means that indicators with more beneficial impacts on the overall estimation of similarity between 2 approaches are assigned higher weights, and vice versa (Gun et al. 2017). Exploring the reasons of the existence of the 3 indicators we realized that the most important indicator is indicator 2 that compares the geometry of the clusters. This is an indicator that by itself is sufficient to examine the similarity of 2 clusters at a specific degree. The other 2 indicators are complementary, thus confirming the similarity of the compared clusters. Weights are ratio based:

Indicator 1 and 3 → 25%

Indicator 2 → 50%

Regarding the aggregation the widespread additive method that is going to be used is the weighted arithmetic mean. That means, sum up the normalized values of sub-indicators to form a final index. Figure 61 shows an example of how the weighted indicators works to format a final index for a polygon

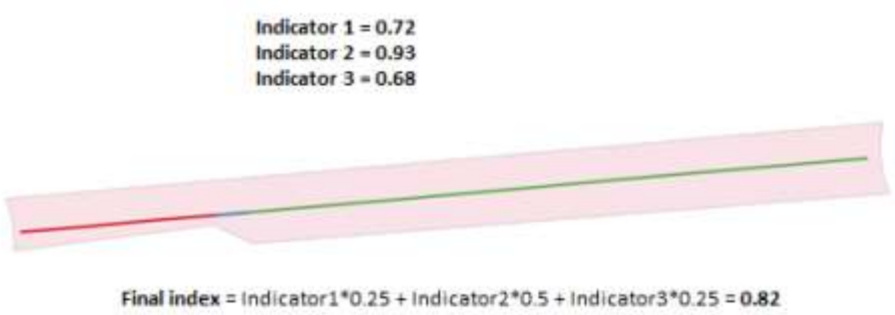

Indicator 1 = 0.72
Indicator 2 = 0.93
Indicator 3 = 0.68
Final index = Indicator1*0.25 + Indicator2*0.5 + Indicator3*0.25 = 0.82

Figure 61 Based on the values of indicators of clustering approach for this polygon, we realize that clusters show quite similar results in their geometry with the clusters of ground truth. Nonetheless, the other 2 indicators indicate that number of clusters and width statistics are not that similar. Based on our weighted approach indicator 2 gets the highest weight and the final index for this polygon is influenced more by the value of indicator 2. If we hadn't followed an weighted approach the final index would have been 0.77 while now is 0.82

vii. Finally, by combining all the separate indexes of the tested polygons, we result to a final score for each approach (Figure 62). Based on the final scores that each approach has for the different categories of the tested polygons (as they defined in step ii), we derived some conclusions on which approach suits best which case.

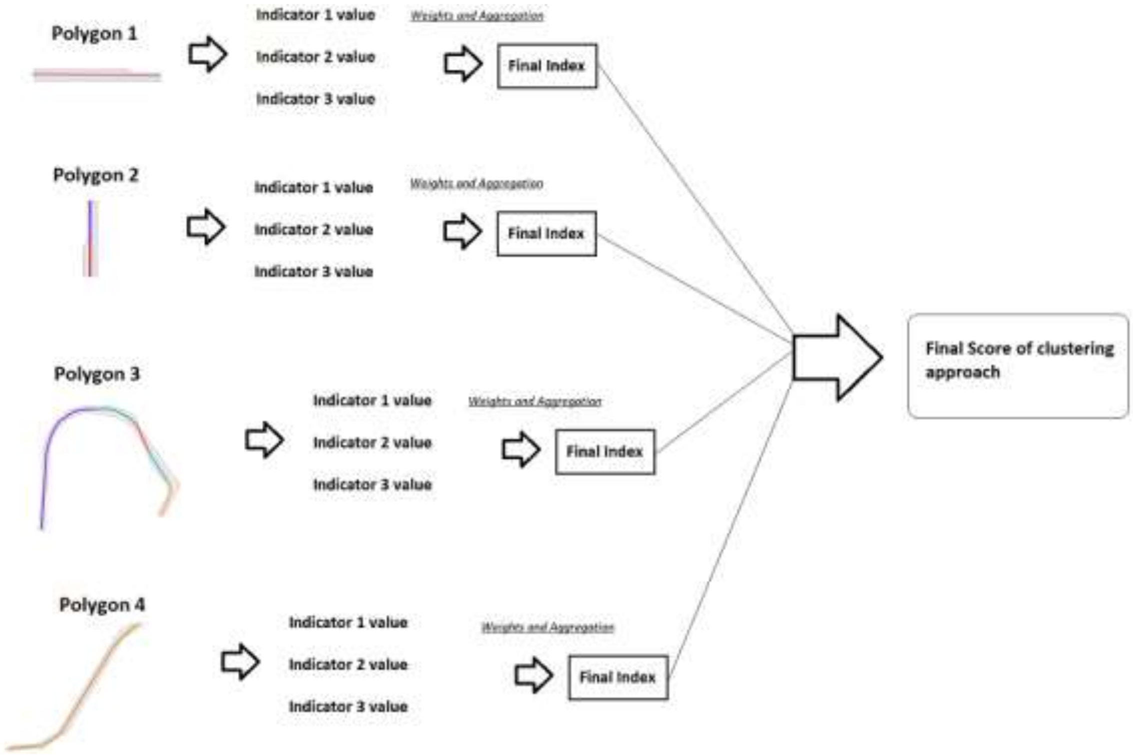


Figure 62 Overall process of result to a final score for a clustering approach.

6 Results & Analysis

In this chapter, various results are presented. Results of initial methodology, standardization process, intersection identification and width clustering approach are analyzed in this section.

6.1 Initial methodology

Before expanding the functionality of the existing methodology that works with vector data, I tested its results. The main goal was to check whether the width measurements resulting from initial methodology correspond to truth width values of roads. To test this, I used 100 manual width measurements as ground truth. As Figure 63 displays, for defining those measurements aerial imagery was used. They are located into 2 different cities (48 measurements in Helsinki and 52 measurements in Poznan). The results are summarized in Table 4.

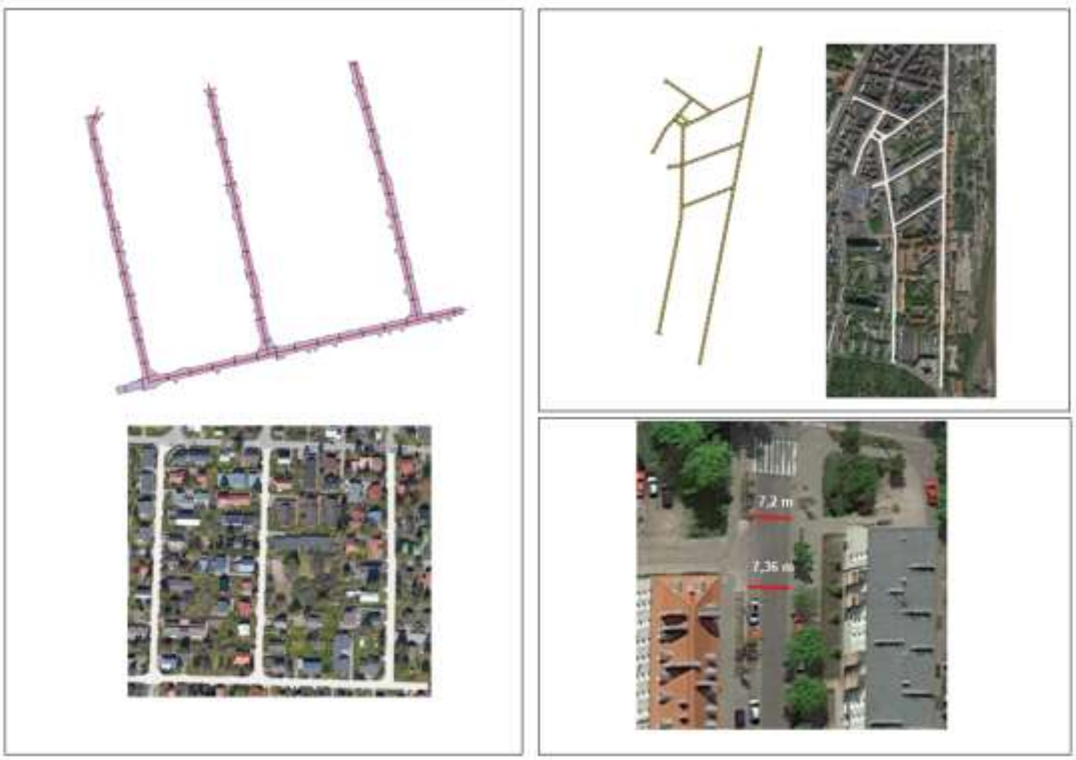


Figure 63: 100 measuring lines was defined as ground truth in order to compared with measuring lines result from initial methodology (top right). 48 measuring lines correspond to roads located in Helsinki (Left), 52 measuring lines correspond to roads located in Poznan (right)

Dataset	Type of measuring	Number of measuring lines (m)	Mean Length of measuring lines (m)	Std around mean (m)	Median length of measuring lines (m)	Range (m)
Helsinki	Initial methodology	48	7.01	1.44	6.9	11
Helsinki	Ground truth	48	7.43	1.3	7.2	8.99
Poznan	Initial methodology	52	7.91	0.71	8.08	4.9
Poznan	Ground truth	52	7.77	0.6	7.97	4.01

Table 4: Comparison of ground truth measurements with measurements resulting from initial width estimation methodology.

From the results summarized in table 4 we can claim that initial methodology results into measuring lines pretty similar to our ground truth measurements.

6.2 Standardization of road vector data

The results of the standardization process of road polygons based on Toronto modelling are presented in this section. First, I investigated how road polygons of four different datasets are affected. For each dataset I defined a tested area corresponding to a radius of 3500 meters around the city center and then, the standardization methodology was applied. The main goal was to explore how the original polygons are affected by the standardization process. Details about the statistics of the polygons before and after the standardization are contained in Appendix A.

Finally, the main goal of standardization process was the creation of a unique and simple polygon to represent each intersection. Thus, 431 intersections that can be found in 2 different datasets (Helsinki, Poznan) were used as ground truth. I checked if an intersection polygon based on Toronto modelling was created for those intersections.

6.2.1 Toronto

Original polygons

The tested area corresponds to 3500 meters around the city center of Toronto, Canada (Figure 64). Many road polygons can be found in the urban area of Toronto (7527 in total). The dataset follows explicit modelling of intersections with one polygon to represent each intersection. More details regarding modelling approach that is followed in this dataset can be found in § 2.1.3. The total number of intersection polygons is 1664 which correspond to 22.1% of the total polygons. The general shape of the polygons is quite regular. The streets in general seem to be rather large. The mean area of the road polygons is around 568 m2 with a large standard deviation around that value (1520 m2). There is a variety in the shape and size of the road polygons with a few pretty small/big polygons.

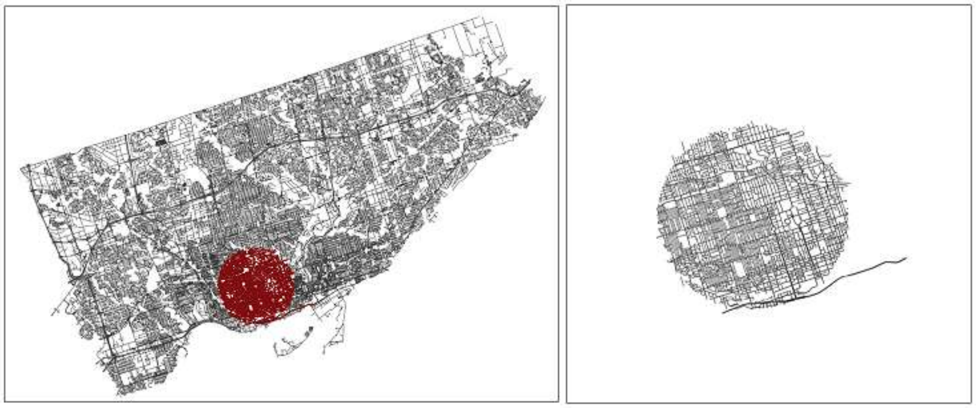


Figure 64 Tested area 3500m around a point at the city center of Toronto

Standardized polygons:

In general, since our methodology is based on the initial modelling of Toronto, the output of the newly created road polygons is quite similar with the initial dataset. The shape of the most polygons is well-preserved. Small changes regarding the size of the polygons can be found between the 2 datasets. As it turns out reasonably, the changes in the values of the mean/median area of the polygons are rather small (0.25% and 12% reduction, after the polygons recreation). The changes in the total number of polygons and in the total number of the intersection polygons are also small (0,2% and 5,3% increase respectively). The **most notable change** is the big reduction of the standard deviation around the mean area of the polygons from 1520 to 688 (54.7% reduction). Figure 65 illustrates the before and after polygons at a part of the tested area.

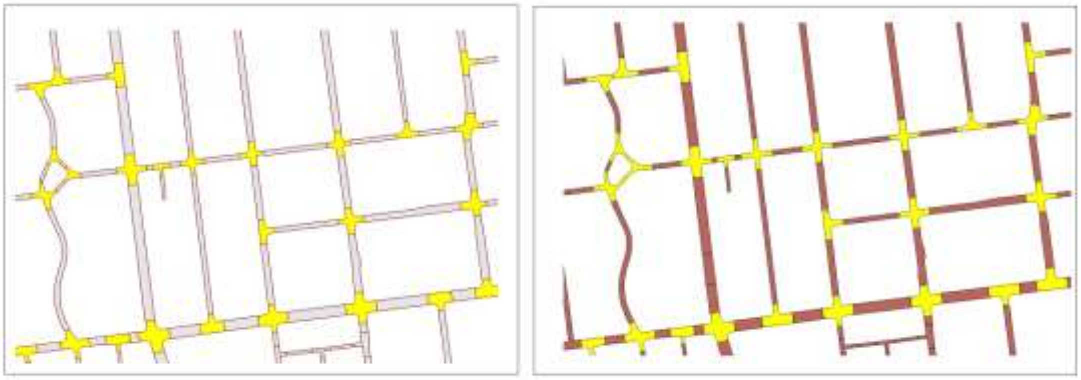


Figure 65 Road polygons of Toronto dataset, before (left) and after (right) standardization process. Since Toronto modelling was the prototype for this standardization the before and after results are quite similar

6.2.2 Helsinki

Original polygons

The tested area corresponds to 3500 meters radius around the city center of Helsinki (Figure 66). There are 5367 road polygons in total. The main road is prioritized by taking most of the intersection area. Thus, there are no intersection polygons. The mean area of the road polygons is around 380 m2 with a large standard deviation around that mean (1313 m2). It is worth to say that there is a big difference between mean area and median area values (median area is 9.3 m2). In this dataset parking spots are modelled either as a part of the road polygon that they belong or as a separate small polygon next to the road polygon that they belong (Figure 67). 2789 polygons correspond to parking polygons and have area less than 15m2. This justifies the big difference between mean and median area values.

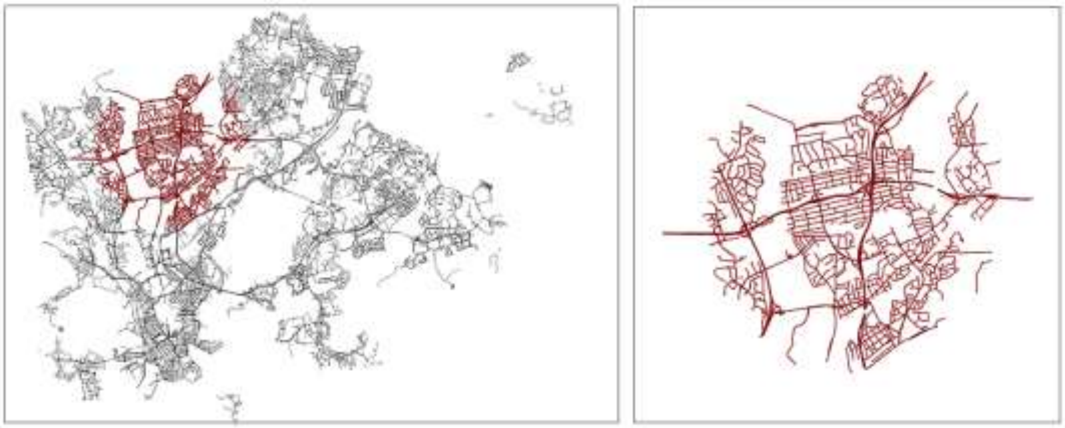


Figure 66 Tested area 3500m around a point at the city center of Helsinki

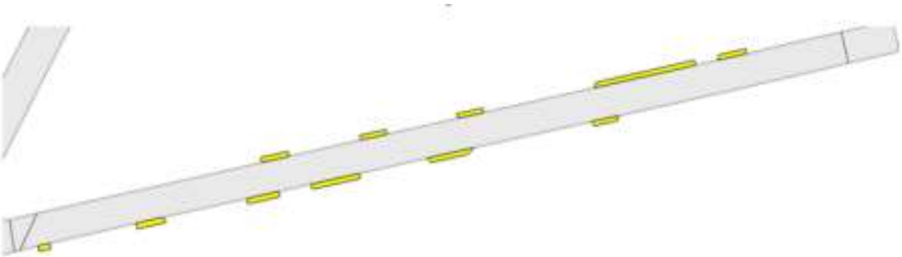


Figure 67 Parking spots modelled explicitly in dataset of Helsinki

Standardized polygons:

Since in our input dataset there is no intersection explicit modelling, the changes with the dataset that results after the standardization of the road polygons are essential. In total, 1043 new intersection polygons are created Those new polygons correspond to 19% of the total polygons. The most notable changes are : i) A quite large reduction in the standard deviation around the mean area value (20.5%) and a big change in the median area value (from 9.3m2 to 41.1 m2). The latter might be affected by the small polygons that are modelled in the initial dataset as mentioned before. These small polygons are removed after the standardization process. Figure 68 illustrates an example of the result of road polygons re-creation at a part of the city of Helsinki.

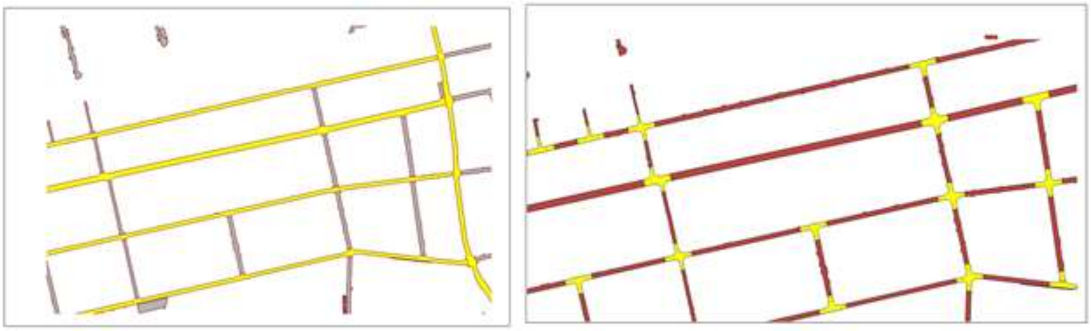


Figure 68 Road polygons of Helsinki dataset, before (left) and after (right) standardization process. Since Toronto modelling was the prototype for this standardization and the initial polygons of Helsinki follow a different strategy (no explicit intersection modelling) the change are notable. Newly created intersection polygons are present after the standardization process

6.2.3 Poznan

Original polygons

The tested area corresponds to 3500 meters radius around the city center of the city of Poznan (Figure 69). There are 2569 road polygons in total. This is a rather small number compared to the 3 other datasets that were examined. By looking at the tested area, we can realize that indeed the road polygons are more sparse in this dataset. Moreover, we observed that in the dataset can be found some polygons that do not correspond to road polygons. Their shape is rather large, and no centerline is passing through them. From aerial imagery it turns out that these polygons correspond to parking spaces in the city of Poznan. Unsimilar to Helsinki dataset, Poznan models the parking areas that can be found in the city and not the parking spots that are present on a road. Figure 70 shows an example of those polygons. Due to the explicit modelling of parking spaces in the city of Poznan, it is reasonable that the mean area of the polygons is quite large in comparison to the other datasets. Specifically, the mean area of the polygons is 1358 m2. Moreover, the standard deviation around that mean is 3043 m2. Important to mention that the median area value is quite lower and it makes more sense in terms of road polygons area (320 m2). Finally, the intersections are not modelled explicitly so there are no intersection polygons in the initial dataset.

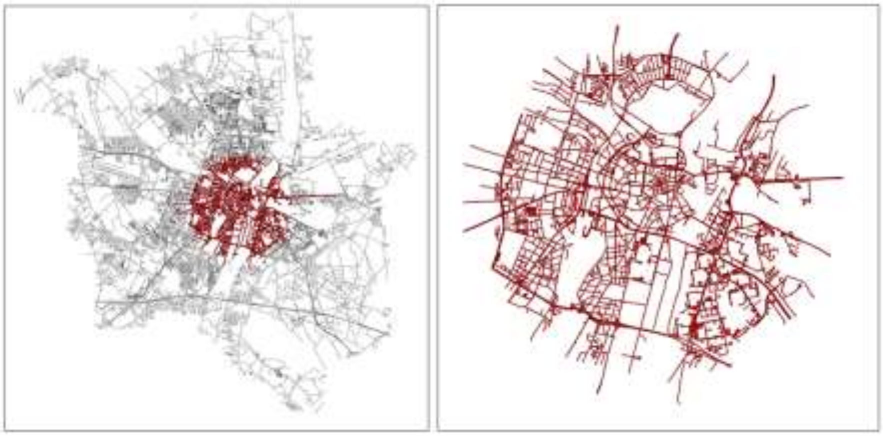


Figure 69 Tested area of Poznan (3500 meters radius around the city center)

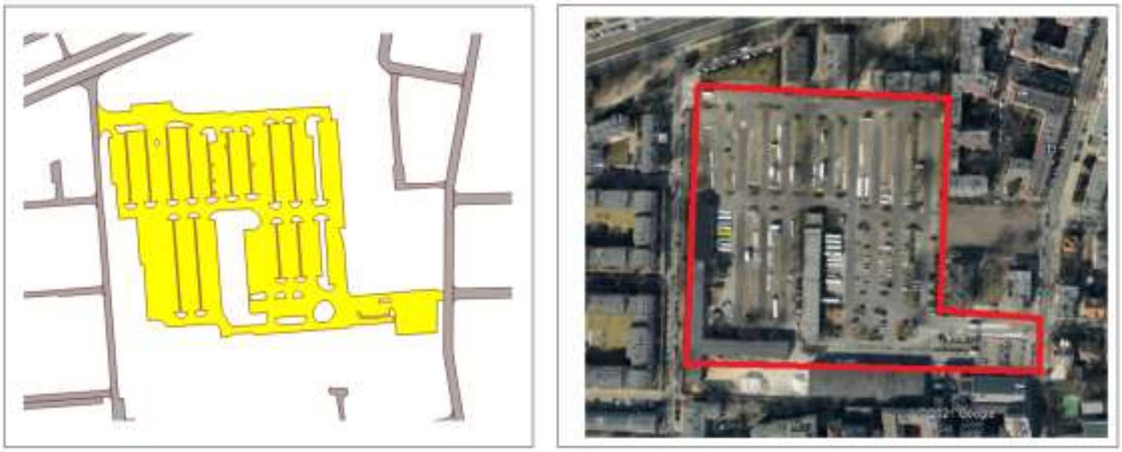


Figure 70 Explicit modelling of parking spaces in the city of Poznan. The polygon in the left image can be found in the dataset. This polygon is rather large with area 4989 m2 and does not correspond to a road polygon. As we can see in the right image it represents a parking space.

<u>Standardized polygons:</u>

As expected, the changes of the road polygons after standardization process are significant. The number of total polygons is increased a lot (almost 90%) while 1128 new intersection polygons are created (28% of the total road polygons). Moreover, a large reduction in the mean area value (47.3%) and to the standard deviation around that mean (60%) was observed. This is due to the removal of parking areas after the standardization process. Since no road centerline passes through those areas they are not identified as road polygons by our methodology. Reasonably turns out that, the number of the pretty large road polygons is significantly reduced after the re-creation (from 19% of the total to 7%). Figure 71 illustrates an example of how the road polygons are standardized in the city of Poznan.

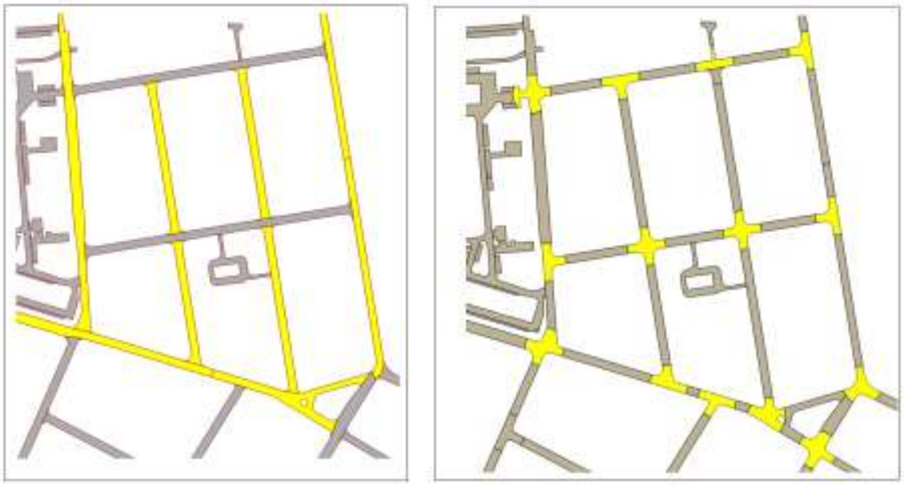


Figure 71 Road polygons of Helsinki dataset, before (left) and after (right) standardization process. Since Toronto modelling was the prototype for this standardization and the initial polygons of Poznan follow a different strategy (no explicit intersection modelling) the changes are notable. Newly created intersection polygons are present after the standardization process

6.2.4 Montreal

Original polygons

The tested area corresponds to 3500m radius around the city center of the city of Montreal (Figure 72). There are 22777 road polygons in total. This a rather large number of road polygons compared with the 3 other dataset that we examined so far. This is due to two reasons. First, the sidewalks are modelled in the initial dataset of Montreal. Secondly, explicit modelling of intersections but in a different way comparing with Toronto exists in this dataset. In Montreal more than 1 polygon correspond to the intersection area. Figure 73 illustrates an example of how cross and T intersections are modelled in dataset of Montreal.

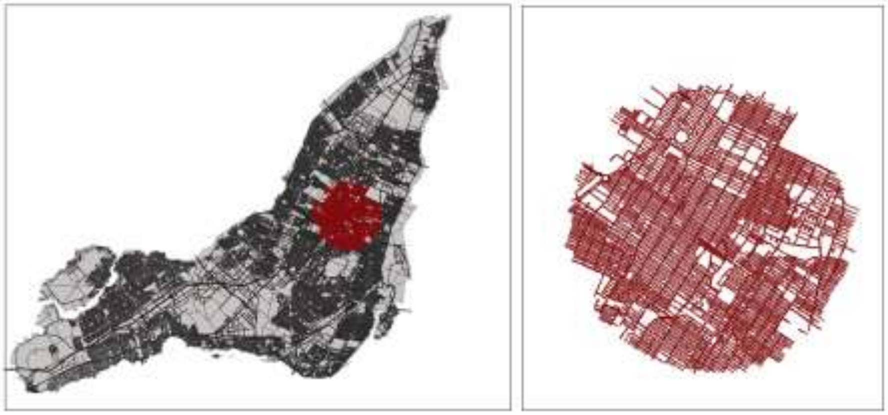


Figure 72 Tested area of Montreal (3500 meters radius around the city center)

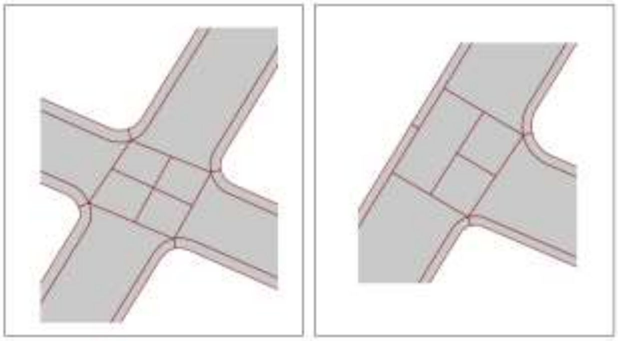


Figure 73 A cross intersection is modelled with 4 different polygons while T intersection is modelled with 3 different polygons in the dataset of Montreal

In general, there number of intersection polygons correspond to the 28,7% of the total road polygons which is a rather large ratio comparing to previous datasets. The mean area of the polygons is 337 m2 with a standard deviation 554,2 m2 around that mean. The median area is quite smaller (118 m2).

Standardized polygons:

In general, a big reduction in the total number of intersection polygons was observed. This is due to the fact that more than 1 polygon was used to represent intersections before, while after the re-creation only 1 polygon is used. Thus, the fact that the number of intersection polygons is also reduced a lot (75%) is reasonable. There is a notable increase in the mean and median area values of the polygons after the re-creation (46% and 87% respectively). This is reasonable as well, since the small intersection polygons are replaced by a bigger intersection polygon. Finally, we have a small increase of the standard deviation around the mean area value (12.7%).

Figure 74 illustrates an example of polygons as they can be found in the initial dataset and after the standardization based on Toronto modelling.

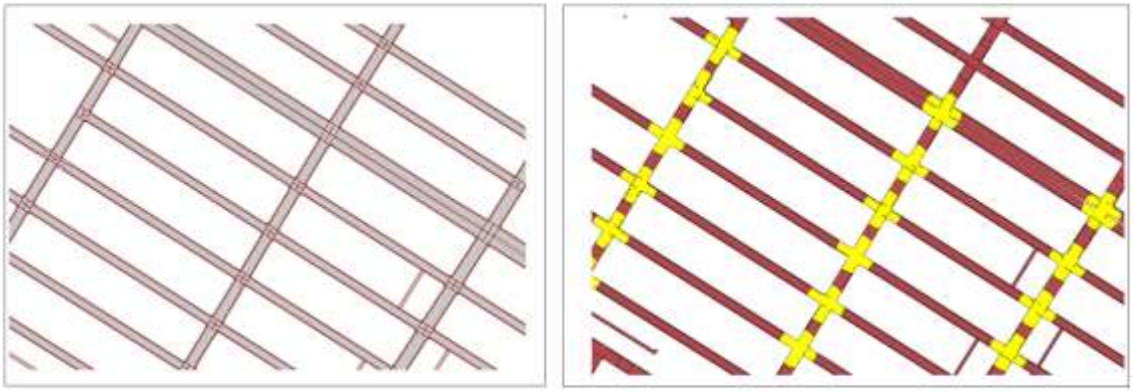


Figure 74 Road polygons of Montreal dataset, before (left) and after (right) standardization process. Montreal also models the intersections explicitly but it follows a different strategy compared to Toronto. Thus, there are notable changes after the standardization of road polygons process.

6.2.5 Intersection polygons creation

The overall purpose of standardization process was to create a unique and simple intersection polygon where an intersection exist. In order to explore how well our approach works I used 431 intersections as ground truth. Out of 431 intersections in total, 398 intersection polygons where created based on Toronto modelling while 33 intersection polygons are completely missing (Figure 75). In particular, 298 intersection were located in Helsinki (Finland) and 133 intersection were located in Poznan (Poland). 290 intersection polygons were created correctly for the dataset of Helsinki (97.3%) while 108 intersection polygons were created correctly for the dataset of Poznan (81.2%).

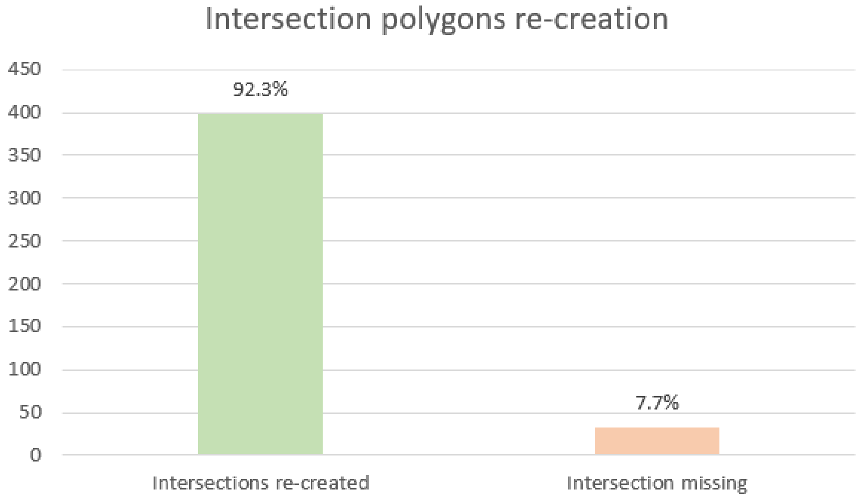


Figure 75 Results of re-created unique and simple intersection polygons in the 2 tested areas. Out of 431 intersections that exist in total, 398 intersection polygons were re-created correctly.

6.2.6 General conclusions

In general we can claim that our goal to standardize modelling of road vector data based on Toronto modelling is achieved. For the ordinary road cases of all the datasets, we achieved the explicit modelling of intersections. The types of the newly created intersection polygons have pretty similar structure to the prototype intersection polygons that can be found in Toronto dataset.

There were some notable changes for all the datasets after the polygon re-creation. As expected, the number of polygons before and after the standardization depends on the initial way of modelling. In datasets that there was no explicit modelling of intersections, reasonably the total number of polygons was increased. On the other hand, in Montreal dataset where an intersection was modelled with more than one polygon, the total number of polygons was decreased.

Important to mention, that in some datasets other features of road networks were modelled explicitly (on-street parking Helsinki, parking areas Poznan). Our standardization methodology does not creates a new polygon for those road features. Thus, it is important to consider that before applying this methodology (risk of losing useful information). Regarding size of polygons before and after standardization, the changes also relying on the initial structure of the polygons. For instance, datasets that include small polygons for modelling special parts of roads (Montreal uses some pretty small polygons to model each intersection) faced a notable decrease in the mean and median areas of the polygons. A common point for 3 out of 4 datasets was the big reduction of the standard deviation around the mean value of polygons area. We need to mention that all the statistics are strongly related with the buffer size that is used for re-creating the intersection polygons. As explained in methodology, the higher the size of the of the buffer the larger the intersection polygons and the smaller the non-intersection polygons that are created.

As already mentioned, our methodology is not applied to motorways. In motorways, road centerlines are modelled in such a way that 3 or more centerlines might meet at point that do not correspond to an intersection. Our methodology will consider those points as intersection nodes since they have a degree higher than 2. Thus, intersection polygons will be created at those points. This will be rather problematic. Figure 76 illustrates an example of what will happened if we try to apply our methodology to motorways.

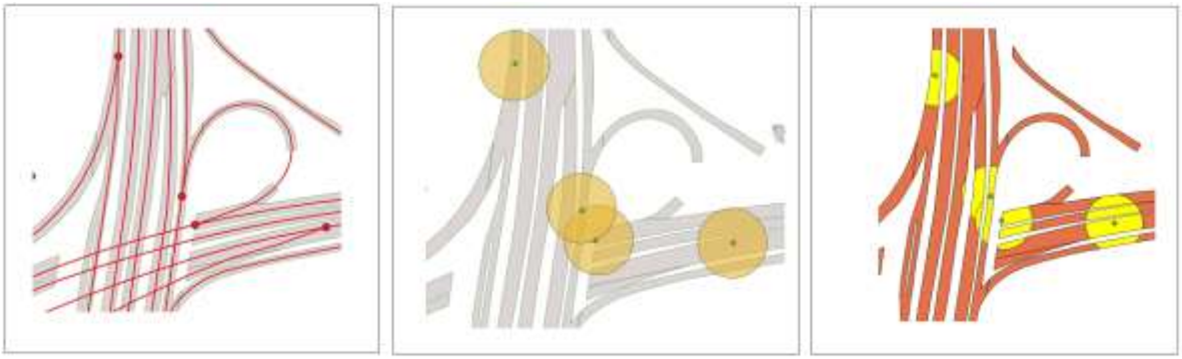


Figure 76 Nodes of motorways that have a degree higher than 2 without being intersection nodes (left). Our methodology will consider those nodes as intersection nodes. Thus buffer will be created (middle). This, will result to undesirable intersection polygons (right).

Although our methodology creates intersection polygons similar to prototype intersections, for some complex intersections where many roads meet or have a rather unusual structure, the results may not always be desirable. Since, sometimes there is no pattern followed for complex intersection by Toronto modelling (which is reasonable since roads are unique and they have a quite complex structure) there is no prototype for some cases of road network. Figure 77 shows some of such examples.

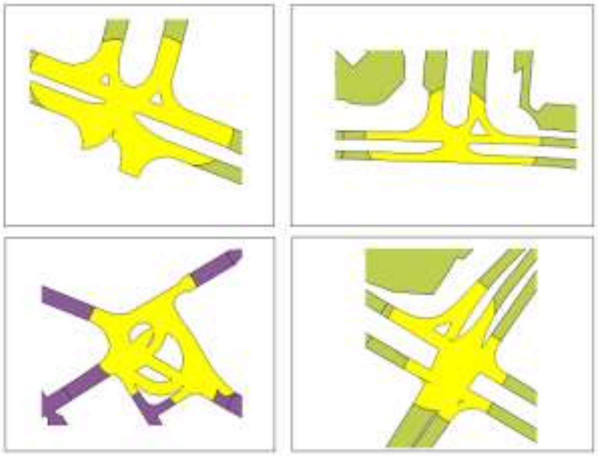


Figure 77 Complex intersections

6.3 Intersection identification

In order to examine and evaluate the identification of the different intersection types I used various ground truth labels. I tested the approach for identifying 3 different intersection types (T, Cross and X) at 2 different sample areas. After standardization, I have defined the number of each intersection type that can be found in the datasets. Then, I examine the results.

T intersections

A sample area in the city of Toronto is tested for methodology of identifying T intersections. The polygons of the area was re-created based on our standardization methodology as explained in § 6.2.

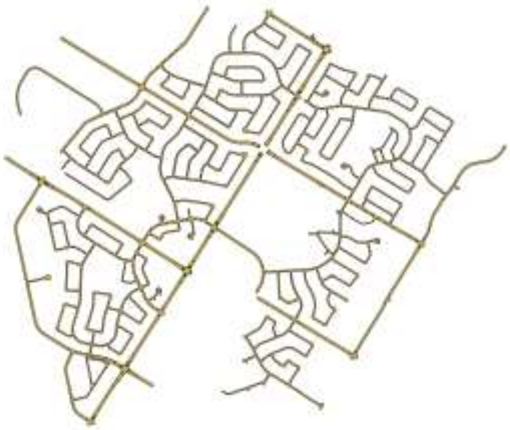


Figure 78 Sample area that used for testing the identification of T intersection polygons

Total Number of T intersections: **298**

Number of T intersections identified correctly: **287 (96.3%)**

Number of T intersections that not identified: **11 (3,7%)**

The 2 main reasons of the non-identification of 11 T intersections are shown in Figure 79. Left, a T intersection has not its usual shape. The wider part of the road polygon is rounded and the measurements in the end/start of the road does not seem to have big different with the rest measurements of the road. Figure 79 right indicates a T intersection that the perpendicular branch has a pretty small width. Thus, not too much measurements lying at that part. The 'wrong' measurements are simply identified as outliers.

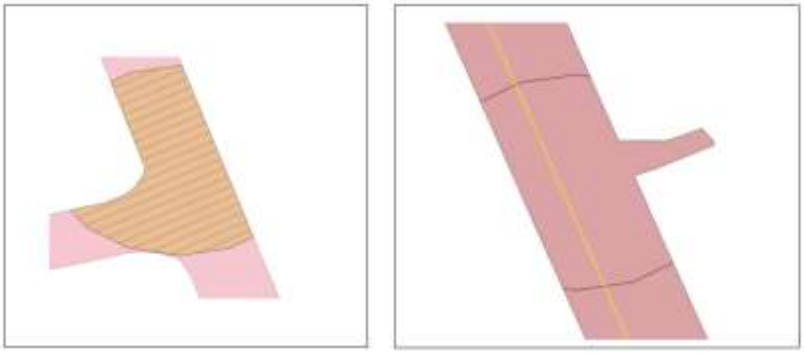


Figure 79 Two typical examples of T intersections that not identified

Number of other road cases that by mistake identified as T intersections: **4**

3 out of 4 cases that by mistake were identified as T intersections correspond to road polygons with round ending (Figure 80 right). 1 case correspond to a road polygon that represents a road turn (Figure 80 left). In those cases, the width measurements in a specific location of the road (in the middle for the turn polygons and in the end for the round ending polygons) are quite larger than the measurements in the rest of the road. Thus, incorrectly identified as T intersections.

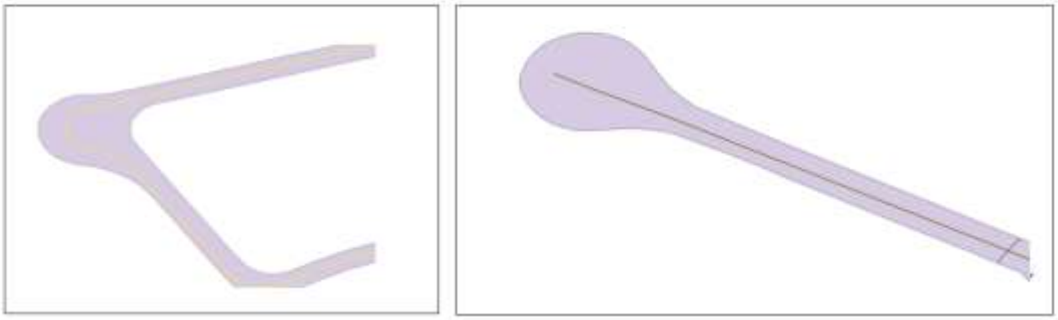


Figure 80 Typical examples of road edges that identified as T intersections

,

Cross/ X intersections

The methodology for identifying these two intersection types is similar (only the last step distinguishing between the 2 types). Thus, they tested in the same sample area for their results

(Figure 81). In addition, this allows us to explore the possible identification of one intersection as a type of the other. The polygons of the sample area re-created based on standardization process developed during this thesis.

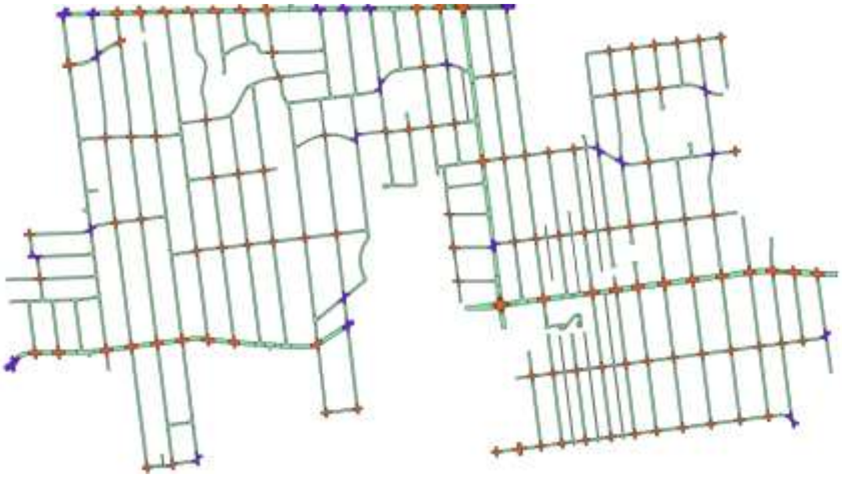


Figure 81 Sample area contain the intersection polygons that identification approach was tested with. Cross intersection polygons in orange color and X intersection polygons in blue color.

Total number of Cross intersections: **108**

Number of Cross intersections identified correctly: **91 (84.3%)**

Number of Cross intersection identified as X intersections: **3 (2.8%)**

Number of Cross intersections that not identified either as cross or as X: **14 (12.9%)**

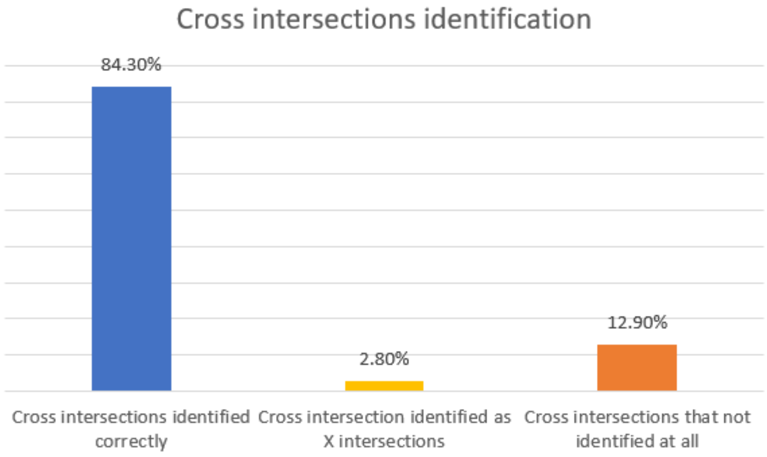


Figure 82 Identification of Cross intersections

Total number of X intersections: **25**

Number of X intersections identified correctly: **20 (80%)**

Number of X intersection identified as Cross intersections: **2 (8%)**

Number of X intersections that not identified either as x or as Cross: **3 (12%)**

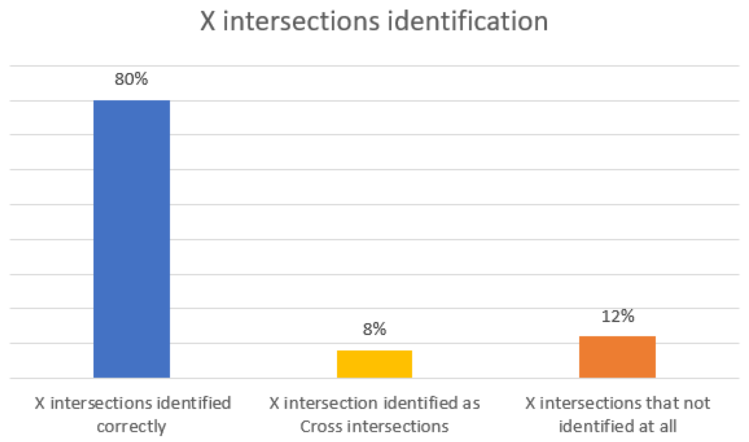


Figure 83

The main reason for the non-identification of a cross or an X intersection was the absence of an X or a Cross polygon. This is due to results of our standardization methodology.

6.3.1 General conclusions

Our methodology was tested for identifying 3 of the main intersection types. In general we can claim that the results were quite promising. Out of 431 intersections used as ground truth, 398 was correctly identified (92.3%). The results were better for T intersections compared to Cross and X intersections. Few polygons were incorrectly identified as intersections. The methodology is still limited on not identified some of the main intersection types (Y, double T are identified as “another type”) while all the intersections where more than 4 roads are met are identified as “another type” intersections.

6.4 Width Clustering

As was discussed in § 4.2, width clustering is expected to have a double positive impact to road safety management application. On the one hand, it will meet the needs of different road users regarding knowledge of significant width changes. In addition, it will add further meaning to the overall process of correlating road accidents with the width of the road. Thus the results of width clustering will be analyzed into 2 different sections. First, I will explore if width clustering can be used to identify significant width changes of roads in terms of different clusters. 2 real-world cases based on the needs of road users and 40 polygons in total were used as ground truth for that process. Then, I will conduct a test into roads of central area of Toronto to find out if clustered centerlines are more representative in terms of their geometry compared to original centerlines.

6.4.1 Identify width changes

First I explored whether width clustering can be used for identifying width changes of roads that are important for the safety of some road users. In particular, 2 real-world cases as they result from the needs of different road user groups (see § 4.2.1) are examined for their results. Main goal of this section is to check whether clustering can produce new roads (in terms of separate clusters) when an important width change in a road occur. The results of 2 different clustering approaches (different parameters see § 5.4.1) are tested and compared with some ground truth clusters (external clustering validation see § 2.4.4) in order to validate which clustering approach suits best which case. More details for the validation approach that is used for this comparison can be found in § 5.4.2.

6.4.1.1 Case 1 – Notable width change

Case 1 correspond to road polygons that show at least one notable change in their width (see § 4.2.1). For this case we made the assumption that a notable change in width is a **change of 2m or more**. Based on this assumption we were able to define our ground truth for each polygon that is selected for evaluating the different clustering approaches. 20 polygons that can be found in dataset of Helsinki and that show some change in their width are selected. It should be noted that the changes in the width of the selected polygons for this case concern changes in the driving space of the road. These changes have been caused only by widening / narrowing of the already existing lanes without any other interference (possible obstacle, tree line, increase / decrease in number of road lanes). Figure 84 shows an example of a selected road polygon for that case and how our ground truth is defined.

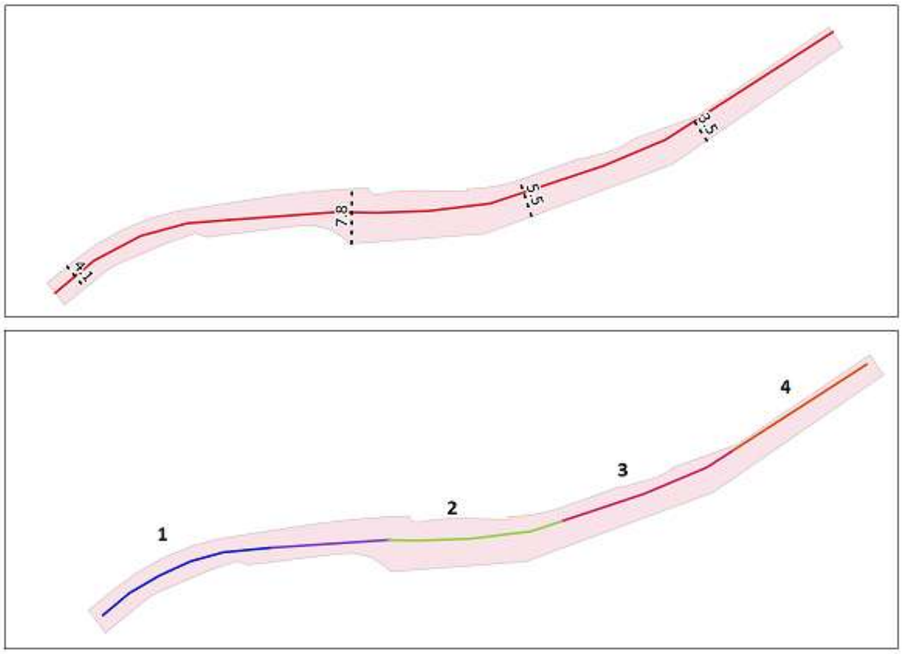


Figure 84 Road polygon with notable changes in its width (Black doted lines shows width at specific locations) and the initial linear representation (Top). The desirable output of clustering methodology based on the assumption that a change over 2m is a notable change in width (bottom). 4 new roads are created based on width of polygon at specific locations.

The selected polygons for this case are further split into some categories. The first separation has been made based on the shape of the polygons into 2 categories: a) curved and b) straight. Then we further subdivide those 2 categories based on the way that the widening/narrowing is taking place. Thus, we define some polygons that contain a clear point (or more than 1) where the change in width takes place while there are others where the change in width takes place in a more gradual pace. Figure 85 indicate some examples of those categories. From the total 20 selected polygons for this case, each polygon category consists of 5 polygons.

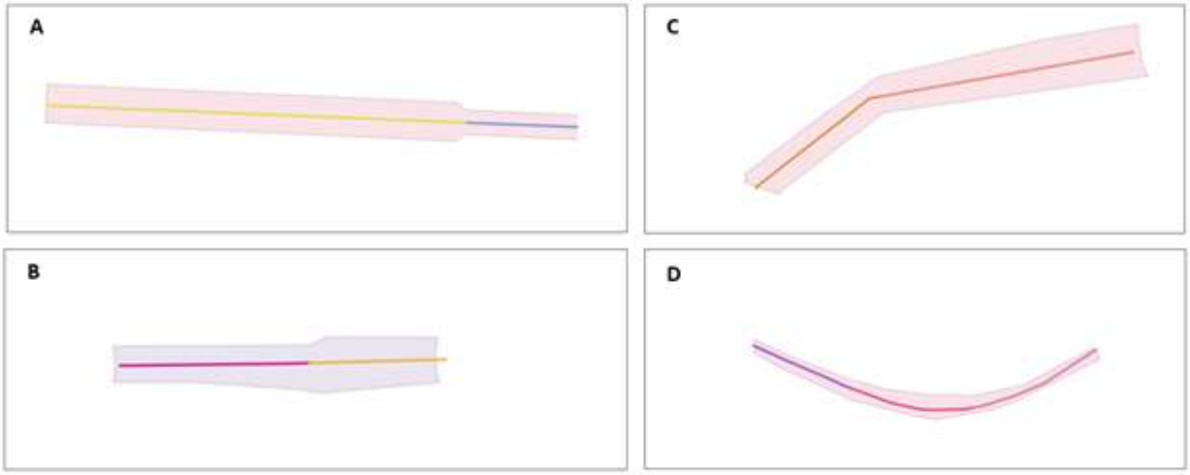


Figure 85 Four categories of selected polygons of this case. Polygons of category A are straight polygons that have a clear point that the change in width takes place. Polygons of category B are straight polygons width gradual change in width. Polygons of category C are curved polygons with clear point of width change and finally, polygons of category D are curved polygons with gradual change in with.

Tow approaches are analyzed for their results for that case. The specific parameters of the 2 approaches are summarized in Table 5.

Approach	Linkage method	Measuring Interval (m)	Distance Threshold (m)
1	Single	5	5.5
2	Average	5	15

Table 5 Different parameters of the 2 approaches that give the most promising results for case 1

The results for each category of the polygons based on the score that each clustering approach gets based on the 3 indicators as they explained in § 5.4.2 are summarized at Table 6 and Table 7.

Approach	Total score Indicator 1	Number of polygons that indicator1 < 0.5	Number of polygons that indicator1 > 0.9	Total score Indicator 2	Number of polygons that indicator2 < 0.5	Number of polygons that indicator2 > 0.9	Total score Indicator 3	Number of polygons that indicator3 < 0.5	Number of polygons that indicator3 > 0.9	Total score
1	**0.8**	1	10	**0.78**	4	11	**0.96**	0	16	**0.83**
2	**0.63**	4	5	**0.59**	4	3	**0.9**	0	18	**0.7**

Table 6

Approach	Polygon Category (5 Polygons)	Total score indicator 1	Number of polygons that indicator1 < 0.5	Number of polygons that indicator1 > 0.9	Total score indicator 2	Number of polygons that indicator2 < 0.5	Number of polygons that indicator2 > 0.9	Total score indicator 3	Number of polygons that indicator3 < 0.5	Number of polygons that indicator3 > 0.9	Final index score
1	A	**0.93**	0	4	**0.96**	0	5	**0.99**	0	5	**0.97**
1	B	**0.87**	0	3	**0.87**	0	3	**0.97**	0	4	**0.89**
1	C	**0.65**	0	1	**0.66**	2	1	**0.94**	0	4	**0.73**
1	D	**0.78**	1	2	**0.65**	2	2	**0.94**	0	3	**0.76**
2	A	**0.75**	0	2	**0.6**	1	1	**0.95**	0	5	**0.71**
2	B	**0.71**	0	2	**0.63**	0	1	**0.93**	0	5	**0.71**
2	C	**0.63**	1	1	**0.64**	1	1	**0.9**	0	5	**0.7**
2	D	**0.45**	3	0	**0.5**	2	0	**0.92**	0	3	**0.59**

Table 7

Results of approach 1

Approach 1 uses the single method as linkage method. That means that it compares the most similar observations between clusters in order to compute the distance. Moreover, the distance threshold for this approach is 5.5m. By analyzing the results of this approach based on the final index that it gets from the aggregation of 3 weighted indicators it can be claimed that is an acceptable approach that gives results that show a pretty high degree of similarity with our ground truth. Final index for this approach is 0.83 (max 1). All the 3 indicators have a similarly quite high total score (0.8, 0.78, 0.96). Thus, we can assume that this approach is equally suitable for determining the number of clusters, for creating clusters of 'correct' geometry and with 'correct' width statistics. If we look into more details, by analyzing Table 7, the results of this approach differs per polygon category. From Table 7, It is apparent that this approach works quite better for the straight polygons in comparison with the curved polygons. For both categories that contain straight polygons (A and B), the final score of this approach is really close to the maximum score. Specifically, for the polygons of category A approach 1 gives almost identical results with our ground truth for all the indicators (final index score is 0.9). Figure 86 indicates the results of approach 1 for the 5 polygons of category A compared with our ground truth for the polygons of that category.

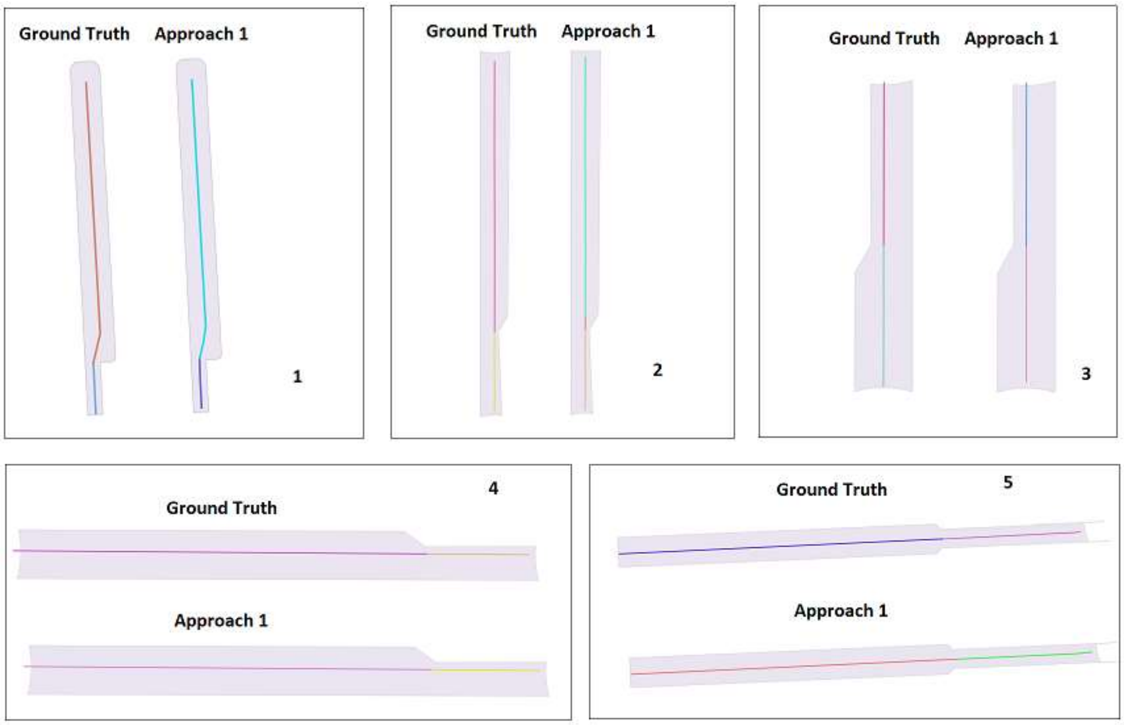


Figure 86 Five polygons of category A. The results of approach 1 is almost identical for all the cases. In case 2 we can notice that approach 1 results to 1 more cluster compared to our ground truth (3 instead of 2). Although this deprives 1 degree of indicator 1, since the cluster is of a very small length it does not affect the other 2 indicators

The category that seems to be the most problematic for this approach is category C (curved polygons with a clear point of width change). The final index for this category is 0.73 which differs a lot from the indexes of categories A and B. Figure 87 shows the results of approach 1 for the 5 polygons of category C compared with our ground truth for the polygons of that category.

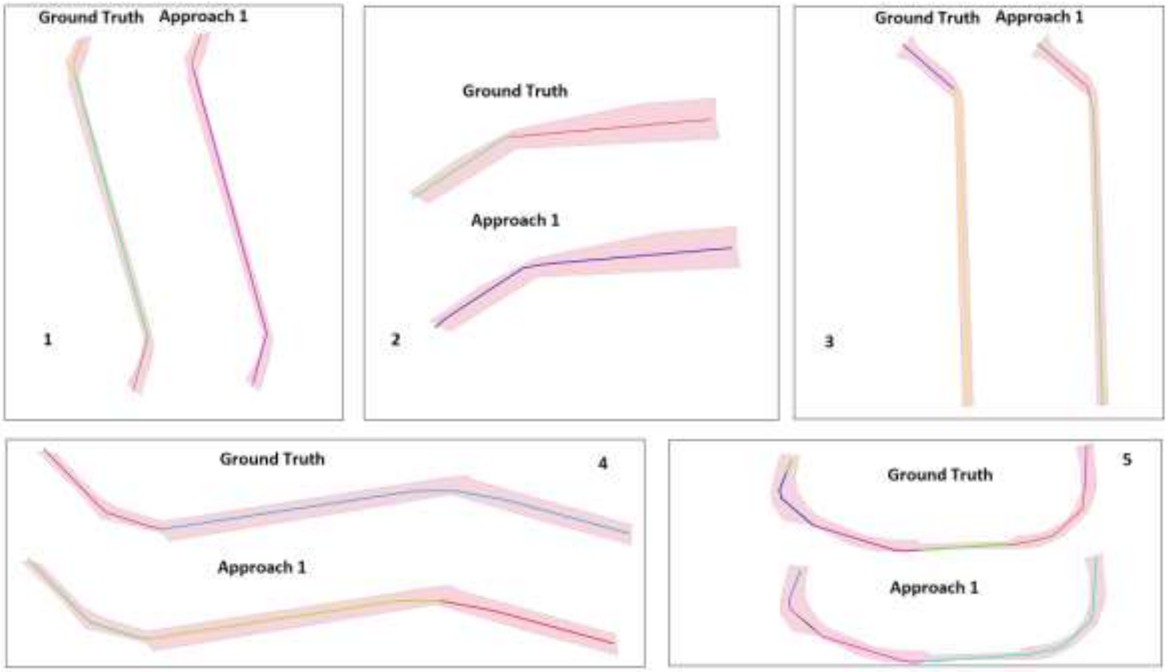


Figure 87 Five polygons of category C. This is the category that approach 1 seems to be give the least good results. In this example, there are some polygons that show similarity (1, 3,4) but they are not identical since either the number of clusters is different or the geometry of them. Moreover polygons 2 and 5 seem to be clustered differently with approach 1. In case of polygon 2, clustering approach do not to recognize the difference in width. Thus it does not defines a second cluster at the wider part of the polygon. In case of polygon 5, clustering approach and ground truth are following different strategy as well.

Results of approach 2

Clustering approach 2 uses average method as a linkage method. Thus, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. Moreover, the distance threshold for this approach is 15m. From a first view in Table 6 this approach seem to be less accurate than the previous one (less close to ground truth). Based on the final index that it gets from the aggregation of 3 weighted indicators it can be claimed that this approach does not follow that similar clustering strategy with our ground truth. The final index of this approach is 0.7. Moreover, we can notice that indicator 3 has quite higher index than the other 2 indicators. Thus, we can assume that the clusters that result from this approach show similarity in terms of width statistics with the ground truth clusters. Indicator 2 which is defined as the most important indicator and it has the highest weight. This indicator get the lowest score among the 3 indicators for this approach (0.59). By analyzing more the results of this approach, we can claim that the total score of indicators is lower than before for all the polygon categories. This approach also gives better results for straight polygons, compared to the curved ones. It is interesting to consider the result for category A. For this polygon category, indicator 1 and 3 seem to have an acceptable score (0.75 and 0.95). From the other hand, indicator 2 gets a quite lower score (0.6). Figure 88 shows the result of this approach for polygons of category A.

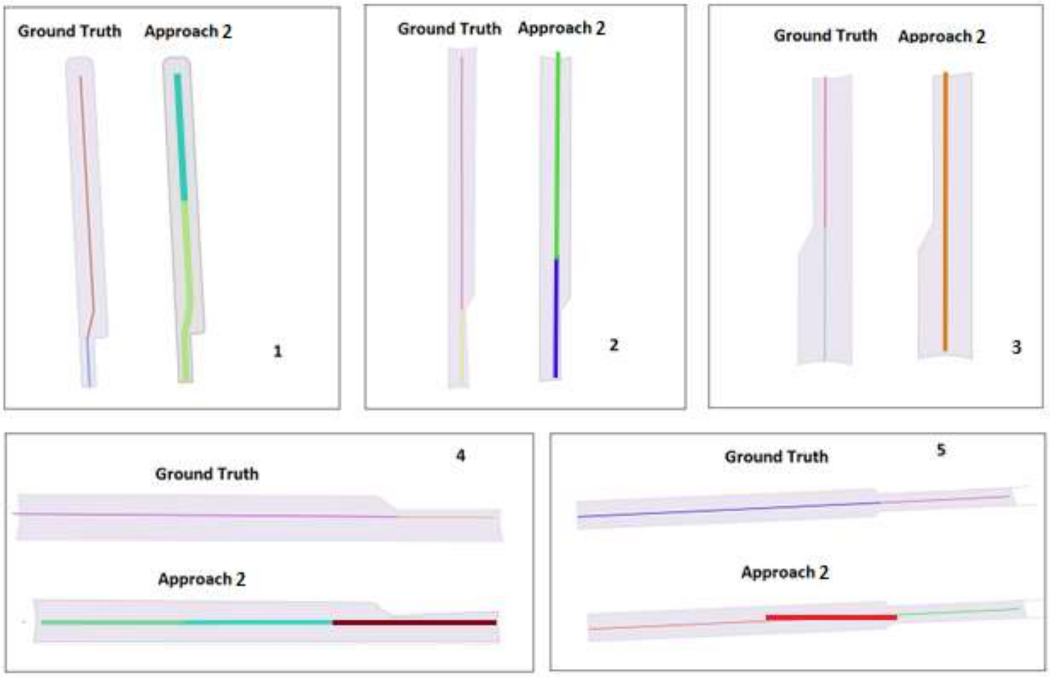


Figure 88 Five polygons of category A. Although the score of indicator 1 and 3 is acceptable (6), the score of indicator 2 is quite low. This, indicates that even if approach 2 results to the same number of clusters with ground truth and those clusters have similar width statistics, the new roads are quite different in practice (different geometry). Since indicator 2 is the one with more impact in the overall score, the final score for this polygon category is 3.25 (far away from 6 that the 2 other indicators are).

The category that this approach gives the less promising results is category D. The final index based on 3 indicators is pretty low (0.59). Thus, we expect that this clustering approach follows completely different strategy than our ground truth in those cases. Figure 89 shows how the 5 polygons of this category were clustered.

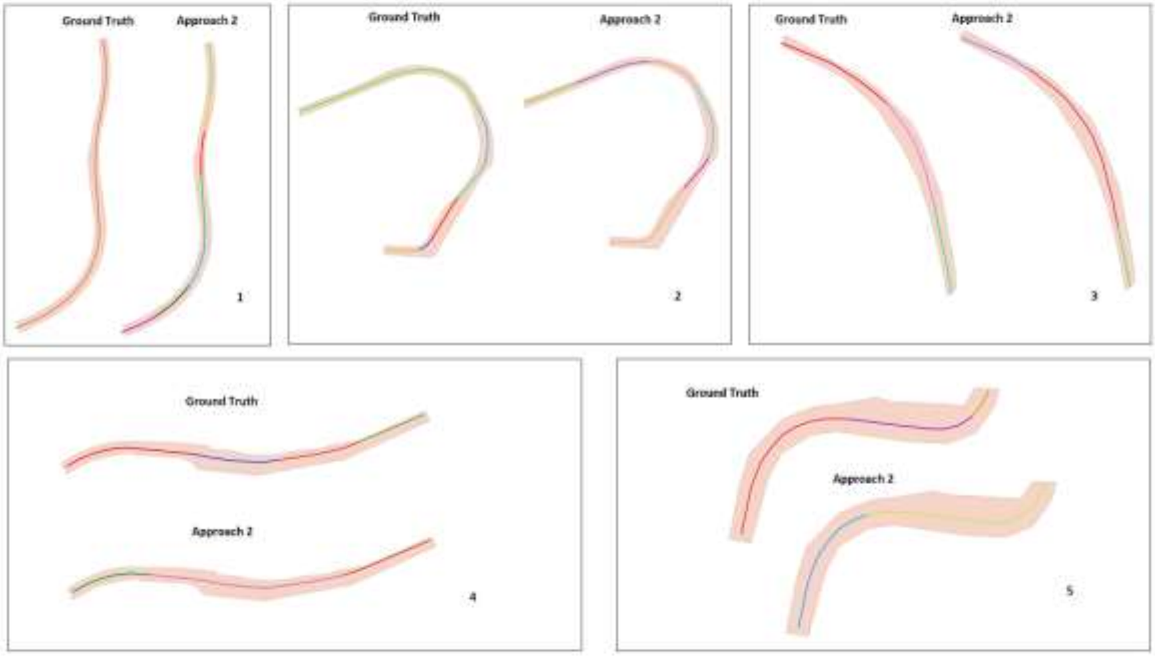


Figure 89 Five polygons of category D. Only for one case (3) the clustering approach results into similar clusters with our ground truth. For the rest approach it follows a different strategy. In some cases (1,4) clustering approach seem to be more sensitive to width changes while in case 5 it seems to be more conservative

Comparison approach1 and approach 2

By looking at the parameters of the 2 approaches we realize that 2 of the parameters that influence the result are different. The 2 approaches are using a different linkage methods. First approach uses single method while the second approach uses average method. Based on linkage methods, we expect the 1rst approach to be less sensitive in width changes. From the other hand, 2nd method has almost 3 times bigger distance threshold. As we already explained, the higher the threshold the less sensitive an approach is to width changes. In practice by looking the results of both approaches we can claim that approach 2 is the more sensitive to width changes and that in general, approach 1 results into clusters that are more similar to our ground truth. All the indicators for all the polygon categories gained a higher score for approach one and by analyzing and visualizing the results we can confirm that in most cases approach 1 has a more desirable output. Moreover, we can claim that both approaches work better with the straight polygons compared to the curved ones. Finally, although in general approach 1 gives better results from approach 2 we could identify a few cases (3 out of 20) where approach 2 results into more similar clusters with our ground truth. Figure 90 shows an example of such case.

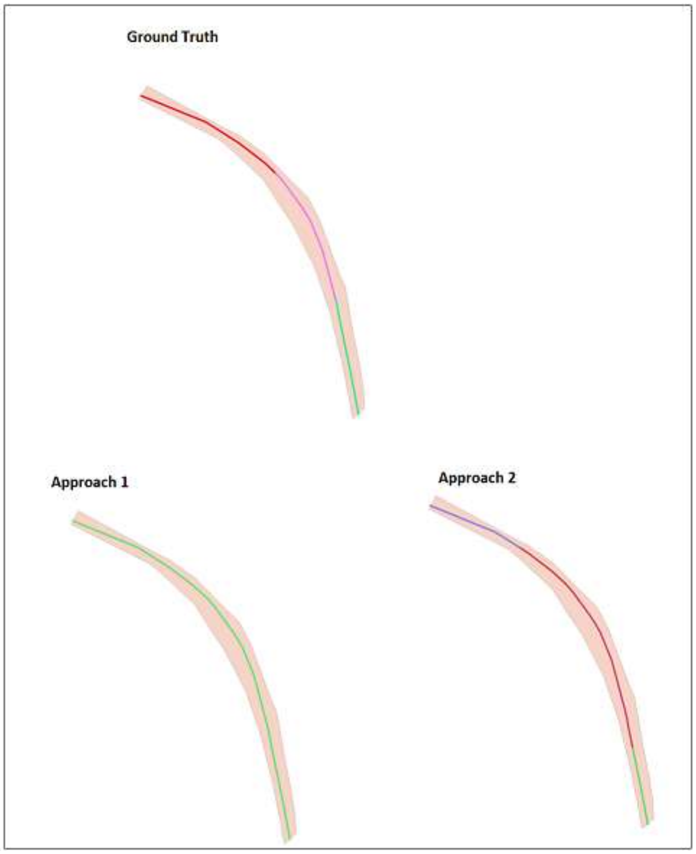


Figure 90 Example polygon of category D where approach 2 has better result than approach 1. Approach 2 seem to be more sensitive in width changes and in that case it seem to work better. Then number of clusters of approach 2 is the same with our ground truth (3) while approach 1 results only to 1 cluster. Moreover, the geometry of the clusters of approach 2 might not be identical to our ground truth but it shows similarity

6.4.1.2 Case 2 On the street parking

Case 2 correspond to road polygons that contain on the street parking. In some datasets the parking spots in the road network are modelled in the same polygon with the drivable space of the road. 20 road polygons that contain on the street parking areas and can be found in the dataset of Helsinki are used for the evaluation of the different clustering approaches. The ground truth is based on the existence or not of parking spots. When a parking spot exists we expect different cluster. Figure 91 displays an example of initial modelling of road polygon with parking spots and our ground truth.

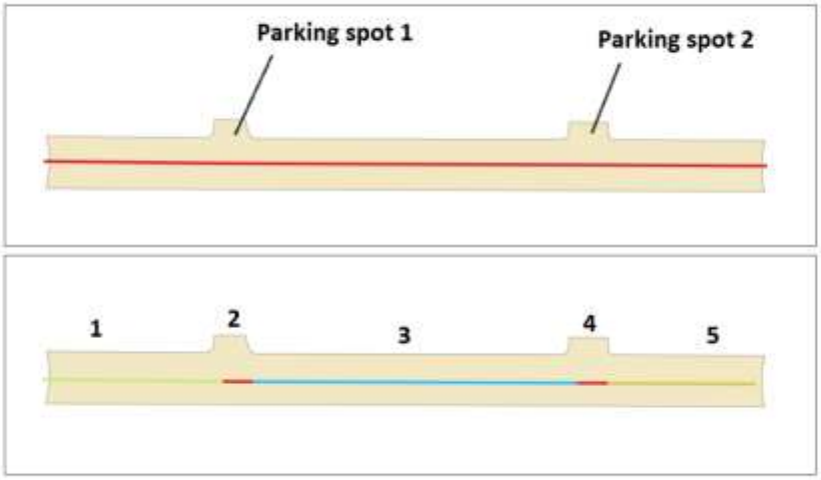


Figure 91 Road polygon that contains 2 parking spots and the initial modelling of one single centerline (Top). Our ground truth that defines a new road cluster every time that a parking spot is present (Bottom). In order to make sure whether there is a parking spot we used satellite images from Google earth (https://earth.google.com/web)

The selected polygons are further split into some categories based on their characteristics. The first distinction is done based on the sinuosity of polygons. Polygons are split into 2 categories: straight polygons and curved polygons. The second factor I used to further split polygons is the size of the parking area that polygon contains. 2 categories can be defined based on that. 1st category is defined for polygons that contain only one or maximum 2 parking spots and the 2nd category corresponds to polygons that contain more than 2 parking spots. Figure 92 shows the 4 categories of the polygons of this case. For the better evaluation of the clustering approaches, each category consists of the same number of polygons (5).

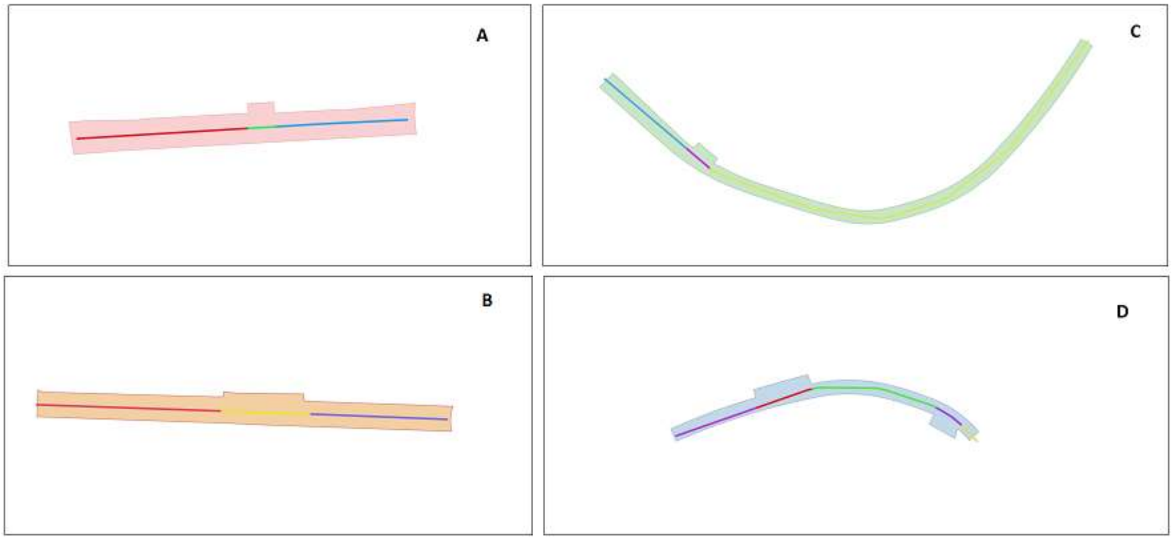


Figure 92 Road polygons of this case are divided into 4 categories. Straight roads with maximum of 2 on-street parking spots (A), Straight roads with more than 2 on-street parking spots (B), Curved roads with maximum of 2 on-street parking spots (C), Curved roads with more than 2 on-street parking spots (D).

Same clustering approaches are analyzed for their results for that case (see Table 5). The results for each category of the polygons are summarized at Table 8 and Table 9.

Approach	Total score Indicator 1	Number of polygons that indicator1 < 0.5	Number of polygons that indicator1 > 0.9	Total score Indicator 2	Number of polygons that indicator2 < 0.5	Number of polygons that indicator2 > 0.9	Total score Indicator 3	Number of polygons that indicator3 < 0.5	Number of polygons that indicator3 > 0.9	Final Index
1	**0.78**	0	10	**0.84**	1	14	**0.94**	0	19	**0.85**
2	**0.57**	5	0	**0.58**	8	3	**0.8**	0	0	**0.63**

Table 8

Approach	Polygon Category (5 Polygons)	Final index indicator 1	Number of polygons that indicator1 < 0.5	Number of polygons that indicator1 > 0.9	Final index indicator 2	Number of polygons that indicator2 < 0.5	Number of polygons that indicator2 > 0.9	Final index indicator 3	Number of polygons that indicator3 < 0.5	Number of polygons that indicator3 > 0.9	Final index score
1	A	**0.85**	0	3	**0.81**	0	3	**0.94**	0	4	**0.85**
1	B	**0.88**	0	3	**0.81**	1	4	**0.99**	0	5	**0.87**
1	C	**0.62**	0	1	**0.86**	0	3	**0.89**	0	4	**0.8**
1	D	**0.8**	0	3	**0.9**	0	4	**0.97**	0	5	**0.89**
2	A	**0.45**	2	0	**0.64**	1	1	**0.8**	0	0	**0.63**
2	B	**0.65**	0	0	**0.72**	1	1	**0.8**	0	0	**0.72**
2	C	**0.53**	2	0	**0.37**	4	0	**0.79**	0	0	**0.49**
2	D	**0.65**	1	0	**0.57**	2	1	**0.8**	0	0	**0.65**

Table 9

Results of approach 1

Approach 1 divides the initial 20 roads into 88 roads (clustered based on width). The approach uses single method as linkage method and distance threshold is 5.5 meters. This approach based on the final index of the 3 indicators seem to have a result quite similar to our ground truth. The final index after weights assignment and aggregation of the 3 indicators is 0.85 in scale of 1. If we further analyze the results based on the polygon categories, we can claim that the approach works similarly well for all the 4 polygon categories. In the previous case (case 1) approach 1 had desirable results for straight polygons but not that desirable results for curved polygons. In contrast to that, in this case the approach gets the highest index score for the polygon category D (curved polygons with more than 2 on-street parking spots). Figure 93 shows the results of approach 1 for this polygon category in comparison with our ground truth labels. As it is obvious from the image the results for this polygon category are quite similar to our ground truth. We can notice that in some cases (2, 3, 4) approach 1 results to some more clusters with a pretty small size (1 or max 2 more clusters). This occurs at the place where the parking spot exists. This is also obvious from Table 8 since indicator 1 (shows the similarity in terms of cluster number) is lower than the 2 other indicators. This, does not seem to affect a lot the final result since these clusters are quite small in size (smaller than 5 meters).

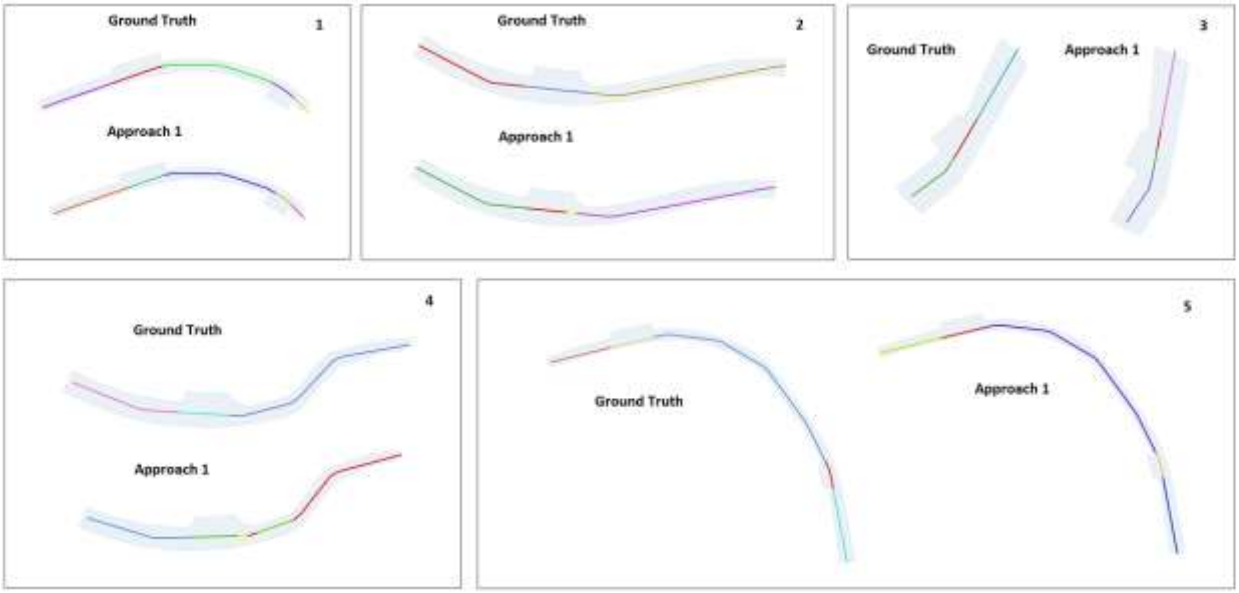


Figure 93 Polygon category D. The polygon category with the highest final index for clustering approach 1.

The polygon category that this approach gets the lowest final index is category C (curved roads with maximum of 2 parking spots). Although the final index of the 3 indicators for polygon category C is 0.8, final index for Indicator 1 is much lower (0.62). This suggests, that there is a difference in the number of clusters between clusters of approach 1 and our ground truth clusters. Figure 94 illustrates the results of approach 1 for polygon category C in comparison with our ground truth labels.

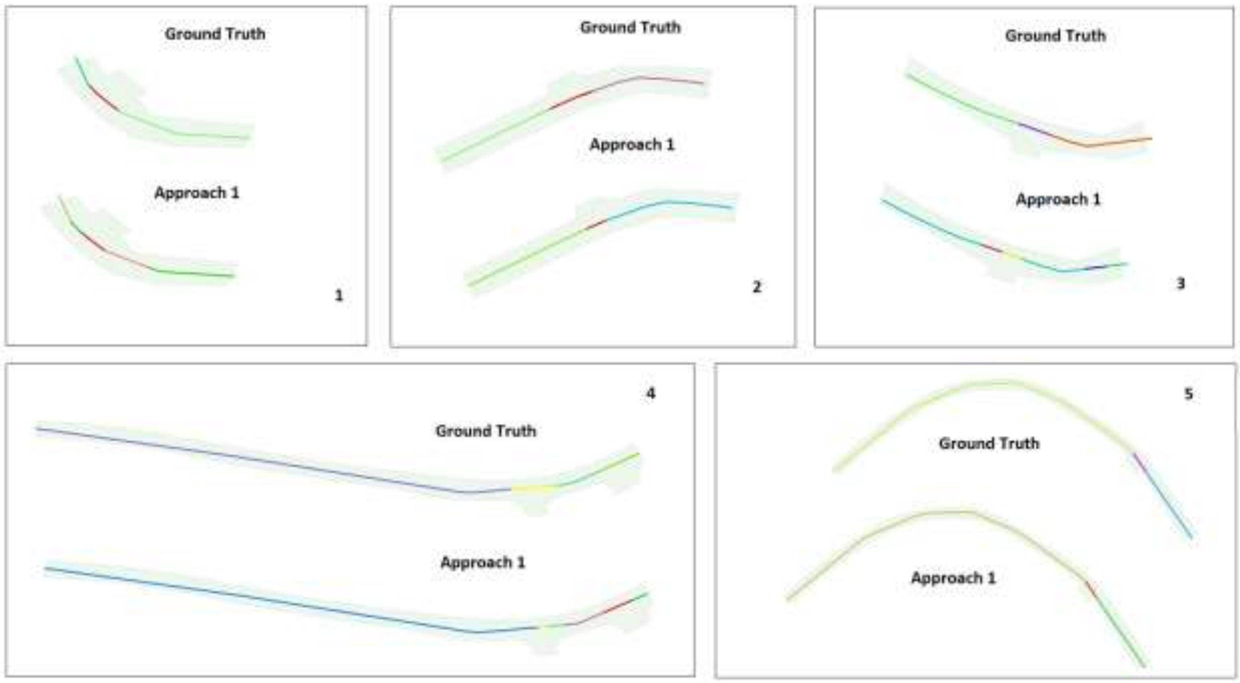


Figure 94 Polygon category C. The polygon category with the lowest final index for clustering approach 1. The clusters of approach 1 show some degree of similarity with the clusters of our ground truth. The main difference lies in the number of cluster which is higher at some cases for the clustering approach 1. 2 out of 5 cases show identical results.

Regarding indicator 2 which is claimed to be the most important indicator, approach 1 gets a final index for all the 20 polygons of 0.84. It is a fairly high value indicating that the geometry of approach 1 clusters has a similar geometry to the clusters of ground truth. Out of 20 polygons, 14 get a value above 0.9 for indicator 2. Only 1 polygon gets a value lower of 0.5 for this indicator.

Results approach 2

Approach 2 divides the initial 20 roads into 54 roads (clustered based on width). The approach uses average method as linkage method and distance threshold is 15 meters. Based on the final index of this approach we can claim that the results is not that similar to our ground truth. The final index is 0.63. If we look into more details, by analyzing Table 9 that indicates the results per polygon category we will notice that this approach has similar results for all the 4 polygon categories. Polygon category B get the highest final index (0.72). All the 3 indicators have higher final index for that category. Figure 95 illustrates the results of approach 2 for polygons of category B (straight polygons with more than 2 parking spots).

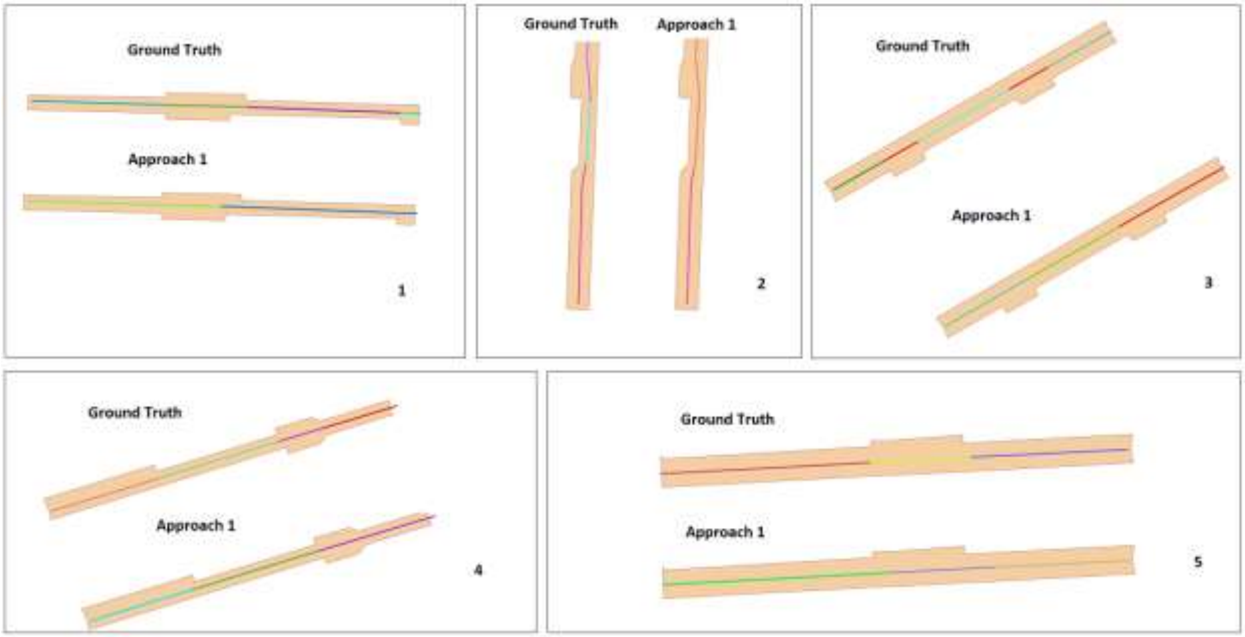


Figure 95 Results of clustering approach 2 for polygons of category B. The final index is 0.72. In most of the cases a different strategy is followed by this approach. In most of the cases the approach 2 resulting clusters seem quite different and the parking spot is not identified as a different cluster (case 1,2,3,4). Only one out of 5 cases seem to have identical results.

Same as approach 1 the polygon category with the lowest final index is the category C. The final index of approach 2 for this polygon category is quite low (0.49) which indicates that a different clustering strategy is followed. Figure 96 illustrates the results of approach 2 for this polygon category.

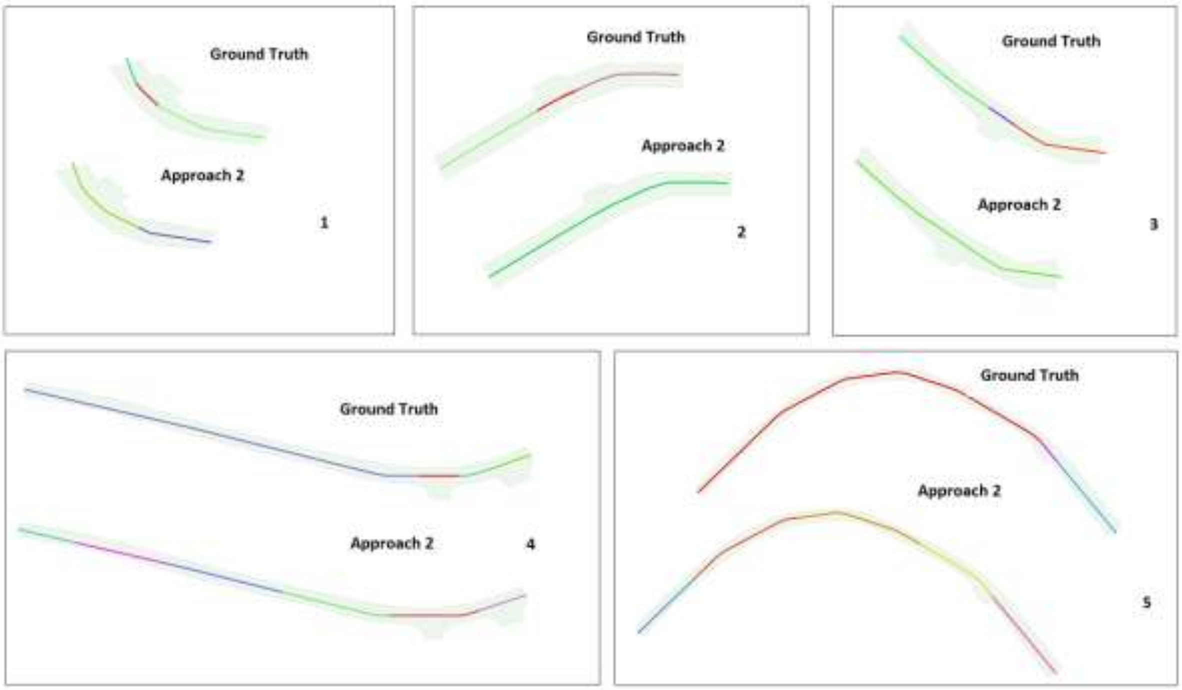


Figure 96

6.4.2 More representative roads

The second reason that lead us to develop clustering methodology, was the argument that clustered centerlines would be more representative in terms of width, compared to original centerlines. In § 4.2.2 an example of why this is expected to happen was presented. Now I will explore this in practice.

Two datasets of the same area (central area of Toronto) will be used to examine if clustering width values are more representative. Figure 97 left, shows an example of how original centerlines can be found in the dataset 1. They extend from one intersection to the other. Clustered centerlines are created based on similar width measurements. Thus, a new centerline is created each time there is a change in width (Figure 97, right). Since intersections is proved that will cause noise (§ 5.3.1), they are identified and removed for the purposes of this test.

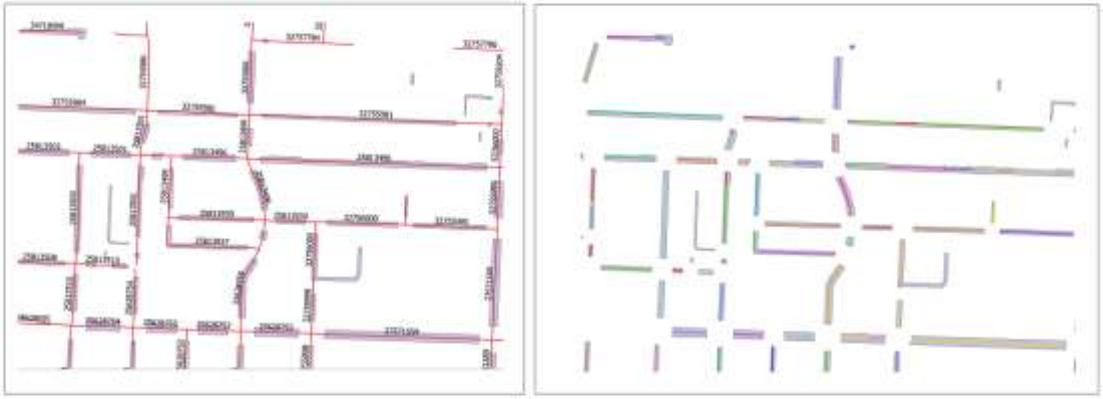


Figure 97 Original road centerlines (left) and clustered centerlines of the same area (right)

Each road is assigned with a mean width value that is computed based on a set of measurements. To check how representative this value is for a road, we can look at the standard deviation around this mean. If the standard deviation is high, this indicates that the measurements differ for the same road. If the standard deviation is low, it means that the measurements are more or less similar. In order to prove that width values are more representative after clustering approach is applied we will compare the change in the mean and median values for the standard deviations of the roads of the 2 datasets. Table 10 shows the values before and after the clustering approach.

Dataset	Number of centerlines	Mean of Standard deviations	Median of Standard deviations
Original centerlines	3296	0.67	0.09
Clustered centerlines	4768	0.08	0.01

Table 10

From the table, it is obvious that standard deviations that roads have around their mean value are reduced after clustering is applied. Especially the mean value of standard deviations is more than 8 times lower for clustered roads. Moreover, as explained in § 2.1.3 roads of Toronto seem to have a quite regular shape and not a lot of width changes occur. Thus, for a city where roads have a more irregular shape those changes might be even higher.

6.4.3 General conclusions

In general, based on the analysis I did I can claim that clustering approach could be used to identify the important width changes of roads. A specific clustering approach appears to have pretty promising results for both real-world cases that were tested. The clustering that result from this approach seem to have a high degree of similarity with the clusters of ground truth. Moreover, from the analysis in § 6.4.2 I came to the conclusion that clustered roads are more representative in terms of width. The standard deviation around mean and median width values showed a significant decrease for the tested roads in the central area of Toronto.

7 Road safety analysis

The purpose of this chapter is to explore whether certain features implemented for this thesis, could affect the process and the result of a road safety analysis. As explained in § 4.1, safety analysis links the number of accidents to the characteristics of the road environment. Its general purpose is to investigate the relationship between the number of traffic accidents and the different characteristics of the road environment that affect this number (see § 4.1). For this thesis and since we work with road width, we will explore whether the correlation between traffic accidents and road width is influenced by the way in which road width is estimated and linked to roads. The width will be estimated in different ways and different datasets will be examined for their correlation with the road accidents for the same tested area. Then, the numerical results will be compared. The main goal of this section is to find out whether the use of certain features of the methodology developed in this research can lead a road safety analysis to different conclusions. It is important to mention that an actual road safety analysis even at a macro-level (see § 4.1) takes into account many factors of road environment that affect the number of accidents. In this section, we are not conducting such an analysis. Our goal is to investigate whether a fundamental re-thinking of the width estimation process is required before performing a safety analysis that correlates road accidents and road widths.

Tested area

A dataset that is provided by the Toronto police service and it can be found at the public safety data portal at: https://data.torontopolice.on.ca/search?collection=Dataset&q=traffic is used. This dataset includes all traffic collisions events where a person was either Killed or Seriously Injured (KSI) from 2006 – 2020 for the city of Toronto. KSI accidents can be considered the most important accidents. Since KSI are major accidents and the police are involved, we can also consider that all KSI accidents are reported. Figure 98 shows how these accidents are spatially distributed in our tested area. From this figure, it is obvious that the spatial distribution of accidents is not equal but most accidents are concentrated in the central area of the city. We can assume that the central area of Toronto concentrates most of the traffic as well, thus it is reasonable that most of the accidents occur there.

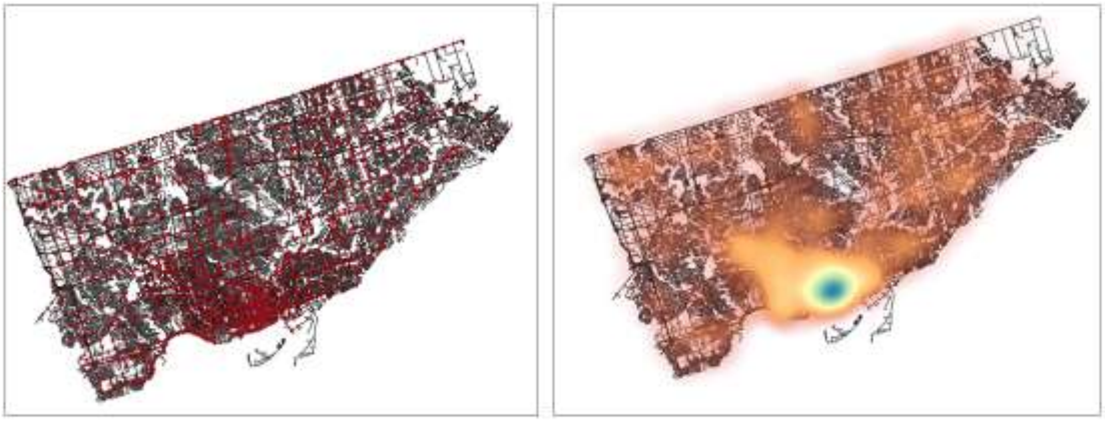


Figure 98 KSI data for the city of Toronto and their distribution. Most of KSI accidents are concentrating at the central area of the town.

Figure 99 shows the tested area that I chose for implementing road safety analysis. Corresponds to a 3500 m radius in the city center of Toronto.

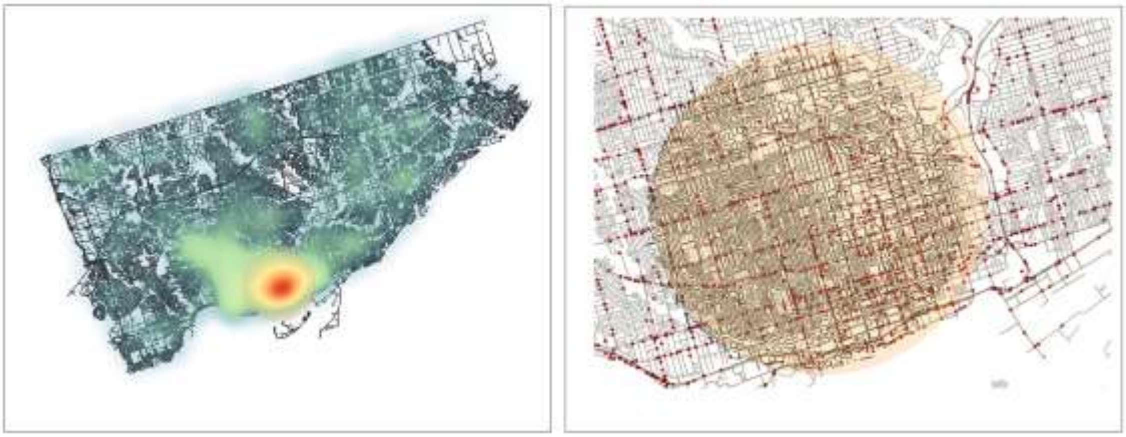


Figure 99 Tested area for examine the relation of road width values and road accidents

Data normalization

Figure 100 shows the relationship between road length and the number of accidents occurring on a road. From this picture we can claim that when the length of the road increases, more accidents occur.

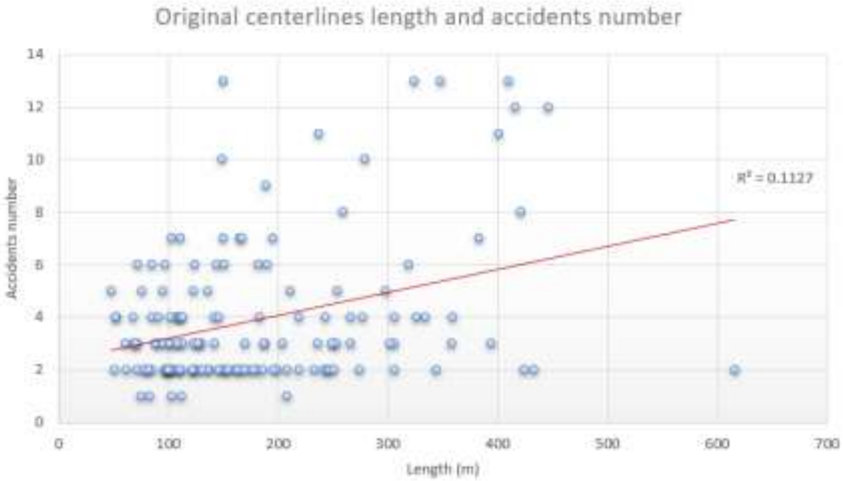


Figure 100 Graph shows the relation between length of a road centerline and the number of accidents that occur. We can claim that while e the road length increases, the number of accidents increases as well.

Thus, before associating the number of accidents to the width of a road, it is reasonable to normalize the data based on the different lengths of the roads. Divide the accidents that occur on a road by the length of the road will lead to a number that indicates the **accidents per meter** for each road.

One more step before conducting our tests is needed. As mentioned before, the tendency of accidents is not the same everywhere in the tested area. Thus, in order to compare the accidents that occur in different roads we need to take into account the overall trend in the area that the road lies. Thus, a grid with 500x500 meter cell size was defined. The cell size is selected to be rather big in comparison with the mean and median length of the road centerlines (details about those statistics are given in next section). This, ensures that most of the roads are passing through one or maximum 2 cells. Then, we count the number of accidents in every cell of the grid and we divide that number by 500. Thus, we result to a number that indicates the **accidents per square meter** for the area that the cell covers. These numbers actually show the trend of accidents in different areas of 500x500 meters.

Now the comparison between the different roads is possible. The final normalized number that indicates whether a road is associated with high or low rate of accidents is defined based on both the length of the road and on the overall accident tendency in the area that the road lies. Figure 101 illustrates an example of a road and how the normalized accident value is computed. The road of the figure is associated with 9 accidents and it has 100 meters length. Thus, the number of accidents per meter for that road is 0.09 (9/100). Then we check the accident per square meter in the grid cell that the road lies. In the area that the grid cell covers, 79 accidents have occurred. Thus, 0.16 accidents per square meter. If we compare these 2 numbers, we derive the conclusion that the road has fewer accidents per meter in comparison to the area that it is located. Therefore, its rate of accidents (normalized final value) is negative (-0.07)

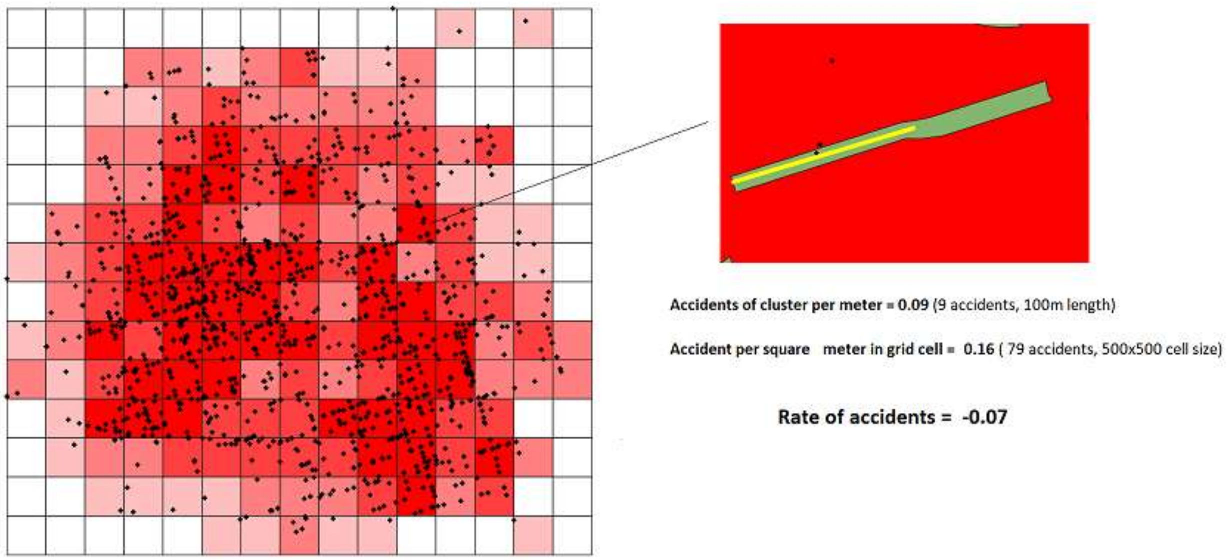

Accidents of cluster per meter = 0.09 (9 accidents, 100m length)
Accident per square meter in grid cell = 0.16 (79 accidents, 500x500 cell size)
Rate of accidents = -0.07

Figure 101

Dataset 1 Original road polygons/centerlines

Width values are estimated for the original road centerlines and final width estimations are linked with them. In order to estimate road width based on methodology described in § 5.1 the original polygons were used. Original road centerlines extend from one intersection to another while original polygons can be either road edge polygons or road intersection polygons (Figure 102).

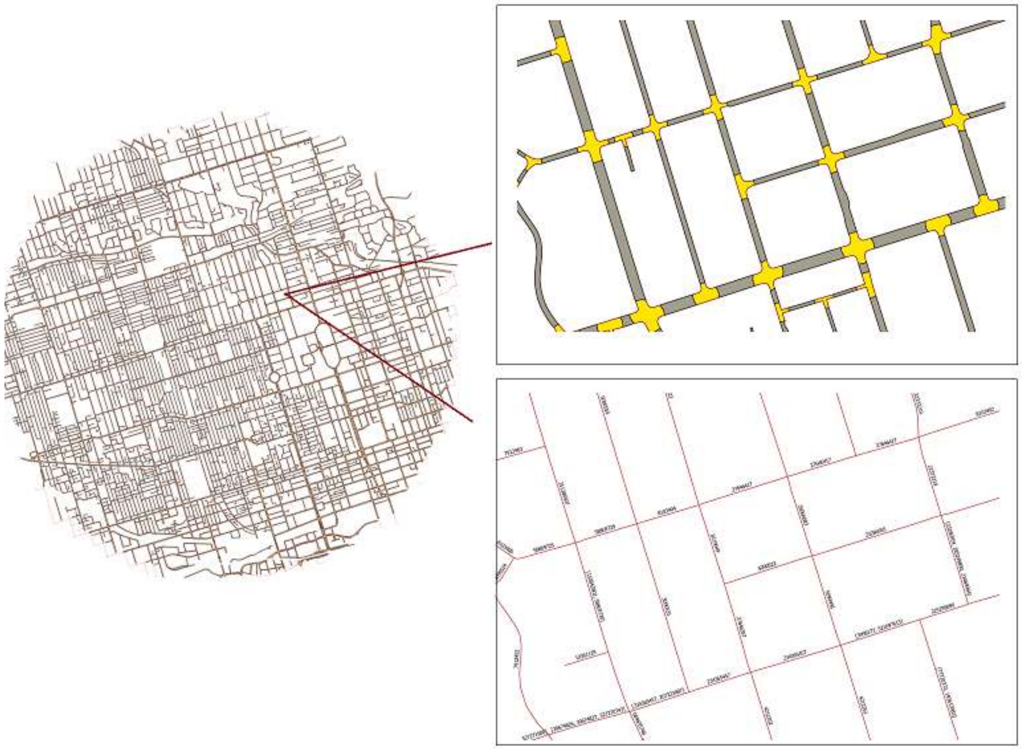


Figure 102 For dataset 1 width is estimated for the original road centerlines (extend from one intersection to another) using the original road polygons

First, some basic statistics of the original road centerlines after the width computation process are generated (Table 11). We can notice that there is a big difference between mean and median width values of the roads. As already mentioned in § 5.3.1 intersection polygons can cause noise to the overall width estimation procedure. By noise, we refer to some ‘wrong’ pretty high width measurements that will be included to the final width estimation. Thus, it is reasonable the mean value (which is influenced more by outliers) to be quite higher than the median width value (also influenced but less).

Total number of centerlines	Mean Width (m)	Median Width (m)	Mean length (m)	Median length (m)
3132	13.49	8.7	125.4	101.6

Table 11

Then, the relationship between mean/median width values of centerlines and normalized accidents is explored. Figure 103 and Figure 104 illustrate this relationship. From the 2 figures we can claim that the correlation between mean/median width values of road and the normalized accidents is rather weak. The wider roads tend to have more accidents but still the coefficient of determination (R^2) is quite low for both mean and median width (0.05 and 0.03 respectively).

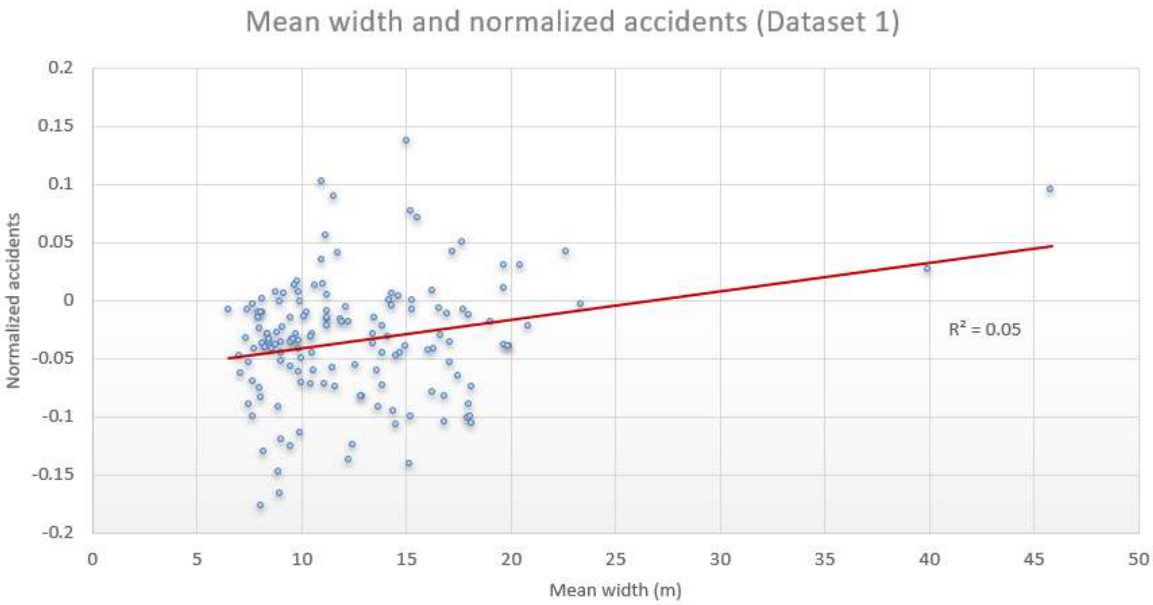


Figure 103

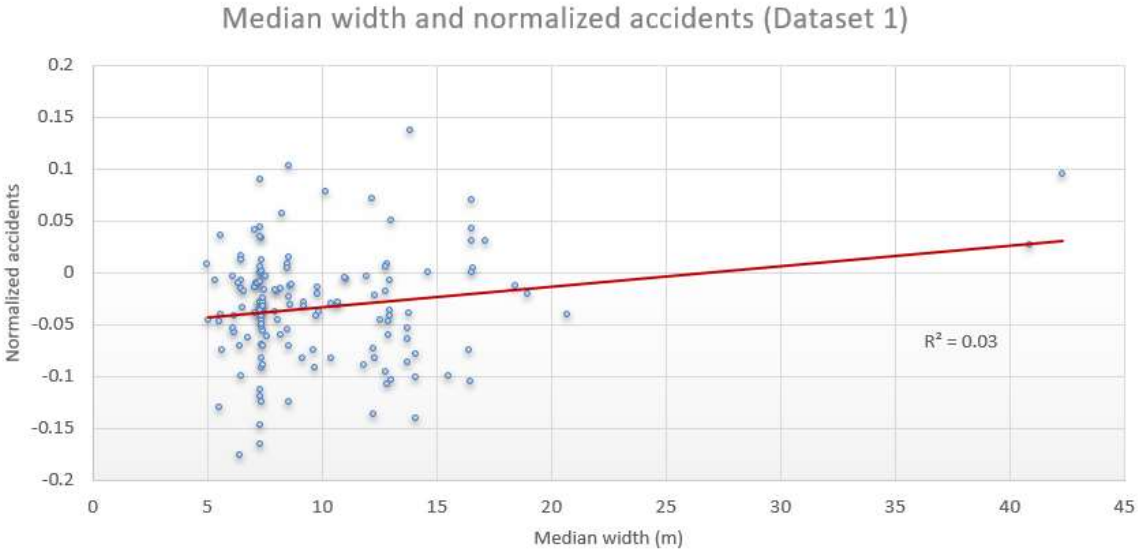


Figure 104

In order to further examine the relationship between normalized accidents and road width, roads were split into 6 categories based on width ranges (7 categories in case of median width ranges, an extra category of roads with width less 7 meters is included). Figure 105 and Figure 106 show how normalized accidents are associated with each category. The 2 box plots, indicate the main statical values for each category. For mean width ranges, while roads that belong in range of 7-9 meters (lowest width range) seem to be associated with the lest normalized accidents, we cannot claim that wider roads are more dangerous. For example, roads belong to range of 13-15 meters seem to be associated with quite less normalized accidents compared to roads of range 11-13 meters and of range 9-11 meters. For median width ranges, it is apparent that road larger than

15 meters are more dangerous compared to narrower roads. Appendix A contains more information on how normalized accidents are associated with roads of different width ranges.

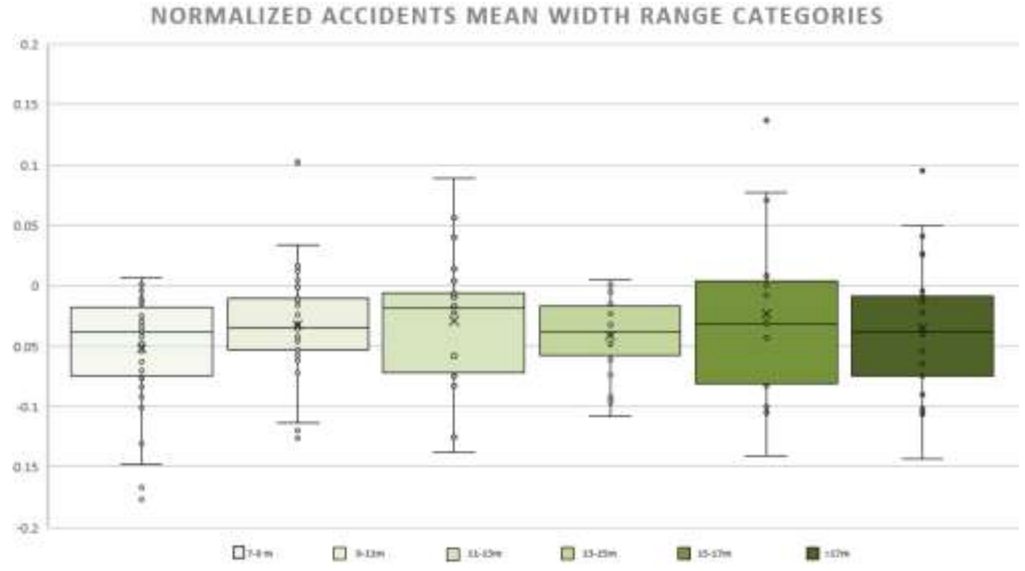


Figure 105

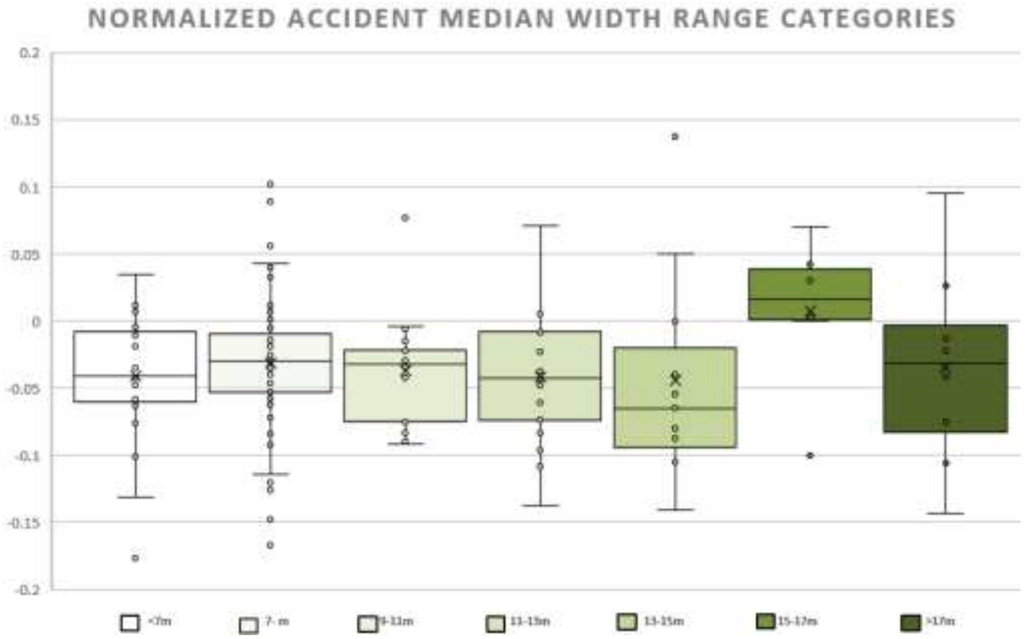


Figure 106

<u>Dataset 2 Original road centerlines-excluded intersection polygons</u>

Width values are estimated for the original road centerlines and final width estimations are linked with them. The difference of dataset 2 is that before estimating the width for the initial centerlines, standardization methodology was applied to the road polygons dataset (§ 5.2). Then, intersection polygons were identified (§ 5.3) and excluded from the overall process. Figure 107 shows the polygons and the centerlines used.

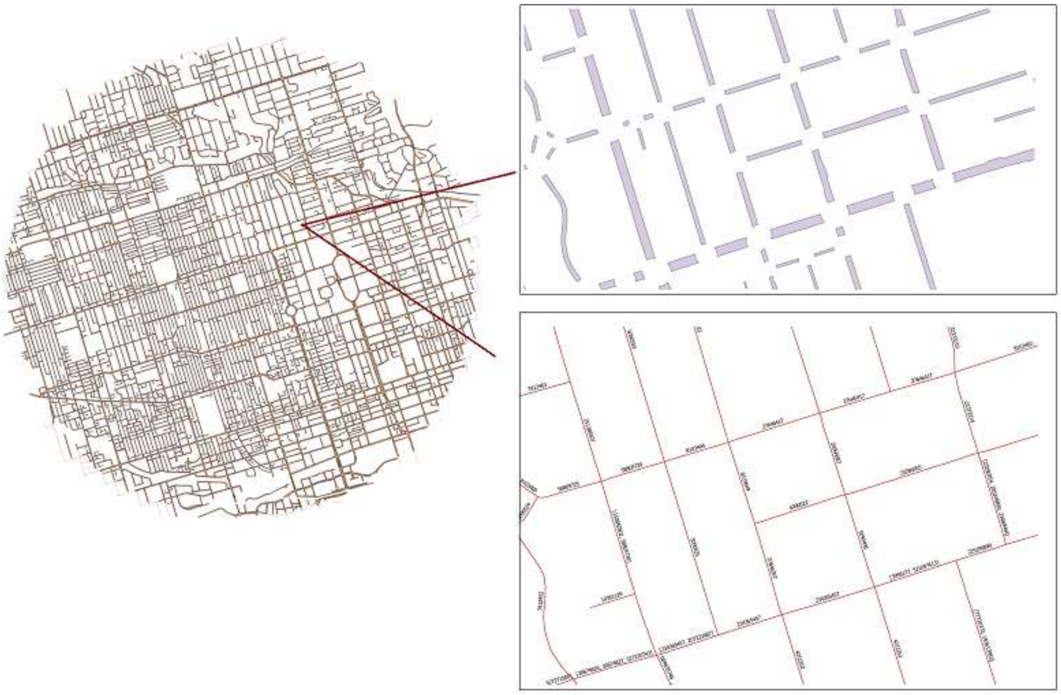


Figure 107 Original road centerlines of Toronto tested area

Basic statistics after width computation for this dataset are generated. More details about the centerlines of this dataset are available in Table 12.

Total number of centerlines	Mean Width (m)	Median Width (m)	Mean length (m)	Median length (m)
2798	9.88	8.3	135.3	104.5

Table 12

Then, the relationship between mean/median width values of centerlines and normalized accidents is explored. Figure 107 and Figure 108 illustrate this relationship. As shown in 2 images, there is no correlation between normalized accidents and road width (either mean or median). The value of the coefficient of determination (R^2) is 0.003 for the mean width values and 0.007 for the median width values.

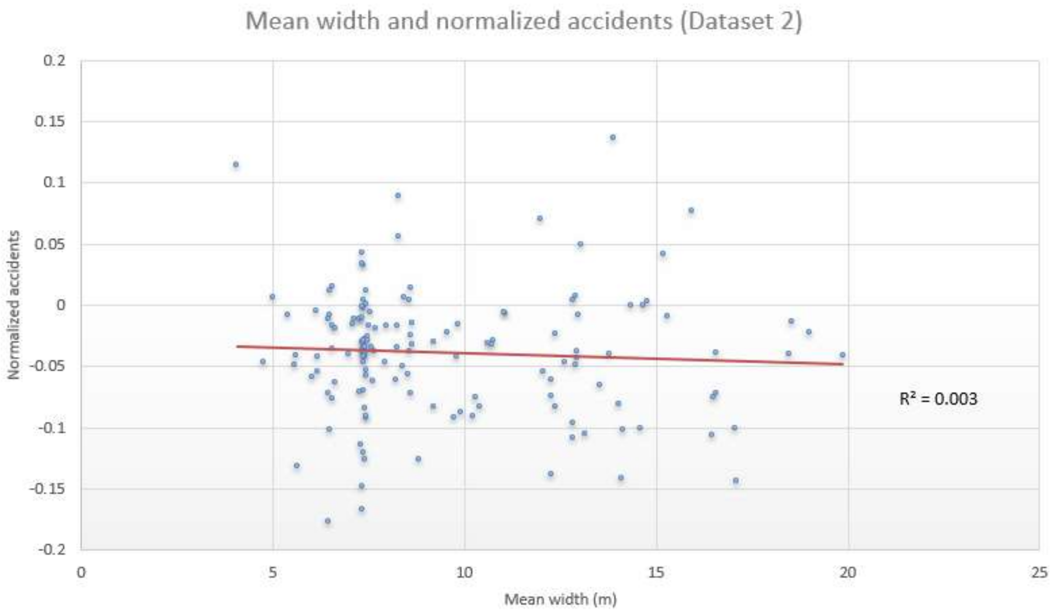


Figure 108

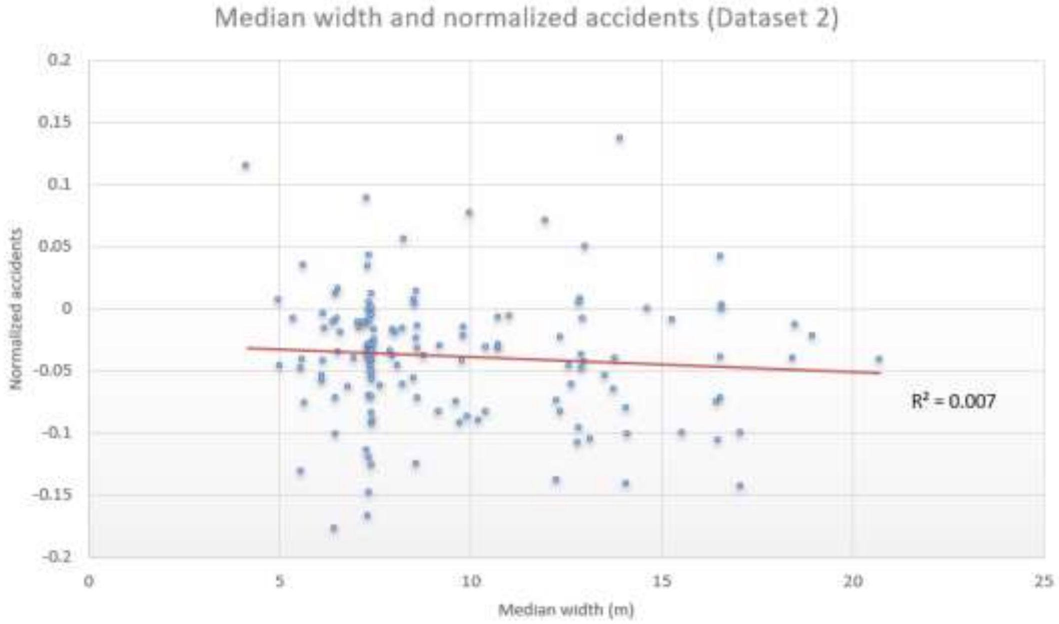


Figure 109

Figure 110 and Figure 111 contain statistical information on normalized accidents in relation to certain categories of mean and median width ranges. It is apparent that there is no relation between road width and normalized accidents. Values of mean and median normalized accidents are independent from the range of widths. For example, while roads with a mean width of less than 7 meters are associated with higher mean and median accident rates compared to the 9-11 meter range, they appear to have lower accident rates compared to roads with a mean width of 15-17 meters.

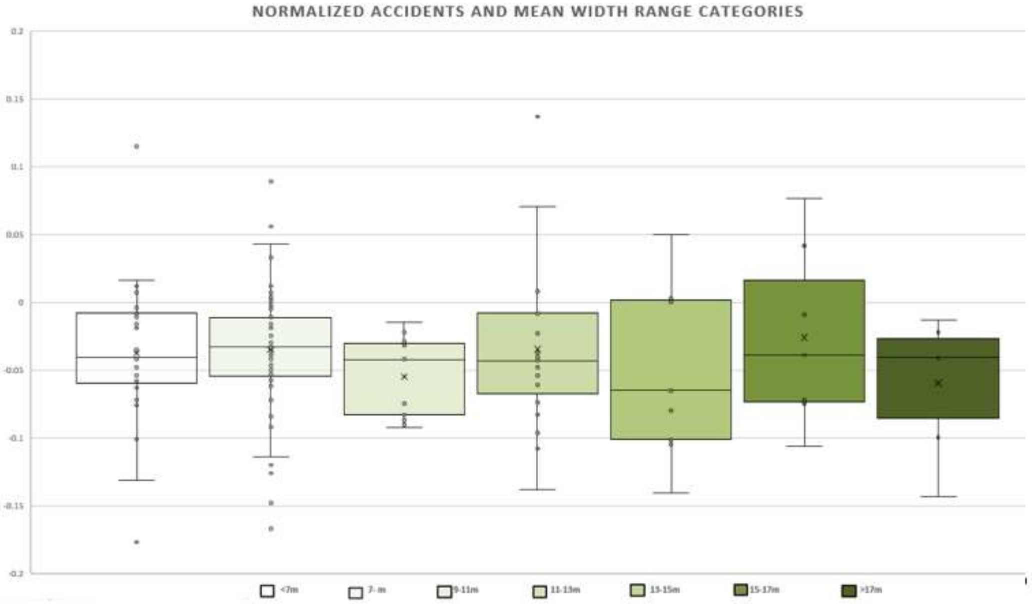


Figure 110

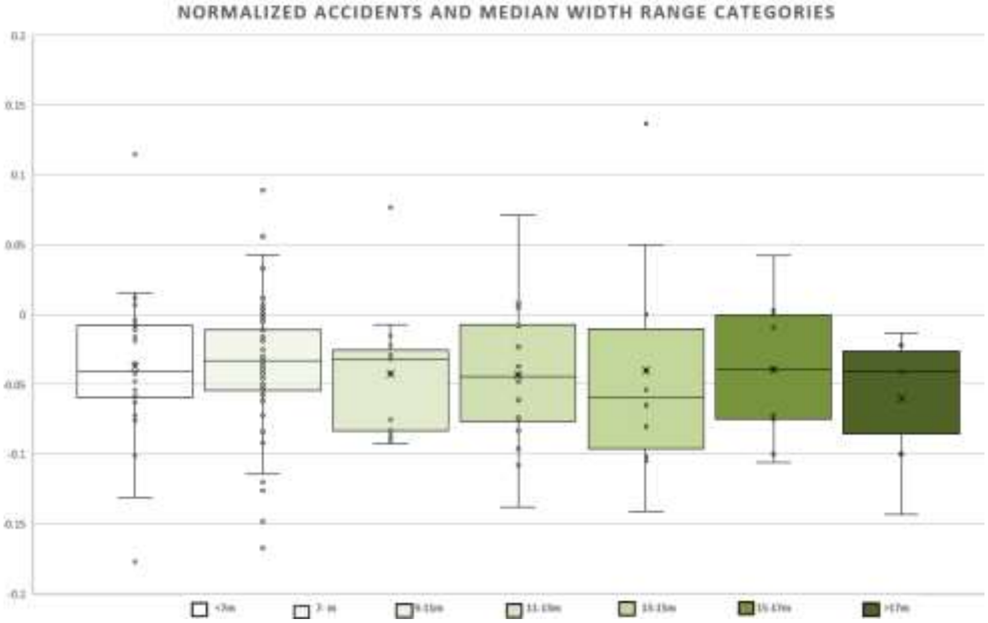


Figure 111

<u>Dataset 3 Clustered centerlines – excluded intersection polygons</u>

Finally, width clustering has been applied. Original centerlines divided into parts with similar width measurements. Width values are estimated and linked to the clustered centerlines. Clustering approach that applied, uses 'single' as linkage method and distance threshold is 5.5 meters. This approach is quite sensitive in width changes of a road but not too sensitive comparing to other approaches that use other linkage methods (average, complete). This clustering approach is tested for its results with 2 real-world cases related to road safety and it produces quite promising results. More details about clustering approaches can be found at § 6.4.1. Figure 112 shows centerlines and polygons used.

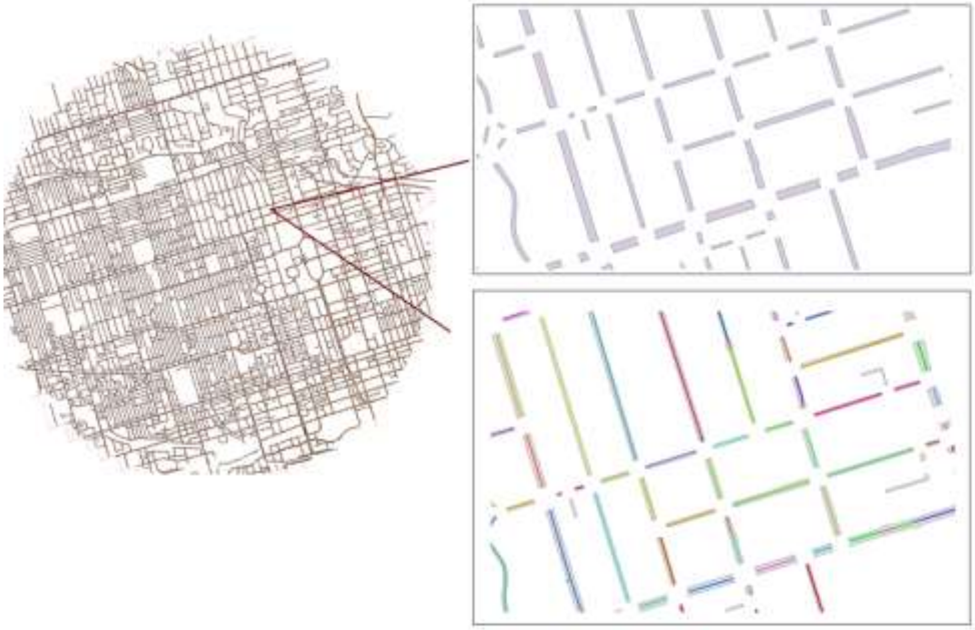


Figure 112 Road centerlines clustered based on width, intersection polygons excluded

Some general statistics for the clustered roads of the tested area are shown in Table 13. Mean and median length is quite reduced, which is reasonable since some of the original roads are 'cut' into smaller parts based on width changes.

Total number of clusters	Mean Width (m)	Median Width (m)	Mean length (m)	Median length (m)
4768	9.9	8.1	49.9	25

Table 13

Then, the relationship between mean and median width values of centerlines and normalized accidents is explored. Figure 113 and Figure 114 illustrate this relationship. The coefficients of determination (R^2) have a value of 0.09 and 0.07 respectively. From the 2 graphs, we can claim that road accidents appear to have a very weak correlation with road width. The wider roads tend to have slightly fewer accidents.

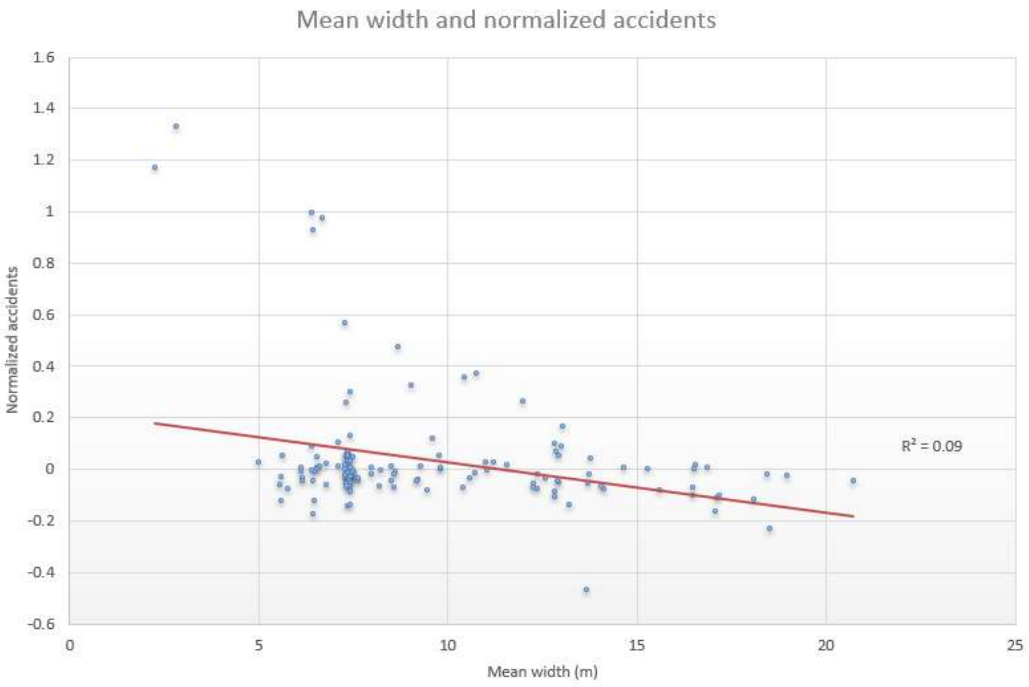


Figure 113

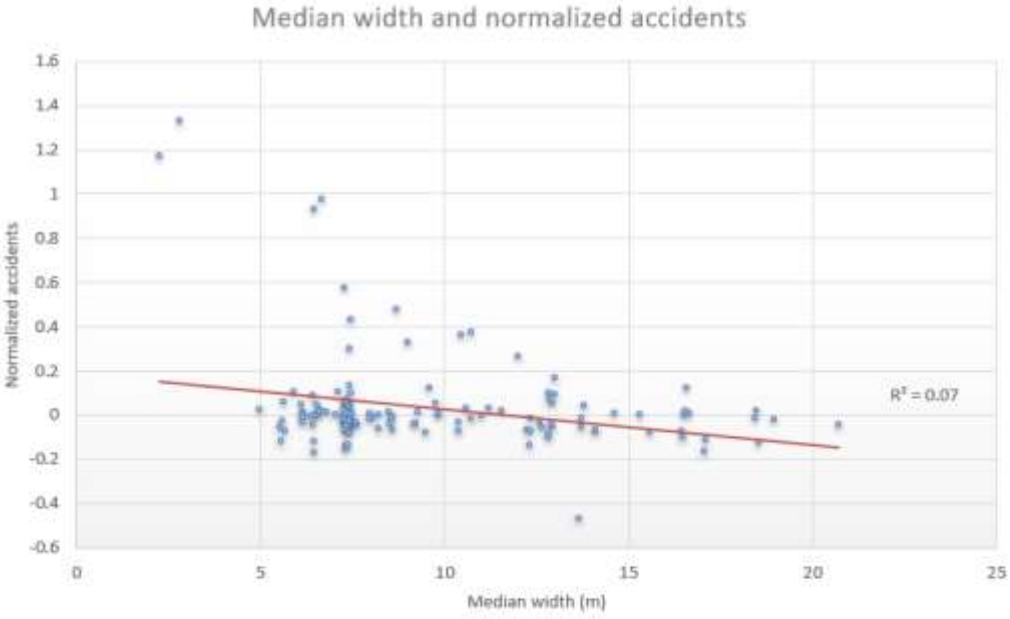


Figure 114

Figure 115 and Figure 116 contain statistical information on normalized accidents concerning certain categories of mean/median width ranges. From the box plots, we can claim that mean and median accident rates seem to decrease when a road extends the 11 meters. But this does not imply a strong correlation, since the values of some narrower road ranges seem to have similar values with the wider ones. For example, 7-9 meters range and 11-13 meters range have quite similar mean and median values of normalized accidents.

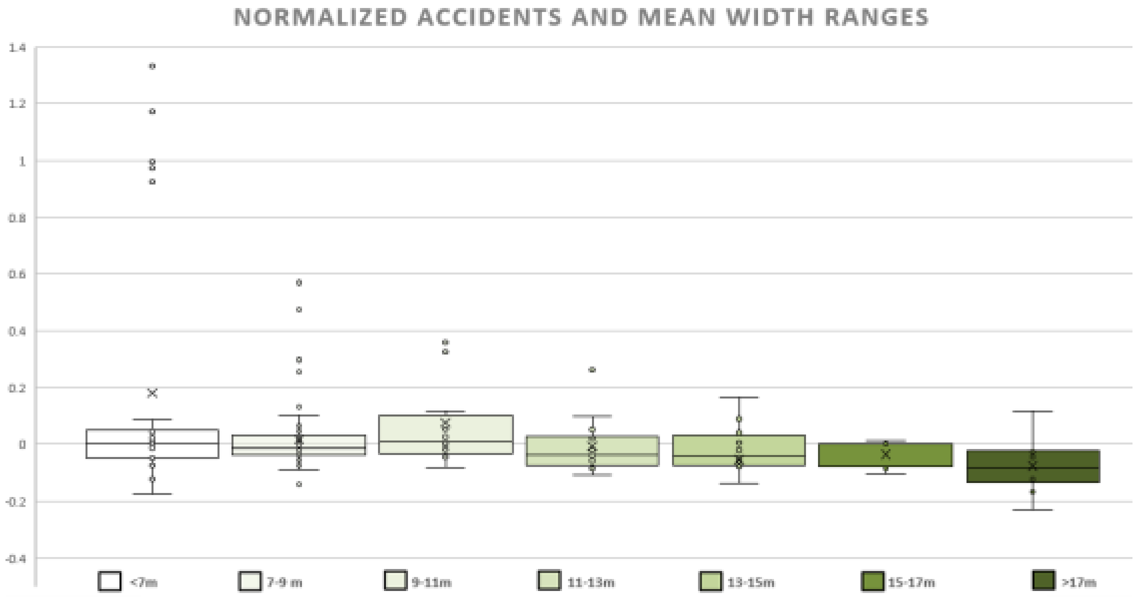


Figure 115

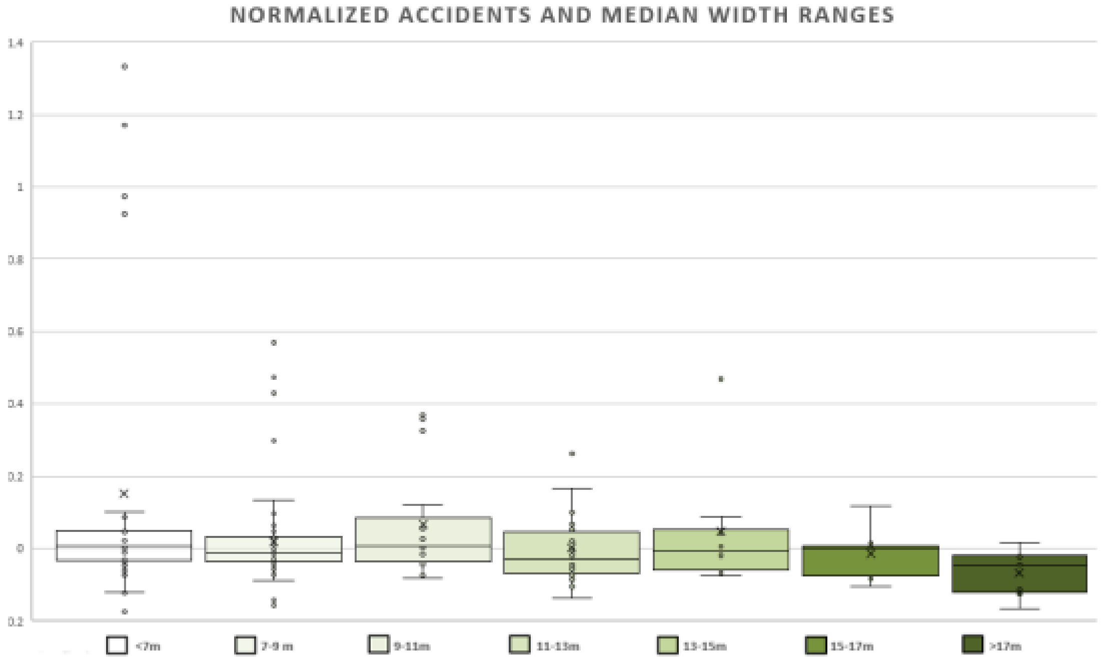


Figure 116

Summary

In general, it can be claimed that certain features implemented for this thesis could affect the result of a road safety analysis that correlates traffic accidents and road width. As aforementioned, traffic accidents are affected by many different factors of the road environment. Traffic conditions, human behavior etc. should be taken into to account, to derive a less biased conclusion about correlation between traffic accidents and road width. The main purpose of this chapter was to explore whether the way that width is estimated and linked with roads can also affect the overall process of a road safety analysis. Based on the numbers obtained from our tests, it can be claimed that a thorough review of the width estimation process is required before examining the correlation with road accidents. Original data led to a conclusion of a weak correlation with wider roads being the most dangerous, while the analysis of the most processed data (dataset 3) led to exact opposite conclusion. Although the coefficient of determination is quite low and does not indicate a clear correlation, there is a huge change in the numbers when different datasets are processed (i.e. from 0.003 to 0.09 for the mean width values when width clustering is applied). More discussion on the results of this analysis and a deeper interpretation of how each feature of the final methodology appears to affect the safety analysis are included in the next chapter.

8 Conclusions

In this final chapter, the separate features implemented during this graduation project and their impact to the overall goal of this thesis are discussed. Then, the research questions are reviewed. Finally, some future work is also recommended.

8.1 Discussion

Vector data for estimating road width

Estimating road width is not a new concept. In § 3.1, other approaches that address this topic were presented. The main difference of our approach lies in the different input used for the estimation. While other inputs such as satellite imagery and LiDAR point cloud can be used to calculate width of a road with a quite good accuracy, road vector data coming from open sources were used as input for this research. This input, like the others, has some drawbacks and some advantages.

The main challenge I faced from choosing that input was related to the high human dependency of vector data. People have to make choices regarding the modelling of roads with vector data. Reasonably turns out that there is no consistency in the way that roads are modelled with vector data. Thus, establishing a generic methodology that works with data from different sources was difficult. Moreover, final width estimation strongly depends on the features that are selected to be modelled with vector data. For example, in this thesis, most of the times road polygons used correspond to the drivable area of a road while centerlines correspond to the geographic center of these polygons. Other approaches are also possible. If the road lanes are modelled with separate polygons, and centerlines correspond to the geographic centers of these lanes, then lane width estimation is also possible. Same with other parts of road networks such as sidewalks, bike lanes, etc. Thus, it turns out that human-made modeling choices strongly influence the outcome of our methodology.

While other inputs appear to have low to zero human dependency, vector data was selected to overcome some of the difficulties that may arise with other inputs. First, the availability of such data. While vector areal representations of roads are rarer than linear representations, there is still some information available for free as opposed to, for example, high-resolution satellite imagery. There are many online services (geo-portals) that provide road polygons and centerlines for free. Some of these open-data were used during this research. In addition, vector data usually has a simple structure. While other inputs (e.g. LiDAR pc) can be quite large and/or complex for proper data analysis by a user unfamiliar with the input, vector data is rather simple. The methodology developed in this project is a user-friendly methodology and could possibly be used by scientists not related to the scientific field of Geomatics. Since the width of the road is important for various applications, the existence of a methodology that can be used by non-experts in this scientific field is desirable.

At this point, it is necessary to mention that vector data is a processed product. It is usually generated either from remote sensing technologies or by measurements in the field. Therefore, the accuracy, as well as the availability of such input, depends on those techniques.

Standardization of road vector data

As aforementioned, the main challenge I faced in implementing a generic approach is related to the fact that there is no unique way of modelling roads with vector data. To address this, a methodology that standardizes road modelling based on a prototype was developed. The modelling approach that followed by Toronto dataset, was chosen as prototype. From the results presented in § 6.3, I can claim that I reached my goal for ordinary road cases.

The main characteristic of this approach, which led me to choose it as a prototype, was the consistent way of modelling the intersections. While other datasets that follow strategies where intersections are explicitly modeled (Den Hague, Montreal) exist, Toronto chooses a simpler way to do it. In particular, Toronto uses a unique and simple polygon to represent each intersection (see Figure 11). Even in cases of more complex intersections or roundabouts, a maximum of 2 polygons is used to represent them. This particularity of this modelling approach is important since it will help us achieve an additional deliverable of this project. A methodology that identifies the location and the type of intersections has been developed. The existence of a simple and unique polygon for the different intersection types enhances this process.

An important factor to consider before standardizing the polygons of a particular area is the buffer size used to create the intersection polygons (see § 5.2, step 6). The size of this buffer strongly affects the final size of the intersection and the non-intersection polygons. For this thesis, a predefined number for buffer size was selected. Other approaches to automatically determine the buffer size are also possible. For example, the size of the buffer could be determined based on the size of the road polygons before standardization. With this approach, different sizes will be used for different areas. This can lead to better results in some cases. For example, if we want to standardize the road polygons of an area where the roads are generally quite short and narrow, the predefined value I used in this thesis might not be suitable. This value can lead to huge intersection polygons compared to the actual size of the intersections.

Intersection identification

Another extra deliverable of this research was the development of a methodology that identifies the location and the type of intersections. The need for this extra feature arose from the hypothesis that intersections would add noise to the overall width estimation process (see § 2.3). Analyzing the results of § 5.3.1. I concluded that the hypothesis was correct. Looking at the data, it can be claimed that intersections could add noise to the width estimation procedure. Moreover, based on my findings, intersections are particularly hazardous for traffic accidents [8]. They have been identified as crash “hot spots” for dangerous driving leading to crashes [56]. Since one of the main objectives of this thesis is to relate the final width estimation methodology with the road safety management application, it seems that special treatment is required in these parts.

Therefore, by exploiting the characteristics of Toronto modelling and by using the areal representations of the newly created intersections, a methodology that identifies the location and the different intersection types was developed. Using some ground truth labels, I tested my identification approach and the results were quite promising (out of 431 intersections 398 were identified and categorized correctly, see § 6.3).

By using this approach I was able to identify and exclude intersections from the final width estimation procedure. Other approaches are also possible for different handling of these parts. For example, instead of removing the whole intersection polygon, an approach that identifies and removes only the ‘incorrect’ measurements would be desirable. Based on the analysis in § 5.3.1, although excluding intersections results in measurements quite closer to ground truth, there is still a slight deviation from it (see Table 2). Therefore, an approach that removes only the outliers that result from intersections and keeps the rest measurements would increase the accuracy of the width estimation process.

In § 3.2 an approach that identifies different intersection types for road safety purposes was presented [61]. The difference with my approach lies in the input used. Wijnands et al. [61], in their approach they used vector data to determine the location of an intersection but not to distinguish between different types of intersections. To do this, they used satellite imagery. As already mentioned, the methodology developed in this graduation project uses data that is freely available, easily accessible, and has a simple structure. Thus, vector data from open sources are used for all the features developed. Another difference is the way that intersection types are defined. Wijnands et al. [61], rather than using predefined types of intersections similar intersections are obtained through unsupervised clustering of similar images. Thus, the final categories of intersection types were the result of this clustering procedure. In contrast to this, I used predefined types based on the main intersection types that can be found in Toronto dataset. While having predefined types observer bias might be included, for this thesis this served well. If a more detailed approach that handles intersections differently was developed then, a reconsideration of the overall process might be needed.

Extra features, extra value?

Along this thesis, two additional features were developed. Standardization of road modelling and intersections identification seemed to be two separate features interconnected and inextricably linked to our overall goal. But let's take a step back and think about other benefits and possible applications that these two deliverables could have. As was discussed, intersections can be seen as a special part of road networks. They can have many different configurations and a rather complex structure. For example, suppose an urban planner wanted to investigate the effect that different types of intersections could have on traffic management. Wouldn't it be an advantage to have a methodology that identifies and distinguishes between different types of intersections? In addition, other methodologies that use vector data and have as their main goal to have a general application have been developed or are about to be developed. Having a road modelling standardization approach could benefit all kinds of applications that need to be applied in different datasets. It is my belief that the side benefits of those 2 features may be even greater than those for which it has been developed.

Width clustering for road safety purposes

In this thesis, an approach that creates new roads based on similar width measurements along their geometry was proposed. I have developed this methodology to be used in practice and specifically to benefit the road safety management application. The overall idea was driven by the argument that road width is not a single numerical value and it can be interpreted differently by the different road users. The correlation of the road width with the respective user or application, is particularly important. After exploring the different needs of the selected application, final width clustering approach was developed. This approach appeared to have a significant 'double' impact on road safety management application.

First, by evaluating clustering for two real-world cases I found that different needs of different road users can be addressed with clustered roads. Pedestrians, cyclists, and other road users seem to be interested in some specific features of a roads. The change in the number of lanes, the existence of on-street parking, the narrow points of the road networks, etc., appear to be quite important for the safety of different users. I examined how clustering approach can provide such information to road users. I argued that if I could identify the places where width change occurs and provide a new cluster (road) with more representative width values it would be beneficial to road safety application. Based on the results of the 2 clustering approaches that were tested in § 6.4.1 (Table 5), I came to conclusion that width clustering can be used to identify important width changes of a road in terms of new clusters. From the 2 clustering approaches tested, a certain approach resulted in clusters with a high degree of similarity to the predefined ground truth clusters for both real-world cases.

Evaluating the results that are present in § 6.4.1, I could suggest that there is no one correct clustering approach. Different cases may have different needs in terms of the sensitivity of a road split into clusters based on width changes. For example, we may need a clustering approach that is quite sensitive and detects even small width changes, while in other cases we might need a more conservative approach that creates a new cluster only when a significant change in width occurs. In general, the results of clustering to produce roads with more detailed width information depicting changes in the width of a road are quite promising. Finally, it is important to mention that for evaluating the results of the different clustering approaches, an external clustering validation method was used (see § 2.4.4). Other methods that use internal data only (check how similar are measurements of same cluster and how unsimilar they are with the measurements of the other clusters), could lead to different results.

The second argument that led me to implement this specific approach was that clustered roads would be more representative in terms of road geometry. By the analysis in § 6.4.2 I derived the conclusion that indeed clustered road centerlines are assigned with more 'detailed' width values in comparison to the original road centerlines. To reach this conclusion, I performed a test on the roads of the central area of Toronto. I examined how the standard deviation around the mean road width value changes before and after clustering. I found that, there is a remarkable reduction in the mean and median values of the overall standard deviations when clustering is applied (88% and 89% respectively). It is important to notice, that roads in Toronto dataset can be characterized by regularity, and the initial mean and median values of standard deviations were already small (0.6 and 0.09 respectively). If the test was performed for a different area where the roads are more irregularly shaped, the clustered roads may result in an even greater reduction in standard deviation values.

Throughout this project other possible approaches of how roads can be divided and linked with width estimations were examined (see § 2.2). In particular, four approaches were analyzed, but there are many more. The main drawback of width clustering approach is that the original road centerlines are manipulated (see Table 1). This includes the risk of losing important information that is linked with those original lines. The easiest solution for this is to keep track of the original centerlines by using an id. This id could be assigned as a foreign key to the new clustered roads that result from each centerline. Then, the linkage of clusters with the higher level centerline that they 'belong' is possible. Moreover, with our approach we are passing the final width estimation only to the linear road representation (centerline). While for the purposes of this thesis this is enough, other applications might require the width estimations to be assigned to the areal representations as well.

Road safety analysis

The main purpose chapter 7 was to explore whether certain features implemented for this thesis could affect the result of a road safety analysis. Three different datasets corresponding to different processing levels were used for the safety analysis performed for the central area of Toronto. Those datasets led us to different results regarding the relation between road width and road accidents. The difference between the 3 datasets lies in the way the width was estimated and linked to the roads. 1st dataset correspond to 'raw' data. Original polygons/centerlines were used. From the analysis based on that dataset, a tendency of wider roads to have slightly more accidents was observed. However, the coefficient of determination (R^2) is pretty low (0.05 and 0.03 for mean and median width values respectively). Thus, we cannot claim that there is a clear correlation there.

For the 2nd dataset, road polygons have been manipulated before being used for width estimation. Standardization methodology and intersection identification methodologies have been applied to the original polygons. Then, intersection polygons have been removed from the width estimation process. The tendency that was observed previously no longer exists. From Figure 108 and Figure 109 it is obvious that there is no correlation between width and accident concentration. The distribution of points appears to be completely random. Coefficient of determination (R^2) showed a big reduction for both mean and median width values (from 0.05 to 0.003 and from 0.03 to 0.008 respectively).

Finally, a 3rd dataset was used. Now, another feature that was implemented during this thesis was used. Width clustering methodology (§ 5.4) has been applied to original centerlines. Safety analysis using the roads of the 3rd dataset also led to a different conclusion. The wider roads seemed to have slightly fewer accidents. The overall trend now is exactly the opposite of what was observed with the analysis based on the 1st dataset. However, as in the first case the coefficient of determination (R^2) is pretty low (0.09 and 0.07 for mean and median width values respectively). Therefore, the correlation between width and accidents still seems to be weak.

While for all three cases the correlation between traffic accidents and road width seems to be very weak (or no correlation at all), we can observe some interesting changes in the numbers and distribution of accidents:

Intersection removal → Tendency changes

When we processed the original data and we performed the safety analysis excluding the intersections, we observed a big reduction in coefficient of determination. Even if the correlation is quite weak with the 1st dataset, a trend was observed. Although it is not necessary that wider roads are more dangerous, when intersections are included in the overall process, the probability of a road over 15 meters being associated with more accidents is increased (see Figure 103 and Figure 104). This tendency of wider roads is no longer present when the intersection polygons are removed. Both for datasets 2 and 3 this trend no longer exists. Especially based on the analysis of 3rd dataset, the tendency is exactly the opposite.

In § 5.3.1 we explored how intersection polygons can add noise to the width estimation process. In particular, existence of intersections lead to some 'incorrect' large width measurements that influence the result of width estimation process. Thus, the roads seem to be wider in general (see Table 2). Since the measuring lines are defined based on the length of the centerline (measuring interval that 'cuts' the line every x meters, see § 5.1), it turns out that shorter roads are affected more by the existence of intersection polygons than longer roads. For smaller roads, incorrect measurements will be a larger part of their total measurements. Mean width value of a 330 m long road is less affected by the intersection polygons than the mean width value of a road with a total length of 69 m. For more details on this, a test was performed in § 5.3.1. Table 3 contains data on how the width values of 150 roads found in Toronto are affected based on their length. From this table, it can be argued that roads under 100 meters experience a quite larger change in their width compared to roads over 300 meters.

In our case, out of 160 roads that are related with at least one KSI accident, 63 roads have length less than 100 meters. 84 roads have length between 100 and 300 meters and only 13 roads are over 300 meters. Therefore, a big percentage of the roads that concentrating accidents are less 100 meters long (39.4%). Based on the analysis made in § 5.3.1 (Table 3), shorter roads seem to be affected a lot by intersections existence. The final width values may differ by more than 5 meters from the actual road width. Therefore, the change in the trend makes sense. Many of the short roads (less than 100 meters) that are associated with KSI accidents may at first have been considered much wider than they actually are. After removing intersections, some roads that were previously assigned with a width value of more than 15 meters now, could be considered as narrow roads. They may have been assigned a width value of even less than 10 meters. Figure 117 illustrates an example of a short road and of a long road and how they are affected by the existence of intersections.

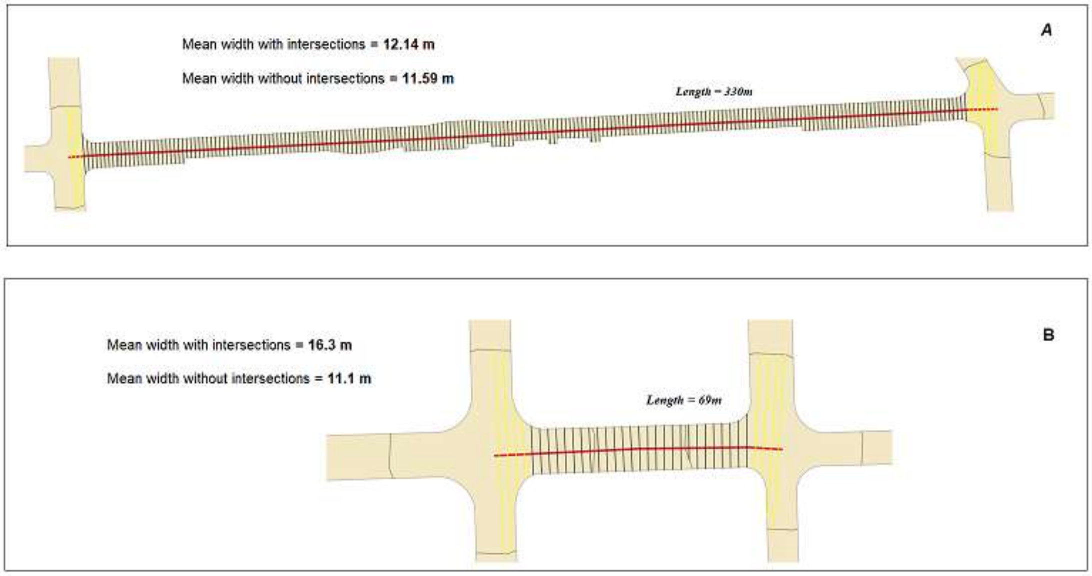

Mean width with intersections = 12.14 m
Mean width without intersections = 11.59 m
A
Length = 330m
Mean width with intersections = 16.3 m
Mean width without intersections = 11.1 m
B
Length = 69m

Figure 117 The mean width estimation for a long road (A) is changing only by 0,5 meters after removal of intersection polygons. For a shorter road (B) of length less than 70 meters, the final width estimation differs more than 5 meters

Width clustering applied → Correlation is improved

Another interesting observation is the improvement of the correlation of the examined phenomena after the width clustering approach was applied. Figure 118 indicates the correlation between traffic accidents (normalized) and the mean width of roads for the 3 datasets. Looking at those 3 graphs, we can observe that the distribution of points in the 3rd graph seems to have a more canonical shape compared to the other 2. While in the first 2 cases distribution of points seem to be random, in the 3rd case it appears like two types of data are included. There are the outliers with extreme values and the rest of the data that seem to follow a more canonical distribution. Moreover, the outliers seem to be reduced significantly. The improvement of the correlation can be claimed also based on the large change of numbers. Coefficient of determination changed from 0.03 (1st dataset) and 0.003 (2nd) to 0.09 after width clustering was applied (3 times higher compared to the 1st dataset that 'raw' data were used).

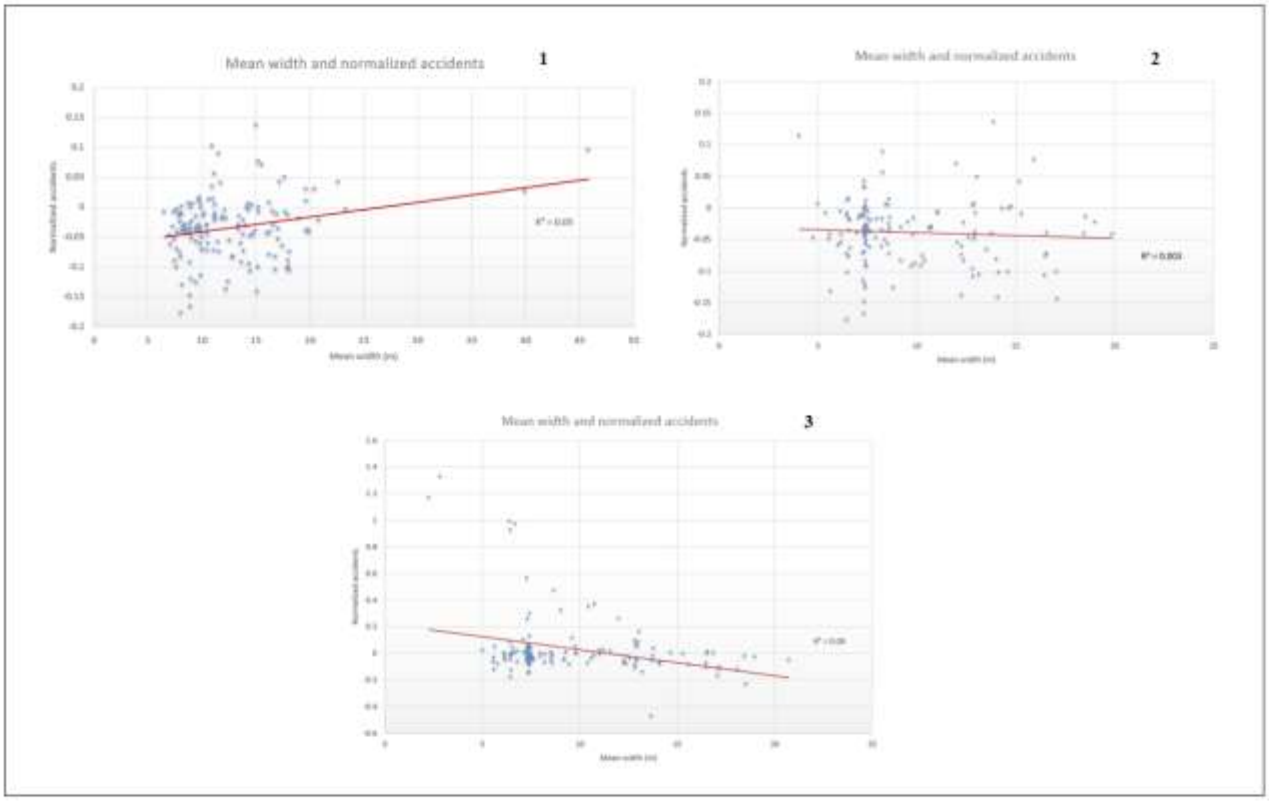


Figure 118 Correlation of mean width values of roads and normalized accidents for the 3 different datasets

The fact that some important factors that should be taken into to account to derive a less biased conclusion for road safety are missing (human behavior, traffic conditions etc.), combined with the significant improvement in correlation when the more ‘detailed’ roads used, made me think that further research is needed. Although the final values are still low enough to support the correlation between road width and road accidents, it made me think that a thorough review of the width estimation process is required before examining the correlation with road accidents.

<u>*Geographical implications*</u>

Along this thesis, road vector data from several datasets have been explored and manipulated. One of the goals of this graduation project was to improve existing methodologies in such a way that a more general use is possible. Thus, for the formulation of our methodology and for the evaluation of our results, road vector data from different open-source datasets was used. One of our main considerations should be that the different datasets used, also represent different geographical areas. Although, some of the differences of road polygons/centerlines among different datasets are related to the different modelling strategies followed, there are also some differences related to the different structure of the actual roads of two areas. For example, it is likely that the streets of one city tend to be wide, long and have a regular shape, while in another city, streets show many geometric changes in their length and are generally smaller and narrower. Therefore, our methodology and our results may have been influenced by these specificities of the geographical areas used to guide our decisions and evaluate our results. It is important to mention that different implications might arise from different geographical areas in the same city as well. For example it might be that the streets of the center of a city have quite different structure from the streets of the suburbs.

But let's discuss it with an example. In § 5.4.2 we defined our evaluation approach for our clustering methodology. Before defining the different indicators we explored the relation between length of a road and width changes that may occur. By using 50 polygons of the central area of Helsinki and 50 polygons of the central area in Toronto, we claimed that the longer a road is the more possible to occur a change in its width. This conclusion, although correct for the cases under consideration, there may be cases where the correlation of the length of a road with the changes in its width does not make any sense. Figure 119 shows 200 road polygons that can be found at a suburb of the city of Montreal. The polygons seem to have very regular shape. If we had used those polygons to explore the relation between width changes and road length, then our conclusions might be different. Figure 120 summarizes the results for those polygons.

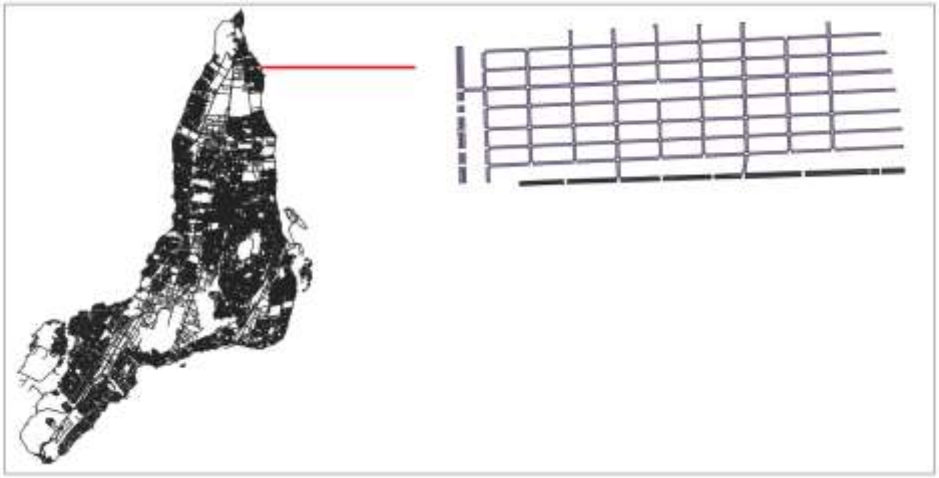


Figure 119 Polygons of a suburb area at the city of Montreal. This polygons are quite regular with not many changes in their geometry

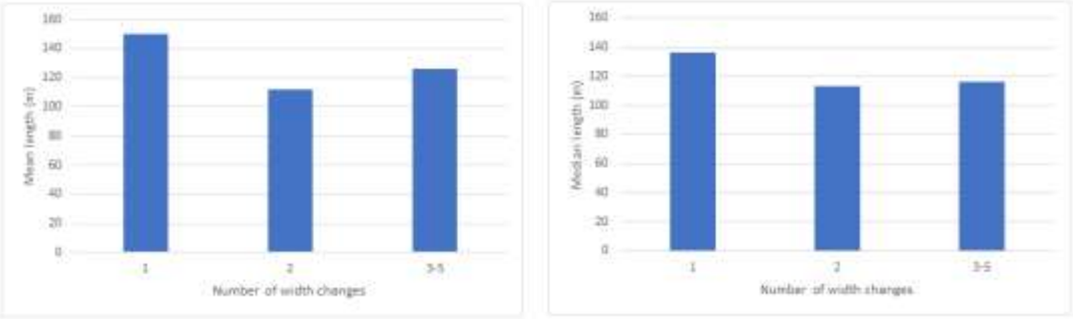


Figure 120

From the figure, it becomes apparent that the claim that the longer a road is, the more likely it is a change in its width to occur, is not valid anymore. But, this claim played a significant role in our evaluation approach. Thus, if we were to apply our clustering methodology to a specific place we might need to rethink of some parameters.

Let us now consider another example. But this time the example is related to streets of different areas of the same city. In § 5.3.1 we have discussed the impact that intersections might have in width estimations. There, we showed an example of how mean and median width values are influenced by the by the presence or absence of intersection polygons during the estimation process. For a selected area, we computed width with and without intersection polygons. Our findings led us to the conclusion that intersections can have a rather big impact on the width estimations. Although for most the datasets this conclusion is valid, there might be cases where intersections do not affect the result of width estimation that much. Figure 121 shows road polygons of 2 different areas of the city of Helsinki. One area corresponds to a central area of Helsinki (left) while the other corresponds to an area outside of the city center (right). The different parts of the city have the same area but their road polygons are showing some differences. First, the roads seem to be more dense in the central area. The total number of polygons in that area is 135 while in the area outside of the city center 75 road polygons can be found. Moreover, at the central area more intersection polygons are present. Finally, some large intersection polygons such roundabouts are present in the city center while such intersections are missing from the non-central area.

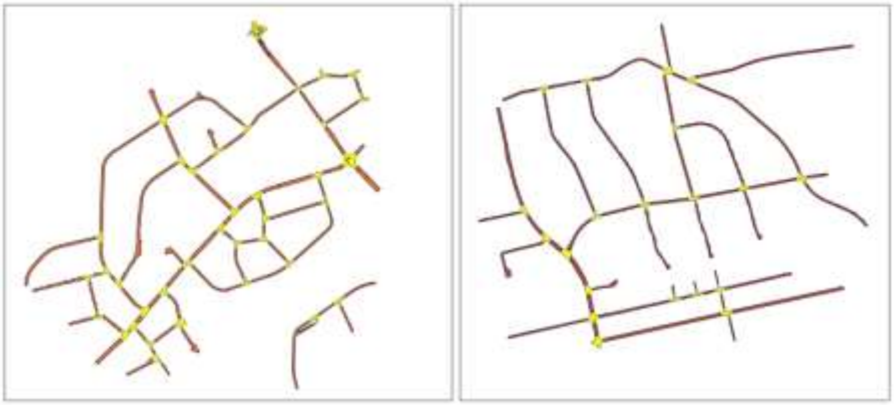


Figure 121

Table 14 summarizes the impact of intersection polygons in width mean and median values of road polygons for the 2 different areas. As results from the table, intersection polygons seem to have a big impact in the central area, since the mean and median value have a difference over 1 meter with and without intersections. On the other hand, the impact of intersection polygons does not seem to be that big for the other part of the city. The mean width value changes less than 0.5 meters while median width value changes only 0.25 meters.

City part	Measuring approach	Mean width (m)	Standard deviation from mean (m)	Median width (m)
Central	Include intersection polygons	10.14	2.03	9.95
Central	Exclude intersection polygons	8.544	1.32	8.91
Non-central	Include intersection polygons	8.92	1.36	8.45
Non-central	Exclude intersection polygons	8.46	1.17	8.2

Table 14

From the 2 examples that discussed above, we can derive the conclusion that geographical areas can lead to different implications. The methodology that is developed in this thesis has a generic nature and the decisions that have been made are justified based on datasets of different geographical areas.

8.2 Research Overview

Purpose of this section is to review the research questions I have defined for this thesis in § 1.1. For each question I provide a short answer that is supported by evidence that is available in previous sections of this thesis. First, the several sub-questions will be answered in order to help me answer the main-question.

Sub-question 1: ***"How road width can affect the safety of different road users?"***

Answer:

During this project, road safety management was sub-divided into 3 categories based on the different road user groups. Cyclists, pedestrians, and motor-vehicle drivers coexist on the same roads. The relation of their safety with road width was examined separately. I have found that different users have different needs, but there are also some common points.

As it comes from the literature, width changes that occur along the geometry of a road usually indicate features that are particularly important for the safety of the various road users. The existence of narrow points, on-street parking spots, median strips, changes in the number of road lanes, etc. are some of the characteristics of road geometry that appear to be particularly essential for the safety of different road users.

Moreover, many studies correlate the width of the road with the number of road accidents. While other factors also seem to be important (traffic conditions, human behavior, etc.), road width is considered by many researchers to be one of the main factors in the road environment that affects the total number of accidents on a road. The main reason for this may be that road width has a crucial effect on self-selected speed on road sections. Thus, a notable change in road width leads to reduced/increased self-selected speed, which would result in a reduced/increased crash risk. Therefore, width changes seem to affect the safety of all the various road users.

Sub-question 2: ***'How can road vector data be standardized in such a way as to benefit the development of a generic methodology for estimating road width?"***

Answer:

In this thesis, I have presented a methodology for standardizing road vector data based on Toronto modelling. During this project, I manipulated vector data from different sources. Thus, I explored some different road modelling approaches in depth. Moreover, I realized that the biggest challenge for width estimation methodology lies in the different approaches that followed for modelling the intersections. Thus, I choose a modelling approach that explicitly models those special parts of road networks. Moreover, another benefit of this approach is that modelling of intersections is done in a simple and consistent way. Other datasets were also found to model intersections explicitly but Toronto uses just a single (max 2 in case of most complex intersections such as roundabouts) polygon for each intersection.

Thus, I used the special characteristics of this approach in order to establish a more robust methodology for width estimations. After standardizing vector data based on Toronto modelling intersection polygons could be identified. For the purposes of our approach, I exclude intersection polygons from the overall process of width estimation since they appear to cause noise. Other possibilities for the different manipulation of those parts are possible as well.

Sub-question 3: ***“In what way original roads could be divided to benefit road safety management application?”***

Answer:

As aforementioned, roads with vector data can be seen in different ways. While each dataset follows a certain modelling approach and uses some original polygons/centerlines to represent each road, other possible approaches exist. During this research, different ways that roads could be divided with vector data have been discussed. In § 2.2 four possible approaches of how original road representations could be divided and linked with width estimations are explored. Each approach has some benefits and some drawbacks.

The main goal of the analysis made in § 2.2 was to result in a way of dividing the roads that could be used to benefit road safety. The width clustering approach appeared to be the most suitable unit of measure approach based on the needs of my selected use case. In § 4.2 the positive impact that this approach of dividing original road centerlines is expected to have on road safety was explained. The overall goal of this approach was to divide the original road centerlines based on width changes that occur along road geometry. This could be used for the identification of some special features of roads (narrow spots, on-street parking, etc.) that appeared to be essential for the safety of some road user groups. In addition, this will result in new, more representative roads in terms of width. Therefore, it will add meaning to the overall process of correlating road width with traffic accidents.

Based on the analysis I did (see § 6.4.1) width clustering could be used to identify specific road features in terms of new clusters. Two real-world cases have been examined and a certain clustering approach seems to have quite promising results for both. Moreover, looking at the results presented in § 6.4.2, I came to the conclusion that the newly clustered roads are indeed more representative in terms of road geometry. The standard deviation of new roads around their width values is significantly reduced compared to the original roads. Therefore, it makes sense to reconsider the correlation of width and road accidents with the use of more detailed roads.

Sub-question 4: ***“How do the different aspects of the final width estimation methodology affect the process and result of a road safety analysis?”***

Answer:

Different features were implemented during this project to reach the main goals. In chapter 7, some of those features were checked for their influence on the process and the result of a road safety analysis that correlates traffic accidents with road width. First, the standardization of vector roads and the intersection identification methodologies were explored. From the results presented in chapter 7, by using those features in the width estimation process, the results of the safety analysis are different. While with the original dataset the wider roads appear to have slightly more accidents when intersection polygons are created, identified, and removed from the width estimation process, this tendency is no longer present. In particular, the coefficient of determination faces a huge reduction, which indicates that there is no correlation at all between width and accident concentration A hypothesis as to why this tendency of wider roads changes is given in the previous section.

Finally, the width clustering approach was tested for its influence on the safety analysis. When the more detailed roads are used, the correlation is significantly improved. There is a remarkable increase in the value of the coefficient of determination that examines the correlation between accidents and width (more than three times higher compared to ‘raw’ data). In addition, the distribution of points now seems to have a more canonical shape, while the outliers faced a notable reduction.

While in any case, the correlation between road width and accidents seems to be weak (or there is no correlation at all), some interesting observations are made as we further process the data.. Therefore, it can be argued that the different aspects of the final methodology have an impact on a road safety analysis.

Main Question: ***"How road width estimations can be derived from vector data to benefit road safety management application?***

Answer:

In this research, I have examined different approaches on how width estimations can be derived, stored and linked to road vector data. While a methodology that derives width estimations from vector data exists (Hoffmans W. [25]), no emphasis has been placed on the application-dependency of road width. In order to find a way in which width estimations could potentially benefit the road safety application, I first examined its relationship to road width. A methodology was then developed based on this relationship.

An approach that new road centerlines are created based on similar width measurements was developed. This methodology appears to have a double positive impact on road safety management. On one hand, it allows the identification of some special width changes that occur along roads in terms of new separate centerlines. In addition, it adds more meaning to the accidents-width correlation process since it creates new, more geometrically representative roads.

Throughout this research, some other features were developed. Those features developed to serve as auxiliary features for achieving the main objectives. The standardization of road modelling and the intersection identification methodologies have helped to make the overall width estimation process from vector data more robust and generally applicable.

8.3 Future Work

In this section, I provide a list of some interesting topics for follow-up research. It is my belief that, it is legitimate to look further at some aspects of the work presented in this thesis.

Handle intersections differently

Intersections appeared to be particularly important aspect of this thesis. While they seem to greatly influence the width estimation process, in this thesis they are just excluded from the overall process. A different handling of those parts is desirable. Based on the analysis in § 5.3.1, although excluding intersections results in measurements quite closer to ground truth, there is still a slight deviation from it. Therefore, an approach that removes only the outliers and keeps the rest measurements would increase the accuracy of the width estimation process.

Developing a methodology that finds only the "wrong" measurements associated with the existence of intersection polygons is made easier by using the approach to identify different types of intersections. If the shape of the intersection polygon is known in advance, then the user knows where to look for the "wrong" measurements. For example, if we are about to estimate width for a cross intersection we know that the outliers will lie in the middle of the polygon. Therefore, the extension of the methodology developed in this project to identify as many types of intersections as possible is also desirable.

Standardization based on different prototype

For the purposes of our project, Toronto modelling approach served well as a prototype for the road vector standardization process, but some alternatives may be interesting to explore. In particular, the choice of this approach as a prototype was guided by the feature of explicit intersection modeling. Especially by the simplicity in the way that intersections are modelled which makes the overall standardization process easier to implement. Other approaches might be harder to implement but they might come with some advantages. For example, having an approach that models intersections in a more detailed way (more than one polygon) might be desirable. Especially if more emphasis has been given to intersections in general. Such an approach can be found in the dataset of Montreal[46].

Width clustering for other applications

Width clustering approach that has been developed, was driven by the needs of road safety management application. However, other applications could benefit from this approach. Based on the analysis I did, I came to the conclusion that clustered roads are more representative in terms of road geometry. So, wouldn't it be preferable for a forensic study investigating the correlation between road width and crime to use more representative roads?

Moreover, based on the results of chapter 7, it can be claimed that width clustering had an impact on the outcome of the safety analysis performed for central area of Toronto. As discussed, the correlation between the examined phenomena experienced a significant change when the more detailed roads were used. The question logically arises: how can the width clustering approach affect the process and the result of other applications? Or, for example, what would be the result of the safety analysis if all the necessary factors influencing road accidents were taken into account? Would clustering also increase the correlation or perhaps its effect would be less in these cases? Such questions are desirable to be answered and a possible way to do this is to use width clustering of roads for purposes of other applications.

Explore application dependency of other geometrical characteristics

Finally, other geometrical characteristics of the roads could be examined for their application dependency. Road length, sinuosity, and other values might add extra benefits to different applications if they are examined for their possible interpretations. So far, I haven’t thought about the width of the road in the way I did during the implementation of this graduation project. Thus, reasonably turns out, that a fundamental review of other road characteristics might benefit various applications.

A| Road polygons standardization statistics

Toronto	Polygons phase	Number of polygons	Initial type of intersection modelling	Number of intersection polygons	%intersection polygons	Mean area of polygons (m^2)	Standard Deviation from mean (m^2)	Median area of polygons (m^2)	Road polygon with max area (m^2)	Road polygon with min area (m^2)
	Original	7527	Explicit modelling of intersections	1664	22.10%	568.7	1520	365	86700	12
	Standardized	7541	Explicit modelling of intersections	1752	23.20%	567.3	688	410	24105	0.02
	Before-After	+14 after (0,2% increase)	-	+88 after (5,3% increase)	+1.1% after	-1.4 after (0.25% reduction)	-832 after (54.7% reduction)	-45 after (12.3% reduction)	-62595 after	-11.98 after

Helsinki	Polygons phase	Number of polygons	Initial type of intersection modelling	Number of intersection polygons	%intersection polygons	Mean area of polygons (m^2)	Standard Deviation from mean (m^2)	Median area of polygons (m^2)	Road polygon with max area (m^2)	Road polygon with min area (m^2)
	Original	5367	No explicit modelling of intersections	0	0%	388	1313	9.3	21015	0.05
	Standardized	5468	Explicit modelling of intersections	1043	19%	381	1044	41.1	21015	0.001
	Before-After	+101 after (1.9% increase)	-	+1043 after	+19% after	-7 after (1.8% reduction)	-269 after (20.5% reduction)	+31.8 after (342% increase)	0	-0.046 after

Montreal	Polygons phase	Number of polygons	Initial type of intersection modelling	Number of intersection polygons	%intersection polygons	Mean area of polygons (m^2)	Standard Deviation from mean (m^2)	Median area of polygons (m2)	Road polygon with max area (m^2)	Road polygon with min area (m^2)
	Original	22777	Explicit model of the different parts of each intersection	6535	28,7%	337,2	554,2	118,6	6584	0,17
	Standardized	15588	Explicit modelling of intersections	1631	10.50%	492.6	624.6	222.2	5929	0.001
	Before-After	-13311 after (58.4% reduction)	-	-4904 after (75% reduction)	-18.2% after	+155.4 after (46% increase)	+70.4 after (12.7% increase)	+103.6 after (87.3% increase)	-655 after	+0.169 after

Poznan	Polygon phase	Number of polygons	Initial type of intersection modelling	Number of intersection polygons	%intersection polygons	Mean area of polygons (m2)	Standard Deviation from mean (m2)	Median area of polygons (m2)	Road polygon with max area (m2)	Road polygon with min area (m2
	Original	2569	No explicit modelling of intersections	0	0%	1358.2	3043.1	319.7	77837	0.001
	Standardized	4896	Explicit modelling of intersections	1128	23%	715.6	1215.2	397.4	26934	0.001
	Before-After	+2,327 after (90.5% increase)	-	+1128 after	+23% after	-642.6 after (47.3% reduction)	-1,827.9 after (60% reduction)	+77.7 after (24.2% increase)	-50903	0

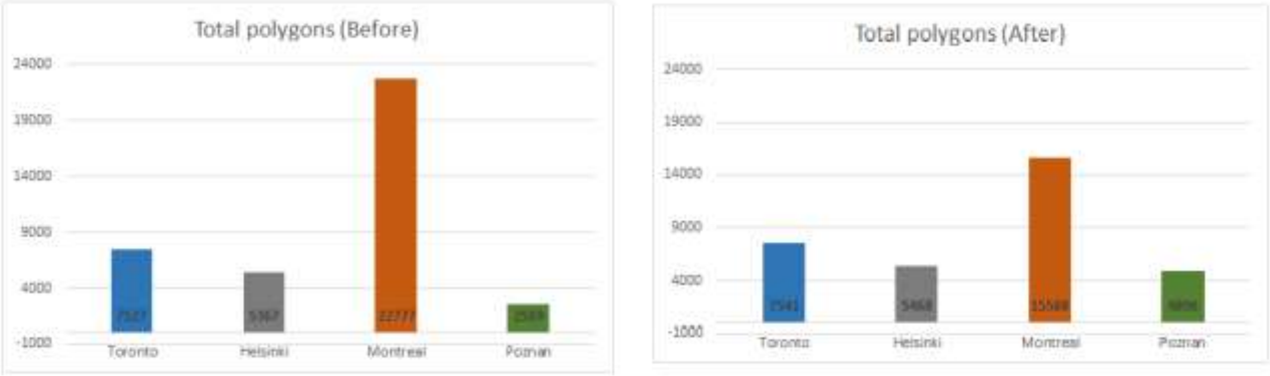


Figure 122 Total number of road polygons before and after standardization process

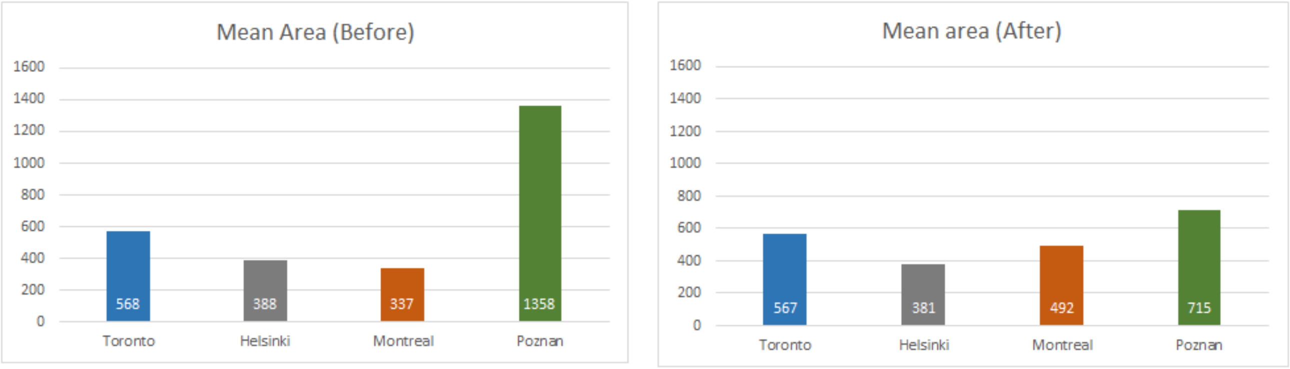


Figure 123 Mean of road polygons before and after standardization process

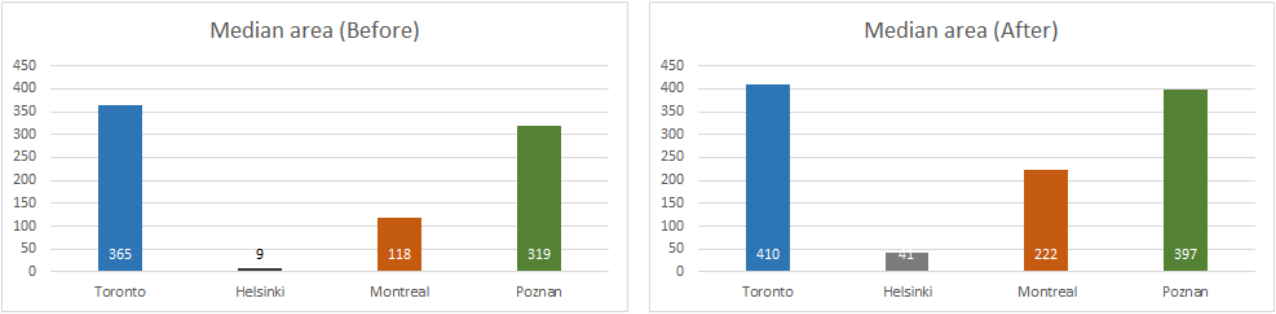


Figure 124 Median area of road polygons before and after standardization process

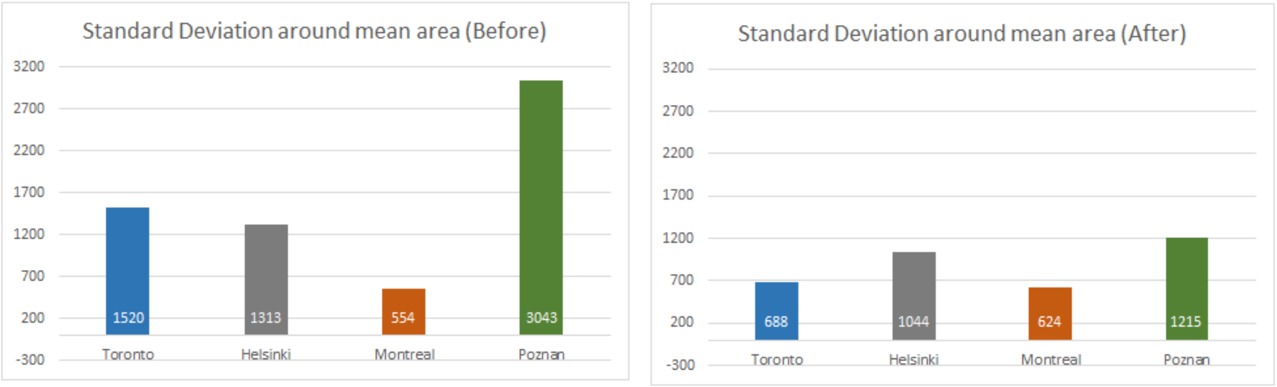


Figure 125 Standard deviation around mean area of road polygons before and after standardization process

B| Normalized accidents and width ranges

Dataset 1 – Relationship of width ranges and normalized accidents statistics

Width range	Minimum	Median	Mean	Maximum
7-9 meters	-0.148	-0.038	-0.051	0.007
9-11 meters	-0.114	-0.034	-0.032	0.034
11-13 meters	-0.138	-0.019	-0.029	0.089
13-15 meters	-0.108	-0.03	-0.04	0.005
15-17 meters	-0.141	-0.03	-0.02	0.077
>17 meters	-0.143	-0.03	-0.034	0.05

Table 15 Mean width ranges and normalized accidents for dataset 1

Width range	Minimum	Median	Mean	Maximum
7-9 meters	-0.11	-0.031	-0.0313	0.043
9-11 meters	-0.092	-0.032	-0.037	-0.004
11-13 meters	-0.138	-0.0415	-0.044	0.07
13-15 meters	-0.141	-0.065	-0.044	0.05
15-17 meters	0.0007	0.016	0.007	0.07
>17 meters	-0.143	0.03	-0.03	0.095

Table 16 Median width ranges and normalized accidents for dataset 1

Dataset 2 – Relationship of width ranges and normalized accidents statistics

Width range	Minimum	Median	Mean	Maximum
< 7 meters	-0177	-0.04	-0.037	0.115
7-9 meters	-0.167	-0.033	-0.035	0.089
9-11 meters	-0.092	-0.042	-0.054	-0.015
11-13 meters	-0.138	-0.043	-0.034	0.137
13-15 meters	-0.141	-0.065	-0.048	0.05
15-17 meters	-0.106	-0.04	-0.026	0.077
>17 meters	-0.143	-0.04	-0.059	-0.013

Table 17 Mean width ranges and normalized accidents for dataset 2

Width range	Minimum	Median	Mean	Maximum
< 7 meters	-0177	-0.04	-0.03	0.115
7-9 meters	-0.167	-0.033	-0.03	0.09
9-11 meters	-0.092	-0.032	-0.04	0.077
11-13 meters	-0.138	-0.044	-0.04	0.13

13-15 meters	-0.141	-0.06	-0.04	0.042
15-17 meters	-0.106	-0.04	-0.04	
>17 meters	-0.143	-0.04	-0.05	-0.013

Table 18 Median width ranges and normalized accidents for dataset 2

<u>Dataset 3 – Relationship of width ranges and normalized accidents statistics</u>

Width range	Minimum	Median	Mean	Maximum
< 7 meters	-0175	0.002	0.18	1.33
7-9 meters	-0.141	-0.014	0.015	0.57
9-11 meters	-0.08	0.085	0.075	0.37
11-13 meters	-0.105	-0.037	-0.005	0.26
13-15 meters	-0.47	-0.038	-0.05	0.16
15-17 meters	-0.104	-0.0001	-0.03	0.01
>17 meters	-0.23	-0.07	-0.07	0.11

Table 19 Mean width ranges and normalized accidents for dataset 3

Width range	Minimum	Median	Mean	Maximum

< 7 meters	-0175	0.008	0.15	1.33
7-9 meters	-0.156	-0.013	-0.018	0.57
9-11 meters	-0.08	0.006	0.06	0.37
11-13 meters	-0.136	-0.028	0.00004	0.26
13-15 meters	-0.07	-0.0007	0.048	0.47
15-17 meters	-0.104	0.0015	-0.015	0.011
>17 meters	-0.16	-0.044	-0.06	0.016

Table 20 Median width ranges and normalized accidents for dataset 3

Bibliography

[1] Adriano M., Maribel, Y., Santos and Sofia Carneiro (2005). Density-based clustering algorithms, DBSCAN and SNN.

[2] Ahmed, Ishtiaque (2014). Road infrastructure and road safety.

[3] Ambros J. (2011). Relationship between road width and safety.

[4] ArcGIS Hub (2021). https://hub.arcgis.com/datasets, Accessed: 20/04/2021.

[5] Beil, Christof & Kolbe, Thomas. (2017). CITYGML and the streets of New York - A proposal for detailed street space modelling. *Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences.*

[6] Brohée, S., van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7, 488 https://doi.org/10.1186/1471-2105-7-488.

[7] Capaldo, Francesco & Nasti, Gennaro. (2012). Analysis of road safety: Three levels of investigation.

[8] Choi, E. H. (2010). Crash factors in intersection-related crashes: An on-scene perspective *Technical Report DOT HS 811 366.*

[9] Claussen, H., Lichtner, W., Heres, L., Lahaije, P., and Siebold, J. (1989). GDF, a proposed standard for digital road maps to be used in car navigation systems. *In Conference Record of papers presented at the First Vehicle Navigation and Information Systems Conference (VNIS '89).*

[10] Davis C. (2020). We Now Have Unprecedented Access to Satellite Imagery. How Do We Turn This Into Action? Available at: https://www.globalforestwatch.org/blog/data-and-research/planet-high-resolution-imagery/

[11] Davies E. (1997). Machine Vision: Theory, Algorithms, Practicalities. Academic Press, 2nd Edition.

[12] Dempsey, Caitlin (2011). "What is GIS." *Retrieved July 1.*

[13] Dr. Michael J. Garbade (2018). Understanding K-means Clustering in Machine Learning.

[14] Egenhofer, M. J. (1993). What's special about spatial? Database requirements for vehicle navigation in geographic space. In SIGMOD *'93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data.*

[15] Elvik, R., Høye A., Vaa T., Sørensen M. (2009). The Handbook of Road Safety Measures, 2nd edition, Emerald, Bradford.

[16] Ewing R., Dumbaugh E. (2009). The Built Environment and Traffic Safety: A Review of Empirical Evidence. *Journal of Planning Literature*.

[17] Fildes M.M, Fletcher M.R and Corrigan J.M (1987). Speed Perception 1: Drivers judgements of safety and speed on urban and rural straight roads, *Report CR54, Federal Office of Road Safety.*

[18] Forgy E. (1965). Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification. *Biometrics, vol. 21, pp. 768-769.*

[19] Freek Boersma (2019): Modelling different levels of detail of roads and intersections in 3D city models.

[20] Gan, Xiaoyu & Fernandez, Ignacio & Guo, Jie & Wilson, Maxwell & Zhao, Yuanyuan & Zhou, Bing-Bing & Wu, Jianguo. (2017). When to use what: Methods for weighting and aggregating sustainability indicators. Ecological Indicators.

[21] Gibbard A, Reid S, Mitche J, Lawton B, Brown E, and Harper H (2004). The effect of road narrowings on cyclists, *TRL Report TRL621, London, Department of Transport.*

[22] Greater Auckland (2010). The importance of road-width. Available at: https://www.greaterauckland.org.nz/2010/12/05/the-importance-of-road-width/

[23] Hauer, E. (1999). Safety in geometric design standards.

[24] Helsinki Region Infoshare - Open data service. https://hri.fi/data/en_GB/dataset/helsingin-kaupungin-yleisten-alueiden-rekisteri. Accessed at: 2020

[25] Hoffmans W. (2018). Divide Width Script BGT. https://github.com/willemhoffmans/bgt_wegbreedte/

[26] Holgado-Barco, B. Riveiro, D. González-Aguilera, and P. Arias (2017). Automatic inventory of road cross sections from mobile laser scanning system. *Aided Civil Infrastructure. Eng., vol. 32, no. 1, pp. 3–17.*

[27] Indika (2011). Difference Between Hierarchical and Partitional Clustering.

[28] Lane PL., McClafferty KJ., Nowak ES. (1994). Pedestrians in real world collisions. *The Journal of Trauma pp. 231-236.*

[29] Lewis-Evans, B. and Charlton, S.G. (2006). Explicit and implicit processes in behavioral adaptation to road width, *Accident Analysis and Preventions. Vol 38, pp. 610-617*

[30] Lee J., and Mannering F. (1999). Analysis of roadside accident frequency and severity and roadside safety management. *Olympia: Washington State Department of Transportation.*

[31] Leung Y., Zhang J. and Xu Z. (2000). Clustering by Space-Space Filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.12, pp. 1396-1410*.

[32] Lun Luo, Yu Zang, Xiaofang Wang, Cheng Wang, Jonathan Li, Sheng Wu & Yuelei Liu (2018) Estimating Road Widths From Remote Sensing Images. *Remote Sensing Letters, 9:9, pp. 819-828,*

[33] Madhulatha T. (2012). An Overview on Clustering Methods. IOSR Journal of Engineering.

[34] Manning D., Raghavan, Prabhakar; Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press.

[35] Mapbox (2019). Mapping guides: mapping for navigation. Available at: https://labs.mapbox.com/mapping/mapping-for-navigation/

[36] Nagesh Singh Chauhan (2019). What is Hierarchical Clustering? Available at : https://www.kdnuggets.com/2019/09/hierarchical-clustering.html

[37] Nedevschi, Sergiu & Danescu, Radu & Marita, Tiberiu & Oniga, Florin & Pocol, Ciprian & Bota, Silviu & Meinecke, Marc-Michael & Obojski, Marian (2009). Stereovision-Based Sensor for Intersection Assistance.

[38] Noland R. B., and Oh L. (2004). The effect of infrastructure and demographic change on traffic-related fatalities and crashes: A case study of Illinois county-level data. *Accident Analysis and Prevention,* pp. 525-32.

[39] Omran, Mahamed & Engelbrecht, Andries & Salman, Ayed. (2007). An overview of clustering methods.

[40] Open Data - City of Toronto. https://open.toronto.ca/dataset/topographic-mapping-edge-of-road/. Accessed at: 2020.

[41] Ordinance Survey (2019). Our history. https://www.ordnancesurvey.co.uk/about/history

[42] Ozbek, M. E., de la Garza, J. M., and Triantis, K. (2010). Data and modeling issues faced during the efficiency measurement of road maintenance using data envelopment analysis. *Journal of Infrastructure Systems.*

[43] Pfitzner, Darius; Leibbrandt, Richard; Powers, David (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems. Springer.*

[44] Portal SIP Poznań. https://sip.poznan.pl/sip/uslugi/get_uslugi/. Accessed at: 2020.

[45] Quartieri, J. & Mastorakis, Nikos & Guarnaccia, Claudio & Troisi, Antonio & D'Ambrosio, Salvatore & Iannone, Gerardo (2009). Road Intersections Noise Impact on Urban Environment Quality

[46] Quebec's collaborative open data hub. https://www.donneesquebec.ca/recherche/dataset/vmtl-voirie-actif# Accessed at: 2020.

[47] Rachel Boba (2001). Introductory Guide to Crime Analysis and Mapping

[48] Ravi R. et al (2020). Lane Width Estimation in Work Zones Using LiDAR-Based Mobile Mapping Systems*, in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 12, pp. 5189-5212.*

[49] Rendón, E., Abundez, I., Gutierrez, C., Zagal, S.D., Arizmendi, A., Quiroz, E.M., & Arzate, H. (2011). A comparison of internal and external cluster validation indexes.

[50] Samuel Langton (2021).The universality of street segments: length and sinuosity.

[51] Schramm, Amy & Rakotonirainy A. (2010). The effect of traffic lane widths on the safety of cyclists in urban areas.

[52] Schramm & Rakotonirainy A. (2007). An analysis of cyclists crashes to identify ITS-based interventions, *15th World Congress on ITS, New York, USA.*

[53] SERBU, Calin & Opruta, Dan & Socaciu, Lavinia. (2015). Ranking the types of intersections for assessing the safety of pedestrians using TOPSIS method. *Leonardo Electronic Journal of Practices and Technologies (LEJPT).*

[54] Sonka M., Hlavac V., and Boyle R., (1993). Image Processing, Analysis, and Machine Vision. Chapman and Hall.

[55] Swift, P., D. Painter, and M. goldstein. (2008). Residential street typology and injury accident frequency. Available at: http://www.newurbanengineering.com

[56] Tay R., & Rifaat, S. M. (2007). Factors contributing to the severity of intersection crashes. *Journal of Advanced Transportation.*

[57] Thomson, R. C. and Richardson, D. E. (1999). The 'good continuation' principle of perceptual organization applied to the generalization of road networks. *In Proceedings of the ICA, Ottawa, Canada, Session 47B.*

[58] Tim Bock (2018). What is Hierarchical Clustering? Available at: https://www.displayr.com/what-is-hierarchical-clustering/

[59] Toronto Police Service – Public Safety Data Portal. https://data.torontopolice.on.ca/pages/ksi Accessed at: 08/2021.

[60] Wei Liu (2013). Enhancing road safety management with GIS mapping and geospatial database.

[61] Wijnands, J.S., Zhao, H., Nice, K.A., Thompson, J., Scully, K., Guo, J., & Stevenson, M.R. (2021). Identifying safe intersection design through unsupervised feature extraction from satellite imagery. *Computer Aided Civ. Infrastructure Eng.,* 36*, pp. 346-361.*

[62] World Health Organization (2013). Pedestrian Safety. A Road safety manual for decision-makers and practitioners. Available at: www.who.int/roadsafety/en/

[63] Xia Z. , Y. Zang, C. Wang and J. Li (2017). Road width measurement from remote sensing images, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, 2017, pp. 902-905.

[64] Zegeer, C. V., and F. M. Council (1995). Safety relationships associated with cross-sectional roadway elements. Transportation Research Record.