SPATIAL SIMILARITY MEASURES WITH APPLICATIONS TO MAP INTEGRATION AND IMPROVING ACCURACY OF MAP DATA SETS

by

MOUSA ALMOTAIRI

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2019

This dissertation is dedicated to my mother, father, wife, and my children (Bayader, Seham, Maher, Jaser, Hashem)...

ACKNOWLEDGEMENTS

I would like to thank my academic advisor Dr. Ramez Elmasri and my PhD committee members Dr. Bahram Khalili, Dr. Leonidas Fegaras, and Dr. Chengkai Li for their interest in my research and for taking time to serve in my dissertation committee.

I am grateful to all the teachers who taught me during the years I spent in school, first in Saudi Arabia and finally in the Unites States. I would like to thank Eng. Anjum Mulla for encouraging and inspiring me to pursue graduate studies in United states.

I wish also to thank my friend Tariq Alsahfi for taking the time to critically evaluate my work and helping me to conduct several experiments.

Finally, I would like to express my deep gratitude to my wife Ashwaq Almotairi who has encouraged and inspired me with a lot of sacrifices. I am extremely fortunate to be so blessed. I am also extremely grateful to my father, mother, brothers, sisters and my children for their encouragement and patience. I also thank several of my friends who have helped me throughout my career especially Dr. Bakur AlQaudi, Eng. Samir Issa, Ahmed Alfaori, Majed Dakhil and Majed Nafea.

December 1, 2019

ABSTRACT

SPATIAL SIMILARITY MEASURES WITH APPLICATIONS TO MAP INTEGRATION AND IMPROVING ACCURACY OF MAP DATA SETS

Mousa Almotairi

The University of Texas at Arlington, 2019

Supervising Professor: Ramez Elmasri

These days we live in a digital era where most societies rely on applications that depend on digital data. One popular type of digital data that is the basis many of applications is spatial data. Road network maps are one of the spatial data sets that are available for many important applications. However, acquisition of Road Network maps is an expensive task in terms of cost and time, not to mention the maintenance and the updating costs on these spatial data sets. In addition, each Road network map is captured for specific applications such as: road navigation, topographic cartography for printing maps, and so on. Thus, each application focuses on some aspects of the real-world while other aspects are ignored or not given sufficient attention. In order to tackle this problem, this dissertation provides a solution to utilize existing Road network maps by integrating them with one another. However, there is a high chance of mismatches between various road maps that represent the same area for many reasons. These reasons include but are not limited to the following: one of the datasets is not up to date; datasets have different names for the same road; the

co-ordinates or features of road segments are not identical in the two maps; and so on. As a result, matching the roads in such datasets with each other is a challenging task. This dissertation introduces a framework that demonstrates methods of how two map datasets for the same area can be matched to each other even though there are some data discrepancies. In addition, it gives an overview of each component of the framework and focuses mainly on the similarity measurements. These measurements are local divergence measurements and global divergence measurement. Local divergence measurements compare two roads from different datasets to each other to see if they are similar or not by deciding if these two roads have a similar shape as well as the same length. On the other hand, global divergence measurement is used in order to ensure that these two roads are similar in the real world with respect to the location, not different roads that happen to be beside each other having similar length and shape. This dissertation discusses several types of applications that could utilize this framework not only for matching different road maps and unifying the information for smart cities usages but also for data enrichment purposes, historical datasets comparison, and ensuring that maps are up-to-date. As an example, we compare a historical map of Tarrant County, Texas roads with a map of the same area after 11 years. We determine which subareas have grown by calculating the percentages of new roads in each subarea, as well as determining which roads have been taken off the map, for example for stadium construction or flood abatement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In this chapter, we first introduce the area of this dissertation research, known as spatial databases and specifically road network maps in section 1.1. Then we talk about the challenges in this field in section 1.2. Our research contributions come next in section 1.3. After that, in Section 1.4, we give an outline of the remaining chapters of the dissertation and how it is organized

1.1 Spatial Database and Road Network Map

With technology revolutions these days, Geographic Information Systems (GIS) field has gotten more attention and a lot of applications nowadays relies on technologies that explore the benefits of GIS data further[1]. A spatial Database is a regular database that has the capabilities to store and query spatial data, which means database that can deal with geometries. Vector data models [2] are used in Spatial databases as they represent the spatial objects as basic shapes- points, lines, or polygons-. Road networks maps is one of the datasets that are stored in spatial databases as they can be represented by one of the geometry basic shapes (lines). There are different types of Road networks map sources. One of these types is the Authoritative Geospatial Data that are owned by the government such as Topologically Integrated Geographic Encoding and Referencing system (TIGER)[3]. Another type is the commercial Geospatial Data that are owned by private entities such as Google maps[4], Bing maps[5], Here maps[6], and so on. Another type of maps that has become more interesting and is the subject for most recent research pa-

pers is crowd-sourced map data, also known as Volunteered Geographic Information (VGI)[7]. Examples of such map datasets are OpenStreetMaps[8] and Wikimapia[9].

Another classification for Road maps datasets is based on the map usage. For instance, car navigation systems are usually using complete and precise information on speed limits or turning restrictions while for creating printed maps some of data sources have been discarded for the sake of readability using cartographic generalization.

1.2 Road Network Maps challenges

Acquisition of road maps is expensive both for the cost as well as the time consumption, not to mention the cost of maintenance and update on these spatial data. In addition as mentioned above, each dataset is captured for specific application and each one focuses on some aspects of the real-world carefully while other aspects are ignored or not given much attention. Therefore, Road maps integration takes place to get the most of such different Road maps with minimum cost and reasonable time. It provides:

- High re-usability: new applications from existing data sets that are not designed for such applications.
- Quality improvement: data set with low quality in some aspects could be improved by integrating it with other data sets with high quality in these particular aspects, finding incorrect captured elements or new updates in the data sets.
- Cost minimization: no need to capture new data from the real world neither maintain or update it.

Thus, the purpose of this dissertation is to present techniques and algorithms for integration of these highly diverse of spatial data sets. It allows to use different sources in a common context and to avoid duplicates and mismatches in an integrated

dataset. Road matching process is the core component of the Road maps Integration. Road matching utilizes the similarity measurements in order to decide if the two roads from different datasets are similar or not. According to [1] the similarity measurements could be one or a combination of the following types: 1) Geometric measures which relate to the location features such as distance[10], shape[11], or area[12]; 2) Topologic measures study the relationships between two near features[13]; 3) Attribute measures evaluate the non-spatial features of a spatial object such as name or ID[14]; 4) Context measures evaluate the geographic context of a feature relative to its neighbourhood[15]; and 5) semantic measures that compare the similarity between concepts[16].

However, matching different datasets, which are represent the same area, together is a difficult task for many reasons. For example, each road maps dataset has some characteristics that differ from others due to the way of using the tools that capture the data and store them and how to represent these data based on its applications [17]. As a result, the road representation via different datasets for the same real road have differences. Even though high technology equipment is used for capturing the data to generate accurate road maps, there are differences that make challenges to compare two roads from different datasets such as the exact coordinates that represent the road. Thus, there is no notion of equality between different road maps coordinates.

The date and time of road maps updates are different among various road maps datasets, which brings another challenge of comparing roads. Some road map datasets are not up-to-date, and there are some new roads or new parts of roads that are not captured by these road maps. Sometimes the roads are expanded from one-way road to two-way road and these changes are missing in some datasets. For example, figure

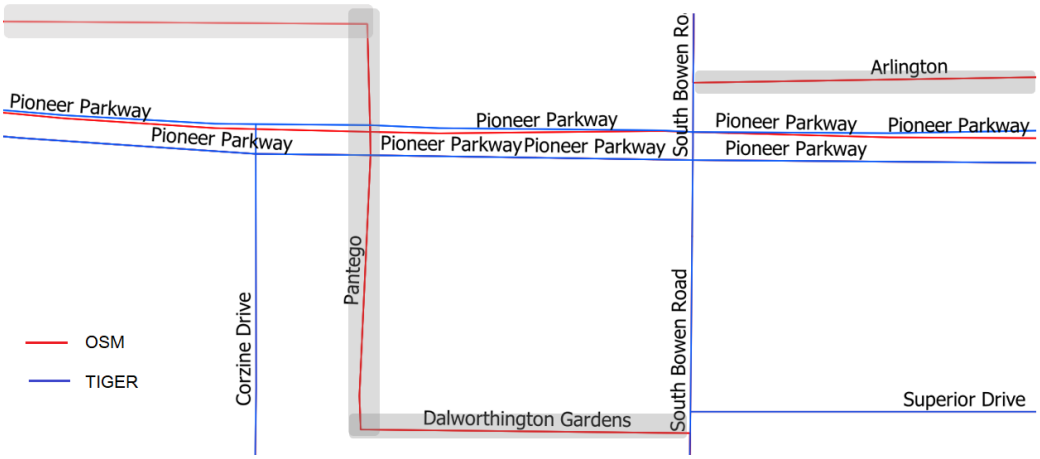


Figure 1.1: OSM has extra information such as cities boundaries while TIGER does not

2.1 in chapter 2 shows a road in OSM that is represented as two-way roads while in TIGER it is represented as a one-way road.

There are cases when the semantic attributes for the same road in the real world are different from one road map dataset to another. These semantic attributes could be the road name, road identifier number, or any other meaningful features that differentiate one road from another. Some road maps use abbreviations instead of whole words such as E for $East$, NW for $Northwest$, or St for $Street$ and so on. These types of differences are easy to solve by string matching algorithms. However, there are cases when there are completely different semantic values for the same road, and such cases complicate the matching process.

As most of our work in the dissertation is focusing in two real datasets, which are TIGER [3] and OSM [8], we are going to discuss the challenges of road matching process by providing some examples that shows the difference of representing a same area of real world. One observation is that OSM has more information other than roads such as the city boundaries. These differences should be taken into consideration so to be avoided to compare them with other physical roads in other road network

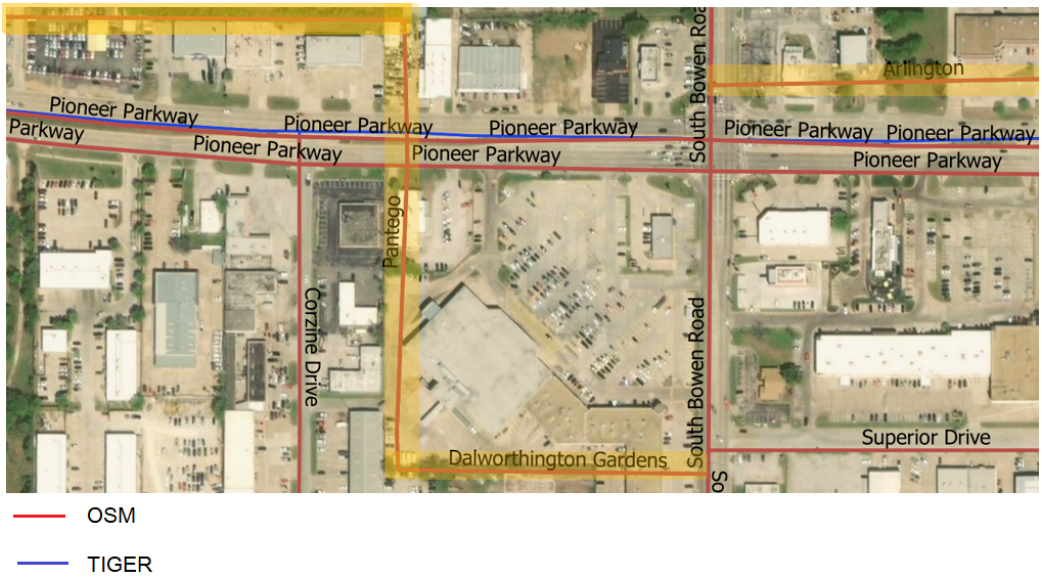


Figure 1.2: Cities' boundaries are not exist in satellite images

maps. For example, OSM has virtual lines represent the city boundaries while the TIGER roads map dataset does not have boundaries. Figure 1.1 shows the boundaries of Arlington, Dalworthington Gardens and Pantego (highlighted in gray color) from OSM (red lines) while TIGER (blue lines) does not have such lines. We can observe in figure 1.2 that there are no boundaries lines in satellite images that means they are virtual lines have same representations of the roads in the datasets.

Another observation is that OSM dataset has information about electric power lines as shown in figure 1.3. Such these information are considered as noisy information from road perspective and they need to be identified and filtered during road similarity matching which is not an easy task. What makes it challenging task is that there is no such attributes that differentiate road lines from other lines (i.e. boundaries lines, power lines and so on).

Semantic attributes could play major role to matching similar roads from different datasets. However, sometimes the names are not identical for different reasons. For instance, naming convention using abbreviation in Tiger such as "S Cooper St"

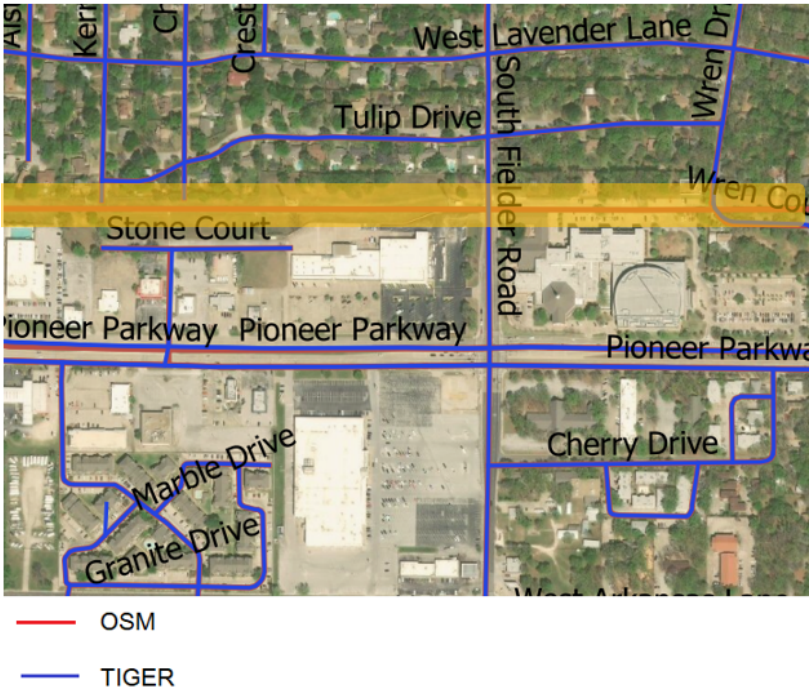


Figure 1.3: Power lines are among the data representations in OSM dataset

while in OSM naming convention is not using the abbreviation and road is named as "South Cooper Street". Such differences are easy to deal with. However, the matter becomes more complicated when there is omission instead of abbreviation like the case in OSM for some roads when the road name contains direction as well as the name, the direction is omitted from the name. For example, a road is named "W Pioneer Pkwy" in Tiger, it has been named "Pioneer Parkway" without mentioning the direction Figure 1.4 shows a plot of part of West Pioneer Parkway and the name of this road in OSM (red fonts) comparing to the name of the road in TIGER (blue font). The challenge becomes more harder when there is completely different name for the same road such as the road in figure 2.1 in TIGER dataset where part of the road called "S Cooper St" and another part is names "FM 157" while in OSM the whole road has been named as "South Cooper Street".

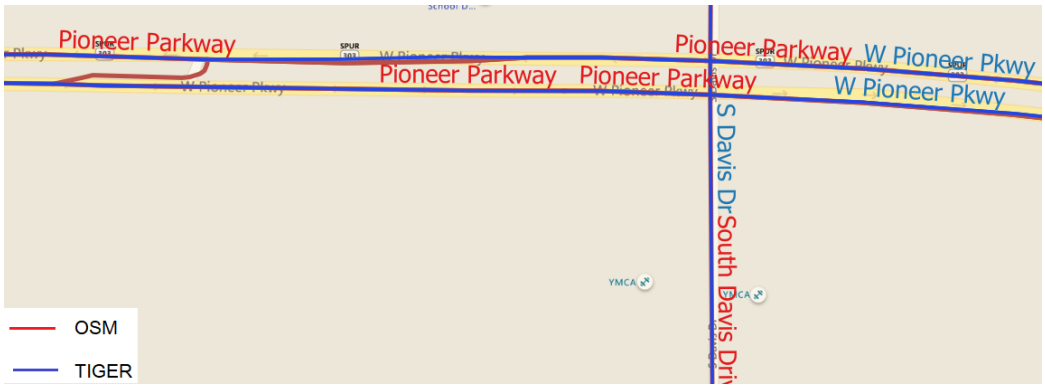


Figure 1.4: W Pioneer Pkwy in TIGER vs. Pioneer Parkway in OSM

One of the main comparison factor between two datasets is the geometric features and most importantly is the location. We should keep into consideration there will be different of capturing the coordinates that represents the same area from different datasets which means we expect slight divergence from different datasets such as "West Lavender Lane" in figure 1.5. Another issue is number of lines in map representing a road. Some roads are represented as a single line (one-way) in one road network map, while another map represents it as two lines (one for each direction of the road). Sometimes, the expansion of the road is not captured from one dataset while another dataset gets the changes and store the information. Usually, this difference occurs in some of the main roads. As an example of this case is "South Cooper Street" as shown in the figures 1.6. Sometimes, the difference could be when some road get extension to be longer to serve additional areas and this extension has been captured by some datasets while the others do not capture this change. For example Center St has an extension in its south side and this extension has been captured by TIGER dataset while OSM is not yet getting this change as shown in figure 1.7.

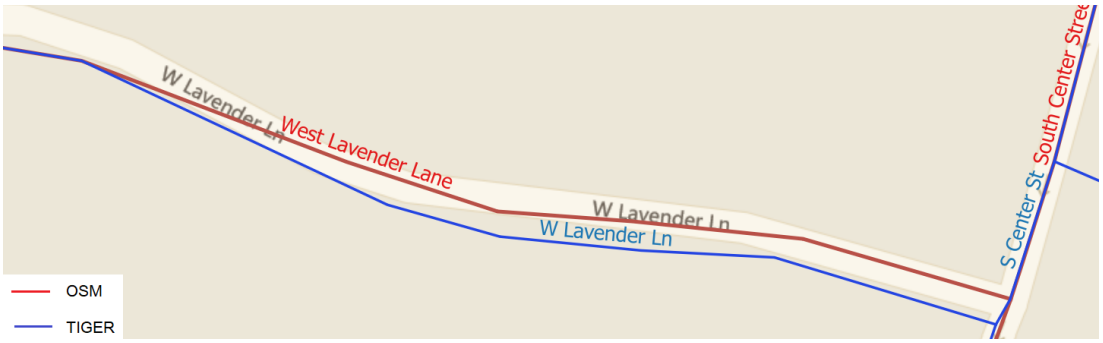


Figure 1.5: Slight differences in representation between TIGER and OSM

1.3 Dissertation Contributions

The contributions of this dissertation are to introduce efficient and scalable Road Map integration system that has the capabilities to match the roads from different Road maps. Additionally, we propose a framework that is confirming the similarity of two roads from different datasets and then highlighting the differences between them such as missing road segments, road name mismatching, and two-way road in one dataset compared with a one-way road in another dataset. Moreover, this dissertation contributes in similarity measurements by introducing a novel methods inspired by Hausdorff distance to confirm if two roads from different maps are really similar to each other or not. We use local divergence measurements that make sure these candidate roads have approximately the same length and also run in parallel to each other, which preserves the shape between them. Confirming the similarity requires also a global divergence measurement to be met that ensures the candidate roads are for the same road, not different roads that happen to be beside each other having similar length and shape.

1.4 Dissertation Organization

In Chapter 2, we introduce our framework for road map networks matching. We name the framework "*TAREEQ*" after the name of "Road" in Arabic Language. The TAREEQ framework has four main components that are: 1) Data Sources, 2) Generate Candidate Roads, 3) Similarity Measurements, which is our topic in chapter 3, and 4) The Results. These components are described in section 2.3. After that, we introduce list of different types of applications that can utilize this framework for different purposes in section 2.5. We sum up our work and contributions in the last section (the conclusion) of each chapter.

In Chapter 3, We introduce a novel approach that is inspired from Hausdorff distance [18] to measure candidate similar roads, which are represented as lines in vector data type [2]. This approach is to confirm if the candidate similar roads from different maps are really similar to each other or not. We introduce local divergence measurements that make sure these candidate roads have approximately the same length and also run in parallel to each other, which preserves the shape between them. Confirming the similarity requires also introducing a global divergence measurement to be met that ensures the candidate roads are for the same road, not different roads that happen to be beside each other having similar length and shape. Moreover, this approach has the capability to identify similar roads when one of the roads has either missing road segments or extra incorrect road segments as shown in the experimental results. Finally, we conclude our work and contributions for this chapter.

One of the application of the proposed framework that have been introduced in chapter 2 is comparing the new version of Road maps with an older version to study the roads changes in the city. We start chapter 4 with introduction in section 4.1 and list different type of usages by comparing historical datasets among each other. After that, we define the problem. Then, we explain the matching process in

section 4.4 by focusing mainly on the second process of TAREEQ framework which is *Generate Candidate*. Next we conduct the experiment in real datasets for two areas 4.5. After that, we compare our outcomes with the state-of-are framework in this field and discuss the outcomes. Finally, we will give a conclusion for this chapter 4.6.

To sum up, Chapter 5 summarizes the contributions made in this dissertation and discuss the future work.

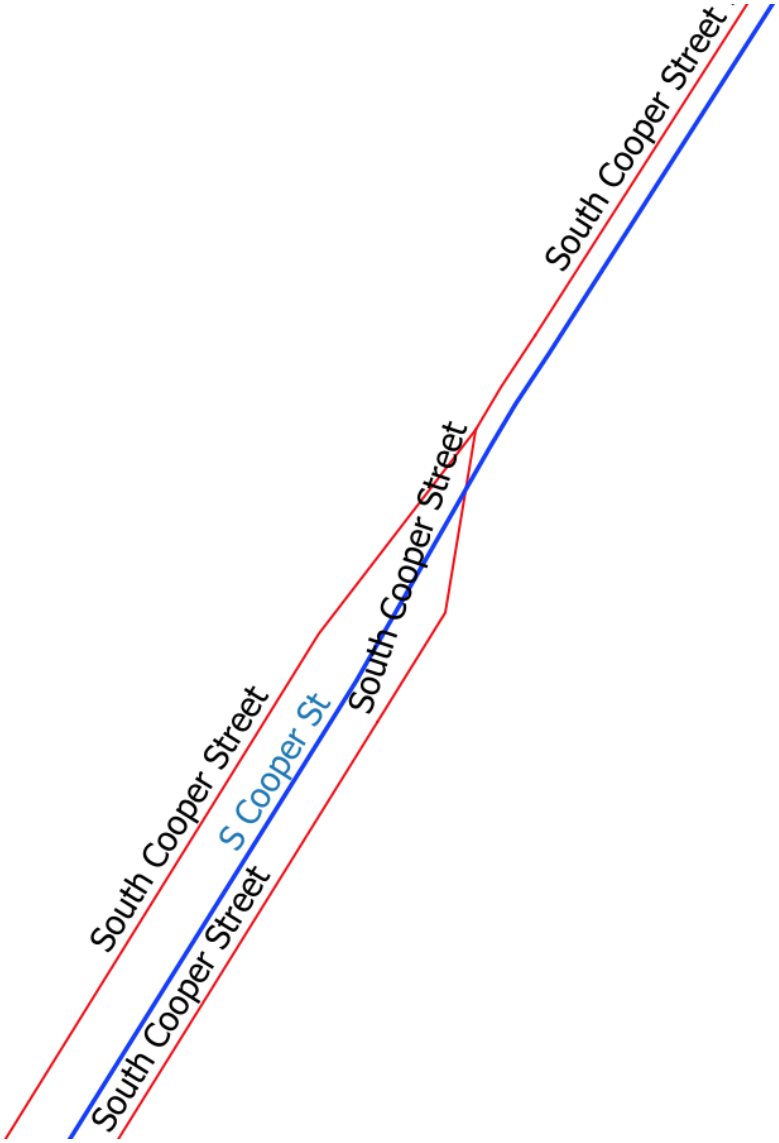


Figure 1.6: One-way representation in TIGER vs. two-way representation in OSM for part of South Cooper Street

Figure 1.7: TIGER captures the extension of Center St while OSM does not

CHAPTER 2

A FRAMEWORK FOR COMPARING AND MATCHING ROADS FROM DIFFERENT SPATIAL DATASETS

In this chapter, after having an introduction in Section 2.1 and considering the challenges of comparing and matching roads in 2.2, we introduce our TAREEQ framework in Section 2.3, which consists of four components: Data sources Preprocessing 2.3.1, Generate Candidates 2.3.2, Similarity Measurements 2.3.3, and Candidate Matching Decision 2.3.4. Then, we discuss the experiments and results on TAREEQ framework 2.4. After that, we list different types of applications that can utilize this framework in section 2.5. We mention the related works in section 2.6. Finally in section 2.7, we conclude our work for this chapter.

2.1 Introduction

Any smart city in the world should have Road Network Maps. These maps have different owners, some of them came from Volunteered Geographic Information (VGI) such as OpenStreetMaps[8], others are produced by governmental authorities like TIGER [3]. Also, there are private companies who create their own maps, for example, Google maps [4] and Bing maps [5]. There are different approaches to create such road maps, for instance, GPS coordinates, moving-objects trajectories, satellite images, and so on. Therefore, there is a high chance that there are differences between these road maps in representing the same road in reality due to different ways of capturing roads coordinates. In addition, the nature of city growth makes these cities change over time (i.e., some new roads are built, others are extended, and some roads

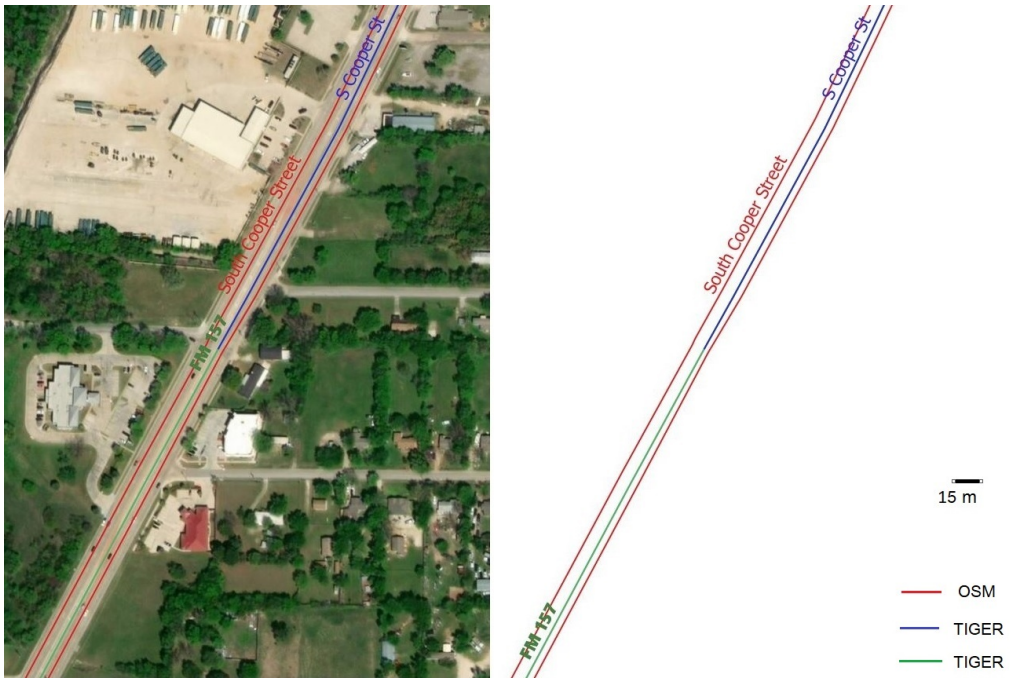


Figure 2.1: Compare roads in OSM (red) vs. TIGER (blue and green)

are permanently closed.) These changes should be captured and stored in road maps. The approaches of capturing the changes and the time of updates are different among various road maps that eventually create differences among these road maps. In order to highlight the differences between road maps, a matching roads process between two road maps should be conducted.

This chapter introduces a general framework of matching the roads. One of the essential components in this framework is the similarity measurements because based on these similarity measurements the framework can decide when comparing two roads from different datasets whether they are similar or not. We are utilizing the Local Roads Divergence Measurements (LRDM) and Global Roads Divergence Measurement (GRDM) [19]. The reason for choosing them is the ability of these measurements to identify the similar roads even if there are missing, or additional, road segments in one of the road map datasets. LRDMs check that two roads from

different datasets have roughly a similar length and these roads keep running in parallel to one another if they represent the same road in the real world, which preserve a similar shape between them. This technique can be performed by computing the gap between the two roads from different datasets after overlaying one of the datasets over another. Affirming the similarity between these two roads requires also passing the GRDM condition after the LRDMs conditions have been met. GRDM ensures that the distances' between the roads' coordinates are within the limits of GRDM threshold. GRDM ensures the two roads are for a similar road in reality and they are not different roads that happened to be adjacent to one another.

The contribution of this work is confirming the similarity of two roads from different datasets and then highlighting the differences between them such as missing road segments, road name mismatching, and two-way road in one dataset compared with a one-way road in another dataset. In addition, this chapter presents applications that get benefits of this framework such as road map corrections, comparing road map with moving objects trajectories (cars, buses, bicycles , and so on), capturing the changes (new roads, permanently closed roads) between two versions of same datasets, and road maps integrations for data enrichment.

2.2 Challenges of Comparing Roads from Different Datasets

Each road maps dataset has some characteristics that differ from others due to the way of using the tools that capture the data and store them and how to represent these data based on its applications [17]. As a result, the road representation via different datasets for the same real road have differences. Even though high technology equipment is used for capturing the data to generate accurate road maps, there are differences that make challenges to compare two roads from different datasets

such as the exact coordinates that represent the road. Thus, there is no notion of equality between different road maps coordinates.

The date and time of road maps updates are different among various road maps datasets, which brings another challenge of comparing roads. Some road maps datasets are not up-to-date, and there are some new roads or new parts of roads that are not captured by these road maps. Sometimes the roads are expanded from one-way road to two-way road and these changes are missing in some datasets. Figure 2.1 shows the road in OSM that is represented as two-way roads while in TIGER it is represented as a one-way road.

There are cases when the semantic attributes for the same road in the real world are different from one road map dataset to another. These semantic attributes could be the road name, road identifier number, or any other meaningful features that differentiate one road from another. Some road maps use abbreviations instead of the whole words such as E for $East$, NW for $Northwest$, or St for $Street$ and so on. These types of differences are easy to solve by string matching algorithms. However, there are cases when there are completely different semantic values for the same road, and such cases complicate the matching process. For example, Figure 2.1 shows $SouthCooperStreet$ in OSM has two different names in TIGER ($SCooperSt$ and $FM157$).

To cope with these challenges, our framework provides a novel techniques that taking into consideration all coordinates that represent the road and compare it with corresponding coordinates in another dataset. By utilizing LRDMs and GRDM the framework can match many-to-many (N:M) road segments, which leads to match roads with missing segments with their pairs that have all road segments from a different dataset. In the following sections, we present our method, which introduces the framework and its components in section 2.3. Next, in section 2.5, we highlight

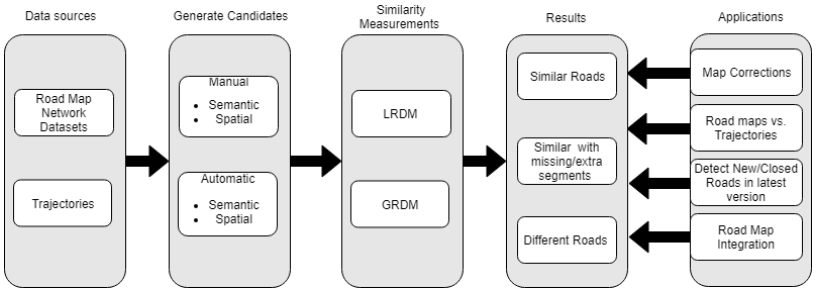


Figure 2.2: TAREEQ Framework to find similar roads between two different road maps

some applications that can benefit from our framework. In section 2.6, we mention the related works have been done for matching similar roads. Finally, we conclude the chapter and highlight the future work based on this chapter.

2.3 The TAREEQ Framework

The proposed framework, which we name it TAREEQ for matching similar roads between two different road maps, works on road maps datasets that are stored as a vector data type. Vector data is a 2-dimensional representation of the world using points, lines, and polygons. These types are composed of coordinates (latitude and longitude) as well as the features that give meanings to the coordinates such as road name, road type, and so on. In a case of the road maps, the roads are usually represented by lines, which is an ordered set of coordinates that define road location. Figure 2.2 shows our proposed framework and its components. This framework has inputs, which are the data sources. Then the process to generate the candidate similar roads. After that, there are similarity measurements that provide the decision about these candidates to determine if they are similar roads or not in reality. The result of matching is one of the three possibilities: the two candidates are similar, they are similar but one of the roads has missing road segments or has additional incorrect

road segments, or they are not similar roads. The following subsections are describing each component from figure 2.2 separately.

2.3.1 Data Sources Preprocessing

This component deals with the inputs to our proposed framework. There are two types of inputs: either Road maps datasets or moving objects trajectories. The framework can take two road maps to match their roads to each other, or it can take one road maps dataset with moving objects trajectories to match the roads with the trajectories. This component has the preprocessing phases that prepare the inputs for the next process. For example, preparing two datasets to have the same scale on the plane, unify the format for the two inputs, using the same spatial reference identifier associated with specific coordinate systems. In this chapter, we used two road maps datasets when applying this framework: 1. OpenStreetMap (OSM): a free road-map open to the public to contribute and build up the data and it belongs to Volunteered geographic information (VGI) category[8]; and 2. Topologically Integrated Geographic Encoding and Referencing system (TIGER): Authoritative Geospatial Dataset, which is produced by the US Census Bureau and it is considered as a professional datasets[3]. Both OSM and TIGER have stored their data as a vector data format.

2.3.2 Generate Candidates Similar Roads

The second component in the framework is the process of generating the candidate similar roads from different datasets. It means these candidates roads from different datasets may be similar and represent the same road in the real world. It needs further verification steps in order to make sure if these two candidates are similar or not. Different techniques can be used in this component in order to generate the

candidates. We have divided them into two main categories: 1. define the candidate similar roads manually, and 2. generate the candidate similar roads automatically. Both categories can use semantic or spatial roads features (attributes) in order to generate the candidates. Semantic features relate to any feature that gives meaning to the road such as Road name and Road type while spatial features identify the geographic location of features and boundaries, such as length and shape. There are techniques that can be used to capture the candidates based on the type of features, for instance, semantic features can generate the candidates that have similar Road Names, and for spatial features, we can get the candidates by using the buffer technique [20, 17].

In this chapter, we generate the candidate similar roads manually using semantic features and particularly road names. Even though road name between two road maps datasets intuitively means that the candidate roads are similar, there are difficulties like the names sometimes are slightly different from one dataset to another such as $SCooperSt$ in $TIGER$ dataset and compare it with $SouthCooperStreet$ in OSM. In addition, sometimes such matching candidate road maps highlight the changes and differences between the two candidates like road extensions that are captured in one dataset and not updated in another one.

2.3.3 Similarity Measurements

The core component of the proposed framework is to determine if the candidates are similar, partially similar, or not similar to each other. This component provides the decisions of the matching similarities between the candidate similar roads. There are several approaches of the similarity measurements, and it will be discussed in related works in section 2.6. This work is utilizing LRDMs and GRDM to define the similarities between two candidate roads. This technique takes the benefits of

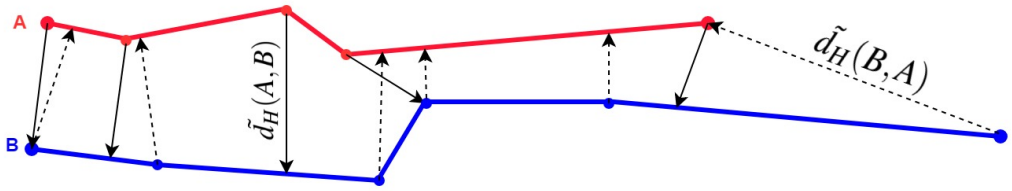


Figure 2.3: Hausdorff distance measures the divergence distances from A to B and from B to A

Hausdorff distance, which is one of the popular approaches to measure lines similarity as a base and tries to find out the abnormalities between a road's representations from different datasets. Figure 3.2 shows an example of Hausdorff distance measuring the divergence distances between lines A and B. It walks through all the points of line segments from the first line's coordinate until the last coordinate. For each point computes the minimum distance between this particular point to the closest point from the corresponding line and vice versa from the points that represent the second line to the nearest point from the first line as the Hausdorff distance is not commutative. However, Hausdorff distance take into consideration only the maximum value of minimum distances. Therefore, if there is one outlier coordinate points of the candidate similar roads, it will affect the decision to confirm the similarity of the candidate similar roads. Moreover, it ignores the other divergence distances values between the candidates, which in our cases have meaning if there is partial similarities between the candidates, i.e., the candidate similar roads are similar in reality but one of them has missing road segments or additional incorrect road segments. As a result, we are using additional measurements, which are the LRDMs and GRDM [19]. They are described in the following subsections.

2.3.3.1 Local Road Divergence Measurements LRDMs

The purpose of these measurements is to find out the similarities between the candidate similar roads in two aspects: shape and length. There are three measurements in this category: 1. Mode value between roads A and B $Mode(A, B)$, 2. Mode Frequency value $Freq(A, B)$, and 3. Local average divergence (Avg_L). The first measurement is the *Mode*, which is the distance value that has the highest frequency of occurrences in the two lists of minimum distances between each road's coordinates and corresponding road $P_{diff}(A, B)$ and $P_{diff}(B, A)$. When creating the list of minimum distances; then we count the most frequent number occurring in the list to indicate the divergence. As in the P_{diff} vectors that are defined as the following:

$$P_{diff}(A, B) = [\min(a_i, B)] \forall a_i \in A$$

$$P_{diff}(B, A) = [\min(b_i, A)] \forall b_i \in B$$

where $\min(a_i, B)$ represents the distance from point a_i in the road (line) A to the nearest part of the line B and this distance could be computed by ordinary straight-line (Euclidean) distance. Also, the distance values are rounded to the closest Integer number (in meter unit). This $Mode$ value is used to determine overall how far apart these two lines are from each other as well as to determine the length similarity between two lines. $Mode$ value can be compared with Hausdorff value d_H, which is the maximum value of the two vectors $P_{diff}(A, B)$ and $P_{diff}(B, A)$, and based on that two possible cases could happen; either $(Mode(A, B) + \tilde{\epsilon}) \geqslant d_H$ or $(Mode(A, B) + \tilde{\epsilon}) < d_H$ where $\tilde{\epsilon}$ is an estimated value of the margin of error of lines locations - the value of the margin error is computed manually and based on the nature of the datasets. In the first case when the single bidirectional Hausdorff is less than or

equal $(Mode(A,B)+\tilde{\epsilon})$, this indicates there is a high possibility that these two lines are identical and have the same shape and length. While if the $(Mode(A,B)+\tilde{\epsilon})$ is less than bidirectional Hausdorff distance, it indicates there is a divergence in one line segment or more from one line or both lines. Furthermore, the two directional Hausdorff distances may provide more information when compared to Mode value. For example, if $\tilde{d_H}(A,B)$ is less than or equal $(Mode(A,B)+\tilde{\epsilon})$ and $\tilde{d_H}(B,A)$ is greater than $(Mode(A,B)+\tilde{\epsilon})$, this indicates that line B has one or more line segments that do not exist in A. This difference can have several explanations from traffic road perspective such as new line segments are captured by B's dataset while missing in A's dataset, or it exists in A's dataset but with different road name, and so on.

The second measurement is the *Mode Frequency* $Freq(A,B)$ and by using this measurement, it can determine the shape similarity between the candidates. *Mode Frequency* counts the frequency of the Mode value occurrences. As there are slight differences (margin errors) between the representations of the road in different datasets, we take a range $[(Mode(A,B)-\tilde{\epsilon}), (Mode(A,B)+\tilde{\epsilon})]$ and count all values in the range. *Mode Frequency* can determine the shape similarity between two lines by comparing its value that represents occurrences with the total number of points in both lists. This comparison can be computed by getting the percentage between *Mode Frequency* and the total number of points in both lists, $Freq(A,B)/(N+M)$ where N and M are the numbers of coordinates for roads (lines) A and B, respectively. Because there is a slight divergence in the road representation from different datasets, it is better to add the three highest occurrences of distance values in the list together. Then we see if the value of summation represents 80% or more of the number of points, which indicates these two candidate lines have an overall similar shape. It is important to mention that these cumulation frequencies should not be apart more than $\pm\tilde{\epsilon}$ from each other.

The third measurement is the Local average divergence (Avg_L) that can be computed from Local road divergence arguments and used to compare with GRDM. This measurement gets the average divergence distance between two candidate similar lines, and it could be computed as the following:

$$Avg_L = \frac{(\sum_{i=1}^{N}[\min(a_i, B)] + \sum_{j=1}^{M}[\min(b_j, A)])}{(N + M)} \tag{2.1}$$

$$\forall a_i \in A \text{ and } b_j \in B$$

2.3.3.2 Global Road Divergence Measurement GRDM

LRDMs determine the similarity between two roads (lines) from the shape and the length perspective. They do not take into consideration the overall datasets divergence. GRDM provides an average of all Local Average divergence between the two datasets that are similar in reality. Therefore, Global divergence is important as sometimes LRDMs can compare two parallel roads and find out they are similar while in the real world they are two different parallel roads happen to run beside each other, even though they have a similar value of $Mode$ and the Frequency $Freq(A, B)$ is bigger than 80%. Therefore, it is important to use GRDM to make sure these roads (lines) are representing the same road in the real world. We can compute Global road divergence value as follows:

$$Avg_G = \frac{(\sum_{i=1}^{w} Avg_{Li})}{w}$$

where Avg_{Li} is the Local Average distance for all similar road pairs and w is the number of those similar road pairs.

Global road divergence can be used to compare the Local road average distance of the candidate similar roads Avg_L with the Global road distance value Avg_G. If the Avg_L is larger than the Avg_G by more than 10%, it indicates that most likely these two candidate roads are different from each other and happens to run parallel in the real world. Otherwise, it is most likely these two roads are similar to each other. The 10% value added to Avg_G is the threshold, and it is manually identified based on the nature of datasets. Based on trial and error, our experiments show that 10% difference splits the similar roads than different roads and this is the reason why we choose 10%. In the case of missing road segments from one road while it is available in the other corresponding road, this measurement can be used to identify the similarity by comparing the $Mode(A, B)$ instead of Avg_L with Global road divergence value Avg_G.

2.3.4 Candidate Matching

Figure 2.4 shows how to use the LRDMs and GRDM in order to determine if the two candidate roads are similar or not. If the three following conditions are met by the candidate similar roads: 1. $Freq(A, B)/(N + M) \geqslant 80\%$, 2. $Avg_L < Avg_G * 1.10$, and 3. $(Mode(A, B) + \tilde{\epsilon}) \geqslant d_H$, that means these two candidates are similar. The first condition $Freq(A, B)/(N + M) \geqslant 80\%$ implies the two roads are running in parallel with an almost fixed distance between them and it means they have a similar shape. The second condition $Avg_L < Avg_G * 1.10$ means they the same road because the distance between them is within the average of other similar roads- to make sure they are not two different roads in the real world that happen to run in parallel. The third condition is to make sure that the candidates have approximately similar length. If this condition is not met while the first two conditions have been met, this indicates these two roads are partially similar, which means one of them has missing segments

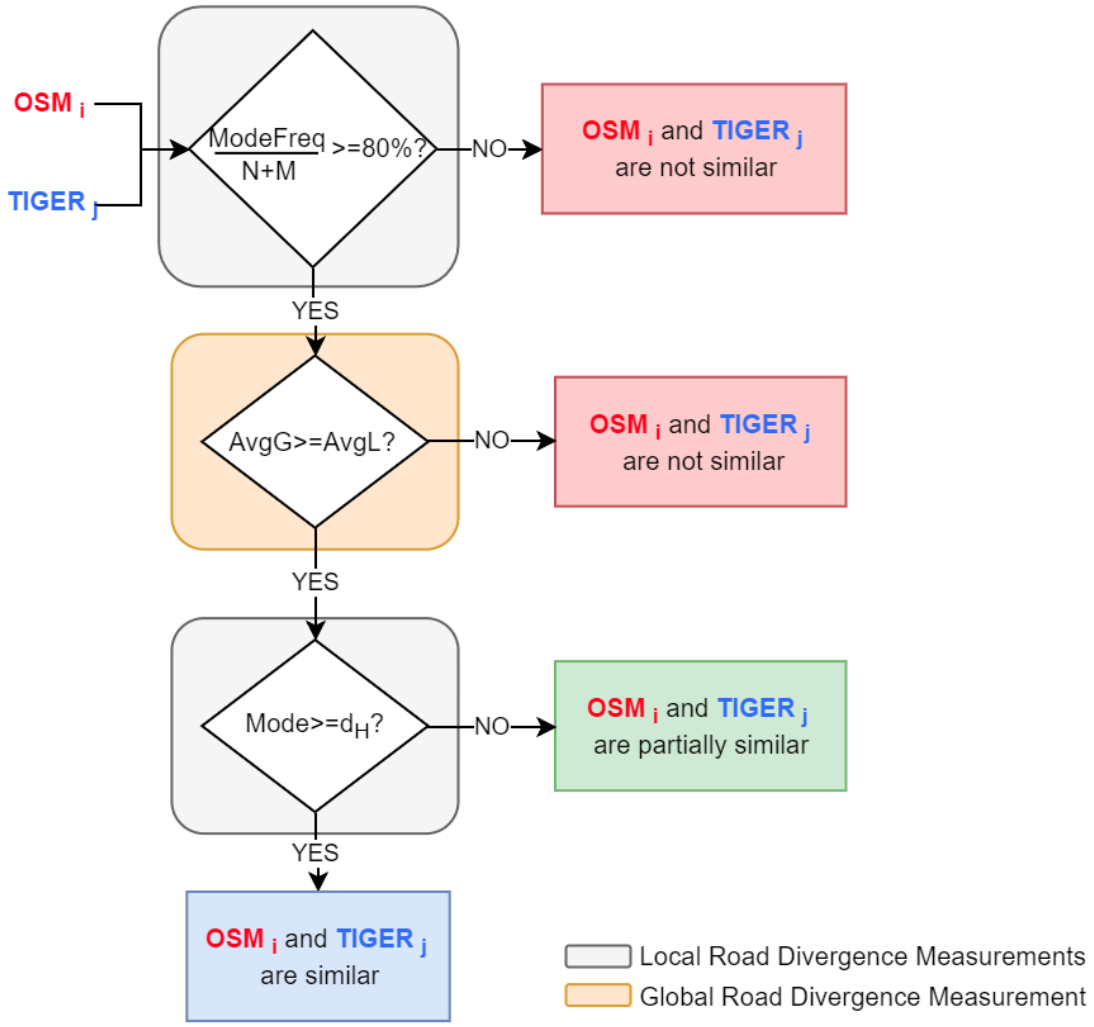


Figure 2.4: Flowchart of how to determine the similarity

of the road for any reason such as different name, new segments of the road have not been captured or simply incorrect additional road segments in another road.

2.4 Experiments and Results

Our TAREEQ framework can work on any application that has more than one road map representing its roads. In addition, TAREEQ framework can work on the real world data, not the synthetic data as most of the time these synthetic data are

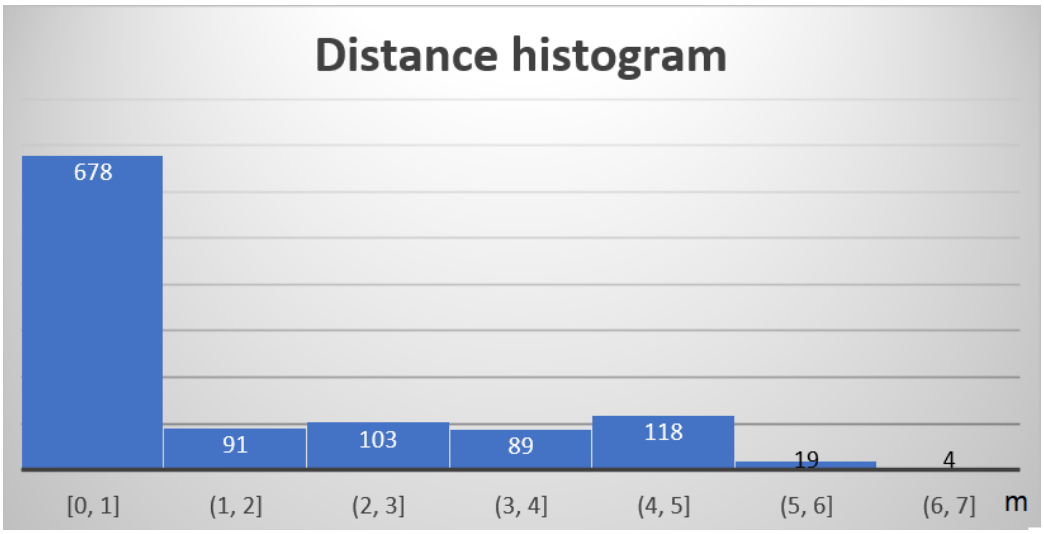


Figure 2.5: Distance Distributions for Pioneer Pkwy

tailored to fit a particular system. As a matter of fact, our datasets are real data, and we have selected OSM [8] and TIGER [3] datasets that represent the roads in Tarrant County, Texas. There are a certain parameters that need to be set up once based on the types of the datasets, such as the error margin parameter $\tilde{\epsilon}$, which we assume it is equal to 7 meters based on the average width of the roads and how far are the similar roads from different datasets to each other. For GRDM value, it is calculated based on training sets from the real data for the correct matching roads; after taking the average of several correct matching roads, the Global road average distance is equal to $Avg_G = 19meters$. Furthermore, this method provides accurate results with relatively long candidate roads as compared to small road segments that could confuse their geometry length with divergence measurements. As we conducted several experiments to test this framework, we list some scenarios below.

2.4.1 Correct Matching between the Candidate Roads

The example is for "Pioneer Parkway" road in Tarrant County. This road stretches more than 17 kilometers. It is represented in TIGER by 434 coordinate points while in OSM this road has 668 points. After computing the distance value for each coordinate to the nearest point of the corresponding candidate road, the following values are the results: $Mode(T, O) = 1$, $d_H = 7$ where T is TIGER and O is OSM, highest $Freq(T, O)_1 = 1102$, Local Average distance $Avg_{L\_Pioneer} = 2$, and global average distance $Avg_G = 19$. There is no need to count the second and third highest frequency as all number of coordinates are falling in the highest frequency category. Figure 3.4 shows the frequency of coordinate points in each meter.

As shown in table 3.1, we check to see whether the measurements conditions have been met or not: 1) Are 80% of roads' coordinates or more running in parallel to each other or not?: $((Freq(T, O)_1 = 1102)/(434 + 668 = 1102)) = 1.00$, which means 100% of candidate roads' coordinates are running exactly in parallel. 2) Are the candidate roads representing the same road in the real world or different parallel road?: $Avg_{L\_Pioneer} = 2 < Avg_G = 19 * 1.10 = 21$. 3) compare $Mode(T, O)$ value with bi-directional Hausdorff distance: $(Mode(T, O) + \tilde{\epsilon}) = 1 + 7 = 8 \geqslant d_H = 7$, therefore, this condition is passed; As a result, all measurement conditions have been met. Therefore, we can say these two candidate roads are matching to each other, and they represent the same road in reality.

2.4.2 Partial Similarity between the Candidate Roads

Figure 2.1 shows a case when the candidate roads are partially similar, which means there are road segments are missing from one dataset while they are available in another. This case is happening for different reasons: capturing new updates while others are not updated, adding additional road segment that is incorrect and does

Table 2.1: Comparisons between TIGER and OSM for Pioneer Pkwy

Measurement	Value
$Mode(T \Leftrightarrow O)$	1
$Freq(T \Leftrightarrow O)_1$	1102
$Freq(T \Leftrightarrow O)_2$	0
$Freq(T \Leftrightarrow O)_3$	0
Avg_L	2
d_H	7
Avg_G	19

Table 2.2: Comparisons between TIGER and OSM for South Cooper Street

Measurement	Value
$Mode(T \Leftrightarrow O)$	7
$Freq(T \Leftrightarrow O)_1$	546
$Freq(T \Leftrightarrow O)_2$	231
$Freq(T \Leftrightarrow O)_3$	31
Avg_L	9
d_H	289
Avg_G	19

not exist in the real world, missing road segments due to having different road name or naming road segments wrongly. For this scenario, we have "South Cooper Street" that spans for more than 14 km, and it is represented in TIGER by 207 coordinates while in OSM, it is represented by 648 coordinates. Table 3.2 has the computation values for these two candidate roads.

Figure 3.6 shows the distributions of the coordinates based on the distances. For first measurement condition, $(546+231+31)/(207+648)) = 0.95$, which indicates that 95% of coordinates are running in parallel to corresponding candidate road; thus this measurement condition has passed. Regarding the second condition, which is the global road divergence, it compares local road average distance to the global road average distance and sees if the local average distance is within the global average

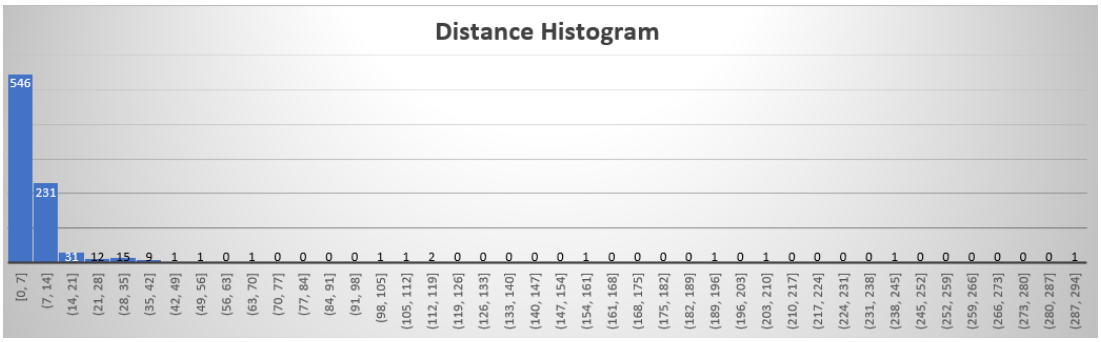


Figure 2.6: Distance Distributions for South Cooper Street

distance range or not: $Avg_{L\_SCooper} = 9 < Avg_G = 19 * 1.10 = 21$ it is shown this condition has been met. For the third condition: $(Mode(T, O) + \tilde{\epsilon}) = 7 + 7 = 14$, which is very small comparing to the bi-directional Hausdorff distance $d_H = 289$ so, this condition is not met. From the above results, the third condition is not met while the first and second conditions of local divergence and global divergence are met. It indicates the two candidate roads are similar, but there is a missing road segment from one dataset while it exists in the other or additional incorrect road segment in one dataset. In this case and after investigation, figure 2.1 shows that TIGER dataset has missing road segment of "South Cooper Street" due to renaming some road segments with another name "FM 157".

2.4.3 Candidate Roads are not Similar

In this scenario, it shows where the global divergence plays its role to decide if the roads are for the same road or for different roads that happen to be in parallel. Such scenarios frequently happen when matching the candidate roads based on spatial features only, i.e., semantic features like road name are not taken into consideration to generate candidate similar roads in the second component of the proposed framework. The experiment conducted in this category is getting the road called "Janann Ave"

Table 2.3: Divergence measurements for Janann Ave (TIGER) vs. Marlee Lane(OSM)

Measurement	Value
$Mode(T \Leftrightarrow O)$	83
$Freq(T \Leftrightarrow O)_1$	10
$Freq(T \Leftrightarrow O)_2$	0
$Freq(T \Leftrightarrow O)_3$	0
Avg_L	82
d_H	84

Figure 2.7: Roads have same length and shape but they are not similar

from TIGER and match it with OSM road called "Marlee Lane". These two roads span for more than $500m$. It generates ten coordinates points in total. The values of local divergence measures are computed and presented in the table 3.4.

These results can pass the LRDMs conditions. It shows that the candidate roads maintain constant distance apart from each other $Freq(T, O)_1 = 10/10 = 1.0$ thus all coordinates are running in parallel to the corresponding road from the other dataset, which meets the first condition. The second condition Global road divergence takes place here: $Avg_L = 82 > Avg_G * 1.1 = 21$, which indicates that these candidate roads are representing different roads and they are not similar in the real world even though they have similar shape as shown in Figure 3.9 that is depicted from a satellite image.

2.5 Some Applications of the Proposed Framework

Our proposed framework can be used in different applications. These applications can solve different types of problems such as the categories of applications listed in figure 2.2. Some of them could be used manually to inspect one or a couple of roads in order to capture the changes like new expansions to the roads or permanently closed. Other types of applications work on entire datasets, which require automatic matching for data enrichment such that we have one accurate road maps dataset but without Point of Interests (POIs) for example, while the other dataset has the POIs but it is not accurate in term of roads information, and it is not up-to-date. Below are examples of different types of applications that can utilize the proposed framework.

2.5.1 Highlight Road Differences for Road Maps Corrections

By matching two different road maps datasets, our proposed framework has the abilities to find out the differences between the road's representations from different datasets:

- Identifying if the road that has two ways (two-line representation) in parts of the road and other parts are not updated and became a single line in one dataset while the other datasets capture all two-lines parts. This application can utilize the directed Hausdorff distance and compare the distances for the road from one dataset to another dataset and vice versa. If there are consecutive coordinates that have distances within 30 meters, this indicates the corresponding road has a single line instead of two-line representation.
- Using Semantic features: It identifies if there are parts of the road having a different name (or without a name at all) than the remaining parts of the road. This scenario could be identified by the proposed framework. When candidates reach similarity measurements component, it will give partial similar due to

some road segments have different road name, and also the process of generating the candidates using semantic features does not capture any road segments with different name. Figure 2.1 shows that some road segments of *SCooperSt* have a different road name in TIGER.

- Identifying missing road segments when one of the road map datasets is outdated and did not capture new changes. Sometimes, the problem is not with the dataset that has missing road segments when it compares with different dataset; the other dataset has incorrect road segments that do not exist in the real world. Also, this could be captured by the partial similarity between the two road map datasets.

2.5.2 Matching Road Maps with Moving Objects Trajectories

Identifying the mismatching between two road maps network is the first step. Next step is to determine which dataset has the correct information and is up-to-date [21]. Most of the times this process is complicated. One solution is to match the differences with moving objects (such as cars and buses) trajectories and check if there are trajectories that pass through these differences and the road map that has road match with these trajectories is considered as a correct map. As the trajectories are stored as a vector data, our proposed framework can deal with them.

2.5.3 Compare the Road Map with its Old Versions

This is a straight forward road matching to capture the updates between two versions of the same road map dataset. This application provides interesting information by identifying road map changes over time (new roads or road extensions, permanently closed roads, a one-way road turns to be a two-way road, and so on). In addition, it gives overall pictures of where is the growth in the city and by comparing

multiple versions of the road map dataset, we can get a prediction of the pace of changes over time-based on the differences between datasets versions.

2.5.4 Road Maps Integrations

Acquisition of new road map dataset is expensive for the matter of the cost, not to mention the cost of maintenance and update. In addition, each road map dataset is captured for a specific application such as road navigation, topographic cartographic for printing map, and so on. Therefore, each one focuses on some aspects of the real-world carefully while other aspects are ignored or not given much attention. Thus road map integration is coming in order to provide new applications from existing datasets that are not designed for such applications and quality improvement. Our proposed framework can integrate two road maps utilizing LRDMs and GRDM. It has the capabilities to match N:M road segments, which leads to match roads with missing segments with their pairs, that have all road segments, from a different dataset. This could be done by automating the process of generating the candidate similar roads and going through all the roads in the road map dataset to match them with corresponding roads in different road map dataset.

2.6 Related Works

Finding the differences and matching similar roads methods from different datasets have been the interests of several papers [1] for different types of applications such as road maps integration and data enrichment, resolving data discrepancies in semantic attributes like road ID, road name, and so on. Several methods have been introduced to identify similar roads such method in paper [20] by Safra et al. that is using endpoints of road segments matching. This approach is matching small straight road segments, and it did not take into consideration more than one road segment

for a matching. Also, it could not help to identify similar roads with additional road segments. Other methods are utilizing buffer for road or part of it in one dataset and find the roads from another dataset that fit entirely inside the buffer [22, 17]. Such methods can fetch candidate roads. However, the buffer technique does not determine if the candidate roads are similar in shape or not; neither specify if it can confirm the matching if one road in one dataset has additional road segment than the road representation in another dataset. In [23] and [24], the road comparisons are begun with road intersection (point-matching) to the next road intersection and then use the topology matching of the polyline to identify matching segments. Finally, they measure the average distance between roads. This method of comparing the intersections considers significant points in the road. However, it does not pay attention to the points that change the shape of a road, and they are not intersections.

Matching similar roads process is utilizing polylines matching process as the road maps datasets are stored as vector data. Therefore, there are well-known measures are used such as Euclidean distance, which it has been widely used in the Geospatial field [25]. It is a simple distance measure process between two points and its value is calculated by determining the average distance between corresponding points. However, it may not give enough information such as how much the polylines are similar to each other. There is a Hausdorff distance method, which is a way to express the spatial similarity between two polylines [26]. It chooses the largest values among minimum distances from one set to another. However, this method takes only into consideration the maximum value of these distances and ignores all other points' distances. Therefore, outliers play a major role to make similar roads different in the results. Moreover, there is another measure called Fréchet Distance [27]. Fréchet distance measures the maximum distance between two oriented lines, which takes into consideration the location and the order of points along the polylines. It has the same

drawback as Hausdorff distance since it considers only the greatest distance value. In addition, Fréchet distance has a high cost of computation and complexity [28]. One of the interesting measures is called Dynamic Time Warping (DTW) distance that has been originally used in automatic speech recognition [29], which selects the optimal alignment between two time-series. However, DTW can lead to dramatically different results due to the sampling process [30].

2.7 Conclusion

This chapter presented a framework that is beneficial to smart cities and other applications that use maps. Our framework could be adapted to be utilized by various applications. It takes advantages of existing multiple road maps then matching them together to find the differences between them and produce more accurate road maps. The main two components in this framework are generating the candidate similar roads and similarity measurements. In this chapter, we have utilized semantic features manually to generate the candidate. Then we have used the measurements to confirm if the candidate similar roads are similar or not by using local divergence measurements that make sure these candidate roads have an approximately same length and these roads run in parallel beside each other, which preserve the shape between them. Confirming the similarity also requires global divergence measurement condition to be met that ensures the candidate roads are for the same road in reality and they are not different roads that happened to be beside each other. Also, our method can identify partially similar roads when one of the roads has either missing road segments or additional incorrect road segments.

CHAPTER 3

USING LOCAL AND GLOBAL DIVERGENCE MEASURES TO IDENTIFY ROAD SIMILARITY IN DIFFERENT ROAD NETWORK DATASETS

In this chapter, we start with an introduction about identifying the road similarity and the difficulties of how to measure the roads' similarity in Section 3.1. Then, we mention the motivation of this chapter in section 3.2. After having an overview of previous works in section 3.3, we define our method of measure the similarity of two roads in Section 3.4. In Subection 3.4.1, we provide an overview about the Hausdorff distance as our method inspired by this measure. Then, we Define our methods in 3.4.2, which are Local Divergence Measurements 3.4.2.1 and Global Divergence measurement 3.4.2.2. In Section 3.5, we list different scenarios of experiments and provide a real example. In Section 3.6, we summarize and conclude the research contribution presented in this chapter.

3.1 Introduction

GIS Road network maps can be represented in a raster format or a vector format. Vector data deal with coordinate systems and its representations could be points, lines, or polygons, while raster data is a cell-based data where each cell has information from sources such as aerial and satellite imagery [31]. This chapter focuses on road network maps, which are represented as vector formats that deal with object coordinates. Most recent vector road network maps have good accuracy from the road's coordinates (positions) perspective that form the roads segments representing

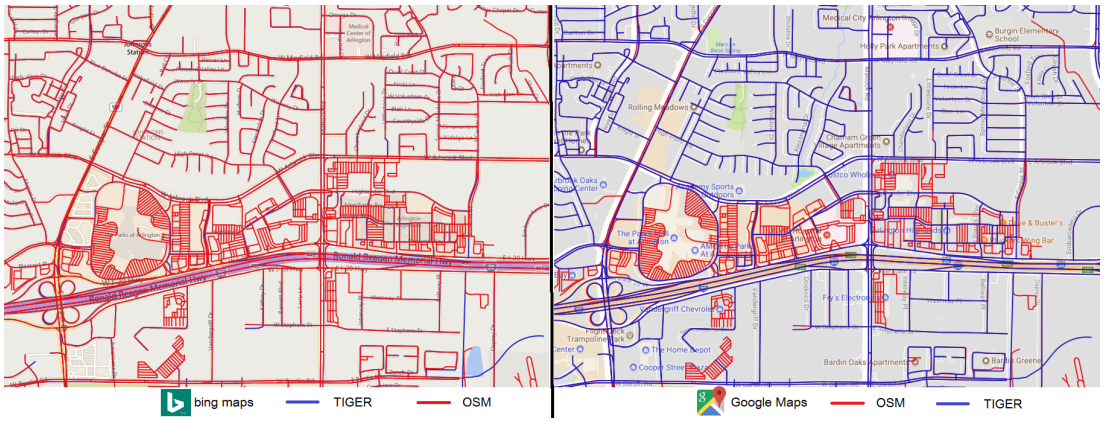


Figure 3.1: OSM over TIGER over Bing Maps (left) and TIGER over OSM over Google Maps (right)

the roads. These maps may not have exactly the same coordinates, but they have minor divergences and are generally close enough to tell they belong to the same road with matching processes. Figure 1 shows an example of four road network maps for the same area with relatively minor divergence. The left picture has the OSM [8] (red lines) and TIGER [3] (blue lines) maps on top of BING maps [5] and in the right picture shows as well OSM and TIGER on the top of Google maps [4] for the same area. However, there are data discrepancies between road network map datasets for various reasons. For example, specific roads may have different names in different maps, or other datasets may have roads that do not exist anymore. Some road maps are not up-to-date, which leads to missing new roads or new road segments. These data discrepancies lead to errors in different applications such as navigation services, data integration, missing/incorrect locations, and more. Therefore, matching the similar roads could help to point out such discrepancies, and determine new roads or missing ones as well as correcting other errors.

This chapter contributes in confirming candidate similar roads from two datasets (OSM and TIGER) to determine if they are really similar or not by using local di-

vergence measurements and global divergence measurement. We define candidate roads as similar or not by making sure they have similar length and shapes, and these roads are located in the same locations. Local divergence measurements ensure these candidate roads have approximately the same length and these roads run in parallel to each other, which preserves the same shape between them. This method could be done by measuring the distances between points of each pair of candidate road segments after overlaying one dataset on top of the other. Confirming the similarity also requires global divergence measurement to be met certain value after the local divergence measurements have been met their conditions. Global divergence measurement ensures the candidate roads are for the same road in the real world and they are not different roads that happened to be beside each other. There are several ways to find the candidate similar roads such as utilizing semantic attributes such as road name or road ID, or using geometric attributes and computing metric features to determine the candidates. Therefore the approach of this chapter is to find a way to confirm if the candidate roads are similar or not. Our method can also identify if there are missing road segments or extra incorrect road segments in one of the datasets by checking if they satisfy all conditions except the *Mode* comparisons (see section 3.4). This novel method utilizes geometric methods for computation that requires the similarity of geometric coordinates for both datasets in order to make these datasets overlay on top of each other.

This chapter is organized as follows: in section 3.2, we provide our motivations to do this work. Then in section 3.3, we describe the related works have been done for matching similar roads. In section 3.4, we define road divergence mathematically and how to compute the local and global divergence measurements. Next in section 3.5, we conduct the experiments to test this method and discuss the results. Finally, we conclude the outcomes and highlight feasible future work based on this chapter.

3.2 Motivation

Many of cities in the world have several road network datasets from different sources that represent the roads for these cities. Most of the times these datasets have data discrepancy among each other such as different names from different datasets for the same road, missing road segments, and so on. As a result, it requires after running the matching process to find out a way to confirm if the candidate similar roads are similar in the real world or not. Road similarity matching can be done either automatically by utilizing the road's spatial or semantical features, or manually by searching for specific roads by any semantic road features such as road name or road ID. The results of this matching process are the candidate similar roads. It is not always the case that the candidate similar roads are really similar. As a result, identifying the actual similar roads from the list of candidate similar roads is essential. Criteria like similar length and parallel road segments can be used. In addition, there are cases where the candidate roads are similar, but one of them has missing road segments or has extra road segments that are mistakenly included in a specific road while in the real world they do not exist.

The contributions of this chapter are to identify if these roads are really similar or not in the real world as well as to determine missing road segments or incorrect extra road segments in one dataset than another. Therefore the method of this chapter is to find a way that can confirm if the candidate roads are similar or not. One of the applications for this method is comparing the road map dataset with its old versions to identify new roads and direction of existing roads' expansions. Also, in this application, determination of new roads can be detected as extra road segments exist in the newer version and also, a closed road can be caught that if the extra road segments come from the older version.

3.3 Related Works

Matching similar roads methods from different datasets have been the focus of several papers [1] for various applications: road networks integration and data enrichment, resolving data discrepancies in semantic attributes such as road name, and more. Different methods have been used to identify similar roads such methods utilizing buffer for road or part of it in one dataset and find the roads from the other dataset that fit entirely inside the buffer [22, 17]. These methods can fetch candidate roads. However, the buffer technique does not determine if the candidate roads are similar in shape or not; neither does it have the capability to confirm the matching if one road in one dataset has extra road segment than the road in another dataset. Other method [20] by Safra et. al is using endpoints of road segments matching. This approach is matching small straight road segments and it did not take into consideration more than one road segment for a matching. Also, it could not help to identify similar roads with extra road segments. In [23] and [24], the matching process is started with road intersection (point-matching) to the next road intersection and then use the polyline topology matching. At the end, they measure the average distance between roads. This method of comparing the intersections considers significant points in the road. However, it does not pay attention to the points that change the shape of a road and they are not intersections.

In general, matching similar roads method is considered as polylines matching and there are well-known measures used such as Euclidean distance and such measure has been widely used in the Geospatial field [25]. Euclidean distance is the simple distance measure between two points. Euclidean distance value is calculated by measuring the average distance between corresponding points. However, measuring the distance between polylines using Euclidean distance may not give enough information such as how much the polylines are similar to each other. Also, there is a Hausdorff distance

which is a way to express the spatial similarity between two polylines [26]. This method chooses the largest values among minimum distances from one set to another. However, it takes into consideration the maximum value and ignores all other points' distances. Therefore, outliers play a major role to make similar roads different in the results. Moreover, there is another measure called Fréchet Distance [27]. Fréchet distance measures the maximum distance between two oriented lines which takes into consideration the location and the order of points along the polylines. It has the same drawback as Hausdorff distance since it considers only the greatest distance value. In addition, Fréchet distance has a high cost of computation and complexity compared to the Hausdorff distance[28]. One of the interesting measures is called Dynamic Time Warping (DTW) distance that has been originally used in automatic speech recognition [29] which selects the optimal alignment between two time-series. However, DTW can lead to dramatically different results due to the sampling [30].

3.4 Mathematical Definitions of Road Divergence

In order to compute the similarity between two roads that represent the same road in the real world but they have slight difference due to how their road's coordinates were captured, there is a need to find out the points coordinates that represent the road segments. These points have coordinates information, longitudes and latitudes. Based on this information, road similarity could be measured between two roads.

There are various ways to measure the divergence between two roads as discussed in section 3.3. One approach is to use Hausdorff distance (section 3.4.1), but this approach is not enough to confirm the similarities between two candidate similar roads. Therefore other divergence measures (local and global divergence measurements) are introduced in this chapter that can give more information to decide whether the can-

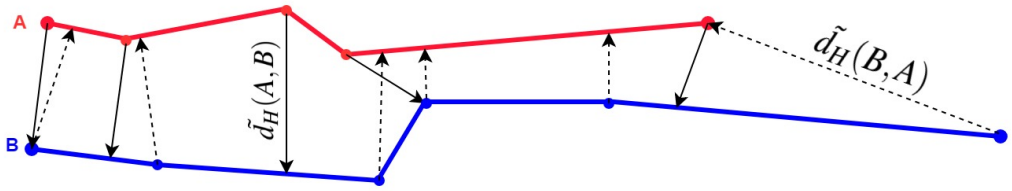


Figure 3.2: Hausdorff distance is not commutative

didate roads are similar or not. These two measurements are discussed in section 3.4.2.

3.4.1 Hausdorff Distance Between Two Candidate Roads

Our method is inspired by Hausdorff distance method [26] which is one of the well-known approaches to measure lines similarity. This method helps to find out the abnormalities between a road's representations from different datasets since roads are represented as lines in the vector datasets. Therefore, we use sometimes in this chapter term "line" to refer to the "road". It basically goes through all points of line segments from the first line and for each point computes the minimum distance between this point and the closest point from another line. After that, the same approach will be repeated from other direction, i.e., from the points that represent the second line to the nearest point from the first line. The Hausdorff distance is a single value defined by finding the largest value of distances that were computed from those points (i.e., the maximum of the minimum distances).

Let us suppose there are two lines A and B. Each line is consisting of list of connected points that represent the line or part of the line (line segment) such that: $A = \{a_1, a_2, ..., a_n\}$ and $B = \{b_1, b_2, ..., b_m\}$

The one-sided Hausdorff distance from A to B is defined as:

$$\tilde{d_H}(A, B) = \max_{a \in A} P_{diff}(A, B) \tag{3.1}$$

where $P_{diff}(A, B)$ is a list of minimum distances between A's points to the line B:

$$P_{diff}(A, B) = [\min(a_i, B)] \forall a_i \in A$$

where $\min(a_i, B)$ represents the distance from point a_i to the nearest part of line B and this distance could be computed by ordinary straight-line (Euclidean) distance as in the following equation:

$$||a - b|| = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$

where a_x, b_x and a_y, b_y represent the longitudes and latitudes of the point for points a and b respectively.

Note the Hausdorff distance is not commutative, which means $\tilde{d_H}(A, B) \neq \tilde{d_H}(B, A)$. Figure 3.2 shows a case where there is a big difference between values of $\tilde{d_H}(A, B)$ and $\tilde{d_H}(B, A)$ because line A is shorter than line B and the distances are short when they are computed from A's points to nearest point of B. On the other hand, if the distances are measured from B the far endpoints on B have longer distance value than the maximum distance from any point on A to B. Therefore, it is

important to compute both directions. As a result, bidirectional Hausdorff distance between A and B can be found as follows:

$$d_H = \max(\tilde{d_H}(A, B), \tilde{d_H}(B, A)) \tag{3.2}$$

3.4.2 Other Road Divergence Measurements

In order to have more accurate findings, this chapter introduces other measurements in order to determine the similarity of two lines. These measures have important indications for global divergence, which take into consideration the dataset overall, and for local divergence relates to the specific two lines' characteristics.

3.4.2.1 Local Divergence

There are two metrics- Mode value $Mode(A, B)$ and its Frequency value $Freq(A, B)$- can be used in order to determine the divergence between two lines. The first metric is the *Mode* which is the distance value that has highest frequency of occurrences in the two lists $P_{diff}(A, B)$ and $P_{diff}(B, A)$. When creating the list of minimum distances, the numbers are rounded to the closest meter; then we count the most frequent number occurring in the list to indicate the divergence. Recall that the P_{diff} vectors are:

$$P_{diff}(A, B) = [\min(a_i, B)] \forall a_i \in A$$

and

$$P_{diff}(B, A) = [\min(b_i, A)] \forall b_i \in B$$

This $Mode$ value is important to determine overall how far apart these two lines are from each other. $Mode$ value can be compared with Hausdorff value from equation 3.2 and based on that two possible cases could happen; either $(Mode(A, B) + \tilde{\epsilon}) \geqslant d_H$

or $(Mode(A,B) + \tilde{\epsilon}) < d_H$ where $\tilde{\epsilon}$ is an estimated value of the margin of error of lines locations -a relatively small value that is computed manually-. In the first case when the single bidirectional Hausdorff is less than or equal $(Mode(A,B) + \tilde{\epsilon})$, this indicates there is a high possibility that these two lines are identical and have the same shape and length. While if the $(Mode(A,B) + \tilde{\epsilon})$ is less than bidirectional Hausdorff distance, it indicates there is a divergence in one line segment or more from one line or both lines. Furthermore, the two directional Hausdorff distances may provide more information when compared to Mode value. For example, if $\tilde{d_H}(A,B)$ is less than or equal $(Mode(A,B)+\tilde{\epsilon})$ and $\tilde{d_H}(B,A)$ is greater than $(Mode(A,B)+\tilde{\epsilon})$, this indicates that line B has one or more line segments that do not exist in A. This difference can have several explanations from traffic road perspective such as new line segments are captured by B's dataset while missing in A's dataset, or it exists in A's dataset but with different road name, and so on.

The second metric that can be used to measure lines divergence locally is *Mode Frequency* $Freq(A,B)$, which counts the frequency of the Mode value occurrences. When counting the frequency of the mode, we take a range $[(Mode(A,B) - \tilde{\epsilon}), (Mode(A,B) + \tilde{\epsilon})]$ and count all values in the range. *Mode Frequency* can detect the similarity between two lines by comparing its value that represents occurrences with the total number of points in both lists. This comparison can be computed by getting the percentage between *Mode Frequency* and the total number of points in both lists, $Freq(A,B)/(N+M)$ where N and M are the numbers of coordinates points for sets A and B, respectively. Because there is a slight divergence in the road representation from different datasets, it is better to add the three highest occurrences of distance values in the list together. Then we see if the value of summation represents 80% or more of the number of points, which indicates these two candidate

lines have overall similar shape. It is important to mention that these cumulation frequencies should not be apart more than $\pm\tilde{\epsilon}$ from each other.

There is one more metric that can be computed from local divergence arguments and used to compare with global divergence measurement. This metric is a local average divergence between two candidate similar lines. Local average divergence (Avg_L) for two lines could be computed as the following:

$$Avg_L = \frac{(\sum_{i=1}^{N}[\min(a_i, B)] + \sum_{j=1}^{M}[\min(b_j, A)])}{(N+M)} \tag{3.3}$$

$\forall a_i \in A$ and $b_j \in B$

3.4.2.2 Global Divergence

Local metrics deal with only two specific roads (lines). It does not take into consideration the overall datasets divergence. Global divergence provides an average of all local divergence between the two datasets. Global divergence is essential as sometimes local divergence measurements can compare two parallel lines and find out they are similar while in the real world they are two different parallel lines, even though they have a similar value of $Mode$ and the Frequency $Freq(A, B)$ is bigger than 80%. Therefore, use global divergence to make sure they are the same road in the real world. We compute global divergence as follows:

$$Avg_G = \frac{(\sum_{i=1}^{w} Avg_{Li})}{w}$$

where Avg_{Li} is the local average distance for all similar road pairs and w is the number of those similar road pairs.

Global divergence can be used to compare the local average distance of the candidate similar roads Avg_L with the global average distance value Avg_G. If the Avg_L is larger than the Avg_G by more than 10%, it indicates that most likely these two candidate roads are different from each other and run in parallel in the real world; such as an example in subsection 3.5.4. Otherwise, it is most likely these two roads are similar to each other. The 10% value added to Avg_G is the threshold and it is manually identified based on the nature of datasets. Based on trial and error, our experiments show that 10% difference splits the similar roads than different roads and this is the reason we choose 10%.

This comparison comes as a second step after the candidate similar road pairs have passed the local divergence test. In the case of missing road segments from one dataset while it is available in the other dataset, this measurement can be used to identify the similarity by comparing the $Mode(A, B)$ instead of Avg_L with global divergence value Avg_G. Moreover, it can predict the locations of the missing segments.

Two candidate roads can be similar if the three following conditions are met: 1. $(Mode(A, B) + \tilde{\epsilon}) \geqslant d_H$, 2. $Freq(A, B)/(N + M) > 80\%$, and 3. $Avg_L < Avg_G * 1.10$, which means the overall distance between the two candidates similar roads is within $Mode(A, B)$ value. While $Freq(A, B)/(N + M) > 80\%$ means the two roads are running in parallel and consequently have similar shape, and $Avg_L < Avg_G * 1.10$ means they are most likely to be the same road because the distance between them is within the average of other similar roads- to make sure they are not two different roads in the real world that happen to run in parallel. If the first condition is not met while the others have been met, this could mean that there is a high possibility these two roads are similar but one of them has missing segments of the road for any

reason such as different name, new segments of the road have not been captured or simply incorrect extra segments.

3.5 Experimental Results

In order to examine our method, we use two road-map categories: Volunteered (crowdsourced) geographic information (VGI) category is represented by OpenStreetMap (OSM)[8]. The second category is Authoritative Geospatial Data which is represented by Topologically Integrated Geographic Encoding and Referencing system (TIGER)[3]. The datasets we have are for Tarrant County, Texas: 1. OSM: which is a free road-map open to the public to contribute and build up the data; 2. TIGER Road map dataset that is produced by the US Census Bureau. TIGER is considered as a professional road map network. In order to apply the method in this chapter, first, we have to identify the candidates of similar road pairs either manually by specific road name for example or automatically such as road maps integration. In this chapter, the experiment has run manually by identifying the candidate similar roads semantically using road name and then applying the divergence measurements on these roads to figure out if they are similar and are representing the same road in the real world or not. Measuring of a distance has been approximated to the nearest integer value of meter unit as the fraction of meter does not have a significant impact since the coordinate is captured as a point in the lane which has a width of up to 3.6 meters [32]. Estimating $\tilde{\epsilon}$ -which is the margin of error of candidate roads' locations- is based on two factors: 1) The average width of the road and 2) how far is the similar roads from different datasets to each other. Based on that, the distance of similar roads from two datasets (OSM and TIGER) is within 4 meters as well as almost all the roads have at least two lanes, one for each direction, so any coordinates point to any location within the road's width (7.2 meters) is considered from the

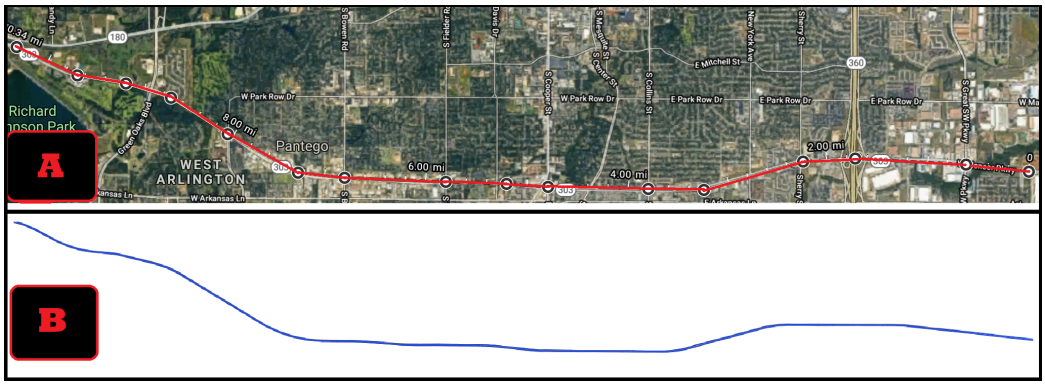


Figure 3.3: Pioneer Parkway road Satellite image (A) vs. Tiger Dataset (B)

road. Therefore, whichever is the bigger value from these two factors is going to be assigned as a value of the margin of error $\tilde{\epsilon}$. In this case, the average road's width is bigger and therefore $\tilde{\epsilon}$= 7. Regarding the global divergence measures after taking the average of several correct matching roads, the global average distance is equal to $Avg_G = 19$. Furthermore, this method provides accurate results with relatively long candidate roads as compared to small road segments could confuse their geometry length with divergence measurements.

3.5.1 Candidate roads are similar

This subsection shows an example of candidate roads that are similar in the real world and have met all three measurements conditions. The example is for "Pioneer Parkway" road in Tarrant County. This road stretches more than 17 kilometers. Figure 3.3 shows a satellite image from Google maps [4] for "Pioneer Parkway" road as well as the plot of TIGER coordinates. It has been represented in TIGER by 434 coordinate points while in OSM this road has 668 points. After computing the distance for each coordinate point to the closest point of the corresponding candidate road, the following values are the results: $Mode(T, O) = 1$, $d_H = 7$ where T is TIGER

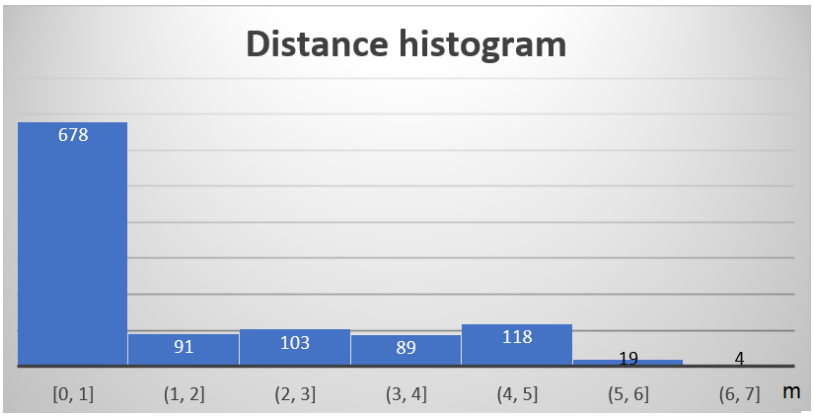


Figure 3.4: Pioneer Pkwy distance histogram between Tiger and OSM

and O is OSM, highest $Freq(T,O)_1 = 1102$, local average distance $Avg_{L\_Pioneer} = 2$, and global average distance $Avg_G = 19$. There is no need to count the second and third highest frequency as all number of coordinates are falling in the highest frequency category. Figure 3.4 shows the frequency of coordinate points in each meter.

Now, we see whether the measurements conditions have been met or not: 1) compare $Mode(T,O)$ value with bi-directional Hausdorff distance: $(Mode(T,O)+\tilde{\epsilon}) = 1+7 = 8 \geqslant d_H = 7$, therefore, this condition is passed; 2) Are 80% of roads' coordinates or more running in parallel to each other or not?: $((Freq(T,O)_1 = 1102)/(434+668 = 1102)) = 1.00$ which means 100% of candidate roads' coordinates are running exactly in parallel. 3) Are the candidate roads representing the same road in the real world or different parallel road?: $Avg_{L\_Pioneer} = 2 < Avg_G = 19*1.10 = 21$. As a result, all measurement conditions have been met. Therefore, we can say these two candidate roads are matching to each other and they represent the same road in the real world. As it is shown in figure 3.5 after we get a close look at TIGER plot and OSM plot these roads are not exactly identical and there is some divergence between them. However, these small differences between them are considered as the margin of errors (section 3.4.2.1) and it does not impact the overall outcome.

Table 3.1: Local and global divergence measures' values for Pioneer Pkwy

Measurement	Value
$Mode(T \Leftrightarrow O)$	1
$Freq(T \Leftrightarrow O)_1$	1102
$Freq(T \Leftrightarrow O)_2$	0
$Freq(T \Leftrightarrow O)_3$	0
Avg_L	2
d_H	7
Avg_G	19

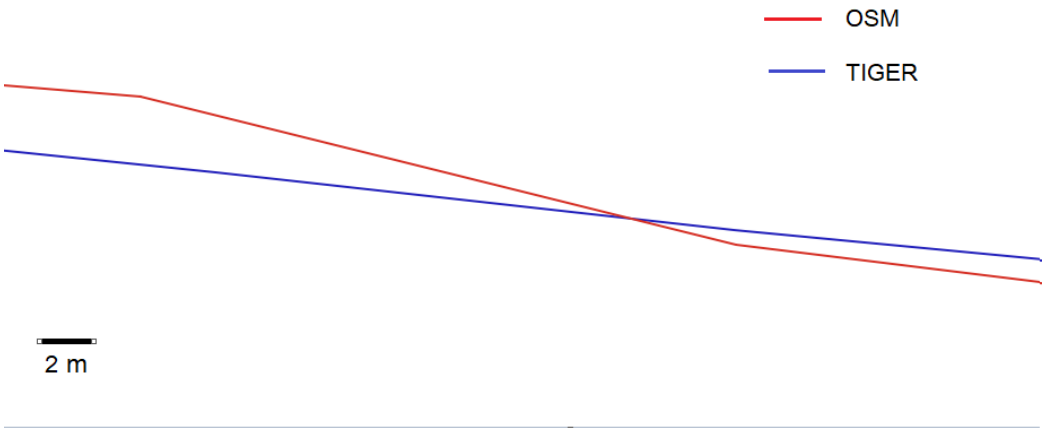


Figure 3.5: TIGER and OSM plots part of Pioneer Pkwy

This case is the optimal scenario but in the real world with different representations, this is not always the case. Sometimes, it represents the same road but with some missing road segments from one dataset or extra incorrect road segments in another such as the next scenario in section 3.5.2.

3.5.2 Candidate roads are similar but there are some road segments missing/extra in one of the dataset

There are some roads that have different representations in different datasets due to many reasons; for example, some datasets capture new updates while others

Table 3.2: Local and global divergence measures' values for South Cooper Street

Measurement	Value
$Mode(T \Leftrightarrow O)$	7
$Freq(T \Leftrightarrow O)_1$	546
$Freq(T \Leftrightarrow O)_2$	231
$Freq(T \Leftrightarrow O)_3$	31
Avg_L	9
d_H	289
Avg_G	19

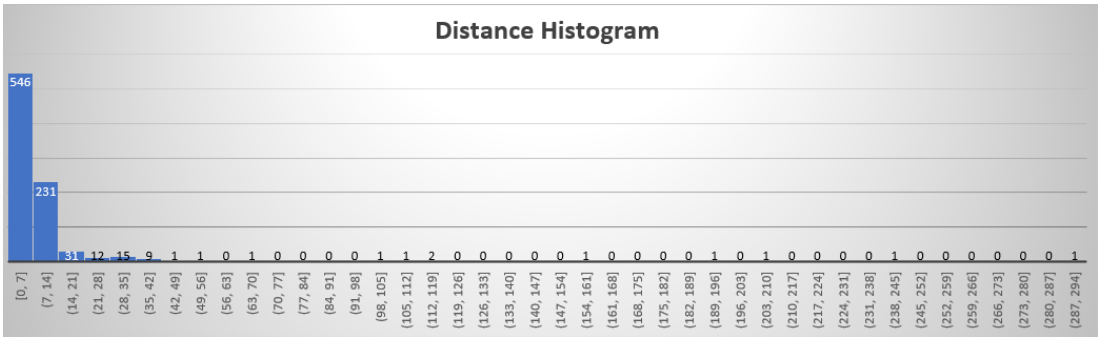


Figure 3.6: South Cooper Street distance histogram between Tiger and OSM roads

are not updated. Sometimes there is some extra road segment that is wrong and does not exist in the real world. In some cases, road segments are missing due to having different road name or sometimes it is renamed in different road segments especially adjacent road segments. Here is the example for "South Cooper Street" which spans for more than 14 km and it is represented in TIGER by 207 coordinates while in OSM, it is represented by 648 coordinates. Table 3.2 has the computation values for these two candidate roads.

Figure 3.6 shows the distributions of the coordinates based on the distance. For the first condition: $(Mode(T, O) + \tilde{\epsilon}) = 7 + 7 = 14$ which is very small comparing to the bi-directional Hausdorff distance $d_H = 289$ so, this condition is not met. For second measurement condition, $(546 + 231 + 31)/(207 + 648)) = 0.95$ which indicates

that 95% of coordinates are running in parallel to corresponding candidate road; thus this measurement condition has passed. Regarding the third condition which is the global divergence, it compares local average distance to the global average distance and sees if the local average distance is within the global average distance range or not: $Avg_{L\_SCooper} = 9 < Avg_G = 19 * 1.10 = 21$ it is shown this condition has been met. Note, sometimes if the first condition is not met and the candidate roads are not relatively long, it is better to use $Mode(T, O) + \tilde{\epsilon}$ instead of Avg_L which is met in this case also $Mode(T, O) + \tilde{\epsilon} = 14 < Avg_G = 19 * 1.10 = 21$. In the above result, the first condition of local divergence is not met while the second condition of local divergence and global divergence are met. It indicates the two candidate roads are similar, but there is a missing road segment from one dataset while it exists in the other or extra incorrect road segment in one dataset. In this case and after investigation, figure 3.7 shows that TIGER dataset has missing road segment of "South Cooper Street".

3.5.3 Candidate roads are not similar

The scenario for this case is significant because common mistakes could happen just with swapping the direction of road name such as "North Cooper" instead of "Cooper North" and sometimes the road names are totally different from each other. Such errors could also happen when we make auto road matching depending on semantic attributes only. The experiment gets two candidate roads where TIGER takes road name called "Cooper North" and OSM takes a road name called "North Cooper"; same names but the difference is one has the direction before the road name and the other has it after the road name. The total coordinates for these two roads are 252 coordinates points. Based on that and measurements values in table 3.3, the following are the explanation for these values:

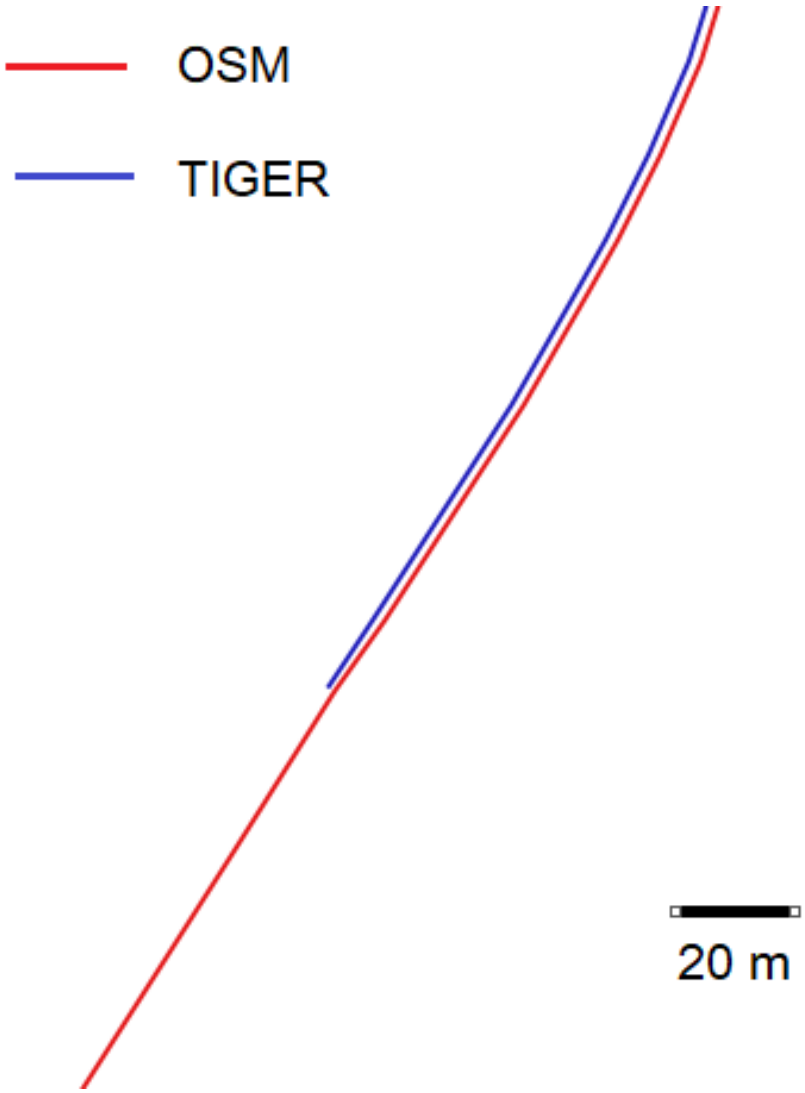


Figure 3.7: S Cooper Street in TIGER has missing road segment

Table 3.3: Local divergence values for North Cooper

Measurement	Value
$Mode(T \Leftrightarrow O)$	7
$Freq(T \Leftrightarrow O)_1$	5
$Freq(T \Leftrightarrow O)_2$	4
$Freq(T \Leftrightarrow O)_3$	4
Avg_L	1165
d_H	3123

Figure 3.8: North Cooper from OSM and Cooper North from TIGER

When applying these values for $Mode(T, O)$ condition in local divergence, it is shown that this measurement condition is not met : $7 + 7 = 14 << d_H = 3123$. To test if the candidate roads are parallel or not: $((5 + 4 + 4)/252) = 0.05$ that is a very small value (5%) to confirm the parallelism between the roads, so, this condition is not passed. For global divergence, this case shows the local average distance is bigger than the global average distance $Avg_{L\_NCooper} = 1165 >> Avg_G = 21$ which does not meet the condition. As local divergence measurements are not met the matching criteria, these two candidate roads are not similar, and there is no need to continue for computing the measurement for global divergence. Figure 3.8 shows the plot of TIGER and OSM and how it is obvious these two roads are not similar in the real world and from a satellite image, too.

Table 3.4: Local divergence values for North Cooper vs. Cooper North

Measurement	Value
$Mode(T \Leftrightarrow O)$	83
$Freq(T \Leftrightarrow O)_1$	10
$Freq(T \Leftrightarrow O)_2$	0
$Freq(T \Leftrightarrow O)_3$	0
Avg_L	82
d_H	84

3.5.4 Candidate roads have similar shape but are not similar

Testing candidates roads to determine if they are apart from each other by a fixed distance along roads plays a significant role in deciding if they are similar roads or not. However, this is not enough because sometimes there are two different roads running in parallel with a fixed distance between them, but this distance is very long and indicates they are different roads in the real world. Here is where the global divergence plays its role to decide if the roads are for the same road or for different roads that happen to be in parallel. Such scenarios frequently happen when matching the candidate roads based on spatial features only. The experiment conducted in this category is getting the road called "Janann Ave" from TIGER and match it with OSM road called "Marlee Lane". These two roads span for more than 500m. It generates ten coordinates points in total. The values of local divergence measures are computed and presented in the table 3.4.

These results can pass the local divergence measurements, since $Mode(T, O) = 83$ and $83 + 7 = 90 > d_H = 84$. So, this indicates there are no missing/extra road segments in one dataset than another. Also, it shows that the candidate roads maintain constant distance apart from each other $Freq(T, O)_1 = 10/10 = 1.0$ thus all coordinates are running in parallel to the corresponding road from the other dataset. At this point, local divergence metrics have been met which could conclude they are

Figure 3.9: Candidate roads have same length and run in parallel but they have a long distance far them apart

similar except we need to make sure if they are similar or just two different roads that happen to run in parallel. Global divergence takes place here: $Avg_L = 82 > Avg_G * 1.1 = 21$ which indicates that these candidate roads are representing different roads and they are not similar in the real world as shown in Figure 3.9 that is depicted from a satellite image.

3.6 Conclusion

This chapter presents a novel method inspired by Hausdorff distance to confirm if the candidate similar roads are similar or not by using local divergence measurements that make sure these candidate roads have approximately same length and these roads run in parallel beside each other which preserve the shape between them. Confirming the similarity also requires global divergence measurement to be met that ensures the candidate roads are for the same road in real words and they are not different roads that happen to be beside each other. In addition, our method can identify the similar roads, but one of the roads has either missing road segments or extra incorrect road segments.

CHAPTER 4

HISTORICAL COMPARISON BETWEEN OLD VERSION OF ROAD MAP DATASETS WITH NEW VERSION

One of the application of the proposed framework that have been introduced in chapter 2 is comparing the new version of Road maps with an older version to study the roads changes in the city. We start with introduction in section 4.1. After that, we define the problem. Then, we explain the matching process in section 4.4 and next we conduct the experiment in 4.5. Finally, we give a conclusion in this chapter 4.6.

4.1 Introduction

Over time everything is changed with slow or fast pace of changing; either we notice it or not. Roads of the cities are also changing over the time for many reasons [28, 33, 34, 35]. For example, new commercial or residential areas need new main and local roads to be constructed to serve these area; or some main roads are extended to serve new areas. Congested roads that have high traffic are gotten changes therefore, such roads are expanded to contain the high traffic and make the traffic flow smooth. From road map perceptive these changes are captured by representing them by two parallel lines -one line for each direction- instead of one-line representation. Another reason of the change in roads are changing semantic attributes such updating the road name or road ID either for misspelling correction or simply changing the attribute for any other reason. Sometimes the existing roads are needed to be removed for some other projects for instance, AT-T stadium which is also known as Cowboys stadium has been built on area that has roads and buildings as shown in figure 4.1 that is taken

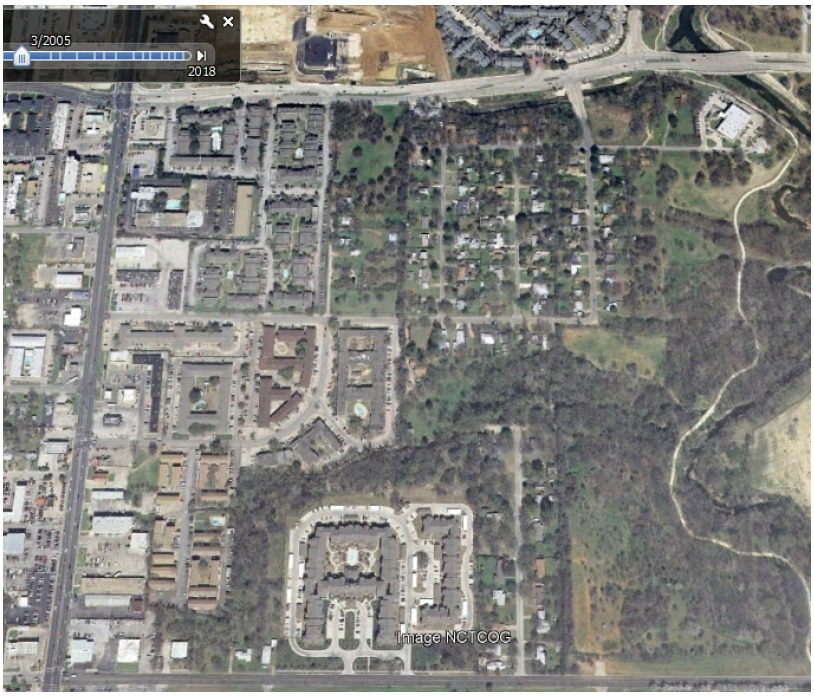


Figure 4.1: The area before building AT-T stadium has roads and buildings

Figure 4.2: Old roads were removed and new roads are built after AT-T stadium is built

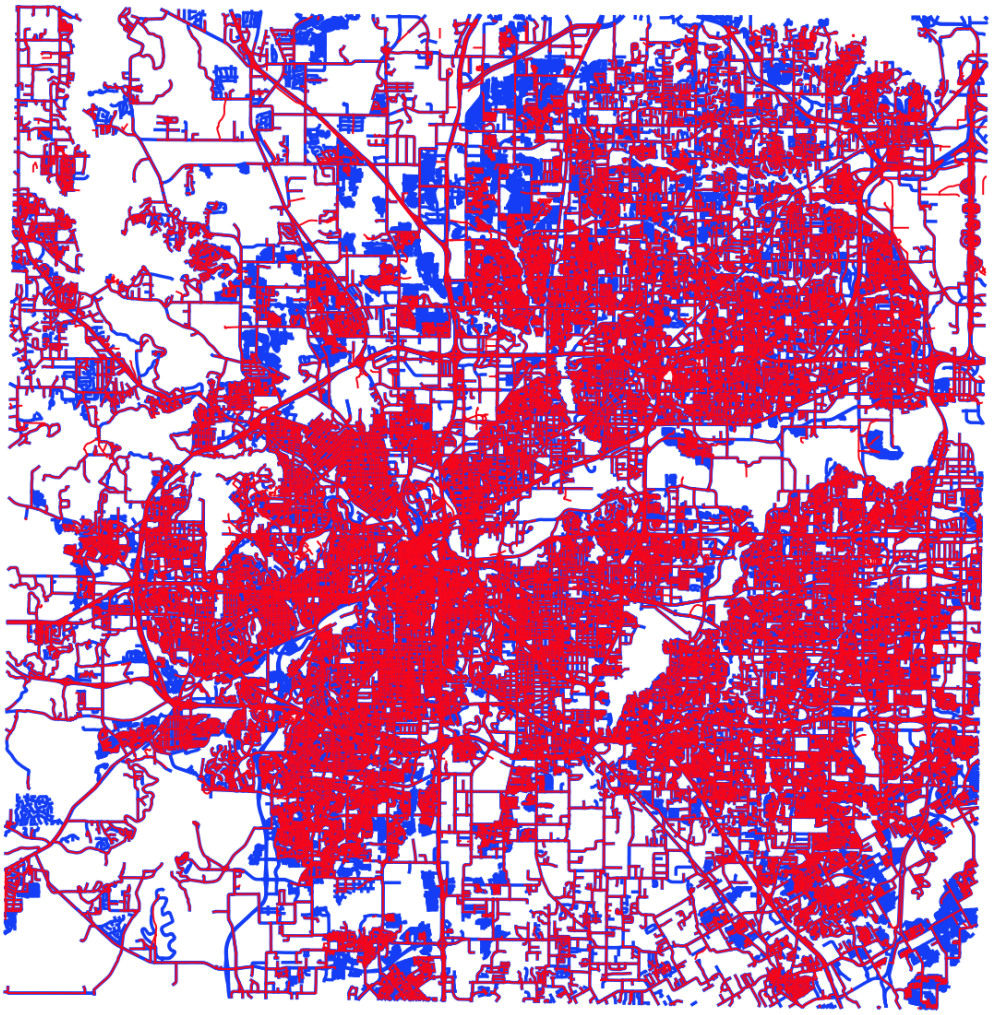


Figure 4.3: 2007 Dataset (red lines) overlay 2018 Dataset (blue lines) for Tarrant County and it shows the new roads in 2018 DS

from [36] and after building the stadium all these roads and buildings are removed and new roads were built as shown in figure 4.2 that is taken from [37]. Occasionally, some roads are shifted and their locations and shapes were changed and usually such cases happen in new areas that are recently built which make eventually changes in road network.

As mentioned above, there are different reasons that change the road network. Therefore, all these changes should be captured in new version of Road Network Map. These maps can be compared with its old versions.

4.2 Motivations and Contributions

One of the TAREEQ framework application that has been introduced in chapter 2 is the ability to matching new version of road map with old road map and making historical comparison between both road maps. This comparison brought several benefits such as keeping track of changes that have happened overtime such as road name was changed, road was permanently closed or removed, road was extended, road was expanded and represented by two-line instead of only one-line, or simply new road was built.

In addition to keep track of the changes, this comparison gives overall picture of where is the city or area growth and we can get a prediction of the pace of changes in future by knowing the pace of changes that have happened between the two maps period. Furthermore, this application can define which region that has the highest percentage of removed roads. Figure 4.3 shows that road map that is updated on 2007 (red lines) overlies road map that is updated on 2018 (blue lines) for Tarrant County and it shows the new roads in 2018 DS. This visual representation does not give clear pictures of the growth percentage for each region neither it gives any idea about the removed roads.

One of interesting findings is simplifying the process of finding errors in updated version of road map. So, the question is that can we count on the most recent version of Road Map and assume it is correct? Are there missing roads that were exist in older version and mistakenly deleted from newer version? Are there in newer version confusing Roads' names such as spelling mistakes, wrong directions that were correct

on older version. This application facilitate the process to highlight the differences between them to take a decision if there are issues with new dataset or not.

The contributions of this work can be summarized as the following: Highlight the differences of historical road maps comparison, Identifying the areas that have gotten changes, Determining the overall changes during the period between the old dataset and new dataset. Finally, Facilitating the investigation of finding the errors by listing the major differences between two historical road map dataset and listing all roads' names changing.

4.3 Preliminaries

In this we are going to define the terminologies we use in this study to avoid any confusion that may happen in section 4.3.1; after that we are going to define the problem formally in section 4.3.2.

4.3.1 Definition

There are some terminologies we use and they may refer to something wrong for the audience, for this reason we try to define them here to avoid any ambiguity.

Definition 1 (Road Coordinates/ Road Points). Road coordinates and Road Points are the same and we use them interchangeably. Road Coordinates are indicate the position of points that forming a line in Road map. It has two values-longitude (X-axis) and latitude (Y-axis).

Definition 2 (Road/ Lines). Road and Lines are the same and it represent a road or part of road that consists of number of coordinates that represents the Road.

Definition 3 (Road segments). is a sub-road that represents part of the road and it also consists of number of coordinates.

Definition 4 (Partial Similar Roads). two roads are partially similar when one of them has extra road segments or the other one has missing road segments.

Definition 4 (Road Extensions). when the road has new road segments connect to the existing road.

Definition 5 (Road Expansions). when the representation of the road has changed from one-line representation in the road map to two-line representation and usually each line represent a direction of vehicle traffic.

Definition 6 (Region). when we have a large area like Tarrant County that has number of cities, this large area can be divided into small areas and each one of them called region.

4.3.2 Problem Definition

We have two version of road network maps for the same area, one is old road network map and the second is the recent road network map. The problem is that: How can we compare same datasets with its old version? Also, How can we make sure missing roads, which were exist in old dataset and no more exist in new version, do not exist anymore in real life or simply the name has changed? In addition, can we get overall picture of the changes and how can we get overall picture of where is the region growth and based on that we can have a sneak peek of the pace of changes over time and predict the changes in future such as the trend of city growth.

TAREEQ FRAMEWORK

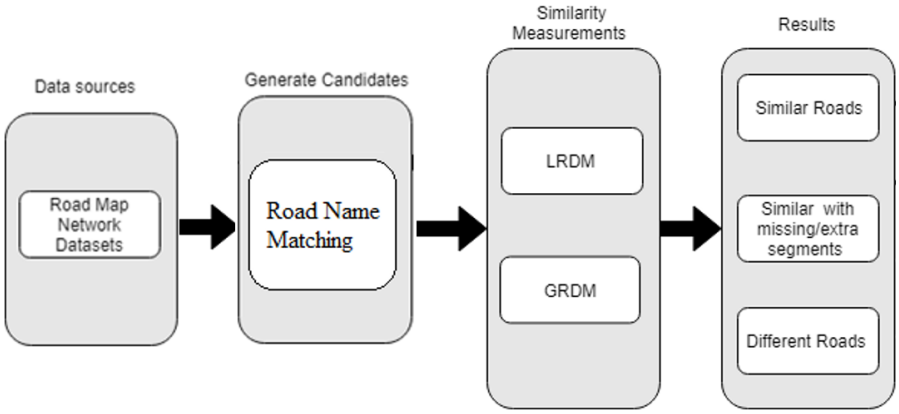


Figure 4.4: TAREEQ Framework

This chapter introduces one of TAREEQ framework application and how the system is built. Then we discuss its results for Tarrant and New York Counties and the quality and performance of the system.

4.4 Map Matching of Old Version vs. New Version (Historical Road Map Matching)

We incorporate our TAREEQ framework in this method of Historical Road Map matching as shown in figure 4.4, which begins with data source preparation. After that, the framework generates the candidate similar roads and it conducts the similarity matching. The later two processes may call each other depends on the certain criteria. Finally, the framework provides the results if the two candidate pairs are similar, partial similar, or different roads. If the decision is partial similar, the framework has the ability to find out what cause this dissimilarity in part of the road. First cause is that if the change is happened because of the road gets extension to

reach further area and it is called road extension; road gets expansion to contain a high traffic and avoid bottleneck, if any, and this is called road expansion which is represented by two-line instead of one-line; or if the new road in different area is built and has a name that is same as name of existing road, then the system will identify it as partial similar with existing of new road that has same name in different area. This section talks about about data source preparation in subsection 4.4.1. After that, in subsection 4.4.2 the main process which is Historical Road Map Comparison that consists of two main methods from TAREEQ framework: generate the candidate similar roads and similarity measurements process. Last subject 4.4.3 is talking about the possible scenarios that TAREEQ framework can determine.

4.4.1 Phase 1:Data Source Preparation

In this phase, we prepare the dataset for the main next phase which is generate candidate similar road and similarity measurements. This phase is important to study the data and how it is constructed and extract the important data. We try to not modifying the data like other methods [38, 39, 40, 41, 42, 43] that they change the roads' coordinates for the sake of their system's results. In this phase we just filter the data that are not related to roads network. It is important to study the data and know which field from first dataset is going to be compared with second data-set's field. Even for the same road map with different version such as the case that we have it when we try to compare 2007 dataset with 2018 dataset. In 2007 dataset, the data has all lines in one dataset and dataset has Hydrography, Rail, Road, and so on while in 2018, the dataset is built to store only roads attributes. Therefore, we work on 2007 dataset to filter and remove all data unrelated to roads before comparison. Even though representing the roads are different in number of records required to store the road's coordinates, our framework has the ability to work with such differences and

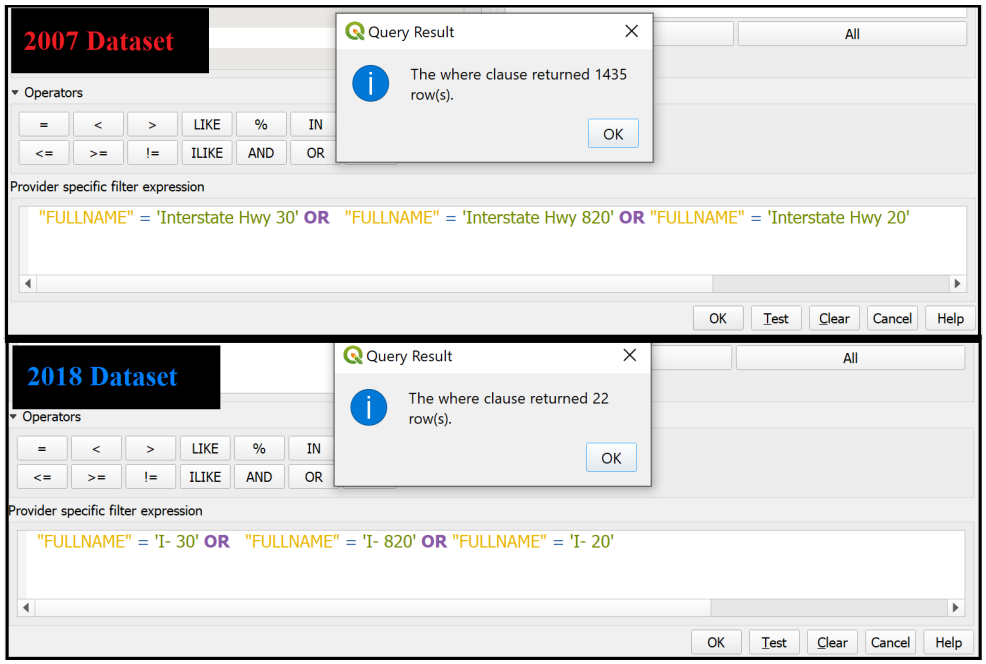


Figure 4.5: Number of records needed to store the information of Interstate Highways in 2007 Dataset compare with 2018 Dataset

it does not need further process to prepare the dataset which differ than other work [44, 31, 45, 20]. For example, if we take all Interstate Highways in Tarrant County, 2007 dataset requires 1435 records while they required only 22 records in 2018 dataset as shown in figure 4.5.

4.4.2 Phase 2: Historical Road Map Comparison: Generate the Candidates and Measure the Similarity

In this phase, we are going to use two processes from TAREEQ framework: "Generate Candidate" and "Measure Similarity" and we will go through this in details. First of all, let us look at the flow chart for the Historical Road Map Comparison Process that is shown in figure 4.6 to get overall idea how it works. Is starts taking the two datasets, i.e. the old dataset and new dataset, we are going to start with generate candidate similarity roads and at the beginning we use semantic attribute

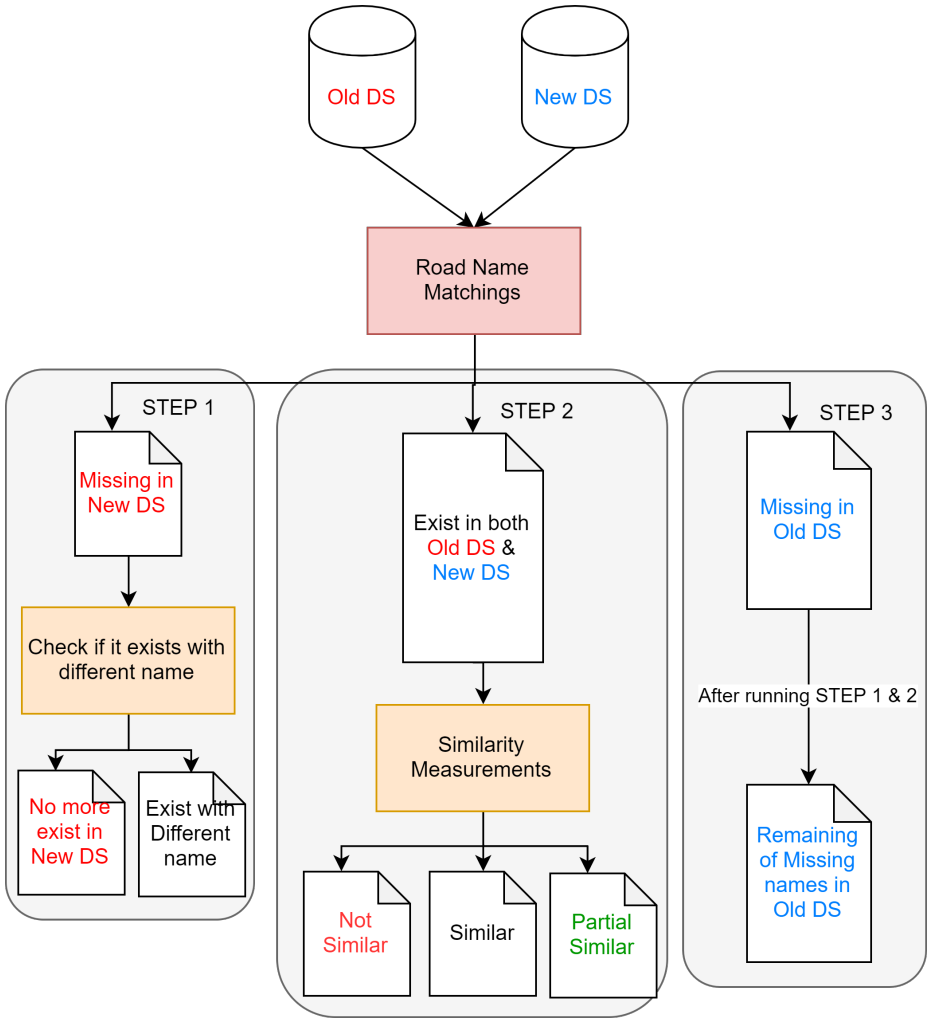


Figure 4.6: Historical Road Map Comparison Process

to find candidate similar roads. The semantic attribute we have used is the name. This process provides three lists: Existing Roads' names in old dataset but missing in new dataset, Existing Roads' names in both datasets, and Existing Roads' names in new dataset and not exist in old dataset.

After we have these three lists, we start with the list that has Existing Roads' names in old dataset but missing in new dataset to do further check if the roads are still exists but with different name. We run the STEP # 1 process as shown in figure 4.6 and we expand this portion of the figure to know exactly how the historical

comparison process do it in figure 4.7. In this subprocess we take first road name that is missing in new dataset. We start the generate candidate process again but with different approach which taking spatial attribute into consideration instead of semantic attribute. We use the buffer technique which we make buffer around the road in old dataset and overlay the new dataset on the top of old dataset and fetch all the roads that are fully or partially inside the buffer. After that we sort the list of candidate similar roads based on how much this candidate road similar in length to the missing road that exist in old dataset and missing in new dataset and also making the sort based on the portion of the candidate road exist in the buffer. In this point we finish the from the generate the candidate process using spatial attribute and start the similarity measurement process. It takes the first candidate roads in the list and measure the similarity with the missing road that exist in old dataset and missing in new dataset and see if they are similar or not. If they are not similar, take the second candidate roads and again make the similarity measurement and see the result. If they are not similar, repeat the last step till either there is similar matching road or the sorted list is empty. If the sorted list is empty that means the missing road that exist in old dataset and missing in new dataset is truly removed in new dataset. Note, if there is partial similarity between the two roads, the loop will continue to see if the unmatched part has matching road with different name or it is removed. Also note that the matching road from new dataset could be from the list of New roads names that exist in new list and missing in old dataset. If so, this list will remove this specific road name.

After step # 1 is processed, the second step is started as it is shown in figure 4.6. It is basically running the similarity measurements process on the list that it has roads names in both datasets; one road name at a time. The possible answers are: the roads that have the same name in the list are similar and identical in both datasets ,

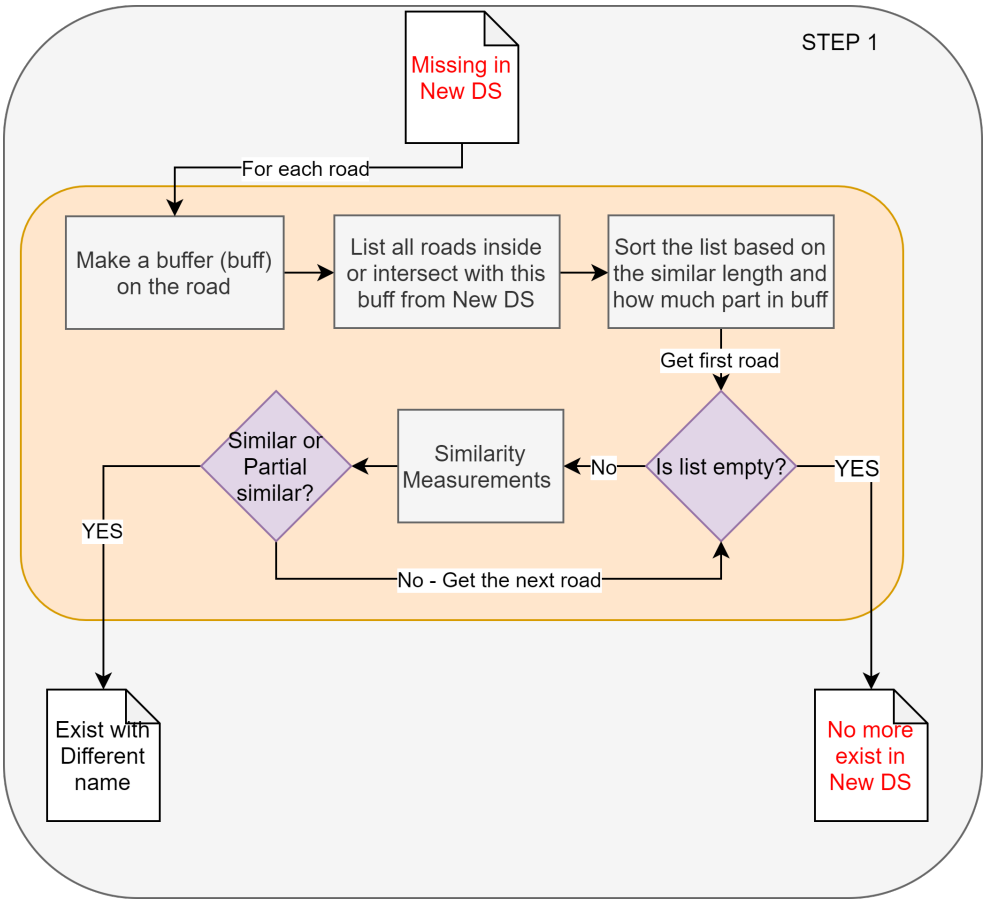


Figure 4.7: Check missing roads names if they exist with different names

the roads are partially similar, or the roads are totally different. If the results of the similarity measurement is either partially similar or different, we took the different portion from the old dataset and call the process of step # 1 again on it to see if it exists in new dataset with different name. It worths to mention that similarity measurements keep track for all dissimilar roads points and its corresponding distance.

After STEP # 2 has finished, we have six lists that are identifying the results of the historical road map comparison process which are: 1) the list of the roads that are similar and there is no changes have been happened to them, 2) the list of the roads that are similar spatially but their names have been changed in new dataset, 3) the list of the roads that are partially similar, and 4) the list of the roads that are

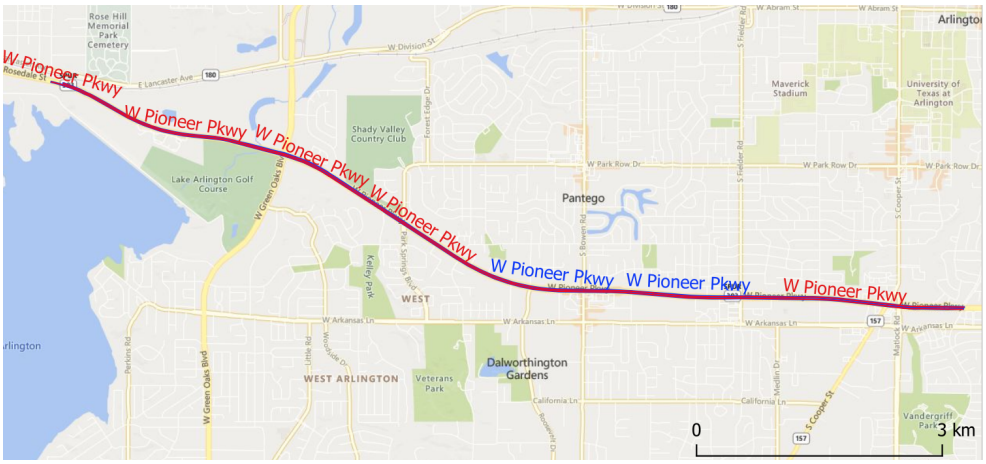


Figure 4.8: Similar Candidate Road

not similar which indicates that the road has been shifted in new dataset, 5) the list of roads that were exist in old dataset and removed in new dataset, 6) and the list of new roads that were captured after old dataset has been created.

4.4.3 Possible Types of Road Similarity Matching

This section is created mainly to highlights the types of partial similar matching as our TAREEQ framework has the ability to differentiate them. Below are the type of road similarity matching:

4.4.3.1 Similar Candidate Roads

It means that the candidate roads are similar and they have the same length and the same shape and located in same position in both datasets. For example, 2007 dataset and 2018 dataset has similar road that is called ”W Pioneer Pkwy”. After running the similarity measurements, the process gave us a decision that the road is similar in both datasets and the figure 4.8 shows that they are indeed similar.

4.4.3.2 Partial Similar Candidate Roads

This means one of the candidate road has one or more road's segments missing or extra when it compares with its candidate pair. There are three types of partial similar candidate roads: 1) Partial similar due to Road Extension, 2) Partial similar due to Road Expansion, and 3) Partial similar due to new road is built in different area.

1. Partial similar due to Road Extension: in this case the old road is gotten extension to serve new area. This makes the new dataset has same road and similar with the old part of the road. The new road segments that built after the old dataset was created do not have match pair in the old dataset such as road 'TX-360' that is shown in figure 4.9. In order to know how we distinguish this type of partial similar, i.e. Road Extension, the framework is looking for gradually increasing in the distance between points consisting the new extension- starting from the coordinates that close to the old part- and the corresponding nearest point in the old dataset. Figure 4.10 shows how the TAREEQ framework know the partial similar case is because of Road Extension. Road Shortening is a special case of Road Extension and TAREEQ framework can identifying this except the extension happened in the road stored in the old dataset. This mechanism of finding the the road extension is using directional Hausdorff distance.

2. Partial similar due to new road is built in different area: sometimes new area has been developed and some of its new roads are gotten names same as existing road in different area. Therefore, it may happens that there are number of different roads in different areas have same name such as the figure 4.11 that shows road name 'Carol Way' in two different area; the old one in Euless city, which is avail-

Figure 4.9: Road TX-360 has exension in 2018 while this part was not there in 2007 dataset

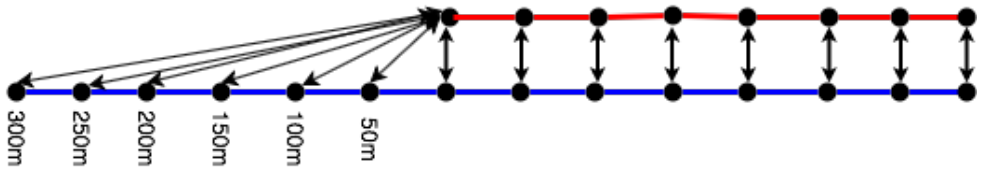


Figure 4.10: The distances between new extension and old road are increasing gradually

able in old dataset and new dataset, and the new one in Newark city that is captured only by new dataset. This case of partial similar roads can be determined when the framework measure the distance between the nearest coordinates from the old road in old dataset to the nearest coordinates for the new road in new dataset and the gap and we define the gap if it is more than 500 meters we consider the new road is a new road in different area. This is the different between road extension and new road in different area. Road Extension the distance is gradually increasing between the old road and its new segments while the new road the distance is very large between the nearest coordinates from new segment to the old road. Figure 4.12 shows there is gap larger than 500 meters between the new road and old road. This case is happening vice versa but we conclude that there is removed road that was exist in old dataset but no more exist in new dataset.

3. Partial similar due to Road Expansion: usually road expansion is happening when there is a need to expand the old road in order to overcome high traffic in this road. Also, this change is captured in most of the road maps by representing the road by two lines running in parallel to each other such as the case in road 'Academy Blvd' that is shown in figure 4.13. Our framework can detect this type of partial similarity. This can be detected by the framework through compare the distances and if there are number of consecutive coordinates have distance bigger

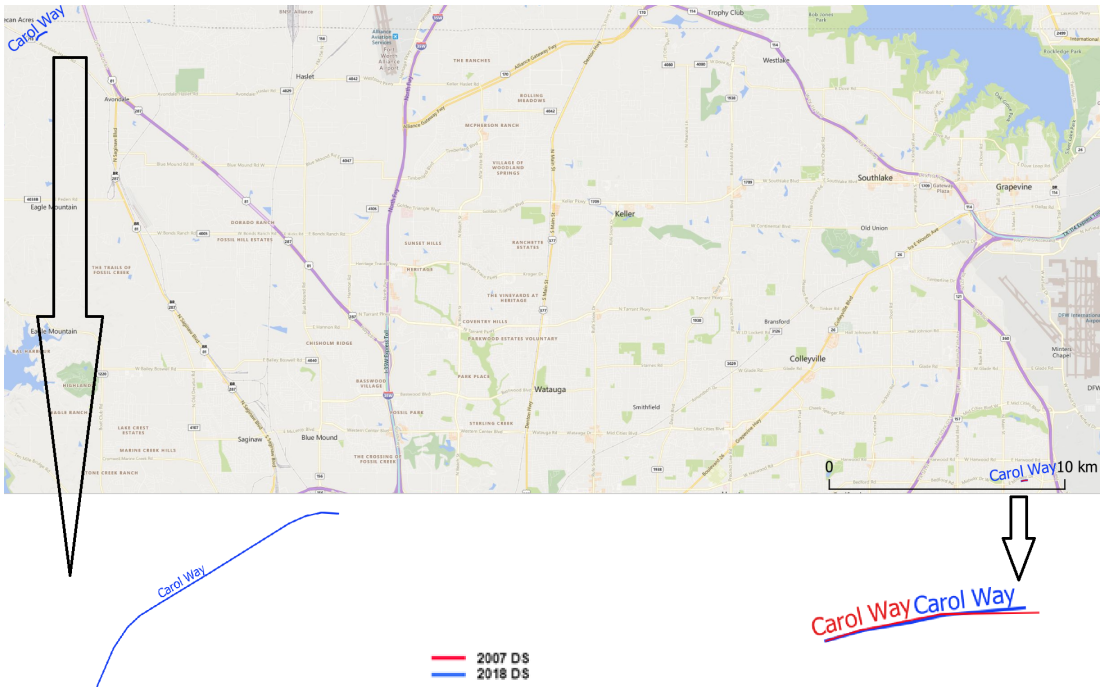


Figure 4.11: New road is built (left) and its name is exist in different area (right)

than the threshold and less than or equal double value of the threshold as shown in figure 4.14. If such case exists that indicates the partial similarity between the two roads is because of the Road Expansion. There are cases that our framework cannot detect the road expansion such if the new expansion road is placed with distance larger that two times of threshold value like in some highway roads or when the old road has removed and its location becomes in the middle of the new two-line road.

4.4.3.3 Candidate Roads are not Similar

When the system try to match the candidate roads and the similarity measurement return the results back that they are not similar and there is no another road with different name has similar matching, we conclude that these two candidate similar roads are not similar. This means the road has been shifted in new dataset and spatial characteristics are not any more the same as before. Such cases usually

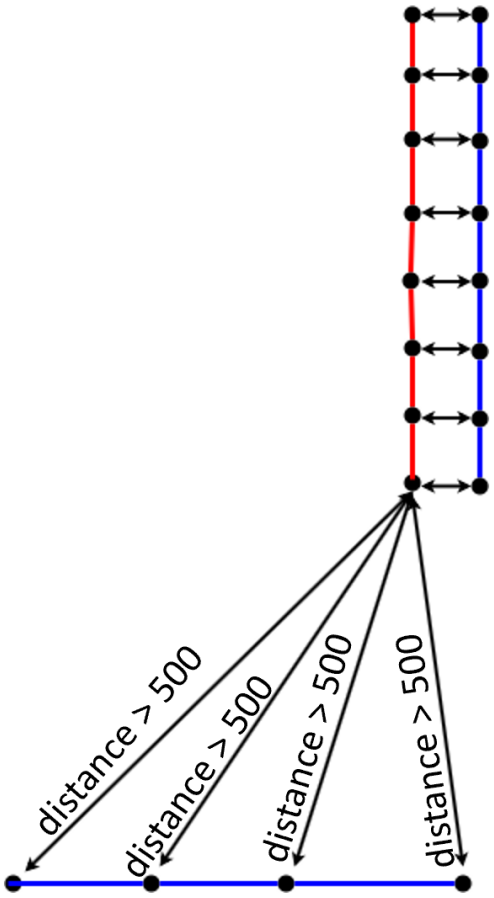


Figure 4.12: New Road in different area Explanation

happens in new area when they remove existing road for the sake of new residential or commercial projects. Figure 4.15 shows the following roads are shifted: Cancun Dr, Sail Fish Dr, and Bertram Dr.

4.5 Experimental Evaluation

We run our experiment in two real-world dataset to test and find the outcomes from these experiments. We choose Tarrant County as it is growing county and there is growth in number of areas inside it and we can have an overlook of the growth trends. We discuss its results in section 4.5.1. Then we have another real-world

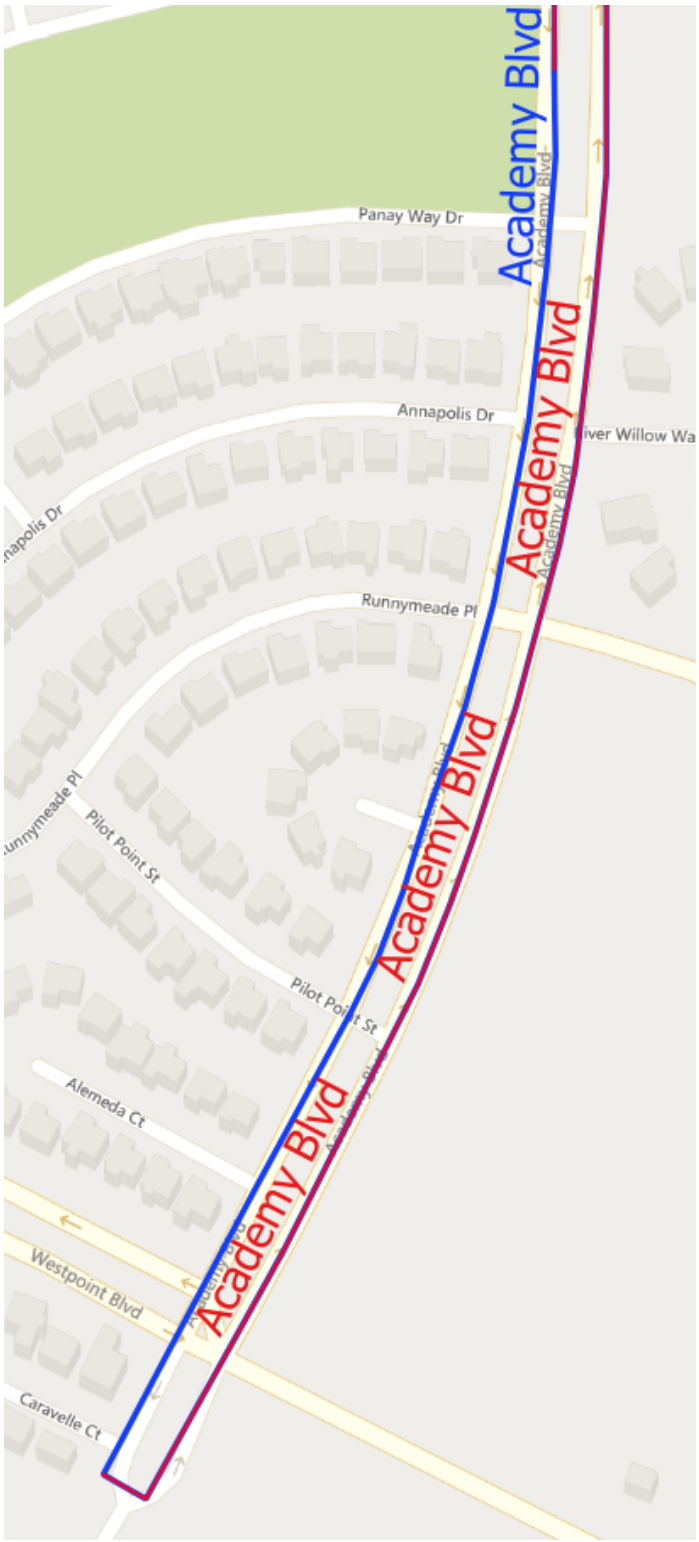


Figure 4.13: Academy Blvd is gotten Road Expansion

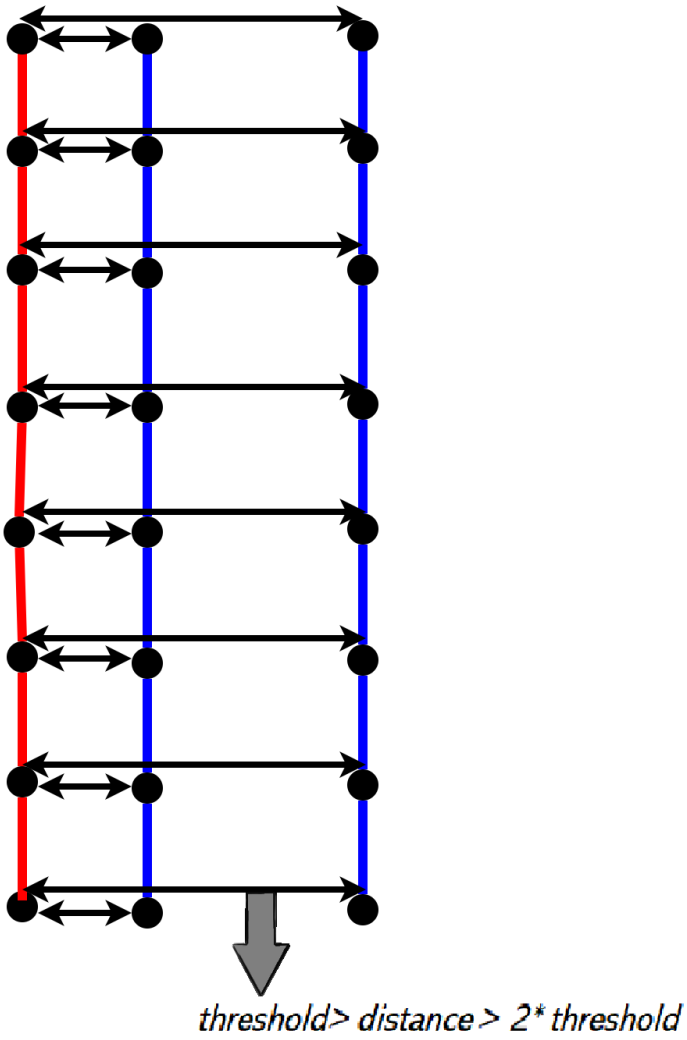


Figure 4.14: The expansion segments run in parallel with the old road

dataset which is about New York County and the reason we take this data is because most of the map matching works take New York County as an experiment data. Therefore, we discuss our results with this data in 4.5.2. After that, in section 4.5.3 we discuss how good is our framework doing in terms of the accuracy and quality of the output and we discuss efficiency of our framework.

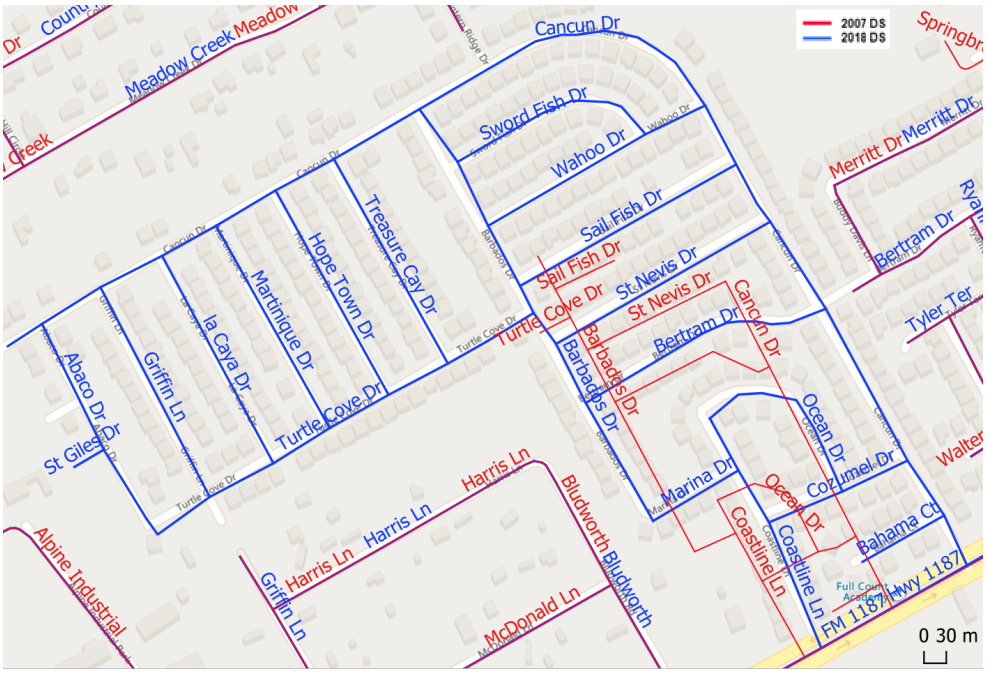


Figure 4.15: Candidate Roads are not Similar and Shifted from their old location

Table 4.1: Tarrant County Datasets basic information

	Dataset # 1	Dataset # 2
Year	2007	2018
Number of Roads	19,007	24,144
Number of Records	104,920	40,671

4.5.1 Tarrant County Experiment

Tarrant County is one of the biggest counties in Texas and it spans in area of 2,238 km^2 [46]. It has at least 20 cities. This area has grown over last 11 years and we would like to study the growth for this area. We have two datasets for Tarrant County from TIGER- Topologically Integrated Geographic Encoding and Referencing that is generated by United States Census Bureau- as a source: 1) 2007 Tarrant County Dataset; and 2) 2018 Tarrant County Dataset. For phase # 1, which is data sources preparation, we have 2007 Datasets for the Tarrant County and this dataset has a lot of unrelated data to the roads. There are Hydrography, Rails, Roads, and so on.

Therefore, we need to prepare this dataset and filter all data that is not related to the roads in order to avoid any confusion could happen from it. After we filter the 2007 dataset we come up with total records 104,920 and these records represent only 19,007 Road names. The other dataset is the 2018 Tarrant County dataset and this new dataset has the most recent update Road Map of Tarrant County. This new dataset has the road data only so it does not need any further action in preparation process. It has 40,671 records that represent 24,144. Table 4.1 shows these number to compare the number for both datasets. As we notice, 2007 dataset has more records to represent less number of roads comparing to 2018 dataset which means the efficiency of memory usage in 2018 dataset is much better than 2007 dataset.Even though the number of records are different between 2007 and 2018 dataset which means the representation may differ than each other, our TAREEQ framework can handle this difference. In the following subsections, we are going to discuss the results of our framework after we run it for Tarrant datasets in 4.5.1.1. Then we are going to discuss how the county was growing in overall pictures then based on regions in 4.5.1.2. After that, we conclude with some highlighted samples of this historical comparison in 4.5.1.3.

4.5.1.1 TAREEQ framework results for Tarrant County

After we run phase # 1 that is the data sources preparation to 2007 dataset and make the dataset is ready and check 2018 dataset and it was ready by itself, we run phase # 2 which is the historical comparison we got the following results as shown in figure 4.16. After we run the initial semantic attribute filtration which is the process of "Generate Candidates", we got 768 road names were exist in 2007 DS and no more exist in 2018 DS. In addition, There are 5,905 new road names show up in 2018 that were not exist in 2007 DS. The remaining road names, which are 18,239 raod names,

are exist in both datasets. This step is very important to speed up the performance for TAREEQ framework. These results are shown as the outputs of "Road Name Matching" process in figure 4.16.

The idea for this historical matching is to go through all old dataset's elements to know if they have matching pair in new dataset. At the end the remaining elements of the new dataset are considered as brand new elements never exist in old dataset. Therefore, we start first with the roads that exist in 2007 and are missing in 2018 and step in STEP # 1. The STEP # 1 process has been explained in section 4.4.2. As mentioned before, this STEP use another type of "Generate Candidate" process which is depending on spatial attribute instead of semantic attribute and then run the "similarity measurements" process on the candidates similar roads to get the results if the missing road exists in new dataset or it is simply removed from the new dataset. Therefore, there are 234 roads are actually not exist any more in 2018 dataset and there are 534 roads are exist in 2018 but with different names. It worths to mention that most of these roads names - 497 road names- come from the list that has new road names in 2018 dataset and they are not in 2007 dataset. Therefore, the list of "Missing in 2007 DS" is reduced by 497 road names.

In STEP # 2, the framework takes the list that has the road names that exit in both datasets, i.e. 2007 dataset and 2018 dataset, and run the "Similarity Measurements" process on them. After we get the initial results, we run the processes in STEP # 1 for each road or segments of a road in 2007 dataset that is not similar with its pair from 2018 dataset with same road name. The later process is important to verify if the road's name was changed or the road has been shifted. Of course after this step the list of "Missing in 2007 DS" will be decreasing if the framework find different road names and after it is running we found 277 new road names from the

list. We get the following results: 69 roads are not similar which means the roads have been shifted; 16,407 roads are similar; and 1,763 roads are partially similar.

After finishing the step # 2, all the remaining road names in the list "Missing in 2007 DS" are considered as new roads that never exist in 2007, which has 5,131 new road names.

As we discuss in section 4.4.3, the partial similar roads has three types that our framework can determine. Based on that and for Tarrant County experiment, we found that there are 586 roads that have missing (Removed) segments and then the new road corresponding to it gets shrinking. On the other hand, there are 605 roads have extended their length to serve new areas.Also, there are 168 whole roads in different area have been completely removed while there are 277 new roads have been built in different areas. The framework finds there are 127 roads have gotten expansion and become two-line representation instead of one-line. Figure 4.17 summarize the findings. It worths to mention that some roads has new extensions and new road segments and expansion, too. However, to avoid the confusion and for the sake of simplicity, we prioritize them as the following: First name the type New/Missing roads if there is new or missing complete roads, second name the type expansion if there is no New/Missing roads and the road gets expansion, finally we name it missing segment or extension if the road has no New/Missing roads nor expansion. We believe that most important is to find New/Missing roads; then find the expansion as they are not a lot.

4.5.1.2 Determine the trends of county growths

Tarrant County is growing county and between year 2007 to 2018 the area has changed and this change can be reflected on the road map. We have done this comparison based on the road's length and not on the number of the road for two

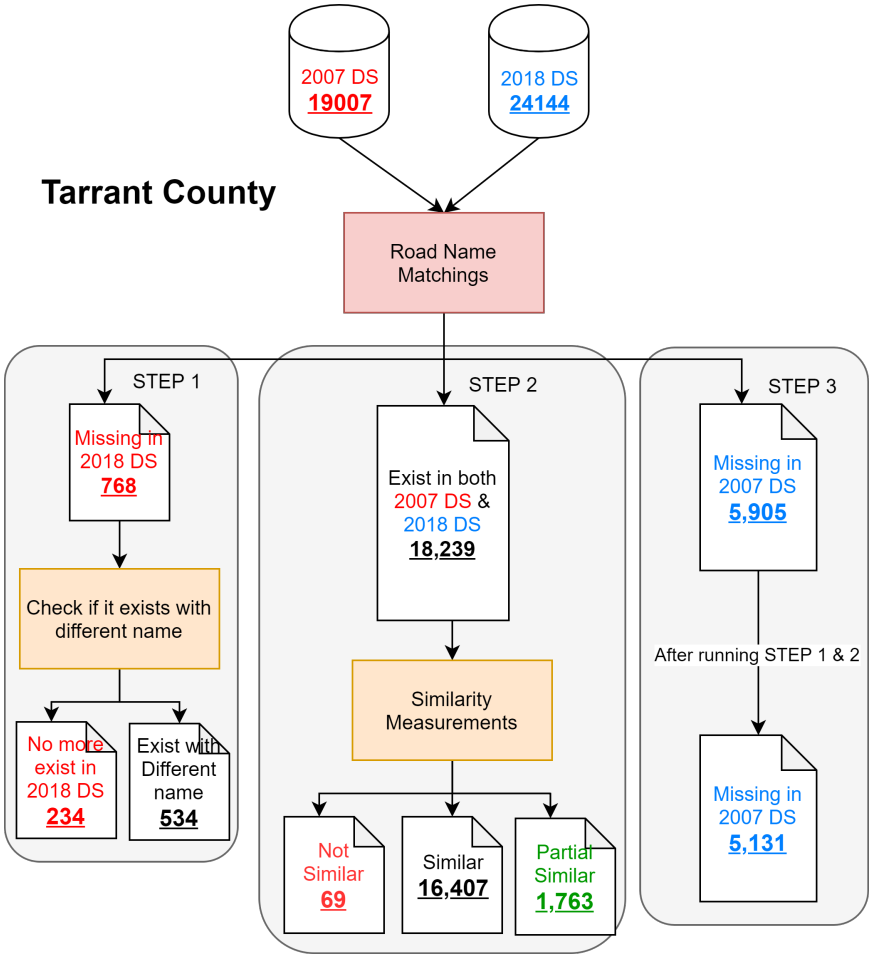


Figure 4.16: Historiacal comparison results for Tarrant County datasets (2007 DS and 2018 DS)

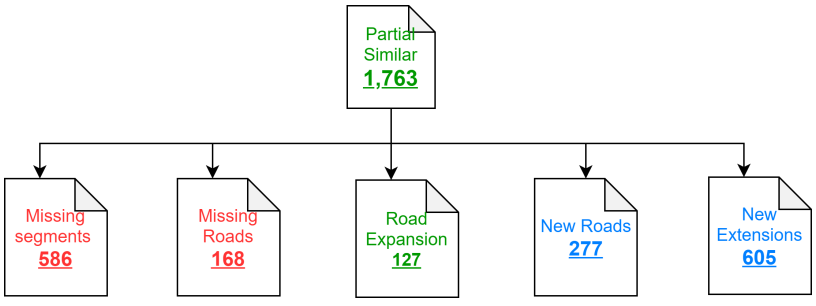


Figure 4.17: Types of partial similarity roads that are found in Tarrant County experiment

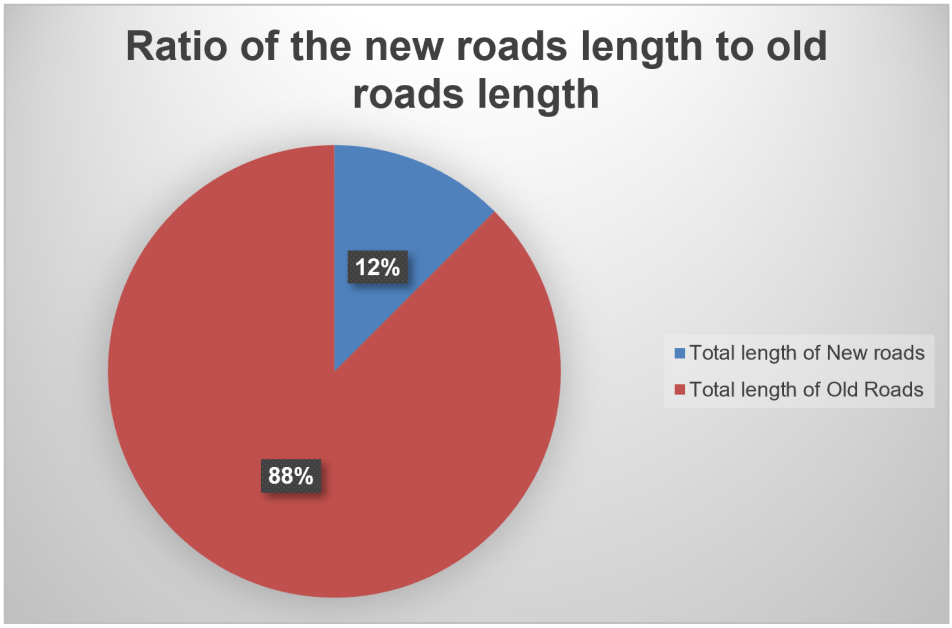
reason: 1) some new roads have small length and others has very large length and if we count by the number, we will make all roads have same length. 2) we cannot calculate by the number of new roads due to partial similarity roads could miss up the results. We start with overall picture of how much the growth of new roads within the period of 11 years. Figure 4.18 shows that overall new roads length represents 12% of the total Road's length and the remaining percentage is for the roads' length in 2007.

Now We would like to divide Tarrant County into regions. We have manually divide the Tarrant County into nine regions that can be named by directions as shown in figure 4.19. After that, we compute the length of the following for each region: new roads and new extensions, old roads, removed roads or road segments, and total current roads. Based on these information we can know the growth in each region and percentage of growth in each region and so on.

Figure 4.20 shows that the north region has major growth comparing to the total number of the current roads' length. After that northwest region then southwest region.

Figure 4.21 and figure 4.22 shows last information in different representation for each region. These two figures give a look of how much represent new roads length in the specific region and how much the old roads' length represent from the current road's length in terms of percentage and roads' length respectively.

Even though the percentage of growth could be high but this does not mean this region gets the major of growth overall Tarrant County. To simplify the idea, Figure 4.23 shows how much the new roads in each region represent from the total new roads all over Tarrant County. Notice the northwest region gets the second region in terms of the growth comparing to existing roads in region. However, these new roads in



Total length of New roads	3015.307049	Total Roads Length	24123.43262
Total length of Old Roads	21108.12557		

Figure 4.18: Ratio of the New Roads' Length to the Old Roads' Length

northwest region represent only 8.1% of total new roads and 6 regions has number to new roads' length greater than northwest region.

In Figure 4.24, we can see both numbers- i.e. percentage of new roads' length to the total roads' length in each region and percentage of new roads' length to the total new roads all over Tarrant County- in one chart to make the comparison.

Now we see how much of removed roads' length represent in each region. In figure 4.25, we can see the removed roads in northwest region represents the highest percentage among all regions when we compare the removed roads' length to the total roads' length inside the region. After that north east region then north region. Figure, 4.26 and figure 4.27 shows the percentage and length of removed roads respectively to the total roads inside each region. When we compare the total removed roads inside each region to the total removed roads all over the Tarrant County, the northeast

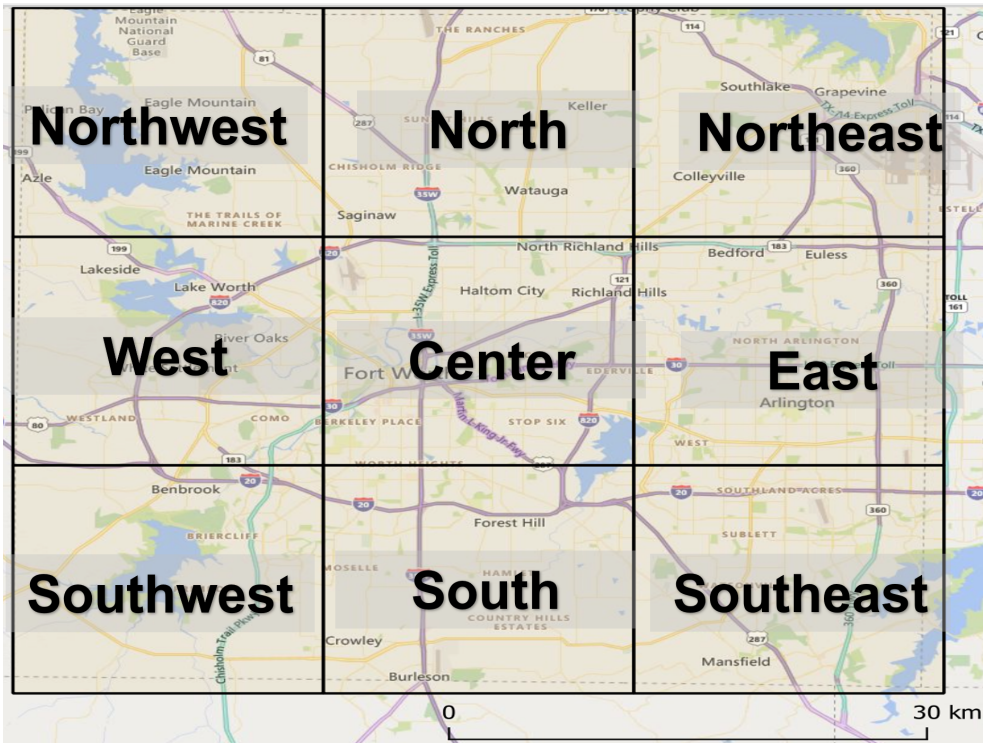


Figure 4.19: Using Grid to divide Tarrant County into 9 regions

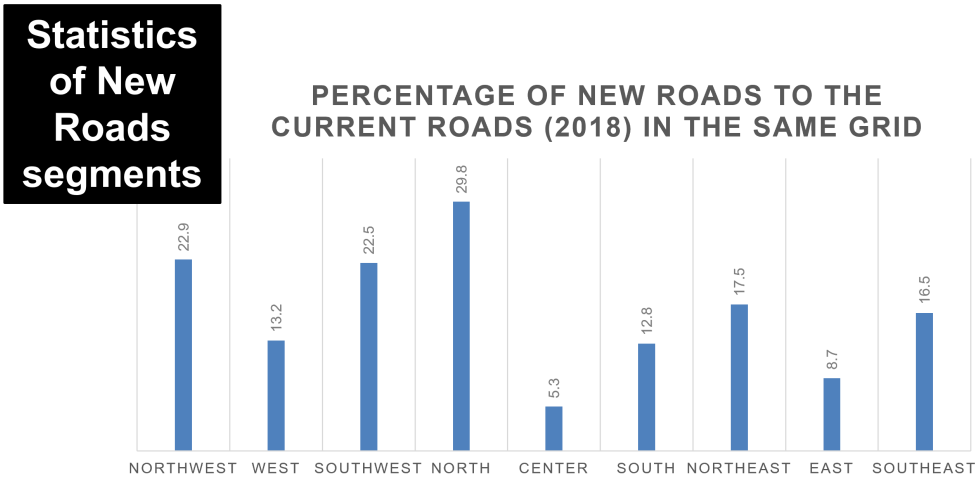


Figure 4.20: Percentage of New Roads Length to the Total roads Length inside each region

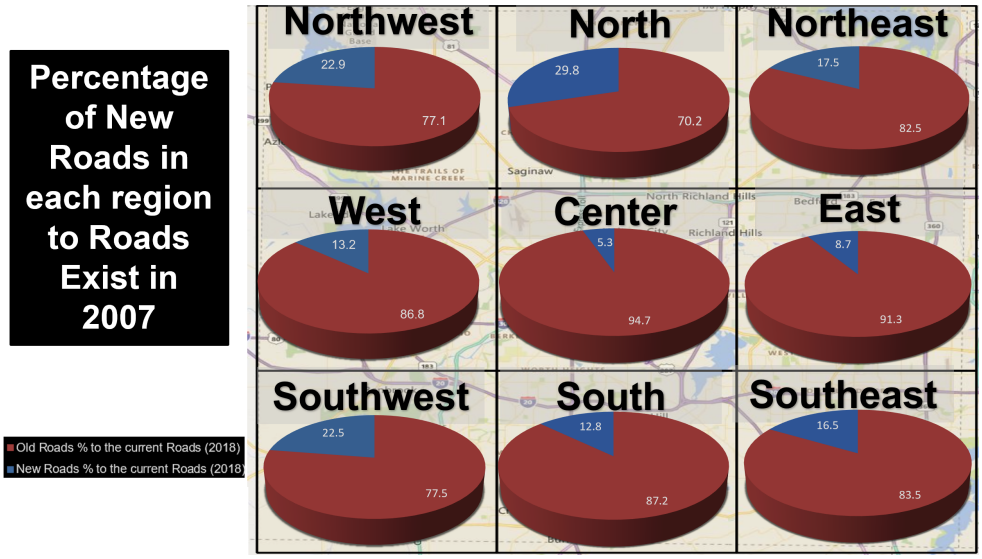


Figure 4.21: Percenatge of new roads length and old roads length for each region

region got the most removed roads compared to other regions same as information depicted on 4.28. For the sake of comparison, figure 4.29 compares percentage of removed roads' length in each region to the total current roads length Vs. percentage of removed Roads' length to the total removed Roads' Length.

Finally, we would like to see if there is relationship between the new roads and removed roads in a region or not. It turns out there is no relationship between these two factors as shown in figure 4.30.

4.5.1.3 Highlight samples of the matching results

As our framework produces the results, we have the ability to draw the controversial cases that let us pay attentions to them. Some these cases are reflecting the real-world but some of them raise a concern if the new dataset is correct or not. Our framework can help on such cases by providing exact locations of that have the differences between old dataset and new dataset.

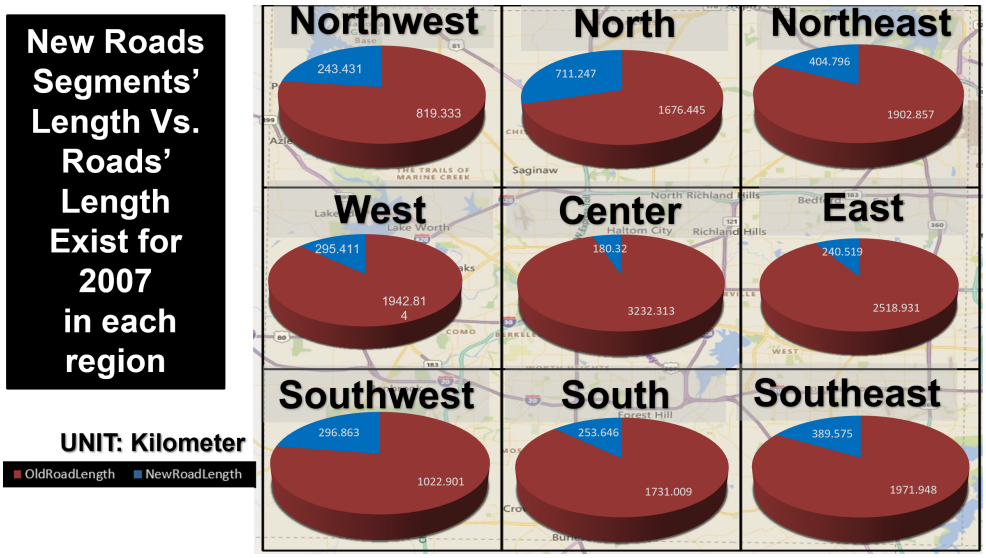


Figure 4.22: New roads' length and old roads' length for each region

Case 1. Removed roads in 2018 dataset that were exist in 2007 dataset. During 11 years period, there were roads that were available and nowadays those roads are not exist anymore for any reason. Our sample is for AT&T stadium- a.k.a. Cowboys Stadium-. The area of the stadium was residential area and there were number of local roads like "Ivy Ln" and "Vine St". After the construction project finished, those local roads are not exist any more and there are new roads are built in different places than the old roads such as "Cowboys Way" and "AT&T Way" as shown in figure 4.31.

Case 2. New Roads are emerged in new dataset that are not available in old dataset. North region has the highest growth for all over Tarrant County, which means there are a lot of new roads are constructed. Figure 4.30

Case 3. One of the two-line Representation for I-820 in 2018 is missing but Exist in 2007. This case shows that there is an issue in part of highway road called "I-820". After doing the "similarity measurements" process, the frame-

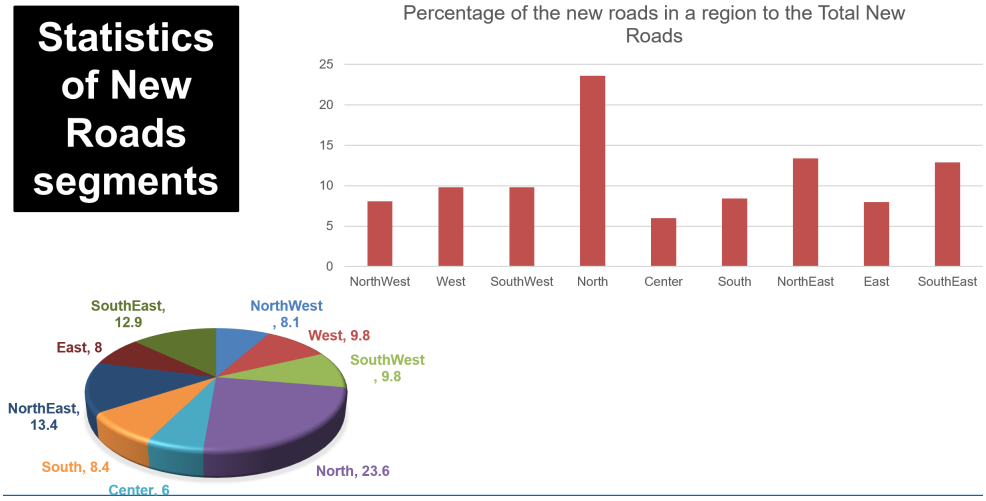


Figure 4.23: Percentage of New Roads in each region to the Total New Roads

work identify that part of I-820 roads has one-line representation instead of two-line representation for the whole road as shown in figure 4.33. The framework identifies this difference after it starts measuring the distance from 2007 DS coordinates for the road to nearest point from the road in 2018 DS. The framework shows that number of coordinates have distances greater than threshold. Figure 4.34 shows the distances that have values greater than threshold. As this part of the road is missing from 2018 DS, then our framework call STEP # 1 to check if it exists with different name or not. It turns out it exist with different name called "Northeast Loop" as shown in figure 4.35

Case 4. Missing Road in 2018 DS while it exists in 2007 and in real-world. This is one of the cases that the road has removed from new dataset while it exist in real-world. TAREEQ framework helps to identify the missing roads in new dataset to verify if they are really removed in the real-world or still exist. Figure 4.36 shows the road "Kaitlyn Ct" is available in 2007 DS and in real-world. However, it is missing in 2018 DS.

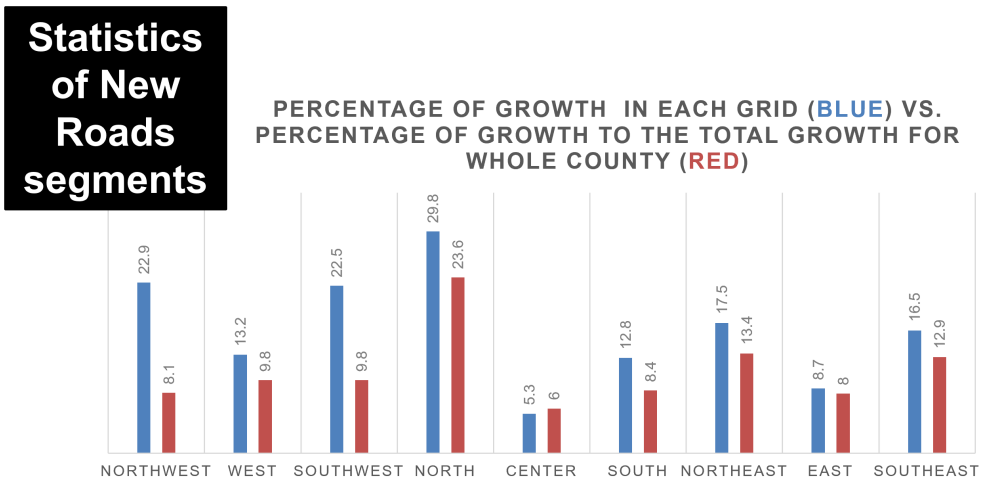


Figure 4.24: Percentage of new roads' length to the total roads' length in each region and percentage of new roads' length to the total new roads all over Tarrant County

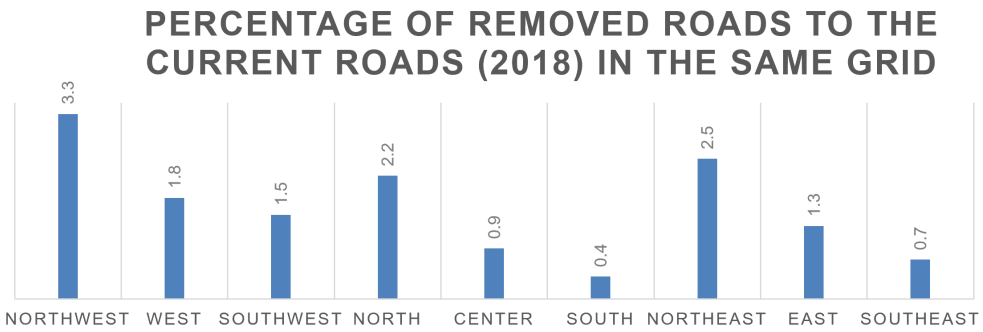


Figure 4.25: Ratio of Removed Roads' length to the total roads' length inside the region

Case 5. Possibly misspelling name or wrong direction in new datasets.

The results show there are cases where there is high potential of misspelling road names such as the case in figure 4.37. The road "Calender Rd" in 2007 DS has two names in 2018 which are: "Calender Rd" and "Callender Rd". Or the direction in the name could be wrong. The example that is shown in figure 4.38 has this type of error where 'S Shadycreek Dr' in 2018 DS is named on Shadycreek Dr that is in the north side. The direction in 2007 DS is correct.

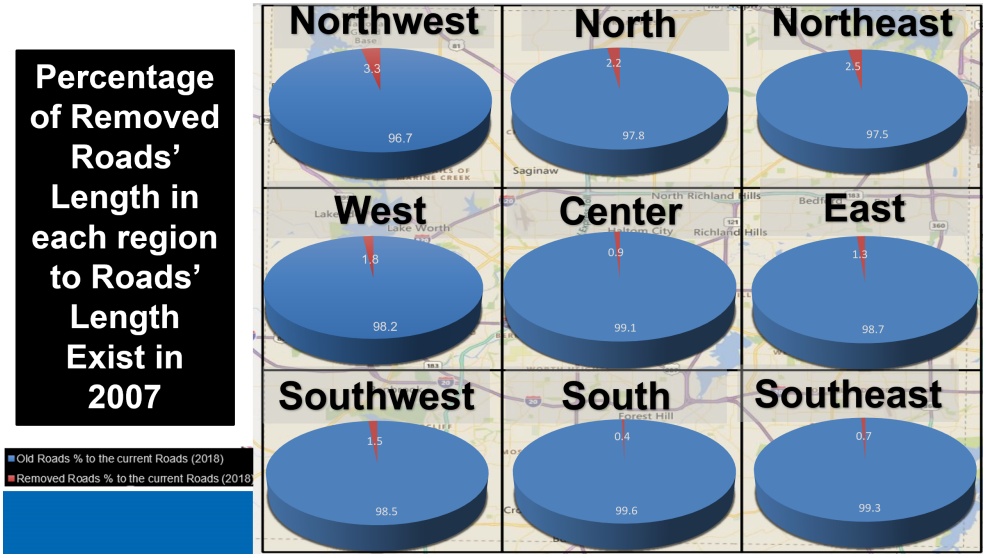


Figure 4.26: Ratio of Removed Roads' length and old roads' length to the total roads' length inside the region

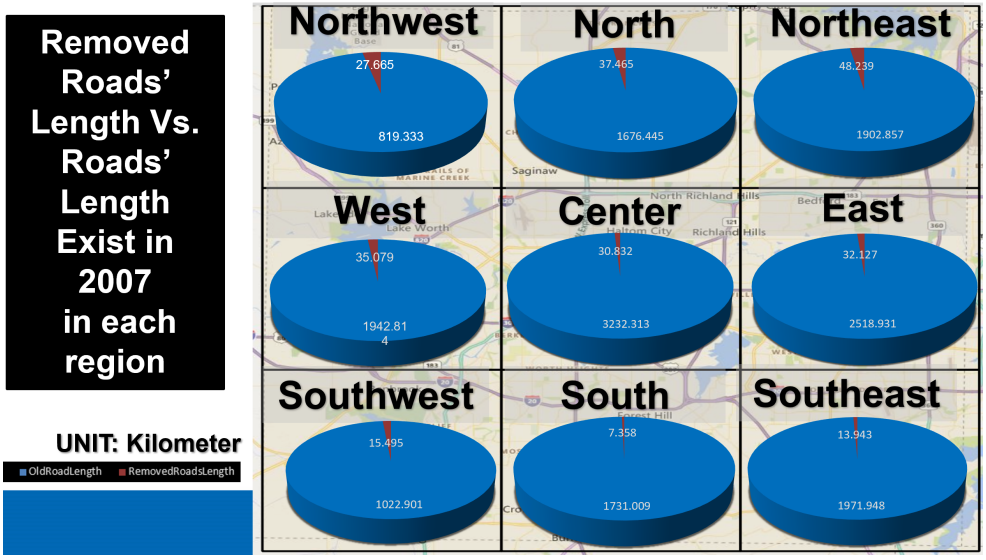


Figure 4.27: Total length of removed roads to the total length of roads in old dataset inside each region

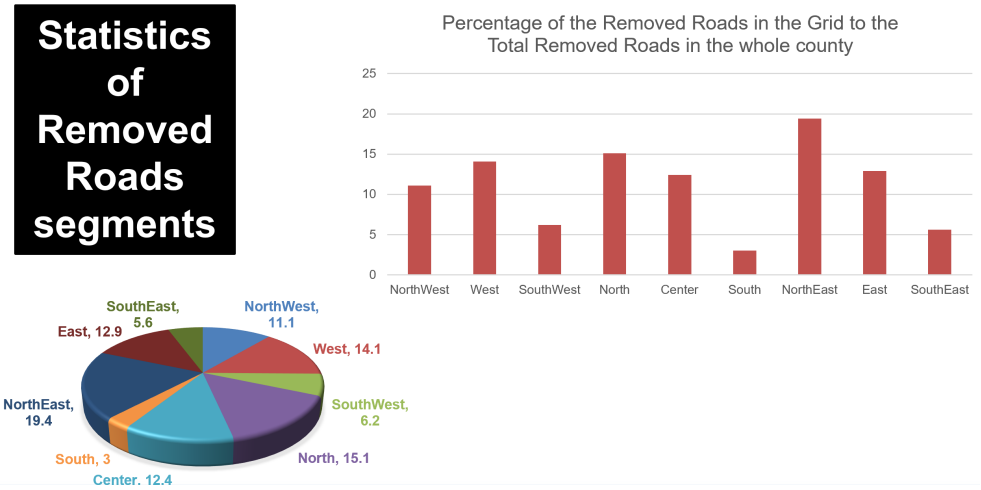


Figure 4.28: Ration of the removed road to the totlal length of removed roads all over the Tarrant County

Table 4.2: New York County Datasets basic information

	Dataset # 1	Dataset # 2
Year	2007	2018
Number of Roads	942	943
Number of Records	16,109	2,058

4.5.2 New York County Experiment

The reason we take New York dataset in the experiment is because most of the works in Road Map Matching took New York County as test data. Therefore, we discuss our results with this dataset and see what is the outcomes. New York county is much less than Tarrant County. Its area is only 87 km^2 [47] compare to 2,238 km^2 [46] for Tarrant County. We have two datasets for New York County from TIGER- Topologically Integrated Geographic Encoding and Referencing that is generated by United States Census Bureau- as a source: 1) 2007 New York County Dataset; and 2) 2018 New York County Dataset. For phase # 1, which is data sources preparation, we have 2007 Datasets for the New York County and this dataset has a lot of unrelated data to the roads. There are Subways, Hydrography, Rails, Roads,

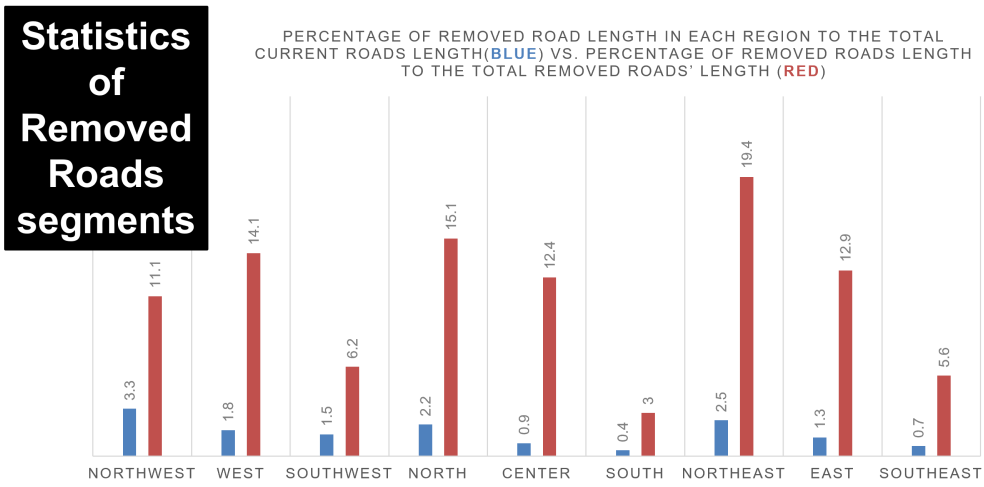


Figure 4.29: Percentage of Removed road length in each region to the total current roads length Vs. Percentage of Removed Roads length to the total Removed Roads' Length

and so on. Therefore, we need to prepare this datasets and filter all data that is not related to the roads in order to avoid any confusion could happen from them. After we filter the 2007 dataset we come up with total records 16,109 and these record represent only 942 Road names. The other dataset is the 2018 New York County dataset and this new dataset has the most recent update Road Map of Tarrant County. This new dataset has the road data only so it does not need any further action in preparation process. It has 2,058 records that represent 943. Table 4.2 shows these number to compare the number for both datasets. As we notice, 2007 dataset has more records to represent less number of roads comparing to 2018 dataset which means the efficiency of memory usage in 2018 dataset is much better than 2007 dataset.

After conducting the experiment on these datasets, we got 366 road names were exist in 2007 DS and no more exist in 2018 DS. In addition, There are 367 new road names show up in 2018 that were not exist in 2007 DS. The remaining road names, which are 576 road names, are exist in both datasets. This step is very important to speed up the performance for TAREEQ framework. These results are shown as the

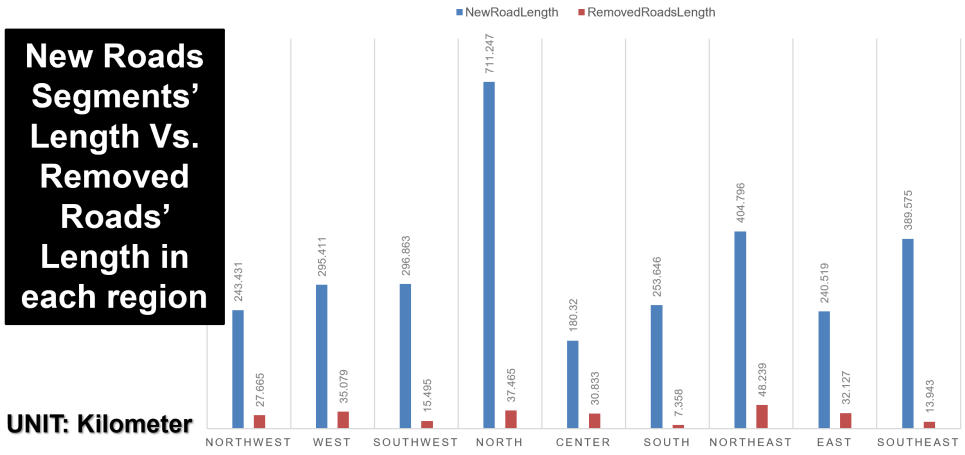


Figure 4.30: Comparison of total length of new roads to the total length of removed roads inside each region

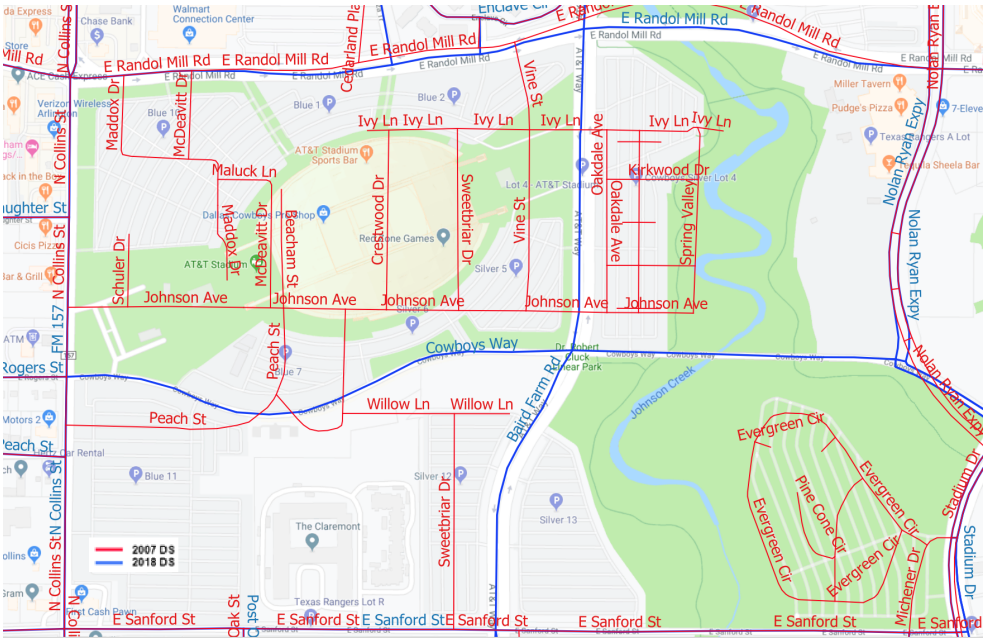


Figure 4.31: Old roads have been removed and new roads are emerged in AT&T Stadium area

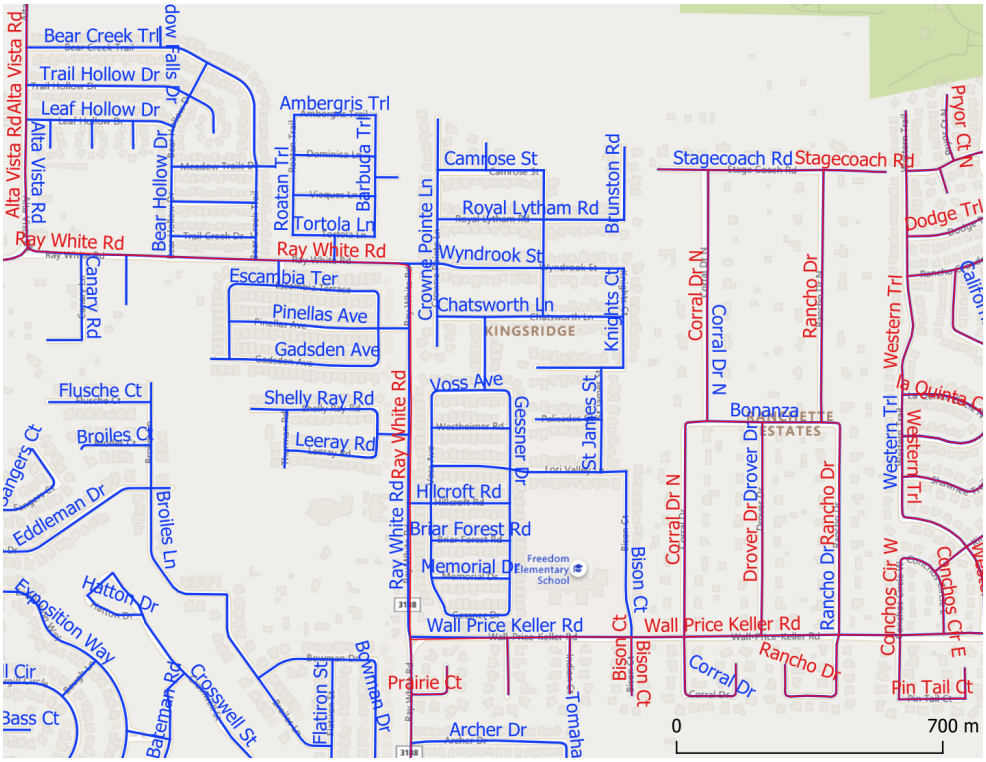


Figure 4.32: New roads are emerged on area that were not developed back on year 2007

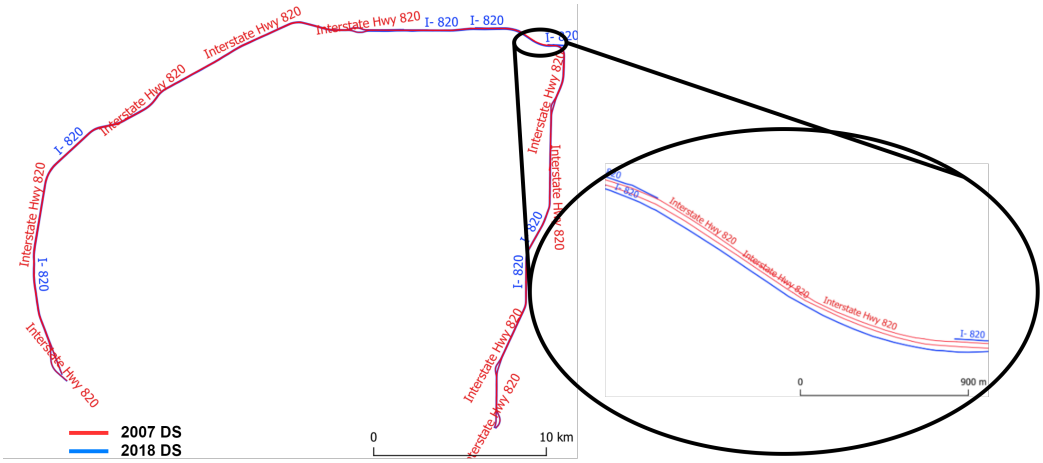
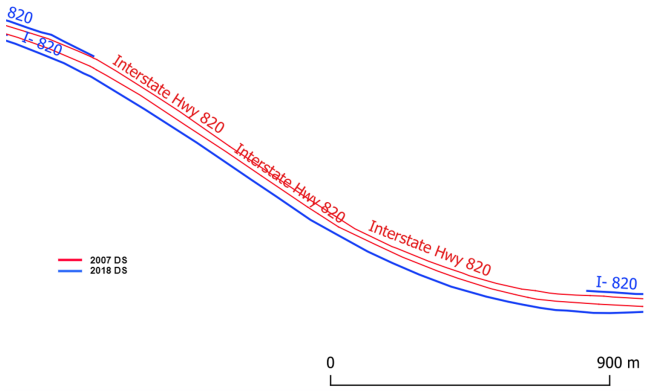


Figure 4.33: Part of road I-820 is missing



	Latitude	Longitude	Distance to nearest point in 2018 I-820 rd
838	32.72968	-97.225	0
839	32.81646	-97.4074	5.57E-10
840	32.81615	-97.4081	0
841	32.83496	-97.2209	42.76341502
842	32.83465	-97.2203	38.46939536
843	32.83422	-97.2192	40.87778862
844	32.83418	-97.2191	42.13721431
845	32.83387	-97.2182	43.48062576
846	32.83371	-97.2177	44.28577565
847	32.83358	-97.2172	44.30113374
848	32.83351	-97.217	42.24397574
849	32.83332	-97.2162	41.27557471
850	32.83319	-97.2155	40.85668919

Figure 4.34: Distances from 2007 DS points coordinates show number of coordinates have distances greater than threshold

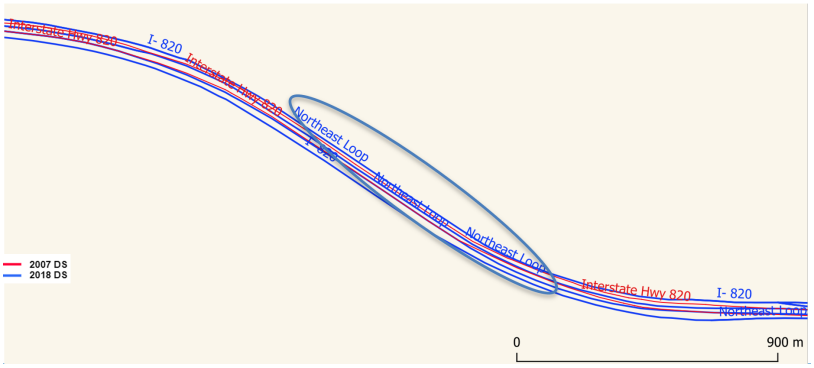


Figure 4.35: The missing part is exist with different name called "Northeast Loop"

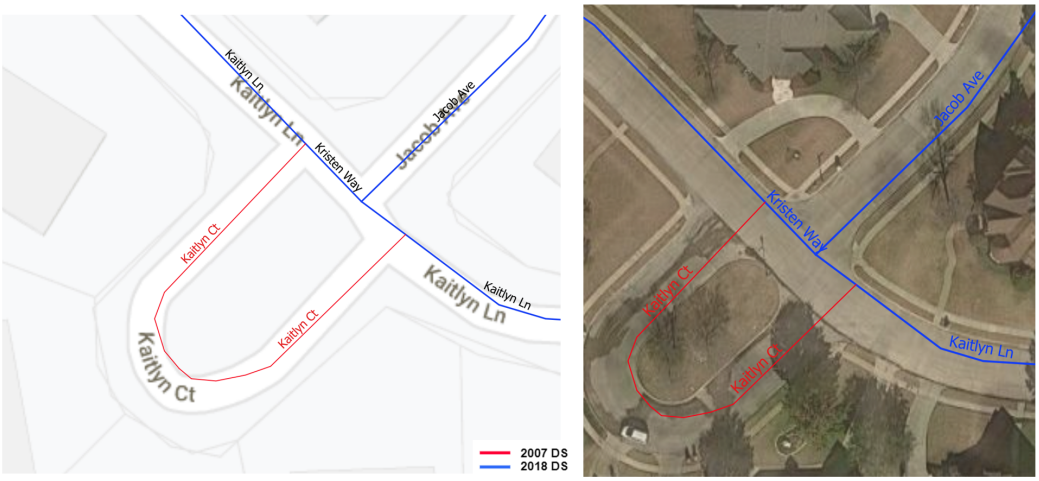


Figure 4.36: Road exist in 2007 DS and in real-life but it is missing in 2018

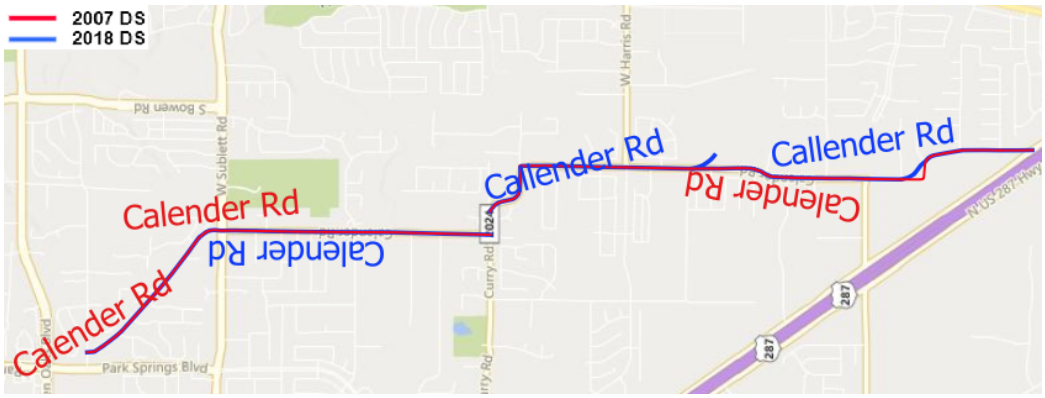


Figure 4.37: There are two names in 2018 DS for the road "Calender Rd" in 2007 DS

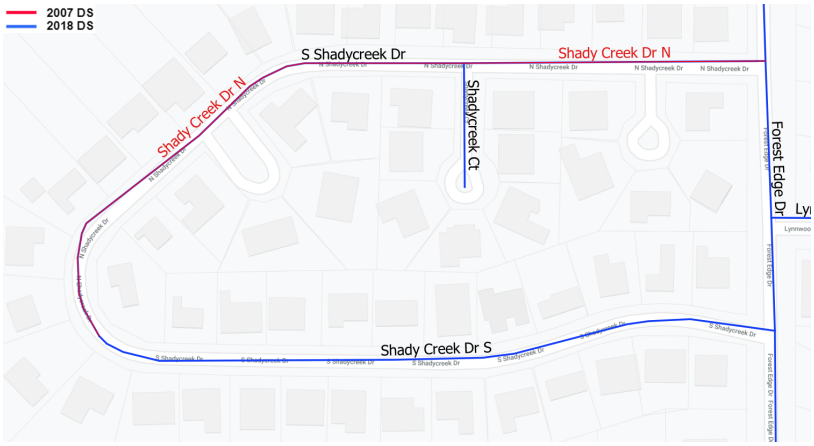


Figure 4.38: 'S Shadycreek Dr' in 2018 DS is named on Shadycreek Dr that is in the north side like in 2007 DS

outputs of "Road Name Matching" process in figure 4.39.

The idea for this historical matching is to go through all old dataset's elements to know if they have matching pair in new dataset. At the end the remaining elements of the new dataset are considered as brand new elements never exist in old dataset. Therefore, we start first with the roads that exist in 2007 and are missing in 2018 and step-in STEP # 1. The STEP # 1 process has been explained in section 4.4.2. As mentioned before, this STEP use another type of "Generate Candidate" process which is depending on spatial attribute instead of semantic attribute and then run

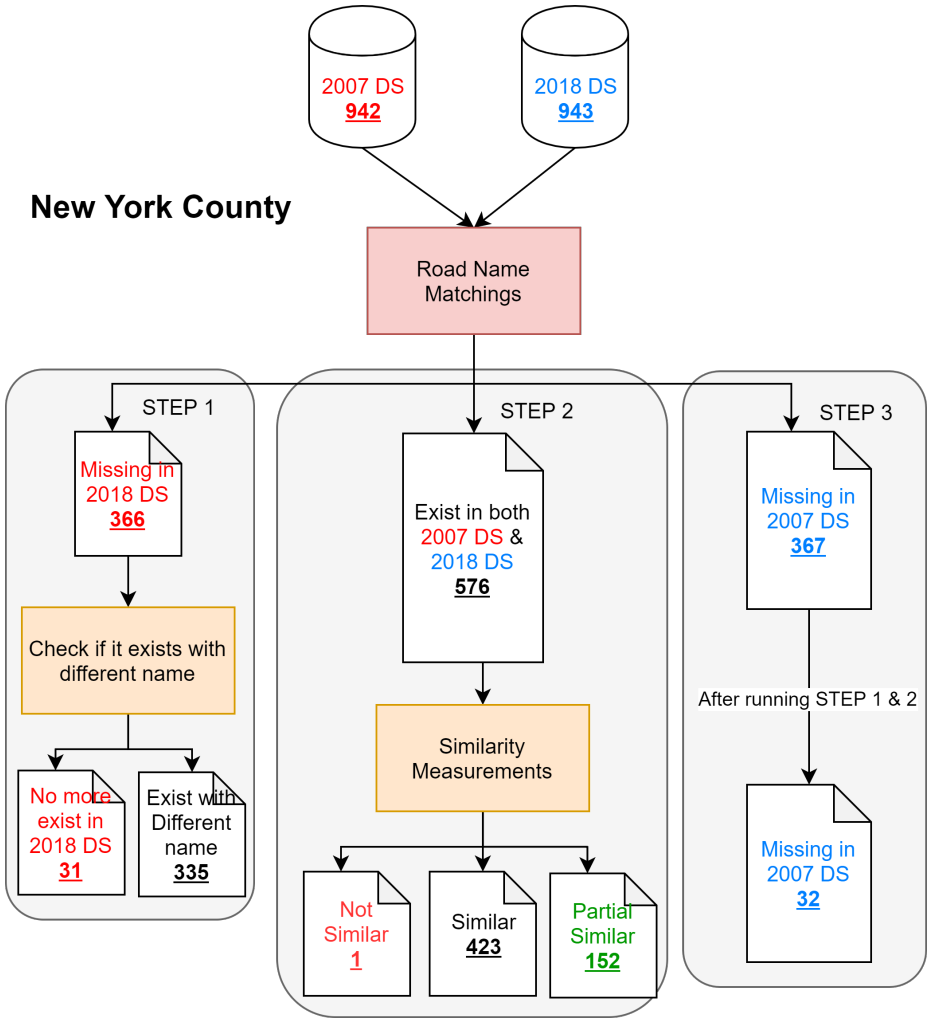


Figure 4.39: The final results of TAREEQ framework experiment on 2007 and 2018 New York datasets

the "similarity measurements" process on the candidates similar roads to get the results if the missing road exists in new dataset or it is simply removed from the new dataset. Therefore, there are 31 roads are actually not exist any more in 2018 dataset and there are 335 roads are exist in 2018 but with different names. It worths to mention that all those roads names - 335 road names- come from the list that has new road names in 2018 dataset and they are not in 2007 dataset. Therefore, the list

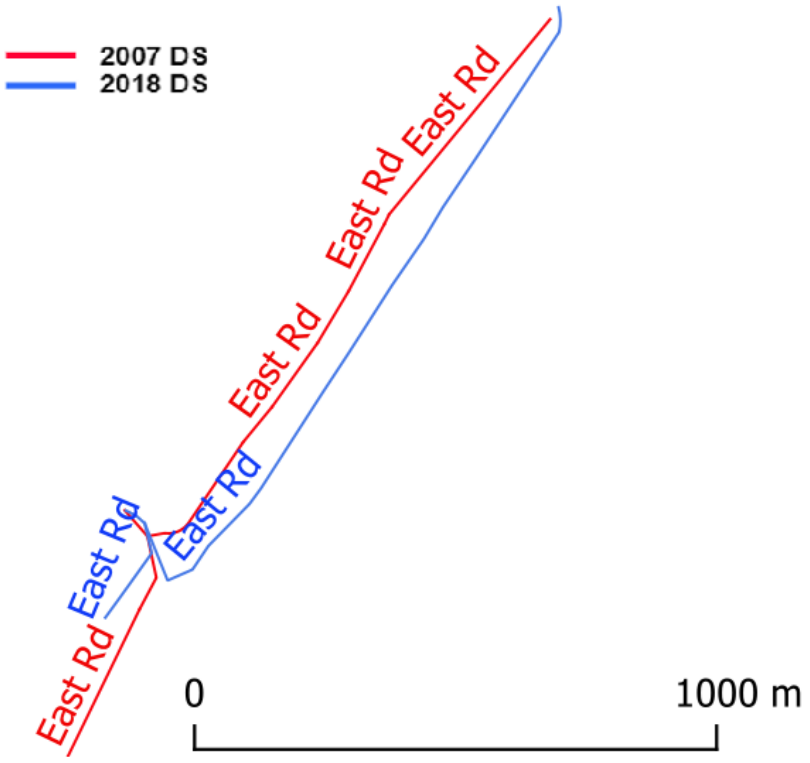


Figure 4.40: 'East Rd' is the only road that has been shifted since 2007

of "Missing in 2007 DS" is reduced by 335 road names.

In STEP # 2, the framework takes the list that has the road names that exit in both datasets, i.e. 2007 dataset and 2018 dataset, and run the "Similarity Measurements" process on them. After we get the initial results, we run the processes in STEP # 1 for each road or segments of a road in 2007 dataset that is not similar with its pair from 2018 dataset with same road name. We get the following results: 1 road is not similar which means the roads have been shifted as it shows in figure 4.40; 423 roads are similar; and 152 roads are partially similar.

After finishing the step # 2, all the remaining road names in the list "Missing in 2007 DS" are considered as new roads that never exist in 2007, which has 32 new road names.

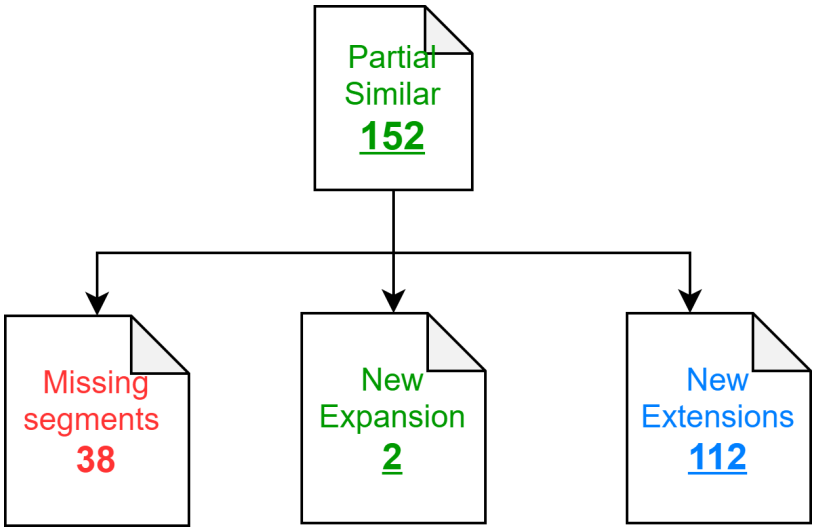


Figure 4.41: Partial similar roads types for New York County

For partial similar roads, we found that there are 38 roads that have missing segments and then the new road corresponding to it gets shrinking. On the other hand, there are 112 roads have extended their length to serve new areas. For new roads, there is no a whole road in different area that has been completely removed and same thing from 2018 dataset there is no new roads in different area. The framework finds that there are only two roads have gotten expansion and become two-line representation instead of one-line, which are F D R Dr and Henry Hudson Pkwy. Refer to the figure 4.41.

We are not talking about the growth trends as the number of new roads and extensions is spread all over the county and especially in the borders of the county. We would like to show two more cases, one where couple of roads have been removed from 2018 dataset that is shown in figure 4.42. The example of removed roads in new dataset where this road is still exist in real-world. We can see that 'Broadway Aly' in 4.43 is no longer available in 2018 DS while it exist in real-world.

Figure 4.42: Roads that are no longer exist in 2018 dataset

Figure 4.43: 'Broadway Aly' is missing in new dataset while it is still available in real-world

4.5.3 Evaluation of the TAREEQ framework (Quality and Performance)

In this section we would like to talk about the quality of the TAREEQ framework and how much is the accuracy that we can get and what is the method that we have use. The first topic is the quality and we discuss it in the following subsection 4.5.3.1. Then we talk about the performance of the framework and what is the overall idea about it in 4.5.3.2.

4.5.3.1 The Accuracy of TAREEQ framework

Most of the road matching studies work manually to evaluate their systems [48, 39, 49, 50, 51, 27, 52, 53]. This evaluation is good when you can manually handle the road map matching process. However, such as our experiment which is conducted on Tarrant County, it seems very difficult and it may takes months to finish the evaluations. Some research use automated evaluation to measure the accuracy. One of the research measure the lengths of matched pair and the length of unmatched roads then see the ratio with respect to the total length of the datasets. One of our work [54] has done experiment that automatically compute the length of matched pair and the length of unmatched roads then provide the quality of the framework based on the given dataset. In this case we assume the new dataset is the ground truth to run the experiment on it.

We choose the North region of Tarrant County to test the quality of our TAREEQ framework. We take new Road Map DS (2018) as the ground truth and calculate the accuracy based on total matched lengths from 2007 Road Map DS to 2018 Road Map DS. We calculate our results to get the values of Precision, Recall, F-Score. Based on our methods the equation of the three metrics as the following:

$$Precision = \frac{Total\_Length\_matched\_pairs}{(Total\_Length\_matched\_pairs + Total\_Length\_unmatched\_roads)}$$

$$Recall = \frac{Total\_Length\_matched\_pairs}{Total\_Length\_ground\_Truth}$$

$$F\_Score = 2 * \frac{Precision * Recall}{Precision * Recall}$$

We prepare two scenarios to test the TAREEQ framework, one when the datasets are not updated and the differences between two datasets exist and the second scenario when eliminate all the differences and we choose only the similar pairs between the two datasets that are gotten from the similarity measurements process:

Scenario 1. All the roads with expansions in new dataset and removed roads in old dataset . We run the experiment and we know the accuracy will not become 100% due to the differences between the datasets. We build a buffer -with TAREEQ framework threshold as the buffer width- around all the roads in ground truth dataset and see how much part of the roads from the old dataset are within the buffer and we compare the length inside the buffer and length outside the buffer and calculate the precision, recall, and F-Score. We have in this scenario:
Total Length of 2007 DS = 1,716,983m
Total Length of Ground Truth (2018 DS) = 2,423,651m
Matched Roads Length = 1,690,343.1

Based on that: Precision = 0.98448447; Recall = 0.69743668; and F-Score = 0.81646584

Our framework have the ability to score high points and fetch all similar roads even if there are differences between the dataset.

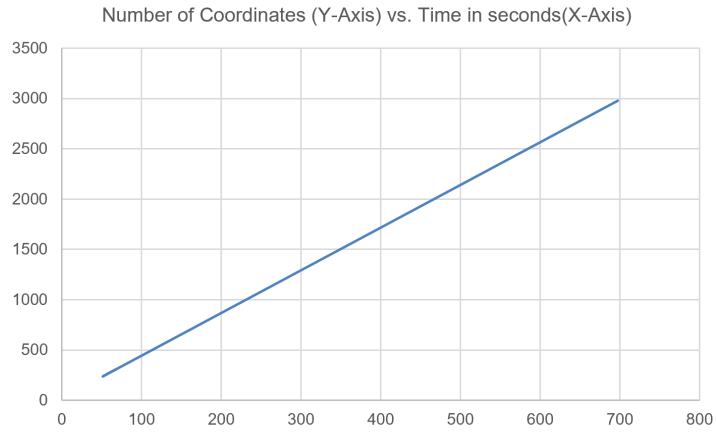
Scenario 2. Filter both datasets to have only the similar roads generated from TAREEQ framework This scenario to verify if the new dataset and old dataset have only the similar roads, the quality should be 100% on Precision, Recall, and F-Score. We eliminate also the partial similar roads. The question is: Does TAREEQ framework produce correct matching when it says these two roads are similar, it is similar in reality. After we conduct this evaluation we have the following:
Total Length of 2007 DS = 347,257m
Total Length of Ground Truth (2018 DS) = 347,257m
Matched Roads Length = 347,257m

Based on that: Precision = 1.0; Recall = 1.0; and F-Score = 1.0

As a result, the values of quality metrics give the same results of TAREEQ framework as all the 2007 DS roads fall inside the buffer of all the roads in ground truth (2018 DS).

4.5.3.2 The Performance of TAREEQ framework

One of main process in TAREEQ framework is "Similarity Measurements" process. This process is run over all coordinates consist the roads in both datasets- i.e. old dataset and new dataset-. Therefore, this process is extensive and consume a lot of time. In performance point of view, our framework can not compete the state-of-the-art. However, other state of the art road matching processes have preprocessing before the matching algorithm. This is to simplifying the matching process to match road segments together. Such preprocessing deleting two-line representation and con-



- Based on personal computer (Intel(R) Core(TM) i7-4600U@2.1GHz)

Figure 4.44: TAREEQ framework has high computation

vert it to single line. Also, our framework study the details to have the ability to determine partial similar roads that most other works do not do it.

Computation cost of the framework is high due to: 1) Similarity measurements process takes into consideration all the coordinates consist the road and match it with whole points consists the candidate pair from corresponding dataset. 2) It generates more details than getting results of matched or unmatched roads. 3) It depends on number of points that represents the roads O(n+m) which gives linear time complexity. Figure 4.44 shows the performance of TAREEQ framework based on the number of coordinates. We use personal computer (Intel(R) Core(TM) i7-4600U@2.1GHz) to conduct the experiment. It consumes around 100 seconds to finish comparing 500 coordinates. Our Tarrant experiment has 868,076 coordinates in both datasets and it takes around 48 hours, 13 minutes, and 35 seconds. It worths to mention that this computation is for 'Similarity Measurements' process with semantic filtration for 'Generate Candidates' process which does not consume time to fetch the results. However, if we consider spatial filtration for 'Generate Candidates', this will consume much more time.

4.6 Conclusion

In this chapter, we have apply TAREEQ framework to make historical comparison between two version of Road Map dataset (old dataset and new dataset). It tries to find missing roads names from new map dataset and indicate the road is either permanently closed or the road is changed its name to different name. Also, it compares all existing roads from old dataset with same roads from new datasets and find out: roads are either similar in length and shape, roads are partial similar- i.e. get extensions in new dataset or complete new roads in different area, road is no more exit in new datasets which means road is permanently closed or road has been shifted. In addition based on framework, we get the results that help to study the growth of the city based on discovering the area that has most new roads by dividing the hole area into 9 main grids. Our framework can help identifying the differences between roads which help to inspect more on datasets to determine if there are mistakes in new dataset or not.Our framework shows the flexibility in terms of how large datasets can be processed through this framework and our Tarrant County dataset is an example of that, though the performance does not compete the state of the art methods.

Future work is to incorporate constructing a road map by moving object trajectories[54], which building new road map network by studying and analyzing moving object trajectories. This will help to identify the errors systematically on new dataset instead of manually investigation.

CHAPTER 5

CONCLUSIONS

5.1 Summary of Contributions

This dissertation is focused on how can we get benefits of existing road maps and enhance each other by finding the matched pairs and identifying the differences for further investigation to determine which is the correct one. The idea of this work is to find feasible solution that could be used to enhance the existing road map or add more features on it by comparing it with other road maps. Acquisition of road maps is expensive in two main perspectives: the cost as well as the time consumption, not to mention the periodically cost of maintaining and updating on these spatial data. In addition to what mentioned above, each dataset is captured for specific application and each one focuses on some aspects of the real-world carefully while other aspects are ignored or not given much attention. Therefore, Road maps matching takes place to get the most of both road map dataset, i.e. the accuracy and up-to-date, with minimum cost and reasonable time.

We introduce our solution by presenting our TAREEQ framework. One of the main components in this framework is the "Similarity Measurements" process because based on these similarity measurements the framework can decide when comparing two roads from different datasets whether they are similar, partial similar, or different. We are utilizing the Local Roads Divergence Measurements (LRDM) and Global Roads Divergence Measurement (GRDM) [19, 33]. The reason for choosing them is the ability of these measurements to identify the similar roads even if there are missing, or additional, road segments in one of the road map datasets. LRDMs check

that two roads from different datasets have roughly a similar length and these roads keep running in parallel to one another if they represent the same road in the real world, which preserve a similar shape between them. This technique can be performed by computing the gap between the two roads from different datasets after overlaying one of the datasets over another. Affirming the similarity between these two roads requires also passing the GRDM condition after the LRDMs conditions have been met. GRDM ensures that the distances' between the roads' coordinates are within the limits of GRDM threshold. GRDM ensures the two roads are for a similar road in reality and they are not different roads that happened to be adjacent to one another.

Our TAREEQ framework shows the flexibility in terms of how large datasets can be processed through this framework and our Tarrant County dataset is an example of that, though the performance does not compete the state of the art methods. However, the ability of the framework to find the partial similar roads between the datasets and it does not identifying they partial similar only, it can determine which type of partial differences cause. Does it because of new road extension, road expansion, or new road in different area.

5.2 Future Work

As we are getting promising results, the future work will have two paths: Matching Road Maps with Moving Objects Trajectories, and Road Maps Integrations. Regarding Matching Road Maps with Moving Object Trajectories, one of the dataset is constructed using moving object trajectories such as buses or cars and the other dataset is road map that is available from other sources such as TIGER, OSM, Google Maps, ... and so on. The advantage of this method is that we can make sure any road constructed from the trajectories is road in real-world and once it is matched

with other road map dataset to get its name for the constructed road map and also to verify and adjust the roads in the other road map dataset.

Road Maps Integration is one of our future work since each road map dataset is captured for a specific application such as road navigation, topographic cartographic for printing map, and so on. Therefore, each one focuses on some aspects of the real-world carefully while other aspects are ignored or not given much attention. Thus road maps integration is coming in order to provide new applications from existing datasets that are not designed for such applications and quality improvement. Our proposed framework can integrate two road maps utilizing LRDMs and GRDM. It has the capabilities to match N:M road segments, which leads to match roads with missing segments with their pairs, that have all road segments, from a different dataset.

Finally, as our framework has high computation cost, we will try to enhance the performance in order to available for on-line applications instead of off-line job. This could be happened if we divide the ”Similarity Measurements” process task into two phase: one to determine the candidate similar roads are similar or not and provide the initial results to get these results fast. The second phase is to go into different candidate pairs again to see if they are partially similar or not and determine the type of partial similar pairs.

REFERENCES

[1] E. M. A. Xavier, F. J. Ariza-López, and M. A. Ureña Cámara, "A survey of measures and methods for matching geospatial vector datasets," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 39:1–39:34, Aug. 2016. [Online]. Available: http://doi.acm.org.ezproxy.uta.edu/10.1145/2963147

[2] S. Academy. Essentials of geographic information systems. [Online]. Available: https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/index.html

[3] T. U. C. Bureau, "Tiger products - geography - u.s. census bureau," 2019. [Online]. Available: https://www.census.gov/geo/maps-data/data/tiger.html

[4] A. Corp. Google maps. [Online]. Available: https://www.google.com/maps

[5] M. Corp. Bing maps. [Online]. Available: https://www.bing.com/maps

[6] H. Company. Here maps. [Online]. Available: https://www.here.com/

[7] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, pp. 211–221, 08 2007.

[8] O. contributors, "Openstreetmap," 2019. [Online]. Available: https://www.openstreetmap.org

[9] W. Contributors. Wikimapia maps. [Online]. Available: http://wikimapia.org/

[10] G. McKenzie, K. Janowicz, and B. Adams, "A weighted multi-attribute method for matching user-generated points of interest," *Cartography and Geographic Information Science*, vol. 41, no. 2, pp. 125–137, 2014. [Online]. Available: https://doi.org/10.1080/15230406.2014.880327

[11] T. Wenjing, H. Yanling, Z. Yuxin, and L. Ning, "Research on areal feature matching algorithm based on spatial similarity," in *2008 Chinese Control and Decision Conference*, July 2008, pp. 3326–3330.

[12] F. van Wijngaarden, J. van Putten, P. van Oosterom, and H. Uitermark, "Map integration—update propagation in a multi-source environment," in *Proceedings of the 5th ACM International Workshop on Advances in Geographic Information Systems*, ser. GIS '97. New York, NY, USA: ACM, 1997, pp. 71–76. [Online]. Available: http://doi.acm.org/10.1145/267825.267844

[13] M. J. Egenhofer and R. D. Franzosa, "On the equivalence of topological relations," *International Journal of Geographical Information Systems*, vol. 9, no. 2, pp. 133–152, 1995. [Online]. Available: https://doi.org/10.1080/02693799508902030

[14] F. Fonseca, C. Davis, and G. Câmara, "Bridging ontologies and conceptual schemas in geographic information integration," *GeoInformatica*, vol. 7, no. 4, pp. 355–378, Dec 2003. [Online]. Available: https://doi.org/10.1023/A:1025573406389

[15] A. Samal, S. Seth, and K. Cueto1, "A feature-based approach to conflation of geospatial sources," *International Journal of Geographical Information Science*, vol. 18, no. 5, pp. 459–489, 2004. [Online]. Available: https://doi.org/10.1080/13658810410001658076

[16] M. A. RodrÃŋguez and M. J. Egenhofer, "Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure," *International Journal of Geographical Information Science*, vol. 18, no. 3, pp. 229–256, 2004. [Online]. Available: https://doi.org/10.1080/13658810310001629592

[17] M. Schäfers and U. W. Lipeck, "Simmatching: Adaptable road network matching for efficient and scalable spatial data integration," in *Proceedings*

of the 1st ACM SIGSPATIAL PhD Workshop, ser. SIGSPATIAL PhD '14. New York, NY, USA: ACM, 2014, pp. 5:1–5:5. [Online]. Available: http://doi.acm.org/10.1145/2694859.2694866

[18] D. Min, L. Zhilin, and C. Xiaoyong, "Extended hausdorff distance for spatial objects in gis," *Int. J. Geogr. Inf. Sci.*, vol. 21, no. 4, pp. 459–475, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1080/13658810601073315

[19] M. Almotairi, T. Alsahfi, and R. Elmasri, "Using local and global divergence measures to identify road similarity in different road network datasets," in *Proceedings of the 11th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, ser. IWCTS'18. New York, NY, USA: ACM, 2018, pp. 21–28. [Online]. Available: http://doi.acm.org/10.1145/3283207.3283214

[20] E. Safra, Y. Kanza, Y. Sagiv, and Y. Doytsher, "Ad hoc matching of vectorial road networks," *Int. J. Geogr. Inf. Sci.*, vol. 27, no. 1, pp. 114–153, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1080/13658816.2012.667104

[21] T. Alsahfi, M. Almotairi, and R. Elmasri, "A survey on trajectory data warehouse," *Spatial Information Research*, May 2019. [Online]. Available: https://doi.org/10.1007/s41324-019-00269-x

[22] Y. Gabay and Y. Doytsher, "An approach to matching lines in partly similar engineering maps," *Geomatica*, vol. 54, pp. 297–310.

[23] S. F. Yerahmiel Doytsher and E. Ezra, "Transformation of datasets in a linear-based map conflation framework," *Surveying and Land Information Systems*, vol. 61, no. 3, pp. 159–169.

[24] J.-H. Haunert, "Link based conflation of geographic datasets," in *Proceedings of the 8th ICA Workshop on Generalisation and Multiple Representation*, 2005.

[Online]. Available: http://www1.informatik.uni-wuerzburg.de/pub/haunert/pdf/HaunertMapGen05.pdf

[25] S. Yuan and D. C. Tao, "Development of conflation components," in *In Proceedings of Geoinformatics, Ann Arbor*, 1999, pp. 1–13.

[26] W. Rucklidge, *Efficient Visual Recognition Using the Hausdorff Distance*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1996.

[27] T. Devogele, "A new merging process for data integration based on the discrete fréchet distance," in *Advances in Spatial Data Handling*, D. E. Richardson and P. van Oosterom, Eds. "Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 167–181.

[28] H. Alt, C. Knauer, and C. Wenk, "Comparison of distance measures for planar curves," *Algorithmica*, vol. 38, no. 1, pp. 45–58, Jan 2004. [Online]. Available: https://doi.org/10.1007/s00453-003-1042-5

[29] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[30] A. Efrat, Q. Fan, and S. Venkatasubramanian, "Curve matching, time warping, and light fields: New algorithms for computing similarity between curves," *Journal of Mathematical Imaging and Vision*, vol. 27, no. 3, pp. 203–216, Apr 2007. [Online]. Available: https://doi.org/10.1007/s10851-006-0647-0

[31] M. Butenuth, G. v. Gösseln, M. Tiedge, C. Heipke, U. Lipeck, and M. Sester, "Integration of heterogeneous geospatial data in a federated database," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 5, pp. 328 – 346, 2007, theme Issue: Distributed Geoinformatics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271607000275

[32] F. H. Administration. Mitigation strategies for design exceptions - safety. [Online]. Available: https://safety.fhwa.dot.gov/geometric/pubs/mitigationstrategies/chapter3/3_lanewidth.cfm

[33] M. Almotairi, T. Alsahfi, B. Alshemaimri, and R. Elmasri, "Challenges of comparing and matching roads from different spatial datasets," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '19. New York, NY, USA: ACM, 2019, pp. 164–171. [Online]. Available: http://doi.acm.org/10.1145/3316782.3316787

[34] S. Williams, A. Waterman, and D. A. Patterson, "Roofline: an insightful visual performance model for multicore architectures," *Commun. ACM*, vol. 52, pp. 65–76, 2009.

[35] D. Xiong and J. Sperling, "Semiautomated matching for network database integration," 2004.

[36] G. LLC., "Google earth pro," *32 44 44.12 N 97 05 27.63 W*, 2005. [Online]. Available: http://www.earth.google.com

[37] G. Corp., "Google earth pro," *32 44 44.12 N 97 05 27.63 W*, 2018. [Online]. Available: http://www.earth.google.com

[38] B. Wan, L. Yang, S. Zhou, R. Wang, D. Wang, and W. Zhen, "A parallel-computing approach for vector road-network matching using gpu architecture," *ISPRS Int. J. Geo-Information*, vol. 7, p. 472, 2018.

[39] L. Li and M. F. Goodchild, "An optimisation model for linear feature matching in geographical data conflation," 2011.

[40] J. Zhang, Y. Wang, and W. Zhao, "An improved probabilistic relaxation method for matching multi-scale road networks," *Int. J. Digital Earth*, vol. 11, pp. 635–655, 2018.

[41] H. Fan, B. Yang, A. Zipf, and A. Rousell, "A polygon-based approach for matching openstreetmap road networks with regional transit authority data," *International Journal of Geographical Information Science*, vol. 30, pp. 748–764, 2016.

[42] Y. Li, "Matching road network based on the structural relationship constraint of hierarchical strokes," 2015.

[43] M. Deng, Z. Li, and X. Chen, "Extended hausdorff distance for spatial objects in gis," *International Journal of Geographical Information Science*, vol. 21, pp. 459–475, 2007.

[44] M. Zhang and L. Meng, "Delimited stroke oriented algorithm-working principle and implementation for the matching of road networks," *Annals of GIS*, vol. 14, pp. 44–53, 2008.

[45] A. Samal, S. C. Seth, and K. Cueto, "A feature-based approach to conflation of geospatial sources," *International Journal of Geographical Information Science*, vol. 18, pp. 459–489, 2004.

[46] T. COUNTY. Tarrant county, texas. [Online]. Available: https://access.tarrantcounty.com/content/dam/main/administration/misc%20docs/Fast%20Fact%20Page%202017.pdf

[47] U. S. C. Bureau. Quickfacts new york county (manhattan borough), new york; united states. [Online]. Available: https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork,US/PST045218

[48] B. Yang, X. Luan, and Y. Zhang, "A pattern-based approach for matching nodes in heterogeneous urban road networks," *Trans. GIS*, vol. 18, pp. 718–739, 2014.

[49] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq, "Challenges of data integration and interoperability in big data," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 38–40.

[50] B. Yang, Y. Zhang, and X. Luan, "A probabilistic relaxation approach for matching road networks," *International Journal of Geographical Information Science*, vol. 27, pp. 319–338, 2013.

[51] S. Yehua, "Research on automatic matching of vector road networks based on global optimization," 2010.

[52] X. Tong, W. Shi, and S. Deng, "A probability-based multi-measure feature matching method in map conflation," 2009.

[53] X. Tong, D. Liang, and Y. Jin, "A linear road object matching method for conflation based on optimization and logistic regression," *International Journal of Geographical Information Science*, vol. 28, pp. 824–846, 2014.

[54] T. Alsahfi, M. Almotairi, and R. Elmasri, "Road map generation and feature extraction from gps trajectories data," in *Proceedings of the 12th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, ser. IWCTS'18. New York, NY, USA: ACM, 2018, pp. 21–28. [Online]. Available: http://doi.acm.org/10.1145/3283207.1113214

BIOGRAPHICAL STATEMENT

Mousa Almotairi was born in Riyadh, Saudi Arabia. He received his Bachelor and Masters' degree in Computer Science from the King Saud University, Saudi Arabia, in 2003 and 2007 respectively. In 2015, he started his studies to pursue his Ph.D at the University of Texas at Arlington. His current research interests include Spatial database integration, and Road similarity.